Geodesic Convex Analysis of Group Scaling for the Paulsen Problem and the Tensor Normal Model

by

Akshay Ramachandran

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Ankur Moitra, Professor
Department of Mathematics
Massachusetts Institute of Technology

Supervisor(s): Lap Chi Lau, Professor
Cheriton School of Computer Science
University of Waterloo

Internal Member: Rafael Oliveira, Assistant Professor
Cheriton School of Computer Science
University of Waterloo

Internal-External Member: Vern Paulsen, Professor
Department of Pure Mathematics and
Institute for Quantum Computing,
University of Waterloo

Internal Member: John Watrous, Professor
Cheriton School of Computer Science and
Institute for Quantum Computing,
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The main results of this thesis are based on the following papers that I have coauthored.

1. [62]: *The Paulsen Problem, Continuous Operator Scaling, and Smoothed Analysis.* Joint work with Tsz Chiu Kwok, Lap Chi Lau, and Yin Tat Lee. 49th Symposium on Theory of Computing (STOC 2018).

2. [63]: *Spectral Analysis of Matrix Scaling and Operator Scaling.* Joint work with Tsz Chiu Kwok and Lap Chi Lau. SIAM Journal on Computing (2021). Preliminary Conference version in 60th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2019).

3. [36]: *Logarithmic sample complexity for dense matrix and tensor normal models.* Joint work with Cole Franks, Rafael Oliveira, and Michael Walter. arXiv preprint arXiv:2110.07583.

I understand that my thesis may be made electronically available to the public.

Abstract

The framework of scaling problems has recently had much interest in the theoretical computer science community due to its variety of applications, from algebraic complexity to machine learning. In this thesis, our main motivation will be two new applications: the Paulsen problem from frame theory, and the tensor normal model in statistical estimation. In order to give new results for these problems, we provide novel convergence analyses for matrix scaling and tensor scaling. Specifically, we will use the framework of geodesic convex optimization presented in Bürgisser et al. [20] and analyze two sufficient conditions (called strong convexity and pseudorandomness) for fast convergence of the natural gradient flow algorithm in this setting. This allows us to unify and improve many previous results [62], [63], [36] for special cases of tensor scaling.

In the first half of the thesis, we focus on the Paulsen problem where we are given a set of n vectors in d dimensions that ε-approximately satisfy two balance conditions, and asked whether there is a nearby set of vectors that exactly satisfy those balance conditions. This is an important question from frame theory [24] for which very little was known despite considerable attention. We are able to give optimal distance bounds for the Paulsen problem in both the worst-case and the average-case by improving the smoothed analysis approach of Kwok et al. [62]. Specifically, we analyze certain strong convergence conditions for frame scaling, and then show that a random perturbation of the input frame satisfies these conditions and can be scaled to a nearby solution.

In the second half of the thesis, we study the matrix and tensor normal models, which are a family of Gaussian distributions on tensor data where the covariance matrix respects this tensor product structure. We are able to generalize our scaling results to higher-order tensors and give error bounds for the maximum likelihood estimator (MLE) of the tensor normal model with a number of samples only a single dimension factor above the existence threshold. This result relies on some spectral properties of random Gaussian tensors shown by Pisier [80]. We also give the first rigorous analysis of the Flip-Flop algorithm, showing that it converges exponentially to the MLE with high probability. This explains the empirical success of this well-studied heuristic for computing the MLE.

Acknowledgements

I would like to express my sincere gratitude to my advisor Lap Chi Lau for a wonderful graduate school experience. His clear-thinking and meticulous nature have had a strong influence on my research. But most of all, I would like to thank him for showing that I was cared for as a person, not just as a researcher.

I would also like to thank Ankur Moitra, Rafael Oliveira, Vern Paulsen, and John Watrous for agreeing to read this thesis and serve on the examining committee.

The research in this thesis was completed in collaboration with: Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Cole Franks, Rafael Oliveira, and Michael Walter. I would also like to once again thank Cole Franks, Rafael Oliveira, and Michael Walter for the zoom meetings we had over this past year, which were great fun.

I would like to take this opportunity to thank Daniel Dadush, Sasho Nikolov, and Nikhil Bansal for being so generous with their time and helping me learn discrepancy theory.

A great big thank you to all my friends and office mates in spirit: Hong, Vedat, Chiu, Alan, Nathan, Amit, Cedric, Sharat, Justin, Madison, Logan, Harry, William, Sifat, Claire, Lily, Leanne, Nolan, and Ryan.

To Amma, Appa, and Patti, thank you for always supporting me and making me feel loved.

And to Dahlia, thank you for filling this past year with support and joy, and helping me keep this all in perspective.

Dedication

Dedicated to my family.

Table of Contents

Chapter 1

Introduction

A key preliminary step in many fundamental linear algebraic problems is to place the given input into an appropriate "normal form" so that downstream algorithms can be applied with better performance. The complexity of finding these preliminary transformations ranges from the very basic (e.g. matrix diagonalization) to intractable (e.g. tensor decomposition). In this thesis, we study two fundamental balance properties of linear algebraic objects and the sets of transformations required to achieve them. Specifically, we are motivated by the Paulsen problem in frame theory, which will be the focus of the first half of the thesis (Chapters 3-5), and the tensor normal model in statistics, which will be the focus of the second half (Chapters 6-9). Our main analytic tool involves the scaling framework of [20], and specifically, a geodesically convex optimization formulation for these problems. Using this perspective, we are able to use ideas from convex optimization to analyze the solution to the Paulsen problem and the maximum likelihood estimator for the tensor normal model. In the following two sections, we discuss known results for these two problems. Then we present the ideas from the scaling framework that are used to prove the main results in this thesis.

1.1 The Paulsen Problem

The first half of this thesis is motivated by the Paulsen problem in frame theory [24].

Question 1.1.1. *Let $U = \{u_1, ..., u_n\} \subseteq \mathbb{C}^d$ be a spanning set of vectors satisfying*

$$\frac{1-\varepsilon}{d} I_d \preceq \sum_{j=1}^{n} u_j u_j^* \preceq \frac{1+\varepsilon}{d} I_d, \qquad \forall j \in [n] : \frac{1-\varepsilon}{n} \leq \|u_j\|_2^2 \leq \frac{1+\varepsilon}{n}. \tag{1.1}$$

Bound the minimum of $\sum_{j=1}^{n} \|v_j - u_j\|_2^2$ over $V = \{v_1, ..., v_n\}$ satisfying Eq. (1.1) exactly:

$$\sum_j v_j v_j^* = \frac{1}{d} I_d, \qquad \forall j \in [n] : \|v_j\|_2^2 = \frac{1}{n}.$$

We point out that this is a different normalization, by a factor d, than normally given in the literature. We choose this normalization in order to make clearer the dependence of the optimal distance bound on the parameters d, n, and ε.

This questions arises from Frame Theory [24], which can be thought of as the study of redundant representations of vector spaces. Doubly balanced frames, those satisfying Eq. (1.1) with $\varepsilon = 0$, are used in coding theory and signal processing for their stability properties. Many of these applications also require frames that satisfy further constraints, such as large pairwise angles or sparsity. Constructions of these frames are difficult and often rely on complicated algebraic structures. On the other hand, there are many simple algorithms to construct frames that approximately satisfy the requirements. For example, by standard matrix concentration results (see [94]), a large enough set of random unit vectors will satisfy Eq. (1.1) for some small ε with high probability. The Paulsen problem asks, for a given ε-doubly balanced frame, whether the conditions in Eq. (1.1) can be corrected without moving too much. And as many of the approximate constructions come from randomly generated frames, understanding the distance bound in the average case is also of interest.

In [49], Holmes and Paulsen studied frames from the perspective of coding theory, and showed that doubly balanced frames were optimally robust with respect to a single erasure. They also showed that Grassmannian frames, doubly balanced frames with large pairwise angles, were optimal for two erasures.

To address the difficulty of constructing these structured frames, the authors of [49] suggested a simple numerical approach: first generate random frames, which approximately satisfy Eq. (1.1), and then correct the conditions. Random frames are good candidates for both of these settings because they are approximately doubly balanced and have large pairwise angles with high probability. One goal of the Paulsen problem is then to validate this numerical algorithm as a simple method of constructing structured frames. The problem is formalized below.

Conjecture 1.1.2 (Paulsen Problem [22])**.** *Let $p(d,n,\varepsilon)$ be the minimum value such that for every ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$, there exists a doubly balanced $V \in \mathrm{Mat}(d,n)$ with $s(V) = 1$ such that*

$$\|V - U\|_F^2 \leq p(d,n,\varepsilon).$$

Then p can be bounded by a polynomial function in d and ε. In particular, this function can be taken to be independent of n.

The optimal bounds for p have been unknown for almost twenty years, despite considerable attention in the frame theory literature. Prior to our work in [62], there were two known partial results on the function p [23] and [16], which showed the bound $p \leq \mathrm{poly}(d,n,\varepsilon)$ when d,n are relatively prime and ε is small enough. These results left open Conjecture 1.1.2, which we positively resolved in [62].

Theorem 1.1.3 (Theorem 1.3.1 in [62])**.** *The distance function can be bounded by $p(d,n,\varepsilon) \lesssim d^{11/2}\varepsilon$, which is independent of n.*

Our new idea was to use scaling algorithms similar to those studied recently in the work of Gurvits, Garg, Oliveira, and Wigderson [38]. In order to carry out this approach, we defined a dynamical system which corrected the balance condition for nearly doubly balanced frames. This dynamical system could then be analyzed using tools from the operator scaling framework studied in [38]. The full proof of [62] required a smoothed analysis argument coupled with an involved convergence analysis of the dynamical system.

Subsequently, in the aptly titled "Paulsen Problem made Simple" [46], Hamilton and Moitra improved the distance bound to $p(d,n,\varepsilon) \lesssim d\varepsilon$, using a totally different and much simpler method. This almost matches the known lower bound $p \gtrsim \varepsilon$, which is shown by simple examples in [23] (see Example A.1.1 in the Appendix).

In this thesis, we give two new results for the Paulsen problem by revisiting the scaling approach. In our first result, we extend the arguments of [62] in order to prove an optimal distance bound for the Paulsen problem.

Theorem 1.1.4. *For any d with $n \gtrsim d$ large enough and $\varepsilon \lesssim \frac{1}{d}$ small enough, the distance function in Conjecture 1.1.2 can be bounded by*

$$p(d,n,\varepsilon) \lesssim \varepsilon.$$

This matches the known lower bound up to constants wherever it applies. The full result is presented in Theorem 4.5.3 and covers a slightly larger range of parameters.

We achieve this improvement by a deeper understanding of the scaling framework, along with refinements of many technical arguments in [62]. An important contribution of this thesis is to connect the approach of [62] for the Paulsen problem to the long line of work on scaling and the Kempf-Ness function in geometric invariant theory [58]. This allows us to re-derive the dynamical system approach in a principled manner, as well as to use powerful tools from convex analysis and algebraic geometry for the analysis.

Our second result is a beyond worst-case distance analysis in the case of random frames. This allows us to prove an optimal distance bound for the case of random frames, which answers the original motivation of the Paulsen problem, and gives a tight improvement of similar results in Theorem 1.12 in [63] and Franks and Moitra [35].

Theorem 1.1.5. *For any $n \gtrsim d$ large enough, if $U = \{u_1, ..., u_n\} \subseteq \mathbb{R}^d$ is generated such that each u_j is independent and uniformly distributed on $\frac{1}{\sqrt{n}} S^{d-1}$, then with high probability U is ε-doubly balanced for $\varepsilon \lesssim \sqrt{\frac{d}{n}}$, and there exists doubly balanced V such that*

$$\|V - U\|_F^2 \lesssim \varepsilon^2.$$

This result validates the numerical approach suggested in [49] to generate doubly balanced frames, and therefore gives a satisfactory answer to the original motivation for Question 1.1.1. As further validation, we follow the approach of [63] and use our distance analysis to give new simple constructions of nearly optimal Grassmannian frames in Theorem 4.4.5. In the final section, we will discuss how scaling is used to show these bounds.

1.2 Tensor Normal Model in Statistics

Covariance matrix estimation is an important task in statistics, machine learning, and the empirical sciences. We consider covariance estimation for matrix-variate and tensor-variate Gaussian data, that is, when individual data points are matrices or tensors. Matrix-variate data arises naturally in numerous applications like gene microarrays, spatio-temporal data, and brain imaging. One significant challenge is that the the dimension of these objects grows as the product of the dimension of the factors, whereas the number of samples available may be much fewer. To get around this issue, we can add a structural assumption to the (unknown) covariance matrix. One natural assumption is known as the tensor normal model ([31]; [99]), and has applications in signal processing and data analysis. Here, we assume $X \in \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ is distributed as a centered Gaussian with covariance $\Theta := \Theta_1 \otimes ... \otimes \Theta_m$, where each $\Theta_a \in \mathrm{Mat}(d_a)$ is a positive definite matrix on $\mathbb{R}^{d_a}$. By adding this structural assumption, the unknown covariance can now be described by fewer parameters, and so we could hope to estimate the covariance matrix with fewer samples.

These are quite natural assumptions for tensor data, and therefore there are many heuristics and algorithms used in practice. One natural solution to this task is known as maximum likelihood estimation. Given a set of samples $X_1, ..., X_n \in \mathbb{R}^D$, the likelihood of estimator Θ is defined as the probability of getting these samples if the true covariance was Θ. The maximum likelihood estimator (MLE) is defined as the parameter $\hat{\Theta}$ which maximizes the likelihood function over all feasible Θ. The quality of the MLE depends on the measure of error that is relevant to the application, and in fact, the MLE does not even have to exist in general. For the tensor normal model, there are known results showing that the MLE converges to the true distribution in the asymptotic setting where the number of samples n goes to infinity [99]. Dutilleul [31] proposed the natural Flip-Flop algorithm to compute the MLE an estimator. This algorithm iteratively updates one tensor factor at a time using the natural Gaussian estimator for that marginal. It has been observed in practice that this algorithm converges to the MLE with high probability, but before our work in [36], there was no known rigorous convergence analysis for this procedure. In this work, we are able to use our analysis of tensor scaling given in Chapter 7 to give high probability error bounds for the MLE as well as a rigorous convergence analysis of the Flip-Flop algorithm for finite samples.

The optimization problem defining the MLE of the tensor normal model is non-convex, so we cannot approach it using standard algorithms. But as we will describe in the following section on the scaling framework, this function is geodesically convex over the set of possible covariance matrices. Further, by some powerful results from the work of Pisier on operator theory [80], the function is geodesically strongly convex with high probability. When there are enough samples for this strong convexity to hold, we are able to bound the error of the MLE.

Theorem 1.2.1. *For samples $X_1, ..., X_n \in \mathbb{R}^D$ from the tensor normal model with unknown covariance $\Theta := \Theta_1 \otimes ... \otimes \Theta_m$, if $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$ for any $\varepsilon \lesssim \frac{1}{\mathrm{poly}(m)\sqrt{d_{\max}}}$, then with high probability, the MLE $\widehat{\Theta}$ satisfies*

$$d_F(\widehat{\Theta}, \Theta)^2 \lesssim Dm\varepsilon^2,$$

for $d_F(A, B) := \|I - B^{-1/2}AB^{-1/2}\|_F$ in Definition 9.1.6. Further, in this event, the Flip-Flop algorithm has linear convergence (in d_F) to the optimizer.

This result should be compared to the known error bounds for general Gaussian covariance estimation for which $n \gtrsim d$ samples are known to be necessary even to estimate the covariance up to constant error. In Section 9.2.7, we show two improved results: first, we are able to reduce the constraint on ε by a factor of $d_{\max}^{\Omega(1/m)}$ which allows us to improve the sample complexity by a similar factor; and second, we are able to prove refined error bounds in the operator norm for each individual factor. We believe that these error bounds should hold as soon as $nD \gtrsim \text{poly}(m) d_{\max}^2$, which would match the standard Gaussian result up to $\text{poly}(m)$ factors. In the final section, we discuss in more detail the connection to scaling and our proof techniques.

1.3 Scaling Framework

In recent years, there has been much interest in problems from the scaling framework [38], [39], [19], [20]. These problems originate in the field of geometric invariant theory, which studies the algebraic structure of group actions on vector spaces. Below, we present some concrete examples of scaling that will be relevant to our main applications discussed in the previous two sections.

One of the simplest problems in this framework is matrix scaling. The goal here is, given a non-negative matrix $A \in \text{Mat}(n)$, to find positive diagonal scalings $L, R \in \text{diag}(n)$ such that $B := LAR$ is doubly stochastic:

$$B\mathbf{1}_n = B^T\mathbf{1}_n = \mathbf{1}_n.$$

This problem has been re-discovered many times throughout mathematics and has been used as a subroutine for a variety of applications in statistics [83], approximation of the permanent [66], and optimal transport [27]. Recently, faster algorithms have been developed for matrix scaling [2], [26] using techniques from convex optimization and fast Laplacian solvers. In fact, in this simple setting, it has long been known that matrix scaling can be solved using a convex formulation.

The next problem we study is frame scaling, which is more directly related to Question 1.1.1. Here, we are given a set of vectors $\{u_1, ..., u_n\} \in \mathbb{R}^d$, and the goal is to find a transformation $L \in \text{Mat}(d)$ and scalars $c_1, ..., c_n$ such that

$$\sum_{j=1}^{n} (Lu_j c_j)(Lu_j c_j)^* = \frac{n}{d} I_d.$$

The frame scaling problem also has a long history in theoretical computer science and mathematics [34], [50], and has been referred to by other names such as the radial isotropic position of vectors [47], and the geometric condition for Brascamp-Lieb inequalities [10]. This is in a sense the next simplest scaling problem after matrix scaling, and this also has known convex formulations [47], though they are slightly less obvious. In the following section, we will show how this problem relates to our solution of the Paulsen problem.

The operator scaling problem is a generalization of both the matrix and frame versions. In this setting, we are given a tuple of matrices $\{A_1, ..., A_K\} \in \mathrm{Mat}(n)^K$, and we want to find $L, R \in \mathrm{Mat}(n)$ such that $\{B_k := LA_kR\}_{k=1}^K$ is doubly balanced:

$$\sum_{k=1}^K B_k B_k^* = \sum_{k=1}^K B_k^* B_k = I_n.$$

The operator scaling problem was defined in the work of Gurvits [45] in the context of the polynomial identity testing question in algebraic complexity, and a simple iterative algorithm was proposed to solve it. Garg, Gurvits, Oliveira, and Wigderson [38] show that this algorithm converges in polynomial time. As a consequence, this gave the first polynomial time algorithm for a variety of problems in algebraic complexity, including a non-commutative version of polynomial identity testing. There are simple reductions showing that both matrix scaling and frame scaling are special cases of operator scaling problems. In the following section, we will briefly describe a generalization of operator scaling to higher order tensors and show how it is applied to our statistical application.

The above examples, as well as the tensor scaling generalization discussed later, are all fundamental linear algebraic problems that have had a wide variety of applications in many areas of mathematics [45], [39], [10]. They also share two important common features that are not immediately visible: the domain is a group of symmetries, and there is an underlying optimization formulation.

These two features motivate the following framework of [20], which gives a unified approach to many scaling problems. It can be shown that the required balance conditions in the scaling problems above can be written as first order optimality conditions for a certain natural function from geometric invariant theory called the Kempf-Ness function (Definition 3.1.6 and Definition 6.2.7). Further, it can be shown that the set of scalings in the problems above can be restricted to subsets of positive definite matrices. Finally, by viewing the domain of positive definite matrices from a particular (geodesic) geometry, we can reveal the underlying convexity of the Kempf-Ness function.

This suggests that we can lift ideas from classical convex optimization to this geodesic setting in order to solve scaling problems. This perspective was a major contribution of [20] and allowed them to give a principled analysis for a variety of known algorithms for the scaling framework. In many instances, they were able to propose new algorithms for previously intractable problems. We discuss these algorithmic results in more detail in Chapter 8.

But these results are given for worst case instances, and do not imply strong enough bounds for our applications (Paulsen problem and tensor normal model). Therefore, an important technical contribution of this thesis is to provide new stronger analyses of matrix and tensor scaling when the inputs satisfy certain special conditions. By using techniques from geodesic convex optimization, we are able to unify and refine the results of [62], [63], [36] to give nearly optimal bounds for scaling in beyond worst-case settings.

Specifically, we will analyze instances when the convex formulation satisfies a certain strong convexity property, or a combinatorial pseudorandom condition, and we show that in these cases, the natural gradient flow algorithm for scaling converges quickly. These analyses allow us to prove much stronger bounds on many parameters of the scaling problem, including distance and condition number bounds on the solution. The special strong convexity and pseudorandom conditions are satisfied by random inputs, as well as in our smoothed analysis setting. These stronger analyses are used to show our main results for the Paulsen problem and tensor normal model, as discussed in the following section.

1.4 Applications of Scaling

In this section, we discuss how frame scaling and tensor scaling arise naturally in our context of the Paulsen problem and tensor normal model, respectively. Then we outline our approaches to use the scaling framework to give strong bounds for these problems.

1.4.1 Solution to the Paulsen problem

In this subsection, we outline the smoothed analysis and scaling approach first used in [62] to give a polynomial distance bound for the Paulsen problem. We then discuss our particular improvements to the approach that lead to optimal distance bounds.

Recall that in Question 1.1.1, we are given a frame that nearly satisfies two balance conditions and would like to transform it into an exactly doubly balanced one. It turns out that it is easy to fix each balance condition individually, and for this simpler problem, optimal distance bounds are well-known in the literature (see Fact A.3.1). This suggests the following natural procedure: alternatively fix each balance condition until both are satsifed. Unfortunately, fixing one condition might destroy the other, and there are examples showing that the above algorithm does not converge to a doubly balanced frame or even converge at all [24]. This procedure also does not come with any meaningful distance guarantees, as the later iterations could take very large steps.

In [62], we thought that the alternating procedure may be taking large steps while moving the frame very little. So we defined a dynamical system on frames that can be viewed as an infinitesimal version of this simple alternating algorithm.

Definition 1.4.1. *Consider the following vector field, defined for each $V \in \mathrm{Mat}(d,n)$ as*

$$\nabla_V := \left\{ \left(d \sum_{j=1}^{n} v_j v_j^* - s(V) I_d \right) v_j + v_j \left(n \|v_j\|_2^2 - s(V) \right) \right\}_{j=1}^{n} . \tag{1.2}$$

where $s(V) := \|V\|_F^2$. Then for input U, dynamical system $\{U(t)\}_{t \geq 0}$ is defined as the solution to the differential equation $\partial_t U(t) = -\nabla_{U(t)}$ with initial condition $U(0) = U$.

Observe that doubly balanced U is a fixed point of the dynamical system in Eq. (1.2). Therefore, we can try to give a distance bound for the Paulsen problem by considering the path length of the flow $U(t)$ until convergence. It turns out that this gradient flow can be profitably understood using the scaling framework. Specifically, by the powerful Kempf-Ness equivalence theorem [58], the dynamical system in Eq. (1.2) is actually a natural gradient flow for the geodesically convex formulation for frame scaling. In [62], we gave a bound on the distance travelled in terms of the convergence of the dynamical system. It turns out that the potential function we used in this analysis can be formally derived from the geodesic convex formulation, though we did not know this at the time. Therefore, in order to give a strong distance bound for the Paulsen problem, it was sufficient to show fast convergence of the potential function for all time.

Unfortunately, this dynamical system does not always converge to a doubly balanced frame. In [62], our solution was to use smoothed analysis, by randomly perturbing input $U \in \mathrm{Mat}(d,n)$ to $V := U + E$, and then applying the dynamical system to this perturbed input. This gave the following distance analysis:

$$\|V(\infty) - U\|_F^2 \lesssim \|V(\infty) - V\|_F^2 + \|V - U\|_F^2 . \tag{1.3}$$

In [62], we used a complicated probabilistic analysis to exhibit a perturbation $V := U + E$ such that $\|V - U\|_F^2 \lesssim \mathrm{poly}(d) \cdot \varepsilon$, and V satisfied a certain combinatorial pseudorandom condition. Then, under this pseudorandom assumption, we were able to prove $\|V(\infty) - V\|_F \lesssim \varepsilon$. Theorem 1.1.3 follows by combining both of these arguments.

In this thesis, we will simultaneously improve and simplify both of these steps using our new understanding of the scaling framework. An important contribution of this thesis is to formalize the connection of the dynamical system approach of [62] to the scaling framework of [20], which allows us to derive much of the distance analysis in a principled way. As our first improvement, we are able to give a new convergence analysis of scaling inputs that satisfy certain strong convexity or pseudorandom conditions. This clarifies and improves our work in [62] and [63], where we proved fast convergence for operator scaling by ad-hoc methods. A key component of our improved analysis is a reduction from frame scaling to the simpler matrix scaling problem, where the optimization formulation is actually convex in the standard Euclidean sense by a simple change of variables. Therefore, we are able to lift tools from standard convex optimization in order to give strong analyses of frame scaling when the input satisfies strong convexity and pseudorandom conditions. This analysis implies our optimal result for the Paulsen problem in the average case, as random frames satisfy these conditions with high probability.

To prove our worst-case result in Theorem 1.1.4, we need to find a perturbation of ε-doubly balanced input $U \in \text{Mat}(d, n)$ so that it satisfies the pseudorandom condition, so that we can apply our improved analysis. Following the smoothed analysis strategy of [62], we take E to be a random Gaussian from a specially chosen subspace, and by choosing the correct parameters, we are able to prove tight distance bounds for both terms in Eq. (1.3).

1.4.2 Error bounds for the Tensor Normal Model

Our results for the tensor normal model follow more directly from the scaling framework. In Section 9.2.4, we reduce the analysis of the optimization problem for the matrix and tensor normal model to the analysis of tensor scaling for standard Gaussian inputs $X_1, ..., X_n \sim N(0, I_D)$. Then we are able to show, using powerful Gaussian concentration results, that these random inputs have very small initial error and satisfy certain strong convexity and pseudorandom conditions.

In Chapter 7, we take a geodesically convex optimization approach to prove strong bounds on the tensor scaling solution for such inputs. With the framework of Chapter 8, we can further view the Flip-Flop algorithm as a natural descent method for geodesic convex optimization, which then implies linear convergence to the MLE via standard results on strongly convex optimization. This gives the first rigorous analysis for the Flip-Flop algorithm and gives an explanation for the empirical success of this algorithm in practice.

The fast convergence results on Chapter 7 can be thought of as a generalization of the results of Chapter 3 on matrix scaling to higher order tensors. In fact, many parts of the analysis are similar in spirit, though they require new ideas to apply to the tensor setting. We believe that this similarity is a useful contribution of this thesis, as we are able to unify and improve the results of [62], [63], and [36]. We point out that many of the known results for operator scaling break down when lifted to higher-order tensors, and in particular even for 3-tensor scaling, there is no known polynomial time algorithm in the worst case. Therefore, the work in this thesis suggests that a preliminary step in understanding the complexity of scaling problems is to find special classes of inputs where the problem is tractable.

1.5 Organization

In Chapter 2, we introduce the necessary preliminary concepts that will be used in this thesis, including basic linear algebra, convex analysis, and concentration inequalities. Then, the first half of the thesis will be devoted to our solution of the Paulsen problem. Specifically, in Chapter 3, we analyze matrix scaling for inputs satisfying strong convexity or pseudorandom convexity. In these cases, we are able to give much stronger bounds on the scaling solution. This will be used in Chapter 4 to give optimal distance bounds for the Paulsen problem in both the worst-case and average-case settings. The proof of the core smoothed analysis argument for the Paulsen problem is deferred to Chapter 5, and will mostly rely on tools from random matrix theory.

The second half of the thesis will build towards our results for the tensor normal model in statistics. In Chapter 6, we will present the scaling framework of [20], especially focusing on the geodesic convex optimization formulation for tensor scaling. Then, Chapter 7 parallels Chapter 3 by giving strong analyses of the tensor scaling problem for strongly convex or pseudorandom tensor inputs. In Chapter 8, we use tools from standard convex analysis to present and analyze natural iterative algorithms for the geodesic tensor scaling setting. Finally, Chapter 9 contains our main results for this application by showing that the random inputs for the tensor normal model satisfy the fast convergence conditions studied in Chapter 7. Therefore, we are able to show sample complexity and error bounds for a natural estimator that are close to optimal, as well as give a rigorous analysis of linear convergence for a natural iterative method to this high-quality estimator.

We conclude this thesis in Chapter 10 with a brief summary and discussion of future directions.

Roadmap: For the Paulsen problem, we suggest reading Chapter 4 to see the promised new results, and then reading the scaling analysis in Chapter 3 followed by the smoothed analysis argument in Chapter 5. For the tensor normal model, we suggest reading Chapter 9 to see the reduction and main results, followed by Chapters 6-8 for our tensor scaling arguments. The reader already familiar with the scaling framework and geodesic convex optimization can skip Chapter 6 and read Chapter 3 and Chapter 7 directly for our new analyses of matrix and tensor scaling, respectively. Chapter 6 contains some interesting background from geometric invariant theory that underlies our geodesic convex formulation and points to many open problems in the scaling framework. The reader interested in geodesic convex optimization can see Chapter 8, which contains the main algorithmic results for tensor scaling.

Chapter 2

Preliminaries

We will write $\mathbb{R}$ for the real numbers and $\mathbb{C}$ for the complex numbers. Given positive integer n, we use $[n]$ to denote the set $\{1, ..., n\}$. We use $\binom{n}{k}$ for the binomial coefficient, and $\binom{[n]}{k}$ for the set of all k-subsets $\{S \subseteq [n] \mid |S| = k\}$. We use $Pr[\cdot]$ for the probability of an event, and $\mathbb{E}[\cdot]$ for the expectation of a random variable. For functions $f, g : \mathbb{N} \to \mathbb{R}_+$, $f(n) \leq O(g(n))$ is used to mean that there is a pair of universal constants $n_0, C > 0$ such that $\forall n \geq n_0 : f(n) \leq Cg(n)$, and $f(n) \geq \Omega(g(n))$ is equivalent to $g(n) \leq O(f(n))$. Similarly, $a \lesssim b$ means that there is an unspecified universal constant C such that $a \leq Cb$, and $a \gtrsim b$ is the same as $b \lesssim a$. A group is a set G with an associative pairwise operation $(\cdot)$, along with identity and inverse elements. The action of group G on set X is defined by a map $\sigma : G \times X \to X$ such that the elements $\{\sigma_g\}_{g \in G}$ under composition are compatible with the group operation $(\cdot)$ on G.

2.1 Linear Algebra

In this section, we present some important concepts from linear algebra: the Spectral Theorem in Section 2.1.3, the Polar decomposition in Section 2.1.8, and some matrix inequalities in Section 2.1.9. The first part of the thesis leading up to the solution of the Paulsen problem (Chapters 3-5) will only use these concepts along with some convex optimization for vector spaces. The more abstract group and algebra perspective covered in the following section will only be used in the second part of the thesis.

2.1.1 Vector Spaces

In this subsection, we formally define vector spaces and inner product spaces. We follow the presentation of Axler [7].

Definition 2.1.1 (Vector Space)**.** *A vector space V over field $\mathbb{F}$ is a set of vectors V along with vector addition $+ : V \times V \to V$, and scalar multiplication $\cdot : \mathbb{F} \times V \to V$ satisfying the following compatibility conditions: commutativity, associativity, additive identity, additive inverse, multiplicative identity, distributive property. We omit the formal definition of these natural properties and refer the reader to Chapter 1 of [7].*

For vector space V, a subset of vectors $\{v_1, ..., v_k\} \subseteq V$ are linearly dependent if there exists $a_1, ..., a_k \in \mathbb{F}$ such that

$$a_1 v_1 + ... + a_k v_k = 0,$$

and they are linearly independent otherwise. A basis of V is a maximal subset of linearly independent vectors $\{v_1, ..., v_d\}$. It can be shown that all bases of V have the same number of elements, known as the dimension of the vector space $\dim(V)$.

For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $\mathbb{F}^d$ is the canonical Euclidean space of dimension d with standard basis $\{e_1, ..., e_d\} \subseteq \mathbb{F}^d$ where $e_{i \in [d]}$ is 1 in the i-th entry and 0 elsewhere. An arbitrary vector space over $\mathbb{F}$ is always isomorphic to $\mathbb{F}^d$ for some $d \in \mathbb{N}$.

Fact 2.1.2. *For vector space V over field $\mathbb{F}$ of dimension $\dim(V) = d$, any choice of basis $\{v_1, .., v_d\}$ induces an isomorphism $V \simeq \mathbb{F}^d$ by the following bijection:*

$$(a_1, .., a_d) \in \mathbb{F}^d \qquad \longleftrightarrow \qquad \sum_{i=1}^{d} a_i v_i \in V.$$

This is injective due to the linear independence of $\{v_1, ..., v_d\}$, and surjective by the fact that it is a basis for V.

$\mathbb{F}^d$ is also equipped with the standard Euclidean inner product

$$\langle x, y \rangle := \sum_{i=1}^{d} \overline{x_i} y_i,$$

where $\bar{\cdot}$ denotes the complex conjugate (if $\mathbb{F} = \mathbb{C}$). This induces the standard Euclidean norm $\|x\|_2 := \sqrt{\langle x, x \rangle}$ on $\mathbb{F}^d$, which measures the length of $x \in \mathbb{F}^d$. By convention, vector $x \in \mathbb{C}^d$ is a column vector, and we use x^* to denote its conjugate transpose row vector

$(x^*)_i = \overline{x_i}$. Similarly, for $x \in \mathbb{R}^d$, we use either of x^*, x^T to denote its transpose row vector (as complex conjugation is trivial on $\mathbb{R}$). Therefore the standard Euclidean inner product between $x, y \in \mathbb{F}^d$ can equivalently be written as

$$\langle x, y \rangle = x^* y.$$

In general, an inner product is used to define lengths and angles in a vector space. An inner product space is a vector space with a Hermitian inner product as defined below.

Definition 2.1.3 (Inner Product)**.** *A Hermitian inner product $\langle \cdot, \cdot \rangle$ on complex vector space V satisfies*

1. *Linearity: $\langle au + bv, \cdot \rangle = a\langle u, \cdot \rangle + b\langle v, \cdot \rangle$ for any $a, b \in \mathbb{C}$ and $u, v \in V$;*
2. *Conjugate symmetry: $\langle u, v \rangle = \overline{\langle v, u \rangle}$ for any $u, v \in V$, where $\bar{\cdot}$ denotes complex conjugation;*
3. *Positive definite: $\langle u, u \rangle > 0$ for any $0 \neq u \in V$.*

If V is a real vector space, then the conjugate symmetry property reduces to symmetry. The induced norm is $\|v\|_2^2 := \langle v, v \rangle$.

A set of vectors $\{v_1, ..., v_k\} \subseteq V$ is orthogonal if $\langle v_i, v_j \rangle = 0$ for every pair $i \neq j \in [k]$, and orthonormal if further $\langle v_i, v_i \rangle = 1$ for every $i \in [k]$. We will often use orthonormal bases of vector spaces to give especially nice ismorphisms between V and $\mathbb{F}^{\dim(V)}$.

2.1.2 Linear Operators

Now that we have defined vector spaces, we can consider the maps between them.

Definition 2.1.4 (Linear Operators)**.** *For vector spaces V, W over $\mathbb{F}$, linear map/ operator/ transformation $A : V \to W$ preserves vector addition and scalar multiplication:*

$$\forall a, b \in \mathbb{F}, u, v \in V : A(au + bv) = a(Au) + b(Av).$$

$L(V, W)$ denotes the set of linear operators $A : V \to W$, and $L(V)$ is used f $V = W$.

If $V = \mathbb{F}^n$ and $W = \mathbb{F}^d$ are the canonical Euclidean vector spaces, then $L(V, W)$ can be identified with $\mathrm{Mat}_{\mathbb{F}}(d, n)$, the set of $d \times n$ matrices over field $\mathbb{F}$. In this case, composition of linear operators is identified naturally with matrix multiplication, inversion is the standard matrix inverse, and the adjoint is the conjugate transpose $(M^*)_{ij} = \overline{M_{ji}}$.

For general vector space V, the identity operator on V is denoted by I_V. For $A \in L(V, W)$ and $B \in L(U, V)$, we have $AB \in L(U, W)$, where we use AB or $A \circ B$ to denote the composition of linear operators.

Linear operator $A \in L(V)$ is called invertible if there is a solution to the equation $BA = AB = I_V$ for some $B \in L(V)$. In this case, $B = A^{-1}$ is known as the inverse of A. It can be shown that A is invertible iff A is injective ($Au = Av \iff u = v$) iff A is surjective $A(V) = V$. The set of invertible linear operators is known as the General Linear group, and is denoted by $\mathrm{GL}(V)$. We will discuss the group structure further in Section 2.2.1.

If U, V are inner product spaces with Hermitian inner products $\langle \cdot, \cdot \rangle$, the adjoint of operator $A \in L(U, V)$ is the unique operator $A^* \in L(V, U)$ satisfying

$$\langle Au, v \rangle = \langle u, A^* v \rangle$$

$\forall u \in U, v \in V$. We will sometimes use A^T to denote the adjoint if the vector spaces in question are real.

Similar to Fact 2.1.2, we can connect abstract linear operators to the familiar matrix multiplication setting.

Definition 2.1.5 (Correspondence between Linear Transformations and Matrices)**.** *Let V, W be vector spaces of dimension $\dim(V) = n$ and $\dim(W) = d$, and consider linear operator $A \in L(V, W)$. Then any choice of bases $\{\psi_1, ..., \psi_n\} \subseteq V$ and $\{\xi_1, ..., \xi_d\} \subseteq W$ induces an isomorphism $L(V, W) \simeq \mathrm{Mat}(d, n)$. Namely, for $A \in L(V, W)$ and $j \in [n]$, let $A\psi_j = \sum_{i=1}^d M_{ij}\xi_i$ be the unique representation in the basis $\{\xi_1, ..., \xi_d\}$. Then $M := \{M_{ij}\}_{i \in [d], j \in [n]} \in \mathrm{Mat}(d, n)$ is the matrix representation of A with respect to these bases.*

Writing $\Xi := \{\xi_1, ..., \xi_d\}$ and $\Psi := \{\psi_1, ..., \psi_n\}$ as the concatenation of vectors gives the following convenient notation for matrix representations:

$$M = \Xi^{-1} A \Psi. \tag{2.1}$$

Matrix representations also allow us to show that the algebra of linear operators (with composition) is isomorphic to that of matrix multiplication as follows: let $A \in L(U, V), B \in L(V, W)$ for vector spaces U, V, W, and consider choice of bases $\Xi \subseteq U, \Psi \subseteq V, \Phi \subseteq W$. Then if we M_A is the matrix representation of A with respect to (Ψ, Ξ) and M_B is the matrix representation of B with respect to (Ξ, Φ), then M_C the matrix representation of $C = BA$ with respect to (Ψ, Φ) is defined as

$$M_C = M_{BA} = \Phi^{-1}(B \circ A)\Psi = (\Phi^{-1}B\Xi)(\Xi^{-1}A\Psi) = M_B M_A. \tag{2.2}$$

$\mathrm{Mat}_{\mathbb{F}}(d, n)$ is also naturally isomorphic to $\mathbb{F}^{d\times n}$, and we use $A \in \mathrm{Mat}_{\mathbb{F}}(d, n) \to \mathrm{vec}(A) \in \mathbb{F}^{d\times n}$ to refer to this isomorphism and $(a \in \mathbb{F}^{d\times n}) \to mat(a) \in \mathrm{Mat}_{\mathbb{F}}(d, n)$ for its inverse.

We will use $\mathrm{Mat}_{\mathbb{F}}(d)$ for square matrices and $\mathrm{diag}_{\mathbb{F}}(d)$ for the subspace of diagonal matrices in $\mathrm{Mat}_{\mathbb{F}}(d)$.

2.1.3 The Spectral Theorem

The simplest linear operator is scalar multiplication applied to the vector space $\mathbb{F} = \mathbb{F}^1$. We can better understand general linear operators by attempting to decompose the action on a given vector space into these simple scalar actions.

To this end, given linear operator $A \in L(V)$ for some vector space V over field $\mathbb{F}$, a non-zero vector $v \in V$ is an eigenvector of A if there is some element $\lambda \in \mathbb{F}$ such that $Av = \lambda v$. In this case, λ is the associated eigenvalue of v, and (v, λ) are an eigen-pair of A. The spectrum of A is the (multi-)set of eigenvalues of A.

The following definition captures those operators which we can understand as a direct sum of simple scalar actions.

Definition 2.1.6 (Diagonalizable Matrix)**.** *Let V be a vector space over field $\mathbb{F}$. Then $A \in L(V)$ is diagonalizable over $\mathbb{F}$ iff V is spanned by a basis of eigenvectors of A.*

If $\Xi := \{\xi_1, ..., \xi_d\} \subseteq V$ is the basis of eigenvectors of operator $A \in L(V)$, then we say A is diagonalized by Ξ. According to Definition 2.1.5, the matrix representation of A in the Ξ basis is $M := \Xi^{-1}A\Xi$, and it can be shown that in this case, M is a diagonal matrix with the spectrum of A on the diagonal.

For any $A \in L(V)$, $A^{k\in\mathbb{N}}$ is well-defined as the repeated application of A k times. For diagonalizable operators, this can be lifted to define arbitrary functions of an operator.

Definition 2.1.7. *Consider scalar function $f : D \to \mathbb{F}$ on domain $D \subseteq \mathbb{F}$ and $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Let V be a vector space of dimension $\dim(V) = d$ and $A \in L(V)$. f can be applied to A if A is diagonalizable with eigen-pairs $\{(\lambda_i, v_i)\}_{i\in d}$ such that $\forall i \in [d] : \lambda_i \in D$. In this case, $f(A) \in L(V)$ is the unique operator with eigenpairs $\{(f(\lambda_i), v_i)\}_{i=1}^d$.*

Not every operator is diagonalizable (consider $A \in L(\mathbb{R}^2)$ defined by $Ae_1 = 0, Ae_2 = e_1$). Below, we further discuss special classes of diagonalizable operators for which more can be said about the eigenvalues and eigenvectors.

Let V be an inner product space, and consider linear transformation $A \in L(V)$. A is called normal if it satisfies $AA^* = A^*A$, and is called self-adjoint if it satisfies $A^* = A$. We distinguish between the real and complex cases, so A is called Hermitian if V is a complex vector space, and A is called symmetric if V is a real vector space. In these cases, we use $\mathrm{H}(V)$ to denote the set of Hermitian operators and $\mathrm{S}(V)$ to denote the set of symmetric operators. If $V = \mathbb{F}^d$, then this corresponds to the set of Hermitian and symmetric matrices, respectively.

The eigenvectors of these classes of operators have some additional structure.

Theorem 2.1.8 (Spectral Theorem 7.9 and 7.13 in [7])**.** *Let V be a complex vector space with Hermitian inner product $\langle \cdot, \cdot \rangle$. Then linear transformation $A \in L(V)$ is diagonalized by an orthonormal basis of eigenvectors over $\mathbb{C}$ iff A is normal ($AA^* = A^*A$).*

If V is a real vector space with symmetric inner product $\langle \cdot, \cdot \rangle$, then $A \in L(V)$ is diagonalized by an orthogonal basis of eigenvectors over $\mathbb{R}$ iff $A \in \mathrm{S}(V)$, i.e. $A^T = A$.

In both cases, the spectrum of A is real iff A is self-adjoint.

Theorem 2.1.8 extends to normal operators that are not symmetric in a real inner product space, but we do not need this more complicated block decomposition.

2.1.4 Trace and Determinant

The trace of a matrix $M \in \mathrm{Mat}(d)$, denoted by $\mathrm{Tr}(M)$, is defined as the sum of the diagonal entries of M. It can be verified that the trace satisfies cyclic property: $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for A and B with appropriate sizes. The trace of a linear operator $A \in L(V)$ is defined as the trace of the matrix representation of A with respect to any basis $\{\xi_1, ..., \xi_d\} \subseteq V$. It can be shown (using the cyclic property) that the trace does not depend on the choice of basis, so this is well-defined. Therefore, if A is diagonalizable, choosing $\Xi = \{\xi_1, ..., \xi_d\}$ to be the basis of eigenvectors with $\{\lambda_1, ..., \lambda_d\}$ the corresponding eigenvalues, we have

$$\mathrm{Tr}[A] = \mathrm{Tr}[\Xi^{-1} A \Xi] = \sum_{i=1}^{d} \lambda_i.$$

The determinant of $M \in \mathrm{Mat}(d)$ is the following polynomial function of its entries:

$$\det(M) := \sum_{\sigma \in \mathcal{S}_d} (-1)^{|\sigma|} \prod_{i=1}^{d} M_{i,\sigma(i)},$$

where $\mathcal{S}_d$ is the set of all permutations on $[d]$, and $|\sigma|$ denotes the parity or signature of a permutation. It can be shown that the determinant is multiplicative $\det(AB) = \det(A)\det(B)$, and this allows us to define the determinant of linear operator $A \in L(V)$ as the determinant of the matrix representation of A with respect to any basis Ξ. For diagonalizable A with $\Xi = \{\xi_1, ..., \xi_d\}$ the basis of eigenvectors and $\{\lambda_1, ..., \lambda_d\}$ the corresponding eigenvalues, we have

$$\det(A) = \det(\Xi^{-1} A \Xi) = \prod_{i=1}^{d} \lambda_i$$

as the terms corresponding to any other permutation vanishes.

In general, $A \in L(V)$ may be diagonalizable over $\mathbb{C}$ even if it is not diagonalizable over $\mathbb{R}$. In this case, the formulas for trace and determinant continue to hold with respect to these complex eigenvalues.

2.1.5 Positive Operators

In this subsection, we consider certain special subsets of Hermitian operators which will be the core of our optimization framework in Chapter 6.

Definition 2.1.9 (Positive Operators)**.** *For inner product space V, transformation $A \in L(V)$ is positive semi-definite if it is self-adjoint and*

$$\forall x \in V : \langle x, Ax \rangle \geq 0.$$

It is positive definite if the inequality is strict for all $x \neq 0$.

The set of all positive definite operators is denoted by $\mathrm{PD}(V)$. We use $A \succeq 0$ to mean that A is positive semi-definite, and $A \succ 0$ to mean that A is positive definite. This definition also defines a partial order on self-adjoint operators as follows: $A \succeq B$ and $A \succ B$ means $A - B$ is positive semi-definite or positive definite respectively.

Our optimization framework in Chapter 6 will involve the following subsets of $\mathrm{PD}(V)$.

Definition 2.1.10. *For inner product space V over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$,* $\mathrm{SPD}(V)$ *denotes the subset of* $\mathrm{PD}(V)$ *with unit determinant. Its corresponding algebra (discussed further in Section 2.2.3) is defined as $\mathfrak{spd}(V) := \log \mathrm{SPD}(V)$, or explicitly as*

$$\mathfrak{spd}_{\mathbb{C}}(V) := \{X \in \mathrm{H}(V) \mid \mathrm{Tr}[X] = 0\} \qquad \mathfrak{spd}_{\mathbb{R}}(V) := \{X \in \mathrm{S}(V) \mid \mathrm{Tr}[X] = 0\}.$$

2.1.6 Isometries

In a general metric space, an isometry is a transformation that preserves the metric. In our setting involving linear operators on inner product spaces, the metric we consider is the Euclidean metric derived from the inner product $\|u - v\|_2 := \sqrt{\langle u - v, u - v\rangle}$.

Definition 2.1.11. *For inner product space V, $\Xi \in L(V)$ is an isometry if it preserves the inner product:*

$$\forall x, y \in V : \langle \Xi x, \Xi y\rangle = \langle x, y\rangle.$$

Equivalently, this can be defined by the equation

$$\Xi\Xi^* = \Xi^*\Xi = I_V.$$

Complex isometries are called unitary transformations and are denoted by $\mathrm{U}(V)$*, and real isometries are called orthogonal transformations and are denoted by* $\mathrm{O}(V)$*.*

If $V = \mathbb{F}^d$, we will use $\mathrm{U}(d)$ and $\mathrm{O}(d)$ as shorthand. Note that the equation $\Xi^*\Xi = I_d$ implies that the columns of Ξ form an orthonormal basis. This implies that isometries capture length-preserving change of basis operations. Further, with this notation, Theorem 2.1.8 characterizes normal and self-adjoint operators as those with real spectrum which can be diagonalized by isometries:

$$\mathrm{H}(V) = \{\Xi \operatorname{diag}_{\mathbb{R}}(d)\Xi^* \mid \Xi \in \mathrm{U}(V)\}, \qquad \mathrm{S}(d) = \{\Xi \operatorname{diag}_{\mathbb{R}}(d)\Xi^* \mid \Xi \in \mathrm{O}(V)\}.$$

Since isometries are clearly normal (as $\Xi\Xi^* = \Xi^*\Xi = I_V$), Theorem 2.1.8 implies that they have an orthonormal basis of eigenvectors. It can further be shown that the spectrum of an isometry is contained in $S^1 := \{\lambda \in \mathbb{C} \mid |\lambda| = 1\}$. Note that in this case, both the real and complex isometries are diagonalizable over the complex field.

2.1.7 Projections

In Chapter 5, we will need to define noise distributions satisfying certain linear constraints. This is accomplished by using the following notion of orthogonal projections, which are a special class of positive semi-definite operators.

Definition 2.1.12 (Projection)**.** *A projection on vector space V is a transformation $P \in L(V)$ such that $P^2 = P$. An orthogonal projection on inner product space V is a projection that is also self-adjoint ($P^* = P$).*

By the Spectral Theorem (Theorem 2.1.8), any self-adjoint orthogonal projection P has an orthonormal basis of eigenvectors. Further, the equation $P^2 = P$ can be used to show that its spectrum is contained in $\{0, 1\}$. Therefore, if $\Xi := \{\xi_1, ..., \xi_k\} \subseteq V$ is any orthonormal basis of the range of P, then the projection can be written as $P = \Xi\Xi^*$.

Given arbitrary $A \in \mathrm{Mat}(d, n)$, the unique orthogonal projection onto the column space of A is defined by

$$A(A^*A)^{-1}A^*. \tag{2.3}$$

If A^*A is not invertible, then we use the pseudoinverse, which is the inverse on the image of AA^* and 0 on the null space. We will mostly deal with the invertible case, so by abuse of notation, we use $()^{-1}$ for both.

2.1.8 The Polar Decomposition

The goal of this subsection is to present the polar decomposition from matrix analysis. This will be useful for our group optimization framework in Chapter 6.

Theorem 2.1.13. *For complex inner product space V, any $A \in \mathrm{GL}_{\mathbb{C}}(V)$ can be uniquely written $A = UP$ for unitary $U \in \mathrm{U}(V)$ and positive definite $P \in \mathrm{PD}(V)$. If V is a real inner product space, then any $A \in \mathrm{GL}_{\mathbb{R}}(V)$ can be written uniquely as $A = OP$ for orthogonal $O \in \mathrm{O}(V)$ and positive definite $P \in \mathrm{PD}(V)$. Further, $P = |A| = \sqrt{A^*A}$ is the unique positive definite square root.*

This statement is the simplest case of the Cartan decomposition [97] from Lie group theory that we discuss further in Section 2.2.3, though we will not need this level of generality for our applications. In Chapter 6, we will lift this polar decomposition to direct sums of certain matrix groups.

2.1.9 Norms and Inequalities

In this subsection, we present some standard norms on vector spaces and matrices, as well as certain inequalities between these norms.

Definition 2.1.14. *For $x \in \mathbb{F}^d$ with $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, then p-norm of x for $p \geq 1$ is defined as*

$$\|x\|_p := \Big(\sum_{i=1}^{d} |x_i|^p\Big)^{1/p}.$$

Note that $\|\cdot\|_2$ is the standard Euclidean norm, i.e. $\|x\|_2^2 = \langle x, x\rangle$ for the standard Euclidean inner product defined above. Further we denote $\|x\|_\infty := \max_{i\in[d]} |x_i|$.

We will use $B_2^d := \{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1\}$ to denote the Euclidean ball, and $S^{d-1} := \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$ for the Euclidean sphere in d-dimensions.

The following well-known result gives duality relations between L_p norms.

Proposition 2.1.15. *Hölder's inequality [48] states that for $x, y \in \mathbb{F}^d$ for $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and any $p \geq 1$,*

$$\langle x, y\rangle \leq \|x\|_p \|y\|_q,$$

where q is the Hölder conjugate exponent satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Note that this generalizes the Cauchy-Schwarz inequality by taking $p = q = 2$.

As a consequence, the p-norms can also be described as (Chapter 2 of [17]).

$$\|x\|_p = \sup_{\|y\|_q \leq 1} \langle y, x\rangle.$$

Similarly, we can define the analogous notion of p-norms for operators. We first present some well-motivated special cases that will be used throughout.

The Frobenius inner product on $\mathrm{Mat}(d, n)$ can be seen as the standard Euclidean inner product on $\mathrm{Mat}(d, n)$: for $A, B \in \mathrm{Mat}(d, n)$,

$$\langle A, B\rangle := \sum_{i=1}^{d} \sum_{j=1}^{n} \overline{A_{ij}} B_{ij} = \langle \mathrm{vec}(A), \mathrm{vec}(B)\rangle$$

where in the first term we are defining the Frobenius inner product, and in the last term we are using the standard Euclidean inner product on $\mathbb{F}^{dn}$. For inner product spaces V, W, the Frobenius inner product on abitrary operators $A, B \in L(V, W)$ is defined as

$$\langle A, B\rangle = \mathrm{Tr}[A^* B].$$

It can be shown that this specializes to the definition above by considering an arbitrary matrix representation of A and B. This inner product naturally induces the Frobenius norm $\|A\|_F^2 := \langle A, A\rangle$.

Another important norm on $L(V, W)$ is the operator norm

$$\|A\|_{\mathrm{op}} := \sup_{v\in V} \frac{\|Av\|_2}{\|v\|_2}, \tag{2.4}$$

where the norms in the numerator and denominator denote the norms induced by the inner products on W and V respectively. More generally, any choice of norms $\|\cdot\|_V$ and $\|\cdot\|_W$ on vector spaces V and W produce the induced operator norm defined on $L(V, W)$ as

$$\|A\|_{V\to W} := \sup_{v\in V} \frac{\|Av\|_W}{\|v\|_V}.$$

Note that $\|\cdot\|_{\mathrm{op}}$ is therefore the operator norm induced by the standard Euclidean norm $\|\cdot\|_2$. By Proposition 2.1.15, we can rewrite this as

$$\|A\|_{\mathrm{op}} := \sup_{w\in W} \sup_{v\in V} \frac{\langle w, Av\rangle}{\|w\|_2\|v\|_2}.$$

Further, it can be shown that if $A \in \mathrm{H}(V)$ or $A \in \mathrm{S}(V)$, then we can restrict the above variational formula to

$$\|A\|_{\mathrm{op}} := \sup_{v\in V} \frac{\langle v, Av\rangle}{\|v\|_2^2}. \tag{2.5}$$

The following definition generalizes $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$, just as the L_p norms on $\mathbb{F}^d$ generalize $\|\cdot\|_2$ and $\|\cdot\|_\infty$.

Definition 2.1.16. *For inner product spaces V, W and $p \in \mathbb{N}$, the p-Schatten norm on $A \in L(V, W)$ is defined*

$$\|A\|_{S_p} = \|A\|_p := (\mathrm{Tr}(A^*A)^{p/2})^{1/p}.$$

This can be extended to to arbitrary $p \geq 1$ by considering the eigendecomposition $A^*A = \sum_{i=1}^{\dim(V)} \lambda_i v_i v_i^*$ according to Theorem 2.1.8 and defining

$$\|A\|_{S_p} = \|A\|_p := \left(\sum_{i=1}^{\dim(V)} \lambda_i^{p/2}\right)^{1/p}.$$

It can be shown that the Frobenius norm $\|\cdot\|_{S_2} = \|\cdot\|_F$ and the operator norm $\|\cdot\|_{S_\infty} = \|\cdot\|_{\mathrm{op}}$ are special cases. This leads to the following operator version of Proposition 2.1.15.

Proposition 2.1.17. *For inner product spaces V, W and $p \in \mathbb{N}$, the p-Schatten norm on $A \in L(V, W)$ is equivalently defined*

$$\|A\|_{S_p} = \sup_{\|B\|_{S_q} \leq 1} \langle A, B \rangle,$$

where q is the Holder conjugate $\frac{1}{p} + \frac{1}{q} = 1$.

As a consequence, the following generalizes the Cauchy-Schwarz inequality to operators:

$$\langle A, B \rangle \leq \|A\|_{S_p} \|B\|_{S_q}.$$

In Section 9.3.1, we will need the following multi-argument generalization of the above inequality for our trace method argument.

Theorem 2.1.18 (Section 1.2 in [90])**.** *For inner product space V and $p \in \mathbb{N}$ with $p \geq 1$, if $A_1, ..., A_p \in L(V)$, then*

$$\left| \mathrm{Tr}\left[\prod_{i=1}^{p} A_i\right] \right| \leq \prod_{i=1}^{p} \|A_i\|_{S_p}.$$

We point out that the above inequality applies only when the number of arguments coincides with the choice of Schatten norm p.

Finally, we present the following extension of the Riesz-Thorin theorem from functional analysis [102] which allows us to interpolate between Schatten norms.

Theorem 2.1.19 (Corollary 3.1 of [61])**.** *Let $\Phi : \mathrm{H}(m) \to \mathrm{H}(n)$ or $\Phi : \mathrm{S}(m) \to \mathrm{S}(n)$ be a linear operator between two spaces of finite dimensional operators such that, for given $p, q \in [1, \infty]$, the operator norms $\|\Phi\|_{p \to p}$ and $\|\Phi\|_{q \to q}$ induced by the Schatten norms S_p and S_q (Definition 2.1.16) are bounded. For any $\theta \in [0, 1]$ and p_θ defined to satisfy $\frac{1}{p_\theta} := \frac{1-\theta}{p} + \frac{\theta}{q}$, the linear operator Φ satisfies*

$$\|\Phi\|_{p_\theta \to p_\theta} \leq \|\Phi\|_{p \to p}^{1-\theta} \|\Phi\|_{q \to q}^{\theta}.$$

This gives us a way to reduce the computation of an entire interval of induced operator norms to just the endpoints of that interval.

2.2 Linear Algebraic Groups and Structure

In this section, we revisit the linear algebraic concepts presented in Section 2.1 from the perspective of Lie group theory. This provides the foundation for our geodesic convex optimization framework for tensor scaling presented in Chapter 6.

2.2.1 Classical Groups

In this subsection, we define some important subsets of operators and consider their group structure. This will be expanded upon in Section 2.2.3, where we will describe Lie groups and their properties.

Recall that $A \in L(V)$ is invertible iff there exists $A^{-1} \in L(V)$ satisfying $A^{-1}A = AA^{-1} = I_V$. The set of invertible operators on vector space V forms a group under composition as shown below.

Definition 2.2.1 (General Linear Group/ Special Linear Group)**.** *The subset of invertible linear operators on V, along with composition as the group operation, is known as the General Linear Group and is denoted by* $\mathrm{GL}(V) \subseteq L(V)$*. The Special Linear Group* $\mathrm{SL}(V) \subseteq \mathrm{GL}(V)$ *is the subgroup of unit determinant operators.*

The prefix S for “special” will henceforth stand for the unit determinant constraint.

For $\dim(V) = d$, $\mathrm{GL}(V)$ is isomorphic to the group of invertible matrices $GL_{\mathbb{F}}(d)$ with matrix multiplication. The isomorphism is given by Eq. (2.2) using the matrix representations in any choice of basis $\{\xi_1, ..., \xi_d\} \subseteq V$. If $V = \mathbb{F}^d$, then we use shorthand $\mathrm{GL}(d)$ and $\mathrm{SL}(d)$ to refer to the above groups.

We can also verify that the isometries described in Definition 2.1.11 are also groups under compositions as, for any $\Xi, \Psi \in \mathrm{U}(V)$,

$$(\Xi\Psi)^*(\Xi\Psi) = \Xi^*(\Psi^*\Psi)\Xi = \Xi^*\Xi = I_V,$$

so that $\Xi \circ \Psi \in \mathrm{U}(V)$ as well. Similar calculations show that $\mathrm{O}(V)$ is also a group. Further, we will often consider the unit determinant subgroups $\mathrm{SU}(V)$ and $\mathrm{SO}(V)$, which are known as the special unitary and special orthogonal groups respectively. These are examples of compact Lie groups, and we discuss this more formally in Section 2.2.3.

While there is not a well-defined group structure for positive definite operators (e.g. $A, B \in \mathrm{PD}(V)$ does not imply $AB \in \mathrm{PD}(V)$), the polar decomposition in Theorem 2.1.13

shows that we can view $\mathrm{PD}(V)$ as the set of equivalence classes of $\mathrm{GL}(V)$ under the action of $\mathrm{U}(V)$. This induced structure motivates the geometry we define in Section 2.2.4, which is very useful for our optimization perspective in Chapter 6.

2.2.2 Torus Groups

In this subsection, we consider some very simple (diagonal) subgroups of $\mathrm{GL}(V)$. These are known as torus subgroups, and are the simplest setting in which we can present some of the ideas of Lie theory that we further elaborate in Section 2.2.3.

Definition 2.2.2 (Torus Groups). *For vector space V and basis $\Xi := \{\xi_1, ..., \xi_d\} \subseteq V$, the set of invertible operators diagonal in the Ξ basis is denoted by $T^\Xi(V)$ and forms a commutative group under composition.* $\mathrm{ST}^\Xi(V) \subseteq T^\Xi(V)$ *is the unit determinant subgroup.*

Note we can simply verify commutativity as follows: for every $A \in T^\Xi(V)$, the matrix representations $\Xi^{-1} A \Xi$ is diagonal as discussed in Definition 2.1.6. In fact, this subgroup is isomorphic to $\mathbb{C}_\times^d$, the direct product of the multiplicative group of complex numbers via the map $A \to \{\lambda_1, ..., \lambda_d\}$ for eigen-pairs $\{(\lambda_i, \xi_i)\}_{i \in [d]}$ of A.

To better understand $T^\Xi(V)$, we consider the simplest setting of $\dim(V) = 1$. In this case, $T^\Xi(V) \simeq \mathbb{C}_\times$, and we decompose any $z \in \mathbb{C}_\times$ into the magnitude and phase $z = \lambda x$ where $x > 0$ and $|\lambda| = 1$. This induces the group decomposition $T^\Xi(V) = T_c^\Xi(V) \times T_+^\Xi(V)$ for general torus groups by applying this component-wise. Note that both of these pieces also have a group structure with the same operation. Further, $T_c^\Xi(V)$ is an example of a compact group, which will be discussed further in Section 2.2.3. We note that the name "torus group" comes from the special case $(S^1)^2$, the geometric torus.

This decomposition $T^\Xi(V) = T_c^\Xi(V) \times T_+^\Xi(V)$ is reminiscent of the Polar decomposition in Theorem 2.1.13, where we decomposed $\mathrm{GL}(V) = \mathrm{U}(V) \times \mathrm{PD}(V)$ into a rotation and a positive component. In fact, when Ξ is an orthonormal basis, this is exactly the restriction of the polar decomposition to the commutative subgroup $T^\Xi(V)$. The following change of variable allows us to further clarify the algebraic structure of $T^\Xi(V)$. Consider the vector space $\mathfrak{t}_+^\Xi := \log T_+^\Xi(V)$, which is isomorphic to the additive group $\mathbb{R}^{\dim(V)}$. Similarly, we consider $\log T_c^\Xi(V) = \sqrt{-1}\mathfrak{t}_+^\Xi$, which is known as the Lie algebra of the torus group $T_c^\Xi(V)$. This gives the decomposition $\log T^\Xi(V) = \sqrt{-1}\mathfrak{t}_+^\Xi(V) \oplus \mathfrak{t}_+^\Xi(V)$, and by commutativity, the exponential mapping gives an isomorphism between $T^\Xi(V)$ and the additive group $\sqrt{-1}\mathfrak{t}_+^\Xi(V) \oplus \mathfrak{t}_+^\Xi(V)$. We will also often consider the unit determinant subsets of these groups, which are denoted by $\mathrm{ST}_\mathbb{C}^\Xi, \mathrm{ST}^\Xi, \mathrm{ST}_+^\Xi$. Their corresponding algebras are denoted by $\mathfrak{st}^{\Xi}, \mathfrak{st}^{\Xi}_{c}, \mathfrak{st}^{\Xi}_{+}$, where we have added a trace constraint, e.g.

$$\mathfrak{st}^{\Xi}_{+} = \{\Xi\Lambda\Xi^* \mid \Lambda \in \mathrm{diag}_{\mathbb{R}}(V), \mathrm{Tr}[\Lambda] = 0\}.$$

As we will show in Section 2.2.3, the exponential mapping is also useful in understanding the polar decomposition for general operators. But in this case the group structure will be slightly more complicated.

2.2.3 Lie Groups and Lie Algebras (Primer)

In this subsection, reconsider the various groups and decompositions discussed above using (a bit of) Lie theory. The results in this subsection are not necessary to understand the work in this thesis, and we present this as background to give some intuition for the geometry discussed in Section 2.2.4. We suggest the reader consult the books [97], [82], [53] (and many others) for a more thorough treatment.

The various optimization problems we study in this thesis concern optimization over Lie groups. These are continuous groups of symmetries which, at every point, locally look like a vector space. The groups of matrices and operators studied above $(\mathrm{GL}, \mathrm{SL}, \mathrm{U}, ...)$ are classical examples of Lie groups, and the corresponding local vector spaces are called Lie algebras $(\mathfrak{gl}, \mathfrak{sl}, \mathfrak{u} = \mathrm{H}, ...)$.

Definition 2.2.3. *A Lie group is a group that is also a smooth manifold. The Lie algebra* $\mathfrak{g} := \mathrm{Lie}(G)$ *of Lie group* G *is the tangent space at the identity* T_eG*.*

We will not formally explain all of the terms in the above definition, instead choosing to focus on some intuitive examples. A smooth manifold is a topological space that, at every point, can be locally and smoothly transformed to a vector space. The vector space associated with the point is called the tangent space, and describes the directions that are "tangent" to the manifold. For concreteness, consider that the Earth we walk on is the surface of a sphere in three dimensional space, where at every point the local neighborhood looks like the Euclidean plane.

Now consider the concrete example $G = \mathrm{U}(d)$. The tangent space at the identity is the vector space of matrices $X \in \mathrm{Mat}(d)$ such that $I_d + tX$ is still unitary for some small (infinitesimal) t. Since the unitary group is characterized by the equation $UU^* = U^*U = I_d$ according to Definition 2.1.11, we can differentiate this to find

$$0 = \partial_{t=0}(I_d + tX)(I_d + tX)^* = X + X^*.$$

This is exactly the set of skew-Hermitian operators, and is known as the Lie algebra of the unitary group, denoted $\mathfrak{u}(d)$. Note that differentiating the constraint $U^*U = I$ at the identity also gives the same equation. We point out that $\mathfrak{u}(d)$ is equivalent to $\sqrt{-1}$ times the set of Hermitian operators $\mathrm{H}(d)$, and we will use this correspondence later to formally describe the polar decomposition.

Lie groups are manifolds such that the group structure is compatible with the set of tangent spaces. Explicitly, if $T_U G$ denotes the tangent space of the unitary group at element U, then we have the following bijection between $T_U G$ and $T_{I_d} G = \mathfrak{u}$:

$$X \in \mathfrak{u} \iff \partial_{t=0}(U + tUX)(U + tUX)^* = \partial_{t=0} U(I + tX)(I + tX)^* U^* = 0.$$

In other words, the tangent spaces are all compatible with the group operation.

In this way, if we consider all of the tangent vectors indexed by a fixed $X \in \mathfrak{u}$, these can be stitched together into a left invariant vector field on the group. This invariance makes the setting of Lie groups especially suited to optimization, as the local structure at every point always looks like the canonical local structure at the identity, and further there is a change of variable operation that reduces this local structure to a vector space.

The flow induced by these left-invariant vector fields is captured by what is known as the exponential map. For our unitary group example, this gives a mapping from the algebra to the group. Any $U \in \mathrm{U}(V)$ is normal and has spectrum contained in S^1 as shown in Section 2.1.6. This means that we can write $U = e^{\sqrt{-1}Y}$ where $Y \in \mathrm{H}(V)$. By this calculation, we have shown that $\mathrm{U}(V) = \exp(\mathfrak{u}(V)) = \exp(\sqrt{-1}\,\mathrm{H}(V))$, and similarly, $\mathrm{O}(V) = \exp(\mathfrak{o}(V)) = \exp(\sqrt{-1}\,\mathrm{S}(V))$.

This global structure is only defined for compact Lie groups, of which $\mathrm{U}(d)$ is an example. On the other hand, for non-compact Lie groups such as $\mathrm{GL}(d)$, there does not always exist a global exponential map which can travel all throughout the group. But in this case, we can use the polar decomposition to describe $\mathrm{GL}(d)$ as the complexification of $\mathrm{U}(d)$. Explicitly, the valid directions at the identity that remain in GL, i.e. the Lie algebra $\mathfrak{gl}$, are exactly $\mathfrak{u} \oplus \sqrt{-1}\mathfrak{u}$. This can be likened to the very simple example of the polar decomposition of the multiplicative group of complex numbers $\mathbb{C}_*$, which can be seen as $\exp(\mathbb{R} \oplus \sqrt{-1}\mathbb{R})$, or the equally simple polar decomposition of torus groups given in Section 2.2.2.

We can also combine the polar decomposition with the exponential map to lift this vector space complexification to the group setting. According to Definition 2.1.9, any $P \in \mathrm{PD}(V)$ is Hermitian, so by Theorem 2.1.8 it can be diagonalized by some isometry Ξ according to Definition 2.1.11. Further, it has strictly positive eigenvalues, so we can write $P = e^X$ for some $X \in \mathrm{H}(d)$ (explicitly, $X := \log P$ by Definition 2.1.7). In fact, the log function is a bijection on strictly positive real numbers, so this shows that $\mathrm{PD}_{\mathbb{C}}(V) = \exp(\mathrm{H}(V))$ and $\mathrm{PD}_{\mathbb{R}}(V) = \exp(\mathrm{S}(V))$.

Using these transformations, we can view the Polar decomposition in Theorem 2.1.13 as a way to write any invertible $A \in \mathrm{GL}(V)$ as the product of two exponentials $A = e^{\sqrt{-1}\cdot Y} e^X$ where $X, Y \in \mathrm{H}(V)$. Earlier, we showed that $\mathrm{U}(V) = \exp(\mathfrak{u}(V))$, for Lie algebra $\mathfrak{u}(V)$. The polar decomposition allows us to lift this to the non-compact group $\mathrm{GL}(V) = \exp(\mathfrak{u}(V)) \cdot \exp(\sqrt{-1} \cdot \mathfrak{u}(V))$, where the Lie algebra is $\mathfrak{gl}(V) = \mathfrak{u}(V) \oplus \sqrt{-1} \cdot \mathfrak{u}(V)$.

This Lie algebra and exponential map structure will be key to the tractability of the group optimization problems studied in Chapter 6. We will use concepts from geometric invariant theory in order to give geodesically convex formulations for the scaling problems we study in this thesis. For background, see the foundational book by Mumford et al. [73], and for a slightly more concrete perspective, see the book of Wallach [97].

Before concluding, we use this perspective of Lie algebras to decompose

$$\mathrm{SPD}_{\mathbb{C}}(V) = \cup_{\Xi \in \mathrm{U}(V)} \mathrm{ST}_+^{\Xi}(V), \qquad \mathrm{SPD}_{\mathbb{R}}(V) = \cup_{\Xi \in \mathrm{O}(V)} \mathrm{ST}_+^{\Xi}(V), \tag{2.6}$$

as well as a similar statement for associated vector spaces as

$$\begin{aligned} \mathfrak{spd}_{\mathbb{C}}(V) &:= \{X \in \mathrm{H}(V) \mid \mathrm{Tr}[X] = 0\} = \cup_{\Xi \in \mathrm{SU}(V)} \mathfrak{st}_+^{\Xi}(V), \\ \mathfrak{spd}_{\mathbb{R}}(V) &:= \{X \in \mathrm{S}(V) \mid \mathrm{Tr}[X] = 0\} = \cup_{\Xi \in \mathrm{SO}(V)} \mathfrak{st}_+^{\Xi}(V), \end{aligned} \tag{2.7}$$

where SPD and $\mathfrak{spd}$ are given in Definition 2.1.10. It will be quite easy to optimize over any single torus group $\mathfrak{st}_+^{\Xi}(V)$, since it is just a vector space of diagonal matrices. Therefore, the above decomposition allows us to reduce optimization on positive definite operators, to vector space optimization. This will be key to the appropriate notion of geodesic convexity that we exploit in Chapter 6 and Chapter 7 in order to analyze tensor scaling problems, which can be thought of as search problems over non-commutative groups.

2.2.4 Calculus for Positive Definite Operators

This subsection will introduce the notion of geodesics on positive definite operators. These geodesics induce a natural geometry on $\mathrm{PD}(d)$ that reveals the underlying convexity of our group optimization problem in Section 6.2.4.

Definition 2.2.4 (Geodesics on Positive Definite operators)**.** *For any two elements $p, q \in \mathrm{PD}(V)$, the curve $\gamma_{p,q} : [0,1] \to \mathrm{PD}(V)$ is defined as*

$$\gamma_{p,q}(t) := p^{1/2}(p^{-1/2} q p^{-1/2})^t p^{1/2} = p^{1/2} \exp(t \log(p^{-1/2} q p^{-1/2})) p^{1/2}.$$

Similary, for $p \in PD(V)$ and $X \in \mathrm{H}(V)$, the map $\gamma_p : \mathrm{H} \to \mathrm{PD}(V)$ is defined as

$$\gamma_p(X) := p^{1/2} e^X p^{1/2}.$$

We will also need the following well-known symmetry properties of these curves.

Fact 2.2.5. *For any two elements $p, q \in \mathrm{PD}(V)$, the curves $\gamma_{p,q}$ and $\gamma_{q,p}$ are related by $\gamma_{p,q}(t) = \gamma_{q,p}(1-t)$ for any $t \in [0,1]$.*

Further, $\| \log p^{-1/2} q p^{-1/2} \| = \| \log q^{-1/2} p q^{-1/2} \|$ for any unitarily invariant norm $\| \cdot \|$.

Proof. The first statement is a well-known fact relating to the so-called weighted geometric means of matrices, and is stated in Lemma 2 of [33].

To show the second statement, we observe that $p^{-1/2} q p^{-1/2} = (p^{-1/2} q^{1/2})(q^{1/2} p^{-1/2})$ is cospectral with $(q^{-1/2} p q^{-1/2})^{-1} = q^{1/2} p^{-1} q^{1/2} = (q^{1/2} p^{-1/2})(p^{-1/2} q^{1/2})$, and therefore, so are their matrix logarithms. Since any unitarily invariant norm depends only on the eigenvalues of the input, this implies that $\| \cdot \|$ takes the same value on $\log(p^{-1/2} q p^{-1/2})$ and $-\log(q^{-1/2} p q^{-1/2})$. □

These curves can be formally derived as geodesics, or shortest path curves, on the Riemannian manifold of positive definite matrices. For a wonderful introduction to the subject see the book by Bhatia [13]. The important thing to note for our purposes is that there is a curve connecting any $p \in \mathrm{PD}(V)$ to any other $q \in \mathrm{PD}(V)$. Further, for any starting point $p \in \mathrm{PD}(V)$, we can parametrize the non-Euclidean set of positive definite matrices by the vector space $\mathrm{H}(V)$ using the exponential curves γ_p.

Another simple derivation of these curves comes from the Lie group structure given in Section 2.2.3. Recall that the Lie algebra of $\mathrm{GL}(V)$ was the vector space $\sqrt{-1}\,\mathrm{H}(V) \oplus \mathrm{H}(V)$, and Theorem 2.1.13 gave a way to write $\mathrm{GL}(V) = \mathrm{U}(V) \cdot \mathrm{PD}(V) = \exp(\sqrt{-1}\,\mathrm{H}(V)) \cdot \exp(\mathrm{H}(V))$. Also, the polar part of $g \in \mathrm{GL}(V)$ is exactly $\sqrt{g^* g}$. Therefore, for any $X \in \mathrm{H}(V)$, the geodesic curve from $p := g^* g \in \mathrm{PD}(V)$ in the X direction is related to the induced curve from $g \in \mathrm{GL}(V)$ in the tangent direction X:

$$(e^{\eta X/2} g)^* (e^{\eta X/2} g) = g^* e^{\eta X} g. \tag{2.8}$$

Note the factor of 2 since the polar part is really the square root of $g^* g$. This is not exactly $\gamma_p(\eta X) = p^{1/2} e^{\eta X} p^{1/2}$ as $p = g^* g$ does not imply $p^{1/2} = g$. But if we consider $\mathrm{PD}(V)$ as the set of equivalence classes of operators in $\mathrm{GL}(V)$ parametrized by their polar part, then this relation becomes more natural.

In Chapter 6 and Chapter 7, we will lift these geodesics to products of simple groups. This will allow us to give an geodesically convex optimization formulation for certain group scaling problems.

2.3 Convex Analysis

In this section, we review some basic definitions and results from convex analysis. We first present the univariate setting, and then discuss its extension to vector spaces. We follow the presentation in [17] and [75]. In Section 6.2.3, we will lift these ideas to the notion of geodesic convexity on positive definite matrices. This will be the key to our analysis of tensor scaling in Chapter 6.

2.3.1 Univariate Convex Functions

Convexity is a natural and very useful property of functions used throughout mathematics.

Definition 2.3.1. *Function $h : \mathbb{R} \to \mathbb{R}$ is convex if any of the following conditions hold:*

1. *(0-th order): $\forall s, t \in \mathbb{R}$ and $\lambda \in [0, 1]$: $h(\lambda s + (1 - \lambda)t) \leq \lambda h(s) + (1 - \lambda)h(t)$;*
2. *(1-st order): if h is differentiable, then $\forall s, t \in \mathbb{R} : h(t) - h(s) \geq h'(s)(t - s)$;*
3. *(2-nd order): if h is twice-differentiable, then $\forall t \in \mathbb{R} : h''(t) \geq 0$.*

We will also often use the following notions of strong convexity at a point.

Definition 2.3.2. *(Twice-differentiable) function $h : \mathbb{R} \to \mathbb{R}$ is α-strongly convex at $s \in \mathbb{R}$ if $h''(s) \geq \alpha$. Equivalently, h is α-strongly convex on the interval $[a, b]$ iff*

$$\forall s, t \in [a, b] : h(t) - h(s) \geq h'(s)(t - s) + \frac{\alpha}{2}(t - s)^2.$$

h is called strictly convex at $s \in \mathbb{R}$ if it is α-strongly convex at s for some $\alpha > 0$.

Most of the functions studied in this thesis are sufficiently differentiable that we can apply any of the above equivalent conditions. Some of the proofs below will use second or third derivatives, and so in the sequel we assume that the input function is thrice-differentiable. The differentiability assumption can often be removed by different arguments, but we choose this presentation as it lifts more naturally to our geodesic setting in Chapter 6.

The following shows the value of convexity, especially for optimization.

Definition 2.3.3. *For function $h : \mathbb{R} \to \mathbb{R}$, $t \in \mathbb{R}$ is a critical point of h if $h'(t) = 0$.*

Lemma 2.3.4. *For convex function $h : \mathbb{R} \to \mathbb{R}$, if $h^* := \inf_{t \in \mathbb{R}} h(t)$, then $h^* = h(s)$ iff $s \in \mathbb{R}$ is a critical point of h.*

Proof. If $h^* = h(s)$, then clearly s is a critical point, as otherwise we could decrease the function by moving to $s - \delta h'(s)$ for some small $\delta > 0$.

The other direction follows simply from the 1-st order condition in Definition 2.3.1, as for any $t \in \mathbb{R}$ we can lower bound

$$h(t) \geq h(s) + h'(s)(t - s) = h(s).$$

□

In fact, if the function is strongly convex, the optimizer is unique as shown below.

Lemma 2.3.5. *Let $h : \mathbb{R} \to \mathbb{R}$ be a convex function that is $\alpha > 0$-strongly convex at the optimizer $t_* = \arg\inf_{t \in \mathbb{R}} h(t)$. Then t_* is the unique optimizer of h.*

Proof. By Lemma 2.3.4, $h'(t_*) = 0$ by optimality. Further, by continuity of the second derivative, there is some non-zero radius $r > 0$ such that $h''(t) \geq \frac{\alpha}{2} > 0$ for all $|t - t_*| \leq r$. Now for contradiction, assume that $h(t) = h(t_*)$ for some $t \neq t_*$. By strong convexity, we can bound

$$\begin{aligned} h(t) - h(t_*) &= \int_{t_*}^{t} h'(s_1) = \int_{t_*}^{t} \Big(h'(t_*) + \int_{t_*}^{s_1} h''(s_2) ds_2\Big) ds_1 \\ &\geq \int_{t_*}^{t} \Big(0 + \int_{t_*}^{\min\{r, s_1\}} h''(s_2) ds_2\Big) ds_1 \geq \int_{t_*}^{t} \frac{\alpha}{2} \min\{r, s_1\} > 0, \end{aligned}$$

where the first two steps were by the fundamental theorem of calculus, in the third step we used $h'(t_*) = 0$ by Lemma 2.3.4 and the fact that $h''(s) \geq 0$ by Definition 2.3.1 of convexity for the inequality, in the fourth step we applied $h''(t) \geq \frac{\alpha}{2}$ for $|t - t_*| \leq r$ as derived above using strong convexity, and the final step was by the assumptions $\alpha > 0$ and $r > 0$. This is our desired contradiction, so t_* is the unique optimizer. □

We can also derive the following approximate version of Lemma 2.3.4.

Lemma 2.3.6. *For α-strongly convex $h : \mathbb{R} \to \mathbb{R}$ and any $s \in \mathbb{R}$, the optimum can be lower bounded by*

$$h^* := \inf_{t \in \mathbb{R}} h(t) \geq h(s) - \frac{|h'(s)|^2}{2\alpha},$$

and the optimizer t_* satisfies $|t_* - s| \leq \frac{|h'(s)|}{\alpha}$.

Proof. We first use strong convexity near s to show the bound $|t_* - s| \le \frac{|h'(s)|}{\alpha}$ for the optimizer. Recall that Lemma 2.3.4 shows that any critical point of h is a global minimum. Therefore we will show that if $|h'(s)|$ is small, then strong convexity implies that there is a critical point $h'(t) = 0$ nearby. We can assume $h'(s) \le 0$ by considering the function $h(-t)$ if necessary. Then for arbitrary $t \in \mathbb{R}$:

$$h'(t) = h'(s) + \int_s^t h''(r) \ge h'(s) + \alpha(t - s),$$

where the first step was by the fundamental theorem of calculus, and in the final step we used $h''(r) \ge \alpha$ by α-strong convexity according to Definition 2.3.2. By continuity of h', this implies that there is some $s \le t \le s - \frac{h'(s)}{\alpha}$ such that $h'(t) = 0$, which by Lemma 2.3.4 means that the optimizer of h is within this range.

To show the lower bound, we use α-strong convexity and calculate, for any $s, t \in \mathbb{R}$,

$$h(t) - h(s) = \int_{t_1=s}^{t} h'(t_1) = \int_{t_1=s}^{t} \Big(h'(s) + \int_{t_2=s}^{t_1} h''(t_2)\Big) \ge h'(s)(t-s) + \frac{\alpha}{2}(t-s)^2,$$

where the first two steps were by the fundamental theorem of calculus, and in the final step we used that $h''(t) \ge \alpha$ by Definition 2.3.2 of α-strong convexity and integrated. Now the result follows by optimizing the quadratic lower bound shown above:

$$h(t_*) = \inf_{t\in\mathbb{R}} h(t) \ge \inf_{t\in\mathbb{R}} h(s) + h'(s)(t-s) + \frac{\alpha}{2}(t-s)^2 = h(s) - \frac{(h'(s))^2}{2\alpha},$$

where the second step was by the lower bound shown above, and in the last step we chose infimizer $t = s - \frac{h'(s)}{\alpha}$. □

Note that in the above proof, we only used strong convexity at points $t \in [s, t_*]$. For our applications, we will study convex functions that are strongly convex only in some neighborhood. Therefore, we rewrite the above statement with these weaker assumptions.

Lemma 2.3.7. *For convex function $h : \mathbb{R} \to \mathbb{R}$ with optimizer $t_* := \arg\min_{t\in\mathbb{R}} h(t)$ and any fixed $s \in \mathbb{R}$, if h is α-strongly convex for all $t \in [s, t_*]$, then*

$$h^* := \inf_{t\in\mathbb{R}} h(t) \ge h(s) - \frac{|h'(s)|^2}{2\alpha},$$

If the optimizer is not known, a sufficient condition is α-strong convexity for $[s \pm \frac{|h'(s)|}{\alpha}]$.

2.3.2 Convex Functions on Vector Spaces

All of these properties lift to convex functions on vector spaces.

Definition 2.3.8. *For vector space V, function $f : V \to \mathbb{R}$ is convex if for every $x, y \in V$, the univariate restriction $t \to f(x + ty)$ is convex.*

f is α-strongly convex in norm $\|\cdot\|$ at point $x \in V$ if, for every $v \in V$, the univariate restriction $t \to f(x + tv)$ is $\alpha\|v\|^2$-strongly convex at $t = 0$ according to Definition 2.3.2.

Convex functions on vector spaces are also optimized at their critical points.

Definition 2.3.9. *For vector space V and function $f : V \to \mathbb{R}$, $x \in V$ is a critical point of f iff*

$$\forall v \in V : \partial_{t=0} f(x + tv) = 0.$$

Equivalently, $x \in V$ is a critical point of f if for every $v \in V$, the univariate restriction $t \to f(x + tv)$ has critical point $t = 0$.

Lemma 2.3.10. *For vector space V and convex function $f : V \to \mathbb{R}$, $f^* := \inf_{v \in V} f(v) = f(x)$ iff x is a critical point of f.*

Proof. If $x \in V$ is a global minimizer of f, then in particular it is the minimizer of every univariate restriction $h(t) := f(x + tv)$. Therefore by Lemma 2.3.4 we must have

$$0 = \partial_{t=0} h(t) = \partial_{t=0} f(x + tv)$$

for every $v \in V$, which is exactly Definition 2.3.9 of critical points.

Conversely, if $x \in V$ is a critical point, then for any $y \in V$, we have that $t = 0$ is a critical point for the univariate restriction $h(t) := f(x + t(y - x))$. Therefore the statement follows from Lemma 2.3.4. □

And strongly convex functions on vector spaces also have unique optimizers.

Lemma 2.3.11. *For vector space V and convex function $f : V \to \mathbb{R}$, if f is $\alpha > 0$-strongly convex at $x_* := \arg\inf_{x \in V} f(x)$, then x_* is the unique optimizer of f.*

Proof. Assume for contradiction that $y \neq x_*$ is also an optimizer. Then the univariate function $h(t) := f(x_* + t(y - x_*))$ is strictly convex at $t = 0$, so Lemma 2.3.5 gives the contradiction. □

The above suggests that a natural algorithm to find the minimizer of a convex function is to follow the direction of steepest descent. To formalize this idea, we will need a choice of inner product.

Definition 2.3.12. *If V has inner product $\langle\cdot,\cdot\rangle$, and f is differentiable, then the gradient of f at point x is the unique element of V satisfying*

$$\forall v \in V : \langle \nabla f(x), v\rangle = \partial_{\delta=0} f(x + \delta v).$$

Therefore, in this case the convexity condition can be written equivalently as

$$\forall x, y \in V : f(x) - f(y) \geq \langle \nabla f(x), y - x\rangle.$$

This suggests a family of minimization algorithms known as gradient methods or first-order methods [75]. Convex functions give a wide class of optimization problems that arise in many applications. The value of convexity, and specifically the first order convexity condition in Definition 2.3.12, is that the gradient at any point defines a halfspace containing the optimizer. Therefore, in order to minimize the function, we can always follow the negative gradient direction and this will intuitively lead us in the direction of the optimizer.

From this perspective, strong convexity means that the function curves strictly away from any tangent hyperplane. This intuitively implies that the negative gradient direction not only makes progress by decreasing the function, but also decreases the size of the gradient. These kind of implications will be helpful for our analysis of the geodesically convex problems in Chapter 6. For further details and formal analyses of various minimization methods, we refer the reader to Section 1.2 of [75].

Finally, we present a standard result on projections in convex analysis that we will use for our distance bounds in Chapter 4.

Lemma 2.3.13 (Lemma 3.1.5 in [75])**.** *Consider vector space V with Euclidean norm $\|\cdot\|_2$. Then for any convex body $K \subseteq V$ and any point $x \notin K$, if $x_* := \arg\min_{z\in K} \|z - x\|_2^2$ is the Euclidean projection, then*

$$\forall y \in K : \|x_* - y\|_2^2 \leq \|x - y\|_2^2.$$

2.4 Quantum Information

In this section, we present some basic definitions from Quantum Information Theory. The tensor scaling problem that we study in Chapter 6 has an equivalent formulation in terms of marginals of quantum states, and so this perspective will give many useful inequalities for our tensor setting. We will follow the presentation of Watrous in [98].

2.4.1 Tensor Products and Quantum Marginals

The tensor product of vector spaces U and V is defined as the the set of linear combinations of formal pairs $u \otimes v$ where $u \in U$ and $v \in V$. If U has basis $\{u_1, ..., u_d\}$ and V has basis $\{v_1, ..., v_{d'}\}$, the tensor product $U \otimes V$ has basis $\{u_i \otimes v_j\}_{i \in [d], j \in [d']}$. This shows $\dim(U \otimes V) = \dim(U)\dim(V)$.

The vector space $U \otimes V$ has operators $L(U, V)$ acting linearly upon it, just as we discussed in Section 2.1.1. By the definition of the tensor product space, any pair $A \in L(U)$ and $B \in L(V)$ has a natural action on $u \otimes v$ with $u \in U$ and $v \in V$ as $(A \otimes B)(u \otimes v) = (Au) \otimes (Bv)$, and this can be extended linearly to the whole space. We emphasize that $L(U, V)$ is not contained in the set of linear combinations of $(A \in L(U)) \otimes (B \in L(V))$. This is in fact some part of the reason for the phenomenon of quantum entanglement.

Below, we collect some simple facts about the tensor product of linear operators.

Fact 2.4.1. *For $X \in \mathrm{H}(U), Y \in \mathrm{H}(V)$ the spectrum of $X \otimes Y \in \mathrm{H}(U \otimes V)$ is $\{x_i y_j\}_{i \in [d], j \in [d']}$, where $\{x_i\}_{i \in [d]}$ and $\{y_j\}_{j \in [d']}$ are the spectra of X and Y, respectively. Consequently,* $\det(X \otimes Y) = \det(X)^{d'} \det(Y)^d$

Given a tensor $x \in \mathbb{R}^d \otimes \mathbb{R}^{d'}$, we will often consider the "flattening" $X := \mathrm{Mat}(x) \in \mathrm{Mat}(d, d')$. The j-th column of X corresponds to the entries $\{x_{ij}\}_{i \in [d]}$. More generally, given $x \in \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \otimes \mathbb{R}^{d_3}$, we can view this as a tuple of tensors $X_1, ..., X_{d_3} \in \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}$ defined entry-wise as $(X_{j_3})_{j_1, j_2} = x_{j_1, j_2, j_3}$. In the following subsections, we will use this correspondence to translate between tensors and quantum states and marginals.

2.4.2 Quantum States and Quantum Maps

We begin with the basic objects of study in Quantum Information.

Definition 2.4.2. *For vector space V, an element $\rho \in L(V)$ is a quantum state if $\rho \succeq 0$ and* $\mathrm{Tr}[\rho] = 1$.

We will often be informal about this definition and call arbitrary $\rho \succeq 0$ a "state" even if it does not satisfy the Tr condition.

Now we consider maps between these objects.

Definition 2.4.3. *For inner product spaces U, V, a quantum map $\Phi : L(V) \to L(U)$ is a linear map that preserves self-adjoint operators. The dual of quantum map Φ is the adjoint map $\Phi^* : L(U) \to L(V)$ under the natural Euclidean inner product:*

$$\langle \Phi_\rho^*(X), Y \rangle = \langle X, \Phi_\rho(Y) \rangle.$$

In Quantum Information theory, the appropriate notion of mappings is defined by quantum channels, which satisfy some additional properties (completely positive, trace-preserving). We do not define these, as we will only use the linear operator interpretation.

2.4.3 Representations of Quantum States and Maps

There are many ways to represent states and channels, each of which emphasize different properties. The following definitions and equivalences are from Prop 2.20 of [98].

Definition 2.4.4. *For vector spaces U, V and tuple of linear operators $A_1, ..., A_K \in L(U, V)$, the associated state representation $\rho_A \in L(U \otimes V)$ is defined as*

$$\rho_A := \sum_{k=1}^{K} \operatorname{vec}(A_k) \operatorname{vec}(A_k)^*.$$

The associated quantum maps $\Phi_A : L(V) \to L(U)$ and $\Phi_A^* : L(U) \to L(V)$ are defined by

$$\Phi_A(Y) := \sum_{k=1}^{K} A_k Y A_k^* \qquad \Phi_A^*(X) := \sum_{k=1}^{K} A_k^* X A_k.$$

In this case, $\{A_1, ..., A_K\}$ is a Kraus representation of Φ_A.

Proposition 2.4.5. *Given inner product spaces U, V, for $\rho \in L(U, V)$ and $\Phi : L(V) \to L(U)$, (ρ, Φ) is an associated pair iff for every $X \in L(U), Y \in L(V)$:*

$$\langle X, \Phi(Y) \rangle = \langle \rho, X \otimes Y \rangle.$$

This gives a bijection between quantum states and quantum maps. If Φ is a quantum channel, then ρ is known as the Choi representation of Φ.

Another helpful perspective is to view quantum operators as matrices acting on vectors $\operatorname{vec}(L(V))$. This allows us to use spectral theory to analyze quantum maps. The following equation provides the translation and can be verified entry-wise:

$$\operatorname{vec}(AXB^*) = (A \otimes \overline{B}) \operatorname{vec}(X).$$

Definition 2.4.6 (Natural Representation)**.** *For vector spaces U, V and tuple of linear operators $A_1, ..., A_K \in L(U, V)$ given as Kraus operators, the natural representation of the channel Φ_A given in Definition 2.4.4 is the matrix*

$$M_A := \sum_{k=1}^{K} A_k \otimes \overline{A_k},$$

where $\overline{A_k}$ denotes complex conjugation if $\mathbb{F} = \mathbb{C}$. Note that this is a map $M_A : L(V) \to L(U)$ by the bijection $Y \in L(V) \to \mathrm{vec}(Y) \in V \otimes V$ and $X \in L(U) \to \mathrm{vec}(X) \in U \otimes U$. Further note the following relations between ρ_A, Φ_A, M_A:

$$M_A(\mathrm{vec}(Y)) = \Big(\sum_{k=1}^{K} A_k \otimes A_k \Big) \mathrm{vec}(Y) = \sum_{k=1}^{K} \mathrm{vec}(A_k Y A_k^*) = \mathrm{vec}(\Phi_A(Y))$$

$$\forall X \in L(U), Y \in L(V) : \langle \mathrm{vec}(X), M_A \, \mathrm{vec}(Y) \rangle = \langle X, \Phi_A(Y) \rangle = \langle \rho_A, X \otimes Y \rangle$$

As shown above, quantum maps can be viewed as bipartite quantum states. In the following, we discuss how to generalize this to multipartite states.

Definition 2.4.7. *Given inner product spaces V, W, the partial trace is the map $\mathrm{Tr}_W : L(V \otimes W) \to L(V)$ defined uniquely by*

$$\forall X \in L(V), Y \in L(W) : \mathrm{Tr}_V[X \otimes Y] = X(\mathrm{Tr}[Y]).$$

For $\rho \in L(V \otimes W)$, the V-marginal is then defined

$$\rho^V := \mathrm{Tr}_W[\rho].$$

Equivalently, the V marginal is the unique operator in $L(V)$ satisfying

$$\forall X \in L(V) : \langle \rho^V, X \rangle = \langle \rho, X \otimes I_W \rangle.$$

In our setting, we will have $\rho \in L(V)$ for tensor space $V = \otimes_{a \in [m]} V_a$. Then for any $S \subseteq [m]$, $V_S := \otimes_{a \in S} V_a$ and the S-marginal $\rho^S \in L(V_S)$ is the $\overline{S}$-partial trace of $\rho \in L(V) = L(V_S \otimes V_{\overline{S}})$.

If $V_a = \mathbb{F}^{d_a}$ is given explicitly, then $\rho_x^S \in \mathrm{Mat}(d_S)$ where $d_S = \prod_{a \in S} d_a$.

This property uniquely determines $\rho^{(S)}$. If ρ is positive definite then so is ρ^S. Moreover, $(\rho^S)^T$ for $T \subseteq S$, and $\mathrm{Tr}[\rho^S] = \mathrm{Tr}[\rho^T]$.

With this perspective for tensors, we can discuss how the flattenings described in Section 2.4.1 relate to marginals. First note that for Kraus operators $\{A_1, ..., A_K\} \in L(U, V)$, the marginals of ρ_A correspond to the operators

$$\rho_A^U = \sum_{k=1}^{K} A_k A_k^* \qquad \text{and} \qquad \rho_A^V = \sum_{k=1}^{K} A_k^* A_k$$

as shown in Definition 2.4.4. More generally, let state $\rho \in L(\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \otimes \mathbb{C}^{d_3})$ be represented by the tuple of tensors $x_1, ..., x_K \in \mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \otimes \mathbb{C}^{d_3}$ as $\rho = \sum_{k=1}^K x_k x_k^*$. In order to express e.g. the d_3 partial trace of ρ, we first view each x_k as a tuple $y_{k,1}, ..., y_{k,d_3} \in \mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ with entry-wise correspondence $(x_k)_{j_1,j_2,j_3} = (y_{k,j_3})_{j_1,j_2}$. This allows us to write

$$\rho^{(12)} = Tr_3[\rho] = \sum_{k=1}^{K} \sum_{j=1}^{d_3} y_{k,j} y_{k,j}^*. \tag{2.9}$$

Further, by Definition 2.4.4, this gives $\{\mathrm{Mat}(y_{k,j})\}_{k \in [K], j \in [d_3]}$ as the Kraus operators of $\rho^{(12)}$.

These translations will be useful in Chapter 7 in order to emphasize different perspectives for tensor scaling inputs.

2.5 Concentration Inequalities

In probability theory, concentration inequalities are used to bound the deviation of random variables from their means. We will heavily rely on such bounds in order to analyze various properties of random inputs to the scaling problems studied in Chapter 5 and Chapter 9.

The simplest such inequality holds for arbitrary non-negative random variables.

Fact 2.5.1 (Markov). *For random variable $X \geq 0$ and any $\theta > 0$*

$$Pr[X \geq \theta] \leq \frac{\mathbb{E}X}{\theta}.$$

Note that this is trivial for any $\theta \leq \mathbb{E}X$.

The above fact is quite elementary, but can be leveraged to prove much stronger inequalities for special classes of random variables. The exponential moment method is one particular approach to stronger concentration, and proceeds

$$Pr[X \geq \theta] = Pr[e^{tX} \geq e^{t\theta}] \leq e^{-t\theta} \mathbb{E} e^{tX}$$

for any $t > 0$, where we applied Markov's inequality in the last step. Therefore, if we have control over the exponential moment, this allows us to optimize over t to give a family of strong inequalities. In the following, we will study several classes of random variables for which we can control the exponential moment and give strong concentration bounds.

2.5.1 Independent and Sub-Exponential Distributions

The standard Chernoff bound is one such consequence involving a sum of independent random variables. We state result for the specialized setting of Bernoulli random variables, as this is all that we require for our analysis in Section 5.1.

Theorem 2.5.2 (Chernoff Bound, Theorem 2.3.1 in [95])**.** *Let $X_1, ..., X_N$ be independent Bernoulli random variables with $\mathbb{E}X_i = p_i$, and denote $\mu = \mathbb{E}\sum_{i=1}^N X_i = \sum_{i=1}^N p_i$. Then for any $t > \mu$, we have*

$$Pr\Big[\sum_{i=1}^N X_i \geq t\Big] \leq e^{-\mu}\Big(\frac{e\mu}{t}\Big)^t.$$

Sub-exponential random variables are another subclass that enjoy strong concentration bounds. The following are standard results from [94], [96].

Definition 2.5.3. *Random variable X with mean $\mathbb{E}X = 0$ is (ν^2, b)-sub-exponential*

$$\forall t \leq \frac{1}{b}: \quad \log \mathbb{E}\exp tX \leq \frac{t^2\nu^2}{2}$$

The definition of sub-exponential distributions allows us to prove strong concentration bounds via the exponential moment method.

Lemma 2.5.4 (Bernstein)**.** *If X is (ν^2, b)-subexponential, then for all $\theta \geq 0$*

$$\mathbb{P}[X \geq \theta] \leq \begin{cases} \exp\left(-\frac{\theta^2}{2\nu^2}\right) & \forall\theta < \frac{\nu^2}{b} \\ \exp\left(-\frac{\theta}{2b}\right) & \forall\theta \geq \frac{\nu^2}{b} \end{cases}$$

Proof. We first proceed by the exponential moment method,

$$Pr[X \geq \theta] = \inf_{t>0} Pr[e^{tX} \geq e^{t\theta}] \leq \inf_{t>0} \frac{\mathbb{E}e^{tX}}{e^{t\theta}},$$

where the last step was by Markov's inequality. Now we can use Definition 2.5.3 on sub-exponential variables

$$\inf_{t>0} \log \frac{\mathbb{E}e^{tX}}{e^{t\theta}} \leq \inf_{0<t<b^{-1}} \frac{t^2\nu^2}{2} - t\theta.$$

The two cases then follow by choosing $t = \theta/\nu^2$ if it is feasible, i.e. $\theta < \nu^2/b$

$$\frac{t^2\nu^2}{2} - t\theta = \frac{\theta^2}{2\nu^2} - \frac{\theta^2}{\nu^2} = -\frac{\theta^2}{2\nu^2}$$

and the boundary $t = b^{-1}$ if $\theta \geq \nu^2/b$ otherwise

$$\frac{t^2\nu^2}{2} - t\theta = \frac{\nu^2}{2b^2} - \frac{\theta}{b} = \frac{1}{b}\left(\frac{\nu^2}{2b} - \theta\right) \leq -\frac{\theta}{2b}.$$

□

We can also control the finite moments of a sub-exponential distribution.

Lemma 2.5.5 (Proposition 2.7.1 in [95])**.** *If both $X, -X$ are (ν^2, b)-subexponential (Definition 2.5.3), then $\forall p \geq 1 : \mathbb{E}|X|^p \leq \frac{\nu^2}{b^2}(bp)^p$.*

Proof. By Definition 2.5.3, $\forall |t| \leq \frac{1}{b}$ we have

$$\log \mathbb{E} \exp tX \leq \frac{t^2\nu^2}{2}.$$

We will use the following simple inequality to prove a bound on moments.

Claim 2.5.6. *For all $x \in \mathbb{R}$ and $p \geq 1$: $|x|^p \leq p^p(e^x + e^{-x})$*

Proof. Since $e^x + e^{-x} \geq 1$ always, the statement is clearly true for $|x| \leq p$. Otherwise, divide both sides by p^p and let $u = x/p$. Then we can show

$$\forall u \geq 1 : \frac{x^p}{p^p} = u^p \leq (e^u)^p = e^x,$$

where the inequality $u \leq e^u$ is due to our assumption $u \geq 1$. The claim follows by a symmetric argument showing $u \leq -1 \implies |u| \leq e^{-u}$. □

Now we bound the p-th moment using the claim to give

$$\mathbb{E}|X|^p = b^p\mathbb{E}|X/b|^p \leq (bp)^p\mathbb{E}(e^{X/b} + e^{-X/b}) \leq (bp)^p\frac{\nu^2}{b^2},$$

where the second step was by the claim above, and the final step was by the subexponential bound on the MGF by Definition 2.5.3. □

In the following subsections, we will apply these general bounds to Gaussian and chi-square random variables.

2.5.2 Gaussian and Chi-square Distributions

Gaussian and Chi-square distributions are some of the most well-studied in probability theory. In this subsection, we will define and give strong concentration bounds for these distributions. We follow the exposition in [96].

Definition 2.5.7. *The probability density function (pdf) of the standard Gaussian distribution $g \sim N(0,1)$ on $\mathbb{R}$ is*

$$f(x \in \mathbb{R}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

It has mean $\mathbb{E}g = 0$ and variance $\mathbb{E}g^2 = 1$.

The multivariate centered Gaussian distribution with covariance matrix $C \in \mathrm{PD}(k)$ is denoted $g \sim N(0,C)$ and has pdf

$$f(x \in \mathbb{R}^k) = \frac{1}{\sqrt{\det(2\pi C)}} \exp\left(-\frac{\langle x, C^{-1}x\rangle}{2}\right).$$

$g \sim N(0,C)$ can be equivalently written $g = \sum_{i=1}^k \sqrt{\lambda_i} g_i u_i$ where the $g_i \sim N(0,1)$ are i.i.d. standard Gaussians and $\{\lambda_i, u_i\}$ are the eigen-pairs of C according to Theorem 2.1.8. As a consequence, $\mathbb{E}g = 0$ and $\mathbb{E}gg^* = C$.

This distribution satisfies the following linear invariance property.

Proposition 2.5.8. *Consider Gaussian random variable $g \sim N(0,C)$ for $C \in \mathrm{PD}_0(k)$. Then for any $A \in \mathrm{Mat}(d,k)$, Ag is also a Gaussian random variable, and in particular, $Ag \sim N(0, ACA^*)$. As a consequence, if $u, v \in \mathbb{R}^k$ are orthogonal ($\langle u, v\rangle = 0$), then $\langle g, u\rangle$ and $\langle g, v\rangle$ are mutually independent.*

We will often need to bound the norm of a random Gaussian vector. Therefore, we introduce the following standard distribution.

Definition 2.5.9. *Let $X = \sum_{i=1}^k g_i^2$ where $g_i \sim N(0,1)$ are independent standard Gaussian variables. Then X is a chi-squared random with k degrees of freedom and is denoted $\chi(k)$. By linearity, the mean is $\mathbb{E}X = \sum_{i=1}^k \mathbb{E}g_i^2 = k$.*

The following is a well known explicit formula for the MGF of chi-square variables.

Fact 2.5.10. *For any $t < \frac{1}{2}$, the moment generating function (MGF) of $X \sim \chi(k)$ is*

$$\mathbb{E}\exp(tX) = (1-2t)^{-k/2}.$$

If $4t \leq 1$, then the MGF of $X - k$ can be bounded by

$$\log \mathbb{E}\exp(t(X-k)) \leq 2kt^2,$$

which implies that both $\pm(X-k)$ are $(4k, 4)$-subexponential according to Definition 2.5.3.

The last statement in the fact above can be combined with Lemma 2.5.4 to give concentration bounds on chi-square variables. But for this explicit distribution, we can rely on the following stronger result.

Theorem 2.5.11 (Laurent, Massart [64])**.** *For $X \sim \chi(k)$ we have tail bounds*

$$Pr[X - k \geq 2\theta\sqrt{k} + 2\theta^2] \leq \exp(-\theta^2), \qquad \textit{and} \qquad Pr[X - k \leq -2\theta\sqrt{k}] \leq \exp(-\theta^2).$$

Before moving onto more complicated distributions in the following subsection, we state an important result showing concentration for the spectrum of Gaussian random matrices. This can be viewed as simultaneous concentration of $\|Gx\|_2^2$ for all directions x, and is proved in using a net argument that is standard in the random matrix literature [94].

Theorem 2.5.12 (Corollary 5.35 of [94])**.** *For $d \leq n$, let $G \in \mathrm{Mat}(d,n)$ be a random matrix with standard Gaussian entries $G_{ij} \sim N(0,1)$ for $i \in [d], j \in [n]$. Then for any $t > 0$,*

$$\sqrt{n} - \sqrt{d} - t \leq \sigma_{\min}(G) \leq \sigma_{\max}(G) \leq \sqrt{n} + \sqrt{d} + t$$

with probability at least $1 - 2e^{-t^2/2}$.

Much of the work of Chapter 5 on the Paulsen problem will be to prove similar spectral results for more complicated Gaussian distributions that arise from our smoothed analysis argument. The key technical results necessary to prove concentration are described in the following subsection.

2.5.3 Hanson-Wright Inequality

We can generalize the results in Section 2.5.2 to more general quadratic forms of Gaussians. We will use standard MGF bounds and the theory of sub-exponential random variables as described in Section 2.5.1.

Theorem 2.5.13 (Hanson-Wright Inequality [81]). *For fixed $A \in \mathbb{R}^{m\times n}$ and $g \sim N(0, I_n)$, consider random variable $\|Ag\|_2^2$. The mean is $\mathbb{E}\|Ag\|_2^2 = \mathbb{E}\langle gg^*, A^*A\rangle = \mathrm{Tr}[A^*A] = \|A\|_F^2$ and we have the following concentration:*

$$\mathbb{P}[|\|Ag\|_2^2 - \|A\|_F^2| \geq \theta] \leq \begin{cases} 2\exp\left(-\frac{\theta^2}{8\|A^*A\|_F^2}\right) & \forall\theta \leq \frac{\|A^*A\|_F^2}{\|A^*A\|_{\mathrm{op}}} \\ 2\exp\left(-\frac{\theta}{8\|A^*A\|_{\mathrm{op}}}\right) & \forall\theta > \frac{\|A^*A\|_F^2}{\|A^*A\|_{\mathrm{op}}} \end{cases}$$

Proof. By the Spectral Theorem in 2.1.8, we can diagonalize $A^*A = \sum_{i=1}^n \lambda_i u_i u_i^*$ with $\lambda \geq 0$. This allows us to write $\|Ag\|_2^2 = \sum_{i=1}^n \lambda_i \langle g, u_i\rangle^2$. By orthogonality of the eigenvectors, the random variables $\{\langle g, u_i\rangle^2\}$ are i.i.d. chi-squared variables with one degree of freedom, so $\mathbb{E}\|Ag\|_2^2 = \sum_{i=1}^n \lambda_i = \mathrm{Tr}[A^*A]$. Further by independence we can separate the MGF as

$$\log \mathbb{E}\exp(t\|Ag\|_2^2) = \log \mathbb{E}\exp(t\sum_{i=1}^n \lambda_i \langle u_i, g\rangle^2) = \sum_{i=1}^n \log \mathbb{E}\exp(t\langle u_i, g\rangle^2).$$

Now we can use Fact 2.5.10 for $\max_i 4|t\lambda_i| \leq 1$ to show

$$\log \mathbb{E}\exp(t(\|Ag\|_2^2 - \mathrm{Tr}[A^*A])) = \sum_{i=1}^n \log \mathbb{E}\exp(t\lambda_i(\langle g, u_i\rangle^2 - 1)) \leq \sum_{i=1}^n 2t^2\lambda_i^2,$$

which shows $\|Ag\|_2^2$ is $(4\sum_i \lambda_i^2, 4\max_i |\lambda_i|) = (4\|A^*A\|_F^2, 4\|A^*A\|_{\mathrm{op}})$-subexponential according to Definition 2.5.3. The theorem follows from the two-sided Bernstien bounds in Lemma 2.5.4. □

In some of our applications, we may not be able to calculate second moments $\|A^*A\|_F^2$. So below, we produce a simple corollary using only first moment information.

Corollary 2.5.14 (Theorem 2.1 in [81]). *For fixed $A \in \mathbb{R}^{m\times n}$ and $g \sim N(0, I_n)$, consider random variable $\|Ag\|_2^2 = \langle gg^*, A^*A\rangle$. The mean is $\mathrm{Tr}[A^*A]$ and we have the following concentration:*

$$\mathbb{P}[|\langle gg^*, A^*A\rangle - \mathrm{Tr}[A^*A]| \geq \theta] \leq 2\exp\left(-\min\left\{\frac{\theta^2}{8\,\mathrm{Tr}[A^*A]\|A^*A\|_{\mathrm{op}}}, \frac{\theta}{8\|A^*A\|_{\mathrm{op}}}\right\}\right).$$

$$\mathbb{P}[|\langle gg^*, A^*A\rangle - \mathrm{Tr}[A^*A]| \geq \eta\,\mathrm{Tr}[A^*A]] \leq 2\exp\left(-\min\{\eta^2, \eta\}\frac{\mathrm{Tr}[A^*A]}{8\|A^*A\|_{\mathrm{op}}^2}\right).$$

Further, the lower and upper tails can be bounded separately by the same term without the leading factor 2.

Proof. The exponents in the two cases of Theorem 2.5.13 match at the boundary:

$$\frac{\theta^2}{8\|A^*A\|_F^2} = \frac{\|A^*A\|_F^2}{8\|A^*A\|_{\text{op}}^2} = \frac{\theta}{8\|A^*A\|_{\text{op}}}.$$

Therefore we can rewrite the bound as

$$Pr\Big[|\langle A^*A, gg^*\rangle - \operatorname{Tr}[A^*A]| \geq \theta\Big] \leq 2\exp\left(-\min\left\{\frac{\theta^2}{8\|A^*A\|_F^2}, \frac{\theta}{8\|A^*A\|_{\text{op}}}\right\}\right)$$

so the probability is always upper bounded by the larger of the two. Continuing with the crude bounds $\|A^*A\|_F^2 \leq \|A^*A\|_{\text{op}} \operatorname{Tr}[A^*A]$, we get

$$\mathbb{P}[|\langle gg^*, A^*A\rangle - \operatorname{Tr}[A^*A]| \geq \theta] \leq 2\exp\left(-\min\left\{\frac{\theta^2}{8\operatorname{Tr}[A^*A]\|A^*A\|_{\text{op}}}, \frac{\theta}{8\|A^*A\|_{\text{op}}}\right\}\right).$$

Now choosing $\theta := \eta \operatorname{Tr}[A^*A]$ gives the second result. □

Since we know that $\|Ag\|_2^2 \geq 0$ always, the lower tail bound becomes trivial for $\theta > \mathbb{E}\|Ag\|_2^2$. In order to get higher probability statements we can use the following bound from [62]. We repeat the proof for completeness.

Lemma 2.5.15 (Fact 4.5.7(3) in [62])**.** *For fixed $0 \preceq A \preceq I_n$ and standard Gaussian $g \sim N(0, I_n)$, if $c \geq 5$ then the quadratic form concentrates as*

$$Pr_{g\sim N(0,I_n)}[\langle g, Ag\rangle \leq e^{-c}\operatorname{Tr}[A]] \leq \exp\left(-\frac{2}{5}c\operatorname{Tr}[A]\right).$$

Proof. Note that $\mathbb{E}\langle g, Ag\rangle = \langle A, \mathbb{E}gg^*\rangle = \operatorname{Tr}[A]$, so the following holds for any $\theta > 0$ and $t > 0$:

$$Pr[\langle g, Ag\rangle \leq \theta\operatorname{Tr}[A]] = Pr[e^{-t\langle g,Ag\rangle} \geq e^{-t\theta\operatorname{Tr}[A]}] \leq e^{t\theta\operatorname{Tr}[A]}\mathbb{E}\exp(-t\langle g, Ag\rangle),$$

where the last step was by Markov's bound applied to $e^{-t\langle g,Ag\rangle}$.

By the Spectral Theorem (2.1.8), we can diagonalize $A = \sum_{i=1}^n \lambda_i u_i u_i^*$. This allows us to bound the MGF by a similar calculation as in the proof of Theorem 2.5.13.

$$\begin{aligned} t\theta\operatorname{Tr}[A] + \log\mathbb{E}\exp(-t\langle g, Ag\rangle) &= t\theta\sum_{i=1}^n \lambda_i + \log\mathbb{E}\exp(-t\sum_{i=1}^n \lambda_i\langle g, u_i\rangle^2) \\ &= \sum_{i=1}^n \left(\theta\lambda_i + \log\mathbb{E}\exp(-t\lambda_i\langle g, u_i\rangle^2)\right), \end{aligned}$$

where the first step was by the definition of Tr, and the last step was by independence of $\{\langle g, u_i\rangle\}_{i\in[n]}$ since $\{u_i\}_{i=1}^n$ are orthonormal eigenvectors. Now we assume $\max_{i\in[n]}(-2t)\lambda_i < 1$ so that the moment generating function is defined and use Fact 2.5.10 to compute $\log \mathbb{E}\exp(-t\lambda_i\langle g, u_i\rangle^2) = -\frac{1}{2}\log(1+2t\lambda_i)$. The main observation is that $\lambda \to -\frac{1}{2}\sum_{i=1}^d \log(1+2t\lambda_i)$ is convex and therefore the maximizer over the convex set $\{\lambda \in [0,1]^n, \sum_{i=1}^n \lambda_i = \mathrm{Tr}[A]\}$ occurs at the boundary where $\lfloor \mathrm{Tr}[A]\rfloor$ entries are 1 and one entry is the remaining fractional part. Therefore we can bound the above quantity

$$t\theta\,\mathrm{Tr}[A]+\log\mathbb{E}\exp(-t\langle g, Ag\rangle) = t\theta\sum_{i=1}^n \lambda_i - \frac{1}{2}\sum_{i=1}^n \log(1+2t\lambda_i) \le \sum_{i=1}^d \lambda_i\left(t\theta - \frac{1}{2}\log(1+2t)\right).$$

Now we want to choose $-2t = 1-\theta^{-1} \ge 0$ for some $0 < \theta \le 1$, for which the moment generating function is well defined as $\max_{i\in[n]} -2t\lambda_i \le -2t < 1$ by our constraint $0 \preceq A \preceq I_n$. Plugging this into the previous probability bound gives

$$\log Pr[\langle g, Ag\rangle \le \theta\,\mathrm{Tr}[A]] \le -\frac{\mathrm{Tr}[A]}{2}(\log(1+2t) - 2t\theta) = -\frac{\mathrm{Tr}[A]}{2}(\log\theta^{-1} - 1 + \theta).$$

Rewriting $\theta = e^{-c}$, the term in the parentheses becomes $\frac{1}{2}(c-1+e^{-c}) \ge \frac{2}{5}c$ for $c \ge 5$. □

This also gives the following simple corollary for the lower tail of chi-square variables.

Corollary 2.5.16. *$X \sim \chi(k)$ can be equivalently written as $X = \langle gg^*, I_k\rangle$ for standard Gaussian $g \sim N(0, I_k)$. By Lemma 2.5.15, for any $c \ge 5$, X can be lower bounded as*

$$Pr[X \le e^{-c}k] \le \exp\left(-\frac{2}{5}ck\right)$$

The final result in this subsection will be a bound on the moments of quadratic forms of Gaussians. This is a specialization of Lemma 2.5.5 and will be used in Chapter 5.

Corollary 2.5.17. *For fixed $A \in \mathrm{Mat}(m,n)$ and $g \sim N(0, I_n)$, for random variable $\|Ag\|_2^2$ we have the following moment bounds for all $p \ge 1$:*

$$\mathbb{E}\|Ag\|_2^{2p} \le \frac{\|A^*A\|_F^2}{\|A^*A\|_{\mathrm{op}}^2}(8p\|A^*A\|_{\mathrm{op}})^p + (2\,\mathrm{Tr}[A^*A])^p.$$

Further, by the simple bound $\|A^*A\|_F^2 \le \|A^*A\|_{\mathrm{op}}\,\mathrm{Tr}[A^*A]$, we have the corollary

$$\mathbb{E}\|Ag\|_2^{2p} \le \frac{\mathrm{Tr}[A^*A]}{\|A^*A\|_{\mathrm{op}}}(8p\|A^*A\|_{\mathrm{op}})^p + (2\,\mathrm{Tr}[A^*A])^p.$$

Proof. Recall in the proof of Theorem 2.5.13 we showed that $\mathbb{E}\|Ag\|_2^2 = \operatorname{Tr}[A^*A]$ and both $\pm(\|Ag\|_2^2 - \operatorname{Tr}[A^*A])$ are $(4\|A^*A\|_F^2, 4\|A^*A\|_{\mathrm{op}})$-subexponential. Therefore we can use Lemma 2.5.5 to prove our moment bound.

$$\mathbb{E}\|Ag\|_2^{2p} \leq 2^p(\mathbb{E}|\|Ag\|_2^2 - \operatorname{Tr}[A^*A]|^p + \operatorname{Tr}[A^*A]^p) \leq \frac{\|A^*A\|_F^2}{\|A^*A\|_{\mathrm{op}}^2}(8p\|A^*A\|_{\mathrm{op}})^p + (2\operatorname{Tr}[A^*A])^p,$$

where the second step was by the inequality $(x+y)^p \leq (2\max\{|x|,|y|\})^p \leq 2^p(|x|^p + |y|^p)$, and the final step was using Lemma 2.5.5. □

2.6 Nets and Approximation Arguments

Throughout Chapter 5 and Chapter 9, we will use standard arguments in order to control the supremum of a set of random variables. This will allow us to generalize the spectral bounds of Theorem 2.5.12 to more complicated distributions. For this purpose, we will need the following standard bounds.

In order to perform a union bound over sets, we need the following standard cardinality bound (see e.g. [90]).

Fact 2.6.1. *For $\beta \in [0, \frac{1}{2}]$ we can bound the binomial coefficient:*

$$\log_2 \binom{k}{\beta k} \leq \beta k(1 - \log_2 \beta)$$

In order to control various operator norms of random matrices, we will perform a standard net argument. As an illustration, say we have a set of random variables $\{X_\xi\}_{\xi \in B}$ and we want to control $\sup_{\xi \in B} X_\xi$. If the set S is finite and we have concentration bounds for each X_ξ, then the result would follow by a simple union bound. But this argument no longer works for infinite B. In this case, in order to show strong bounds for every $\xi \in S$, we first discretize the set to $N \subseteq B$ and perform the union bound over every $\xi \in N$. Then we show that $\sup_{\xi \in N} X_\xi$ approximates $\sup_{\xi \in B} X_\xi$ to give the result.

We use the following standard notions to discretize a unit ball B for such an argument.

Definition 2.6.2. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then given subset $B \subseteq \mathbb{R}^d$, $N \subseteq \mathbb{R}^d$ is called an η-net for B if for every element $v \in B$, there exists a nearby $u \in N$ such that*

$$\|u - v\| \leq \eta.$$

$N \subseteq B$ is called an η-packing of B if $\|u - v\| \geq \eta$ for every pair $u, v \in N$.

The following results will help with union bound arguments over nets and packings.

Fact 2.6.3. *[Lemma 4.10 in [78]] Let $\|\cdot\|$ define a norm on $\mathbb{R}^d$ with unit ball and sphere*

$$B := \{v \in \mathbb{R}^d \mid \|v\| \leq 1\} \qquad \text{and} \qquad S := \{v \in \mathbb{R}^d \mid \|v\| = 1\}.$$

For any $\eta > 0$, let $N_p \subseteq S$ be a maximal η-packing of S and $N_c \subseteq S$ be a minimum η-net for S according to Definition 2.6.2. Then

$$|N_c| \leq |N_p| \leq \left(1 + \frac{2}{\eta}\right)^d.$$

For the specific case of the unit ball of $\|\cdot\|_\infty$ or $\|\cdot\|_{\text{op}}$, we can use the following refined characterization for our discretization.

Fact 2.6.4. *The vertices of the polytope*

$$H := 1_n^\perp \cap B_\infty = \{y \in \mathbb{R}^n \mid \sum_{j\in[n]} y_j = 0, -1_n \leq y \leq 1_n\}$$

are of the form $1_S - 1_T$ for disjoint sets $S, T \subseteq [n]$ with $S \cap T = \emptyset$. In particular, $S, T \in \binom{n}{\lfloor \frac{n}{2} \rfloor}$ if n is odd, and $T \in \binom{n}{n/2}$ and $S = [n] - T$ if n is even. Note that for the even case, these vertices can be rewritten $1_n - 21_T = 21_S - 1_n$.

This lifts naturally to the matrix setting as the vertices of

$$I_n^\perp \cap B_{\text{op}} = \{Y \in \mathrm{H}(n) \mid \mathrm{Tr}[Y] = 0, \|Y\|_{\text{op}} \leq 1\}$$

are of the form $P - Q$ for disjoint orthogonal projections $P, Q \in \mathrm{H}(n)$ with $PQ = 0$, i.e. $Im(P) \cap Im(Q) = \emptyset$. In particular, $Q = I_n - P$ for some $\mathrm{rk}(P) = \frac{n}{2}$ projection if n is even, both P, Q are projections with $\mathrm{rk}(P) = \mathrm{rk}(Q) = \lfloor \frac{n}{2} \rfloor$ n is odd. Note that for the even case, these vertices can be rewritten $2P - I_n = I_n - 2Q$.

The final part of this subsection deals with the approximation part of the argument. Specifically, it shows how to translate bounds on $\sup_{\xi \in N} X_\xi$, where $N \subseteq S$ is some appropriate discretization of the Euclidean sphere, into a bound on $\sup_{\xi \in S} X_\xi$.

Lemma 2.6.5. *For $M \in \mathrm{Mat}(n, d)$, if N is an η-net of S^{d-1}, then*

$$\|M\|_{\text{op}} \leq (1 - \eta)^{-1} \sup_{\xi \in N} \|M\xi\|_2.$$

Note this can be rewritten as $\|M^*M\|_{\mathrm{op}} \leq (1-\eta)^{-2} \sup_{\xi \in N} \langle \xi\xi^*, M^*M \rangle$ for positive semi-definite M^*M.

This can be generalized to non-definite $X \in \mathrm{H}(d)$ matrices as follows:

$$\|X\|_{\mathrm{op}} \leq (1-2\eta-\eta^2)^{-1} \sup_{\xi \in N} |\langle \xi\xi^*, X\rangle|.$$

Proof. We follow the standard approximation argument given in [95].

First consider arbitrary $M \in \mathrm{Mat}(d,n)$, and let $\xi_* := \arg\sup_{\xi \in S^{d-1}} \|M\xi\|_2$ so that $\|M\xi_*\|_2 = \|M\|_{\mathrm{op}}$ by definition of the operator norm as discussed in Section 2.1.9. For shorthand let $\mu := \sup_{\xi \in N} \|M\xi\|_2$. N is an η-net, so by Definition 2.6.2 we can decompose $\xi_* = \xi + \xi'$ for some $\xi \in N, \xi' \in \eta B_2^n$. This allows us to bound

$$\|M\|_{\mathrm{op}} = \|M\xi_*\|_2 \leq \|M\xi\|_2 + \|M\xi'\|_2 \leq \mu + \eta\|M\|_{\mathrm{op}},$$

where the first step was by definition of ξ_*, the second step was by the triangle inequality, and in the final step the first term is bounded by definition of μ as $\xi \in N$, and second term is bounded by the definition of the operator norm and $\xi' \in \eta B_2^n$. The statement follows by rearranging:

$$\|M\|_{\mathrm{op}} \leq (1-\eta)^{-1}\mu = (1-\eta)^{-1} \sup_{\xi \in N} \|M\xi\|_2.$$

The second statement follows by a similar calculation except that the triangle inequality has more terms. For any Hermitian matrix $X \in \mathrm{Mat}(d)$, let $\xi_* := \arg\sup_{\xi \in S^{d-1}} |\langle \xi\xi^*, X\rangle|$ so that $\|X\|_{\mathrm{op}} = |\langle \xi_*\xi_*^*, X\rangle|$ by Definition 2.1.16. Further, for shorthand let $\mu := \sup_{\xi \in N} |\langle \xi\xi^*, X\rangle|$. N is an η-net, so by Definition 2.6.2 we can decompose $\xi_* = \xi + \xi'$ for some $\xi \in N, \xi' \in \eta B_2^n$. This allows us to bound

$$\|X\|_{\mathrm{op}} = |\langle \xi_*\xi_*^*, X\rangle| \leq |\langle \xi\xi^*, X\rangle| + 2|\langle \xi'\xi^*, X\rangle| + |\langle \xi'\xi'^*, X\rangle| \leq \mu + (2\|\xi'\|_2\|\xi\|_2 + \|\xi'\|_2^2)\|X\|_{\mathrm{op}},$$

where the first step was by definition of ξ_*, the second step was by the triangle inequality, and in the final step the first term is bounded due to $\xi \in N$, and the next terms are by the dual definition of the operator norm as discussed in Section 2.1.9. Therefore, using $\|\xi'\|_2 \leq \eta$ and rearranging gives the lemma:

$$\|X\|_{\mathrm{op}} \leq \mu + (2\eta + \eta^2)\|X\|_{\mathrm{op}} \implies (1-2\eta-\eta^2)^{-1}\|X\|_{\mathrm{op}} \leq \mu = \sup_{\xi \in N} |\langle \xi\xi^*, X\rangle|.$$

□

We will also want to bound the smallest singular values of random matrices with a similar strategy. Therefore, we will discretize S^{d-1} and perform a similar approximation argument below for $\inf_{\xi \in S^{d-1}} X_\xi$. The following lemma is helpful to bound well-conditioned matrices.

Lemma 2.6.6. *For $X \in \mathrm{Mat}(n, d)$, if N is an η-net of S^{d-1}, then*

$$\inf_{\xi \in S^{d-1}} \|X\xi\|_2^2 \geq \inf_{\xi \in N} \|X\xi\|_2 - \eta \|X\|_{\mathrm{op}}.$$

Proof. Let $\xi_* := \arg\inf_{\xi \in S^{d-1}} \|X\xi\|_2$ be the optimizer. N being an η-net, so by Definition 2.6.2 we can decompose $\xi_* = \xi + \xi'$ for $\xi \in N, \xi' \in \eta B_2^n$. Letting $\sigma := \inf_{\xi \in N} \|X\xi\|_2$ for shorthand, we can bound the above as

$$\inf_{\xi \in S^{d-1}} \|X\xi\|_2 = \|X(\xi + \xi')\|_2 \geq \|X\xi\|_2 - \|X\xi'\|_2 \geq \sigma - \eta \|X\|_{\mathrm{op}},$$

where the first step was by definition of $\xi_* = \xi + \xi'$, the second step was by the triangle inequality, and in the final step we bounded the first term by definition of σ as $\xi \in N$ and the second term by definition of the operator norm. □

Note that $\|X\xi\|_2 \geq 0$ always, so the above bound is only non-trivial when $\eta < \frac{\inf_{\xi \in N} \|X\xi\|_2}{\|X\|_{\mathrm{op}}}$. This will be useful for random matrices that have small condition number.

Chapter 3

Matrix Scaling Improvement

In this chapter we will study the matrix scaling problem.

Definition 3.0.1. *For matrix* $A \in \mathrm{Mat}(d,n)$*, output diagonal matrices* $L \in \mathrm{diag}(d), R \in \mathrm{diag}(n)$ *such that* $B := LAR$ *is doubly balanced, i.e.*

$$\forall i \in [d] : \sum_{j=1}^{n} |B_{ij}|^2 = \frac{\|B\|_F^2}{d}, \qquad \textit{and} \qquad \forall j \in [n] : \sum_{i=1}^{d} |B_{ij}|^2 = \frac{\|B\|_F^2}{n},$$

or prove that no such scaling exists.

Matrix scaling is an important subroutine in many fields of pure and applied mathematics, and has been rediscovered from a variety of perspectives. Our main motivation is to give an optimal analysis for the Paulsen problem in Chapter 4. We will also generalize the results in this chapter to the tensor scaling setting in Chapter 6 and Chapter 7 using the geodesic convex optimization framework developed in [20]. In Chapter 8, we present some background on algorithms for the matrix scaling problem. In this chapter, we focus on proving strong bounds on the solution for certain classes of inputs.

Overview: In Section 3.1, we formally introduce the matrix scaling problem. We then present a well-known convex formulation for this problem, along with the natural gradient flow algorithm used to solve it. We then give an analysis of the gradient flow algorithm using tools from convex optimization. In Section 3.2, we prove strong bounds on the solution to matrix scaling when the input satisfies a strong convexity assumption. In Section 3.3, we prove even stronger bounds when the input satisfies a combinatorial pseudorandom condition. In Section 3.4, we discuss the quantitative relationship between the above two conditions, and in particular show that the pseudorandom condition implies strong convexity. This is a new result in spectral graph theory which we believe to be of independent interest. Finally, in Section 3.5, we describe the ideas necessary to lift this analysis to the more general frame and operator scaling problems, where the set of scalings come from a non-commutative group. These ideas will be explained in more detail in Chapter 6, where we will fully present the geodesic convex optimization framework for more general scaling problems. The main application of the improved analyses of matrix scaling will be given in Chapter 4 for the Paulsen problem.

3.1 Matrix Scaling and Convexity

In this section, we formally define the specific version of matrix scaling that we study. The main goals of this section are to describe the convex formulation and the gradient flow algorithm used to find the solution.

3.1.1 Matrix Scaling

The original matrix scaling problem concerns non-negative square matrices and has an exceedingly long and varied history (see the survey of Idel's for a detailed exposition [54]). We will study a generalization of this problem, where the input is a tuple of rectangular matrices with arbitrary elements from characteristic zero fields $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. We choose this tuple version so that our results can be lifted to the more general frame and operator settings as discussed in Section 6.3 of Chapter 6.

We first define the quantities of interest for matrix scaling.

Definition 3.1.1. *For tuple $A = \{A_1, ..., A_K\} \in \mathrm{Mat}(d, n)^K$, the size is defined as*

$$s(A) := \sum_{k=1}^{K} \|A_k\|_F^2.$$

For $i \in [d]$ and $j \in [n]$, the row and column sums are defined respectively as

$$r_i(A) := \langle E_{ii}, \sum_{k=1}^{K} A_k A_k^* \rangle = \sum_{k=1}^{K} \sum_{j=1}^{n} |(A_k)_{ij}|^2, \quad c_j(A) := \langle E_{jj}, \sum_{k=1}^{K} A_k^* A_k \rangle = \sum_{k=1}^{K} \sum_{i=1}^{d} |(A_k)_{ij}|^2.$$

In the original matrix scaling problem, the goal is to output a non-negative doubly-stochastic matrix (all row and column sums equal to 1). Below, we define a similar condition for our setting.

Definition 3.1.2. *Tuple $A = \{A_1, ..., A_K\} \in \mathrm{Mat}_{\mathbb{C}}(d,n)^K$ is called ε-doubly balanced if*

$$\frac{s(A)(1-\varepsilon)}{d} \leq r_i(A) \leq \frac{s(A)(1+\varepsilon)}{d}, \qquad \textit{and} \qquad \frac{s(A)(1-\varepsilon)}{n} \leq c_j(A) \leq \frac{s(A)(1+\varepsilon)}{n},$$

for all $i \in [d], j \in [n]$. A is called doubly balanced if the above holds with $\varepsilon = 0$.

We can now rephrase the matrix scaling problem in this language.

Definition 3.1.3 (Matrix Scaling Problem)**.** *Given input matrix tuple $A := \{A_1, ..., A_K\} \in \mathrm{Mat}(d,n)^K$, find non-zero scalings $(L, R) \in \mathrm{diag}(d) \oplus \mathrm{diag}(n)$ such that*

$$LAR := \{LA_1R, ..., LA_KR\}$$

is doubly balanced according to Definition 3.1.2.

Remark 3.1.4. *The original matrix scaling problem of Linial et al. [66] has as input a nonnegative matrix $B \in \mathbb{R}_+^{n\times n}$ and requires positive diagonal matrices $L, R \in \mathbb{R}_{++}^{n\times n}$ such that the scaled matrix LBR is doubly stochastic, i.e. that $(LBR)\mathbf{1}_n = (LBR)^T\mathbf{1}_n = \mathbf{1}_n$. This is equivalent to the $K = 1$ case of Definition 3.1.3 on input $A_{ij} := \sqrt{B_{ij}}$, with the added requirement that the ouptut must have size n.*

In some cases, the only solution to the matrix scaling problem is the trivial scaling $(0,0)$. For a discussion of this failure case, and its combinatorial consequences, see the work of Linial, Samorodnitsky, and Wigderson [66]). For further context on the 0 solution to scaling problems, see the discussion on the Null Cone in Section 6.1.2.

If the input is already close to doubly balanced, we can hope to give a refined analysis of the matrix scaling problem in Definition 3.1.3. The goal of Section 3.2 and Section 3.3 is to define and analyze sufficient conditions for nearly doubly balanced inputs to have scaling solutions that are close to the identity.

Our strategy will be to analyze the natural gradient flow algorithm for a convex formulation of matrix scaling. These concepts will be formally defined in the next two subsections.

3.1.2 Convex Formulation/Kempf-Ness Function

In this section, we will present the convex formulation for matrix scaling. The formulation comes from the work of Kempf and Ness [58] in the context of geometric invariant theory, and we discuss this connection in more detail in Section 6.1.3.

We first simplify the domain of scalings. We can perform a change of variable $(L, R) \to (e^X, e^Y)$ for $X \in \mathrm{diag}(d), Y \in \mathrm{diag}(n)$. Note that the trivial solution $(L, R) = (0, 0)$ is no longer feasible (see Remark 3.1.4), but our focus in this chapter is on sufficient conditions for scaling solutions, so this failure case will not be of concern to us. Next, we observe that the row and column sums in Definition 3.1.2 depend only on the magnitude of entries, so we can ignore the sign and complex phase of scalings and restrict our attention to $(X, Y) \in \mathrm{diag}_{\mathbb{R}}(d) \oplus \mathrm{diag}_{\mathbb{R}}(n)$. (This is an instance of the polar decomposition $\mathbb{C} = \mathbb{R} \oplus i\mathbb{R}$ as discussed in Theorem 2.1.13, and we will revisit this for more general scaling problems in Chapter 6.) Finally, we can assume the normalization $\sum_{i=1}^{d} X_i = \sum_{j=1}^{n} Y_j = 0$ without loss, as the doubly balanced condition is homogeneous. Note that this is equivalent to restricting to unit determinant scalings as $\det(e^X) = \exp(\mathrm{Tr}[X])$. Therefore we can restrict the domain of the matrix scaling problem as follows.

Definition 3.1.5. *The scalings in Definition 3.1.3 can be restricted to subspace*

$$\mathfrak{t} := \{(X, Y) \in \mathrm{diag}_{\mathbb{R}}(d) \oplus \mathrm{diag}_{\mathbb{R}}(n) \mid \mathrm{Tr}[X] = \mathrm{Tr}[Y] = 0\}.$$

We will sometimes use X to refer to its embedding $(X, 0) \in \mathfrak{t}$ by abuse of notation (and similarly for $Y \to (0, Y) \in \mathfrak{t}$).

This vector space can be derived more formally using the perspective of Lie groups and Lie algebras as discussed in Section 2.2.3, and these ideas lift to the more general tensor scaling setting as shown in Chapter 6 and Chapter 7. At this point, we can introduce the Kempf-Ness function which gives an optimization formulation for matrix scaling.

Definition 3.1.6. *For matrix tuple $\{A_1, ..., A_K\} \in \mathrm{Mat}(d, n)^K$, the Kempf-Ness function $f_A : \mathfrak{t} \to \mathbb{R}$ is defined as*

$$f_A(X, Y) := s(e^{X/2} A e^{Y/2}) = \sum_{k=1}^{K} \|e^{X/2} A_k e^{Y/2}\|_F^2 = \sum_{k=1}^{K} \sum_{i=1}^{d} \sum_{j=1}^{n} e^{X_i} |A_k|_{ij}^2 e^{Y_j},$$

where size is given in Definition 3.1.1. The factor 2 is just to remove leading constants for future calculations.

Below, we prove that the Kempf-Ness function gives a convex formulation for matrix scaling by showing (1) the set of doubly balanced scalings are exactly the critical points of f_A, and (2) f_A is convex on its domain. This is actually a general phenomena in the setting of geometric invariant theory, and we discuss this connection to the work of Kempf and Ness [58] in Section 6.1.3. Therefore, in order to analyze the solution of the matrix scaling problem, we can rely on tools from convex optimization. Properties (1) and (2) are verified by the simple derivative calculations below.

We will repeatedly use the following property of the Kempf-Ness function to reduce all calculations to the origin.

Fact 3.1.7 (Equivariance)**.** *The family of Kempf-Ness functions $\{f_A \mid A \in \mathrm{Mat}(d,n)^K\}$ given in Definition 3.1.6 satisfies the following equivariance relation:*

$$f_A(X,Y) = s(e^{X/2}Ae^{Y/2}) = f_{e^{X/2}Ae^{Y/2}}(0,0).$$

We can therefore characterize critical points of the Kempf-Ness function by a straightforward first-order calculation at the origin.

Lemma 3.1.8. *For input $A \in \mathrm{Mat}(d,n)^K$, the (X,Y)-directional derivative of f_A is*

$$\partial_{\delta=0} f_A(\delta X, \delta Y) = \sum_{i=1}^{d} X_i r_i(A) + \sum_{j=1}^{n} Y_j c_j(A).$$

Therefore, doubly balanced scalings of A correspond to critical points of f_A.

Proof. We first expand f_A to calculate the first derivative:

$$\partial_{\delta=0} f_A(\delta X, \delta Y) = \partial_{\delta=0} \sum_{i=1}^{d}\sum_{j=1}^{n} e^{\delta(X_i+Y_j)} \sum_{k=1}^{K} |\langle E_{ij}, A_k\rangle|^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} (X_i+Y_j) \sum_{k=1}^{K} |\langle E_{ij}, A_k\rangle|^2$$
$$= \sum_{i=1}^{d} X_i \sum_{k=1}^{K}\sum_{j=1}^{n} |\langle E_{ij}, A_k\rangle|^2 + \sum_{j=1}^{n} Y_j \sum_{k=1}^{K}\sum_{j=1}^{n} |\langle E_{ij}, A_k\rangle|^2.$$

The first statement in the lemma then follows by Definition 3.1.1 of row/column sums.

For the second statement, note that by the equivariance property in Fact 3.1.7, (X,Y) is a critical point of f_A iff the origin $(0,0)$ is a critical point for $f_{e^{X/2}Ae^{Y/2}}$. Therefore it is enough to show that $A \in \mathrm{Mat}(d,n)^K$ is doubly balanced iff $(0,0)$ is a critical point of f_A.

We first show that if A is doubly balanced, then the derivative vanishes for every direction in $\mathfrak{t}$, which gives the forward implication by Definition 2.3.9 of a critical point. So considering $(X, Y) \in \mathfrak{t}$,

$$\partial_{\delta=0} f_A(\delta X, \delta Y) = \sum_{i=1}^{d} X_i r_i(A) + \sum_{j=1}^{n} Y_j c_j(A) = \sum_{i=1}^{d} X_i \frac{s(A)}{d} + \sum_{j=1}^{n} Y_j \frac{s(A)}{n} = 0,$$

where the first equality is by the first order calculation above, the second equality is because A is doubly balanced (Definition 3.1.2), and the final equality is because $(X, Y) \in \mathfrak{t}$ so $\sum_{i=1}^{d} X_i = \sum_{j=1}^{n} Y_j = 0$.

To show the converse implication, assume A is not doubly balanced so $(X, Y) := (r(A) - \frac{s(A)}{d}, c(A) - \frac{s(A)}{n}) \neq 0$. Note that this vector is in $\mathfrak{t}$, as

$$\sum_{i=1}^{d} r_i = \sum_{k=1}^{K} \sum_{i=1}^{d} \sum_{j=1}^{n} |(A_k)_{ij}|^2 = \sum_{k=1}^{K} \|A_k\|_F^2 = s(A),$$

by definition of the Frobenius norm on $\operatorname{Mat}(d, n)$. The same calculation shows that $\sum_{j=1}^{n} c_j = s(A)$, so $(X, Y) \in \mathfrak{t}$ by Definition 3.1.5. We will show that the derivative in the (X, Y) direction does not vanish, which shows that the origin is not a critical point by Definition 2.3.9. By the derivative calculation above,

$$\partial_{\delta=0} f_A(\delta X, \delta Y) = \sum_{i=1}^{d} \left(r_i - \frac{s}{d}\right) r_i + \sum_{j=1}^{n} \left(c_j - \frac{s}{n}\right) c_j = \sum_{i=1}^{d} \left(r_i - \frac{s}{d}\right)^2 + \sum_{j=1}^{n} \left(c_j - \frac{s}{n}\right)^2,$$

where in the first step we substituted $(X, Y) := (r(A) - \frac{s(A)}{d}, c(A) - \frac{s(A)}{n})$, and the last equality is because $\frac{s}{d} \sum_{i=1}^{d} \left(r_i - \frac{s}{d}\right) = \frac{s}{n} \sum_{j=1}^{n} \left(c_j - \frac{s}{n}\right) = 0$. The above is strictly positive since $(X, Y) \neq 0$, so the origin is not a critical point for f_A. □

We next calculate the second derivative to verify that the Kempf-Ness function is convex according to Definition 2.3.8.

Lemma 3.1.9. *For $A \in \operatorname{Mat}(d, n)^K$ and direction $(X, Y) \in \mathfrak{t}$, the second derivative of f_A at the origin is:*

$$\partial_{\delta=0}^2 f_A(\delta X, \delta Y) = \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{K} |\langle E_{ij}, A_k \rangle|^2 (X_i + Y_j)^2.$$

As a consequence, f_A is convex on domain $\mathfrak{t}$ for every input A.

Proof. The first statement follows by expanding f_A as

$$\partial^2_{\delta=0} f_A(\delta X, \delta Y) = \partial^2_{\delta=0} \sum_{i=1}^{d} \sum_{j=1}^{n} e^{\delta(X_i+Y_j)} \sum_{k=1}^{K} |\langle E_{ij}, A_k\rangle|^2 = \sum_{i=1}^{d} \sum_{j=1}^{n} (X_i+Y_j)^2 \sum_{k=1}^{K} |\langle E_{ij}, A_k\rangle|^2.$$

For the second statement, the equivariance property in Fact 3.1.7 shows that f_A is convex at $(X, Y) \in \mathfrak{t}$ iff $f_{e^{X/2}Ae^{Y/2}}$ is convex at the origin. The first statement in this lemma shows that the second order derivative is always non-negative at the origin for every direction in $\mathfrak{t}$, so f is convex by Definition 2.3.8. □

We can therefore collect the above facts into the following proposition, which shows that the Kempf-Ness function is the desired convex formulation for matrix scaling.

Proposition 3.1.10. *For every input $A \in \mathrm{Mat}(d, n)^K$,*

1. *f_A is convex on domain $\mathfrak{t}$.*
2. *$e^{X/2}Ae^{Y/2}$ is a doubly balanced scaling of A iff (X, Y) is a critical point for f_A.*
3. *$e^{X/2}Ae^{Y/2}$ is a doubly balanced scaling of A iff (X, Y) is a global minimizer of f_A.*

Proof. (1) and (2) are exactly the content of Lemma 3.1.9 and Lemma 3.1.8 respectively. The final item also follows from Lemma 2.3.10, which shows critical points of convex functions are always global minima. □

3.1.3 Gradient Flow

The formulation in Proposition 3.1.10 shows that the matrix scaling solution is an optimizer of the convex Kempf-Ness function given in Definition 3.1.6. Therefore in this subsection we formally define a natural gradient flow which converges to the optimizer of f.

Definition 2.3.12 specifies that for inner product space $(V, \langle\cdot,\cdot\rangle)$, the gradient ∇h of differentiable function $h : V \to \mathbb{R}$ at point $x \in V$ satisfies

$$\forall v \in V : \langle \nabla h(x), v\rangle = \partial_{\delta=0} h(x + \delta v).$$

Any choice of (positive-definite) inner product on $\mathfrak{t}$ will induce a unique gradient vector at each point, and will therefore induce a different gradient flow. We will choose an inner product that corresponds with the scaling properties of $\mathfrak{t}$. This defines a gradient vector field of f_A on $\mathfrak{t}$, and the gradient flow is then defined as the solution to the differential equation produced by this vector field.

Definition 3.1.11 ($\mathfrak{t}$ Inner Product)**.** *For elements $(X, Y), (X', Y')$ in vector space $\mathfrak{t}$ (Definition 3.1.5), we define their inner product as*

$$\langle (X, Y), (X', Y') \rangle_{\mathfrak{t}} := \frac{1}{d} \sum_{i=1}^{d} X_i X_i' + \frac{1}{n} \sum_{j=1}^{n} Y_j Y_j'.$$

The induced norm is $\|(X, Y)\|_{\mathfrak{t}} = \sqrt{\langle (X, Y), (X, Y) \rangle_{\mathfrak{t}}}$.

Similar to Definition 3.1.5, the above inner product is natural from the appropriate scaling perspective. We give further explanation after Definition 7.1.2, where this inner product is lifted to the tensor scaling setting.

With this choice of $\mathfrak{t}$-inner product, we can define the gradient vector at each point.

Proposition 3.1.12. *For input $A \in \mathrm{Mat}(d, n)^K$, and $(X, Y) \in \mathfrak{t}$, the gradient is*

$$\nabla f_A(X, Y) = \left\{ d \cdot r_i(e^{X/2} A e^{Y/2}) - s(e^{X/2} A e^{Y/2}) \right\}_{i=1}^{d} \oplus \left\{ n \cdot c_j(e^{X/2} A e^{Y/2}) - s(e^{X/2} A e^{Y/2}) \right\}_{j=1}^{n}.$$

We will often use shorthand $\nabla_A := \nabla f_A(0, 0)$ and ∇_A^L, ∇_A^R for the left and right parts of ∇_A, respectively.

Proof. First note that $\nabla f_A(X, Y)$ is in fact an element of $\mathfrak{t}$, as $d \sum_{i=1}^{d} r_i = s = n \sum_{j=1}^{n} c_j$ by the calculation in Lemma 3.1.8 . To verify the formula above, we first reduce our calculation to the origin by noting

$$\partial_{\delta=0} f_A(X + \delta X', Y + \delta Y') = \partial_{\delta=0} f_{e^{X/2} A e^{Y/2}}(\delta X', \delta Y')$$

by the equivariance property in Fact 3.1.7. This induces the relation $\nabla f_A(X, Y) = \nabla f_{e^{X/2} A e^{Y/2}}(0, 0)$, so it is enough to verify the formula for the gradient at the origin for every $A \in \mathrm{Mat}(d, n)^K$. For arbitrary $(X, Y) \in \mathfrak{t}$, Lemma 3.1.8 gives

$$\partial_{\delta=0} f_A(\delta X, \delta Y) = \sum_{i=1}^{d} X_i r_i(A) + \sum_{j=1}^{n} Y_j c_j(A) = \frac{1}{d} \sum_{i=1}^{d} X_i (d \cdot r_i(A) - s(A)) + \frac{1}{n} \sum_{j=1}^{n} Y_j (n \cdot c_j(A) - s(A)),$$

where the last equality was because $(X, Y) \in \mathfrak{t}$ so $\sum_{i=1}^{d} X_i \cdot s(A) = \sum_{j=1}^{n} Y_j \cdot s(A) = 0$. Matching this last expression to the definition $\partial_{\delta=0} f_A(\delta X, \delta Y) = \langle \nabla f_A(0, 0), (X, Y) \rangle_{\mathfrak{t}}$ gives the statement. □

This choice of inner product and gradient is well suited to analyze matrix scaling as shown by the following approximate version of Lemma 3.1.8.

Fact 3.1.13. *For ε-doubly balanced $A \in \mathrm{Mat}(d,n)^K$, the gradient satisfies the norm bound*

$$\|\nabla_A\|_{\mathfrak{t}}^2 = \frac{1}{d}\sum_{i=1}^{d}(dr_i - s)^2 + \frac{1}{n}\sum_{j=1}^{n}(nc_j - s)^2 \leq \frac{1}{d}\sum_{i=1}^{d}(s\varepsilon)^2 + \frac{1}{n}\sum_{j=1}^{n}(s\varepsilon)^2 = 2s^2\varepsilon^2,$$

and $\|\nabla_A\|_{\mathfrak{t}}^2 = 0$ iff A is doubly balanced.

Therefore, our goal will be to find sufficient conditions for the optimizer of an approximate critical point to be close to the origin. To show a distance bound on the optimizer, we will follow gradient flow of f_A. Informally, at each time t we would like to move our scalings (X_t, Y_t) infinitesimally in the direction of steepest descent.

Definition 3.1.14 (Gradient Flow)**.** *For input $A \in \mathrm{Mat}(d,n)^K$, the gradient flow of the Kempf-Ness function f_A is the dynamical system $\{(X_t, Y_t) \in \mathfrak{t} \mid t \geq 0\}$ satisfying*

$$(X_0, Y_0) := (0,0), \qquad \partial_t(X_t, Y_t) = -\nabla f_A(X_t, Y_t).$$

This induces a dynamical system on matrices by $A_t := e^{X_t/2}Ae^{Y_t/2}$ with $A_0 = A$. By the equivariance property of Fact 3.1.7, we equivalently have $\partial_t(X_t, Y_t) = -\nabla f_{A_t}(0,0)$.

In this chapter, we reserve t and ∂_t exclusively for time variables, and we use Greek letters for directional derivatives on the vector space (e.g. $\partial_t A_t$ vs $\partial_\delta f_A(\delta X, \delta Y)$), so as not to confuse the domains.

One advantage of using this simple gradient flow algorithm is that certain quantities controlling convergence to the optimum can be analyzed simply. The proposition below gives a principled derivation of Lemma 3.4.2 in [62] while (slightly) simplifying the proof.

Proposition 3.1.15. *For matrix input $A \in \mathrm{Mat}(d,n)^K$ and gradient flow (X_t, Y_t) as in Definition 3.1.14,*

$$\partial_t s(A_t) = \partial_t f_A(X_t, Y_t) = -\|\nabla_{A_t}\|_{\mathfrak{t}}^2.$$

Proof. The first equality follows since $s(A_t) = f_A(X_t, Y_t)$ for all time by Definition 3.1.14 of $A_t = e^{X_t/2}Ae^{Y_t/2}$. To show the second equality, we calculate

$$\partial_t f_A(X_t, Y_t) = \langle \nabla f_A(X_t, Y_t), \partial_t(X_t, Y_t)\rangle_{\mathfrak{t}} = \langle \nabla f_A(X_t, Y_t), -\nabla f_A(X_t, Y_t)\rangle_{\mathfrak{t}} = -\|\nabla_{A_t}\|_{\mathfrak{t}}^2,$$

where the first step is by the chain rule, in the second step we used Definition 3.1.14 of gradient flow on (X_t, Y_t), and in the last equality we again used $A_t = e^{X_t/2}Ae^{Y_t/2}$ so $\nabla_{A_t} = \nabla f_A(X_t, Y_t)$. □

Our analyses in the next two sections will proceed by showing $\|\nabla_{A_t}\|_{\mathfrak{t}}$ decreases exponentially under special assumptions on the input. By Definition 3.1.14 of $\partial_t(X_t, Y_t) = -\nabla_{A_t}$, this allows us to bound the path length of gradient flow, showing (X_t, Y_t) stays close to the origin. Further, Proposition 3.1.15 will be useful in showing strong lower bounds on the objective function f_A in this case.

3.2 Strongly Convex Setting

In the previous Section 3.1, we showed that there is a convex formulation for matrix scaling. In this section, we will analyze the gradient flow given in Definition 3.1.14 when the input satisfies a strong convexity condition. There are many well known techniques which show fast convergence of various descent methods (see e.g. [75]), but these tend to apply to functions that are strongly convex on their whole domain. In Section 3.2.1, we define a notion of strong convexity for matrix inputs. Then, in Section 3.2.2 we show that strong convexity is maintained if all entries of the scaling (X, Y) are small. This leads to a preliminary convergence analysis of gradient flow for matrix inputs A which are sufficiently strongly convex as compared to their initial error $\|\nabla_A\|_{\mathfrak{t}}$. Finally, in Section 3.2.3, we make an important structural observation about matrix scaling which allows us to directly analyze the convergence of the worst error $\|\nabla_{A_t}\|_\infty$ over time. This allows us to show the same fast convergence guarantee with a much weaker requirement on strong convexity. The improvement from Section 3.2.2 to Section 3.2.3 is done by going beyond standard strong convexity analyses and directly considering the ∞-norm of the gradient, which is better suited for analyzing convergence.

In Section 3.3, we will further improve the convergence analysis when the input satisfies a certain pseudorandom condition. In Section 3.5, we will discuss how to lift both of these results to the more general frame and operator scaling problems.

3.2.1 Strong Convexity

In this subsection, we will define strong convexity for matrix scaling and show some preliminary convergence results that follow from standard convex analysis.

Definition 3.2.1. *Matrix tuple $A \in \mathrm{Mat}(d, n)^K$ is α-strongly convex iff f_A is α strongly convex at the origin:*

$$\forall (X, Y) \in \mathfrak{t} : \quad \partial^2_{\delta=0} f_A(\delta X, \delta Y) \geq \alpha \|(X, Y)\|^2_{\mathfrak{t}} = \alpha \left(\frac{1}{d} \sum_{i=1}^{d} X_i^2 + \frac{1}{n} \sum_{j=1}^{n} Y_j^2 \right).$$

Notice that unlike Definition 3.1.2 of doubly balanced matrices, this concept is not homogeneous. Therefore, in general the amount of strong convexity should be compared to the size. A simple motivating example is the all-ones matrix $\frac{1}{dn}J$, which has size 1 and satisfies $\alpha = 1$ strong convexity. In Appendix A.2, we show that this is in fact an extremal example with maximum α/s.

From a graph-theoretic perspective, strong convexity of a matrix tuple A can be related to the graph expansion of the bipartite graph with edge weights $w_{ij} := \sum_{k=1}^{K} |A_{ij}|^2$. We will use this connection to graphs in Section 3.4, where we compare strong convexity and the pseudorandom condition of Section 3.3.

This strong convexity assumption is immediately useful in analyzing gradient flow. The following is a standard result from convex analysis.

Proposition 3.2.2. *If A is α-strongly convex then $-\partial_{t=0}\|\nabla_{A_t}\|_{\mathfrak{t}}^2 \geq \alpha\|\nabla_A\|_{\mathfrak{t}}^2$. In particular $\partial_{t=0}\|\nabla_{A_t}\|_{\mathfrak{t}}^2 \leq 0$ always for matrix gradient flow according to Definition 3.1.14. As a corollary, if A_t is α-strongly convex for all $t \in [0, T]$, then*

$$\|\nabla_{A_T}\|_{\mathfrak{t}}^2 \leq e^{-2\alpha T}\|\nabla_A\|_{\mathfrak{t}}^2, \qquad \textit{and} \qquad \|(X_T, Y_T)\|_{\mathfrak{t}} \leq \frac{\|\nabla_A\|_{\mathfrak{t}}}{\alpha},$$

where (X_t, Y_t) is the solution to gradient flow given in Definition 3.1.14.

Proof. We first show $-\partial_{t=0}\|\nabla_{A_t}\|_{\mathfrak{t}}^2 = 2\partial_{\delta=0}^2 f_A(-\delta\nabla_A)$. This will imply the first statement by strong convexity. Starting from the left hand side, we calculate

$$\begin{aligned}
\frac{-\partial_{t=0}\|\nabla_{A_t}\|_{\mathfrak{t}}^2}{2} &= \langle \partial_{t=0}\nabla_{A_t}, -\nabla_A\rangle_{\mathfrak{t}} = \lim_{t\to 0} t^{-1}\langle \nabla_{A_t} - \nabla_A, -\nabla_A\rangle_{\mathfrak{t}} \\
&= \lim_{t\to 0} t^{-1}\Big(\partial_{\delta=0} f_{A_t}(-\delta\nabla_A) - \partial_{\delta=0} f_A(-\delta\nabla_A)\Big) \\
&= \lim_{t\to 0} t^{-1}\Big(\partial_{\delta=0} f_A((X_t, Y_t) - \delta\nabla_A) - \partial_{\delta=0} f_A(-\delta\nabla_A)\Big) \\
&= \partial_{t=0}\partial_{\delta=0} f_A\Big((X_t, Y_t) - \delta\nabla_A\Big),
\end{aligned}$$

where the first two steps are by calculus, in the third step we used Definition 2.3.12 of the gradient of f so that $\langle \nabla f_{A_t}(0,0), -\nabla_A\rangle_{\mathfrak{t}} = \partial_{\delta=0} f_{A_t}(-\delta\nabla_A)$, the fourth equality was by the equivariance property of Fact 3.1.7 as $A_t := e^{X_t/2}Ae^{Y_t/2}$ and $(X_0, Y_0) = (0, 0)$, and the final step is again by calculus. To show this is equal to the right hand side, we calculate

$$\begin{aligned}
\partial_{\delta=0}^2 f_A(-\delta\nabla_A) = \partial_\delta\Big(\partial_\delta f_A(-\delta\nabla_A)\Big)|_{\delta=0} &= \partial_{\delta=0}\langle \nabla f_A(-\delta\nabla_A), -\nabla_A\rangle_{\mathfrak{t}} \\
&= \partial_{\delta=0}\langle \nabla f_A(-\delta\nabla_A), \partial_{t=0}(X_t, Y_t)\rangle_{\mathfrak{t}} = \partial_{\delta=0}\partial_{t=0} f_A\Big((X_t, Y_t) - \delta\nabla_A\Big),
\end{aligned}$$

where the second step used Definition 2.3.12 of the gradient of f_A at point $-\delta\nabla_A$, the third step was by Definition 3.1.14 of gradient flow, and the final step was by Definition 2.3.12 of the gradient as well as the chain rule.

Therefore, we can show the first statement, $-\partial_{t=0}\|\nabla_{A_t}\|_{\mathfrak{t}}^2 = 2\partial_{\delta=0}^2 f_A(-\delta\nabla_A) \geq 2\alpha\|\nabla_A\|_{\mathfrak{t}}^2$ by Definition 3.2.1 of strong convexity.

Equivalently, $-\partial_{t=0}\log\|\nabla_{A_t}\|_{\mathfrak{t}}^2 \geq 2\alpha$ by the chain rule. This implies the second statement, as

$$\log\|\nabla_{A_T}\|_{\mathfrak{t}}^2 - \log\|\nabla_A\|_{\mathfrak{t}}^2 = \int_{t=0}^{T} \partial_t \log\|\nabla_{A_t}\|_{\mathfrak{t}}^2 \leq -2\alpha T,$$

where the first step was by the fundamental theorem of calculus, and the second was by fast convergence. Therefore, exponentiating both sides gives the result.

The final statement on scaling (X_T, Y_T) is also a consequence of the fundamental theorem of calculus.

$$\|(X_T, Y_T)\|_{\mathfrak{t}} = \left\| \int_0^T -\nabla_{A_t} \right\|_t \leq \int_0^T \|\nabla_{A_t}\|_{\mathfrak{t}} \leq \|\nabla_A\|_{\mathfrak{t}} \int_0^T e^{-\alpha t} \leq \frac{\|\nabla_A\|_{\mathfrak{t}}}{\alpha},$$

where in the first step we used $(X_0, Y_0) = 0$ and $\partial_t(X_t, Y_t) = -\nabla_{A_t}$ as given in Definition 3.1.14 of gradient flow, the second step is by the triangle inequality on $\|\cdot\|_{\mathfrak{t}}$, the third step was by using $\|\nabla_{A_t}\|_{\mathfrak{t}}^2 \leq e^{-2\alpha t}\|\nabla_A\|_{\mathfrak{t}}^2$ as shown above, and the final step was by integration. □

We have shown in Proposition 3.2.2 that if A maintains strong convexity according to Definition 3.2.1 throughout the trajectory of gradient flow, then we have exponential convergence of $\|\nabla_{A_t}\|_{\mathfrak{t}}$, which implies a strong bound on the scaling $\|(X_t, Y_t)\|_{\mathfrak{t}}$. The work of the next Section 3.2.2 is to study how the strong convexity property changes over time.

3.2.2 Maintaining Strong Convexity

It will be difficult to have control over the entire trajectory of gradient flow, so the main work in this subsection will be to prove that if strong convexity is sufficiently large at time $t = 0$, then it remains large throughout gradient flow.

To this end, we observe that small scalings will preserve strong convexity. We define the following measurement of scalings under which convexity is quantitatively robust.

Definition 3.2.3. *For vector space $\mathfrak{t}$, the infinity norm is defined as*

$$\|(X,Y)\|_\infty := \max_{i\in[d]} |X_i| + \max_{j\in[n]} |Y_j|.$$

Similar to Definition 3.1.11, the above norm is related to the natural operator norm of scaling $(X,Y) \in \mathfrak{t}$. We further explanation this choice after Definition 7.1.12, where this norm is lifted to the tensor scaling setting.

This norm gives a way to bound the change in convexity caused by scalings.

Lemma 3.2.4 (Robustness)**.** *If $A \in \mathrm{Mat}(d,n)^K$ is α-strongly convex, then for any $(X',Y') \in \mathfrak{t}$, the scaling $B = e^{X'/2}Ae^{Y'/2}$ is at least $\alpha \cdot e^{-\|(X',Y')\|_\infty}$-strongly convex.*

Proof. We can lower bound each entry of the scaling, i.e. for any $i \in [d], j \in [n], k \in K$:

$$|(B_k)_{ij}|^2 = e^{X'_i}|(A_k)_{ij}|^2 e^{Y'_j} \geq e^{-\|(X',Y')\|_\infty}|(A_k)_{ij}|^2,$$

where we substituted in $B = e^{X'/2}Ae^{Y'/2}$ in the first step, and the last step was by Definition 3.2.3 of the infinity norm $\|\cdot\|_\infty$. Therefore we can lower bound the second-order derivative for arbitrary direction $(X,Y) \in \mathfrak{t}$.

$$\begin{aligned}
\partial^2_{\eta=0} f_B(\eta X, \eta Y) &= \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{K} |(B_k)_{ij}|^2 (X_i + Y_j)^2 \\
&\geq \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{K} e^{-\|(X',Y')\|_\infty}|(A_k)_{ij}|^2 (X_i + Y_j)^2 \\
&= e^{-\|(X',Y')\|_\infty} \cdot \partial^2_{\eta=0} f_A(\eta X, \eta Y) \geq e^{-\|(X',Y')\|_\infty} \cdot \alpha \|(X,Y)\|^2_{\mathfrak{t}},
\end{aligned}$$

where the first and third steps were by the formula in Lemma 3.1.9 for second order derivatives, the second step was by the entry-wise bound derived above, and the lower bound in the last step was by α-strong convexity of A. Since the direction $(X,Y) \in \mathfrak{t}$ was arbitrary, this verifies Definition 3.2.1 showing B is $e^{-\|(X',Y')\|_\infty} \cdot \alpha$-strongly convex. □

Remark 3.2.5. *We show in Appendix A.2 that the factor $e^{-\|\delta\|_\infty}$ is in fact optimal. In Lemma 7.3.11, following Section 3.6 of [63], we prove a weaker robustness statement ($\alpha \to \alpha - O(\delta)$) for the more general frame and operator scaling problems. The difference between these two kinds of robustness, additive vs multiplicative, is discussed further in Section 7.3. While it looks like a small change, this difference was an important impetus for the improvements to the Paulsen problem given in this thesis. This is discussed further at the end of Section 4.2.2.*

Remark 3.2.6. *By a similar calculation, we can in fact show the stronger statement*

$$e^{-\|\delta\|_\infty}\partial^2_{\eta=0} f_A(\eta X, \eta Y) \leq \partial^2_{\eta=0} f_B(\eta X, \eta Y) \leq e^{\|\delta\|_\infty}\partial^2_{\eta=0} f_A(\eta X, \eta Y)$$

for any $(X, Y) \in \mathfrak{t}$. *This result was derived in [26] under the name of second-order robustness with respect to* $\|\cdot\|_\infty$, *and was used to give a nearly-linear time algorithm for the* $K = 1$ *matrix scaling case.*

In the next lemma, we give a two-sided relation between $\|\cdot\|_{\mathfrak{t}}$ and $\|\cdot\|_\infty$. This allows us to combine the bound on $\|(X_t, Y_t)\|_{\mathfrak{t}}$ given by Proposition 3.2.2, with the robustness property of Lemma 3.2.4 with respect to $\|(X, Y)\|_\infty$.

Lemma 3.2.7. *For vector space* $\mathfrak{t}$, *the two norms* $\|\cdot\|_{\mathfrak{t}}$ *and* $\|\cdot\|_\infty$ *are related by*

$$\|(X, Y)\|^2_{\mathfrak{t}} \leq \|X\|^2_\infty + \|Y\|^2_\infty \leq \|(X, Y)\|^2_\infty \leq \left(\sqrt{d}\|X\|_{\mathfrak{t}} + \sqrt{n}\|Y\|_{\mathfrak{t}}\right)^2 \leq (d+n)\|(X, Y)\|^2_{\mathfrak{t}}.$$

Proof. To show that the first inequality, we calculate

$$\|(X, Y)\|^2_{\mathfrak{t}} = \frac{1}{d}\sum_{i=1}^d X_i^2 + \frac{1}{n}\sum_{j=1}^n Y_j^2 \leq \frac{d}{d}\max_i X_i^2 + \frac{n}{n}\max_j Y_j^2 = \|X\|^2_\infty + \|Y\|^2_\infty \leq \|(X, Y)\|^2_\infty,$$

where the first step is by Definition 3.1.11 of the inner product, the third step is by Definition 3.2.3, and the final step is by $a^2 + b^2 \leq (a+b)^2$ for $a, b \geq 0$. To show the other inequality, we calculate

$$\begin{aligned}\|(X, Y)\|^2_\infty = (\max_{i\in[d]}|X_i| + \max_{j\in[n]}|Y_j|)^2 &\leq \left(\sqrt{\frac{d}{d}\sum_{i=1}^d X_i^2} + \sqrt{\frac{n}{n}\sum_{j=1}^n Y_j^2}\right)^2 \\ &\leq (d+n)\left(\frac{1}{d}\sum_{i=1}^d X_i^2 + \frac{1}{n}\sum_{j=1}^n Y_j^2\right) = (d+n)\|(X, Y)\|^2_{\mathfrak{t}},\end{aligned}$$

where the first step was by Definition 3.2.3, the third step was by Cauchy-Schwarz, and the final step was by Definition 3.1.11 of the $\mathfrak{t}$-norm. □

At this point we can show a weak form of our main convergence theorem. The stronger version is given in Theorem 3.2.19, and will follow by a refined analysis that avoids the translation in Lemma 3.2.7 and directly controls convergence of $\|(X_t, Y_t)\|_\infty$.

Theorem 3.2.8. *If A is $\alpha \geq 6\sqrt{d+n}\|\nabla_A\|_\infty$-strongly convex, then $(X_\infty, Y_\infty) := \lim_{t\to\infty}(X_t, Y_t)$ exists and gives a solution to the matrix scaling problem in Definition 3.1.3 on input A.*

Proof. We claim that A_t is $\frac{\alpha}{e}$-strongly convex for all time. So for contradiction, let T be the first time A_T is $\leq \frac{\alpha}{e}$-strongly convex. Since A is α-strongly convex, Lemma 3.2.4 in the contrapositive shows that $\|(X_T, Y_T)\|_\infty \geq 1$. By definition of T, A_t is at least $\frac{\alpha}{e}$-strongly convex for all $t \in [0, T]$. Therefore, we can bound

$$\|(X_T, Y_T)\|_\infty \leq \sqrt{d+n}\|(X_T, Y_T)\|_{\mathfrak{t}} \leq \frac{\sqrt{d+n}\|\nabla_A\|_{\mathfrak{t}}}{\alpha/e} \leq \frac{e}{6} < 1,$$

where the first step is by Lemma 3.2.7, the second is by the strong convexity analysis in Proposition 3.2.2, and the third step is by the assumption $\alpha \geq 6\sqrt{d+n}\|\nabla_A\|_\infty$. This is a contradiction, so the claim is shown.

To show that the limit exists, we can use strong convexity for all time to bound

$$\lim_{T\to\infty} \int_{t\geq T} \|\partial_t(X_t, Y_t)\|_{\mathfrak{t}} = \lim_{T\to\infty} \int_{t\geq T} \|\nabla_{A_t}\|_{\mathfrak{t}} \leq \lim_{T\to\infty} \|\nabla_A\|_{\mathfrak{t}} \int_{t\geq T} e^{-\alpha t/e} = 0,$$

where the first step was by Definition 3.1.14 of gradient flow, and the second was by Proposition 3.2.2 with $\frac{\alpha}{e}$-strong convexity. This implies that $\lim_{t\to\infty}(X_t, Y_t) = (X_\infty, Y_\infty) \in \mathfrak{t}$ exists, and $\nabla_{A_\infty} = 0$ for $A_\infty = e^{X_\infty/2} A e^{Y_\infty/2}$, so A_∞ is doubly balanced by Proposition 3.1.10(2). □

In the remainder of this chapter, we make two kinds of improvements to the above theorem. First, we weaken the assumptions by reducing the ratio $\frac{\alpha}{\varepsilon}$ needed to deduce fast convergence; here α represents strong convexity (or pseudorandomness in Section 3.3) and ε represents the initial error of A. This will be useful for our application to the Paulsen problem in Chapter 4. At a high level, given nearly doubly balanced A, we want to bound the distance to an exactly doubly balanced input B. Our plan will be to perturb A and then apply the fast convergence of matrix scaling. Therefore, the smaller the requirement for $\frac{\alpha}{\varepsilon}$, the less we have to move to find B. Theorem 3.3.10 is our strongest result in this direction and is used to give an optimal bound for the Paulsen problem in Chapter 4.

Our second improvement gives a strengthening of the conclusions of Theorem 3.2.8 by proving bounds on various scaling parameters, e.g. size, $\|(X_\infty, Y_\infty)\|_{\mathfrak{t}}$, $\|(X_\infty, Y_\infty)\|_\infty$. This will be useful for our statistics application in [36]. In this setting, we are given samples from some unknown distribution, and the distance from the scaling solution to the origin represents the error of a particular estimator. For this purpose, we give a strong result in Theorem 3.2.19, and then generalize it to the tensor setting in Chapter 7.

We present a high-level description of the techniques used to improve Theorem 3.2.8. Note that any ε-doubly balanced input $A \in \mathrm{Mat}(d,n)^K$ with size $s(A) = 1$ satisfies $\|\nabla_A\|_\infty \leq 2\varepsilon$ by definition, and $\|\nabla_A\|_{\mathfrak{t}}^2 \leq 2\varepsilon^2$ by Fact 3.1.13. In the proof of Theorem 3.2.8, we only used the condition on $\|\cdot\|_{\mathfrak{t}}$ and translated to $\|\cdot\|_\infty$. In the following Section 3.2.3, we will directly use the fact that A is ε-doubly balanced initially and analyze the change in $\|\nabla_{A_t}\|_\infty$ through gradient flow. This will allow us to replace the $\sqrt{d+n}$ factor loss by a $\log d$ factor, which in turn allows us to weaken the assumption to $\frac{\alpha}{\varepsilon} \gtrsim \log d$. To weaken this assumption further so that it is dimension independent, in Section 3.3, we will go beyond strong convexity and analyze a combinatorial pseudorandom condition. We show in Section 3.4 that this is in fact a strictly stronger condition than strong convexity. We will apply these results to give optimal bounds for the Paulsen problem in Chapter 4.

3.2.3 Monotonicity and Improved Analysis

In the proof of Theorem 3.2.8, we used Proposition 3.2.2 to show a strong bound on $\|(X_t, Y_t)\|_{\mathfrak{t}}$. We then translated this to a bound on $\|(X_t, Y_t)\|_\infty$ in order to show that strong convexity is maintained by Lemma 3.2.4. The inequality $\|(X,Y)\|_\infty \leq \sqrt{d+n}\|(X,Y)\|_{\mathfrak{t}}$ in this translation cannot be improved for general $(X,Y) \in \mathfrak{t}$. In this section, we will consider ε-doubly balanced inputs, where we also have a bound on the infinity norm $\|\nabla_A\|_\infty \leq s(A)\varepsilon$. Therefore in this subsection, we are able to directly bound $\|(X_t, Y_t)\|_\infty$ and prove fast convergence throughout for $\frac{\alpha}{\varepsilon} \gtrsim \log d$.

We accomplish this by a refined analysis of the individual row and column sums through gradient flow. Recall by Definition 3.2.3 that

$$\|\nabla_A\|_\infty = \max_{i \in [d]} \Big| d \cdot r_i - s \Big| + \max_{j \in [n]} \Big| n \cdot c_j - s \Big|.$$

We explicitly calculate the change in these quantities under gradient flow.

Fact 3.2.9. *If A_t follows gradient flow according to Definition 3.1.14, then for any $i \in [d], j \in [n]$, the following formula gives the change in the row and column sums:*

$$\partial_{t=0} r_i(A_t) = \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \Big((s - d \cdot r_i) + (s - n \cdot c_j) \Big),$$

$$\partial_{t=0} c_j(A_t) = \sum_{i=1}^{d} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \Big((s - n \cdot c_j) + (s - d \cdot r_i) \Big),$$

where we use $s = s(A), r_i = r_i(A)$, and $c_j = c_j(A)$ as shorthand.

Proof. We expand Definition 3.1.1 of row sums and calculate

$$\begin{aligned}\partial_{t=0} r_i(A_t) &= \partial_{t=0} \sum_{j=1}^{n} \sum_{k=1}^{K} e^{(X_t)_{ii}} |(A_k)_{ij}|^2 e^{(Y_t)_{jj}} = \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 (-(\nabla_A^L)_{ii} - (\nabla_A^R)_{jj}) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \Big((s - d \cdot r_i) + (s - n \cdot c_j) \Big),\end{aligned}$$

where the first step is by $A_t = e^{X_t/2} A e^{Y_t/2}$ according to Definition 3.1.14, in the second step we used that $(X_0, Y_0) = (0, 0)$ and $\partial_t (X_t, Y_t) = -\nabla_{A_t}$ again by Definition 3.1.14, and the last step is by the formula for gradient given in Proposition 3.1.12. The calculation for the columns is the similar. □

Our plan is to use the above formulas in order to directly analyze $\|\nabla_{A_t}\|_\infty$ instead of resorting to the fast convergence of $\|\nabla_{A_t}\|_{\mathfrak{t}}$ shown in Proposition 3.2.2. We will give two different arguments for the left and right errors, respectively, as the matrix scaling problem is asymmetric with $d \leq n$. This will allow us to prove a stronger bound on $\|(X_t, Y_t)\|_\infty$, which then implies that strong convexity is maintained longer by the robustness result in Lemma 3.2.4.

We first show that the worst row or column error grows very slowly under gradient flow. This was observed in Prop 3.2 of [63] for the more general operator scaling setting.

Lemma 3.2.10 (Monotonicity)**.** *If A_t follows gradient flow, then*

$$\partial_t \max\{\|\nabla_{A_t}^L\|_\infty, \|\nabla_{A_t}^R\|_\infty\} \leq -\partial_t s(A_t) = \|\nabla_{A_t}\|_{\mathfrak{t}}^2.$$

Proof. We prove the statement at time $t = 0$, from which the lemma follows by considering $A = A_t$. We will show that the row or column with the worst error is being pushed towards the average by gradient flow.

First, consider the case when $\|\nabla_A^L\|_\infty \geq \|\nabla_A^R\|_\infty$. The proof below is entirely symmetric in rows and columns, and so the other case $\|\nabla_A^R\|_\infty \geq \|\nabla_A^L\|_\infty$ follows by the same argument. Let $i \in \arg\max_{i' \in [d]} |d \cdot r_{i'}(A) - s(A)|$ be the row with the worst error. We will separate into two cases depending on the sign of the error, so consider the case $\|\nabla_A^L\|_\infty = d \cdot r_i(A) - s(A)$, meaning $r_i(A)$ is larger than average, and we want to show it is decreasing.

$$\begin{aligned}\partial_{t=0} r_i(A_t) &= \sum_{j=1}^{n}\sum_{k=1}^{K} |(A_k)_{ij}|^2 \Big((s - d\cdot r_i) + (s - n\cdot c_j)\Big) \\ &\leq \sum_{j=1}^{n}\sum_{k=1}^{K} d|(A_k)_{ij}|^2(-\|\nabla_A^L\|_\infty + \|\nabla_A^R\|_\infty) \leq 0,\end{aligned}$$

where the first step was by Fact 3.2.9, the second step was by the assumption that i had the worst row error, and the final step was by the assumption $\|\nabla_A^L\|_\infty \geq \|\nabla_A^R\|_\infty$. In the other case when $\|\nabla_A^L\|_\infty = s(A) - d\cdot r_i(A)$, $r_i(A)$ is smaller than average so we want to show that it is increasing. By a similar calculation,

$$\partial_{t=0} r_i(A_t) \geq \sum_{j=1}^{n}\sum_{k=1}^{K} d|(A_k)_{ij}|^2(\|\nabla_A^L\|_\infty - \|\nabla_A^R\|_\infty) \geq 0.$$

Therefore, in both cases (see Remark 3.2.11 for differentiability of max),

$$\partial_{t=0}\|\nabla_{A_t}^L\|_\infty = \partial_{t=0} \max_{i\in[d]} |d\cdot r_i(A_t) - s(A_t)| \leq 0 + |\partial_{t=0} s(A_t)| = \|\nabla_A\|_{\mathfrak{t}}^2,$$

where in the last step we used the fact that $s(A_t) = f_{A_t}(0,0)$ and Proposition 3.1.15 on the change in s.

The other case $\|\nabla_A^R\|_\infty \geq \|\nabla_A^L\|_\infty$ follows symmetrically, so the statement is shown. □

Remark 3.2.11. *The previous lemma bounded $\|\nabla_{A_t}\|_\infty$ by bounding its derivative. Technically, the infinity norm is not always differentiable, but this can easily be made rigorous by following the proof of Prop 3.2 in [63], which used the generalized envelope theorem of Milgrom and Segal (Corollary 4 of [70]) to bound the error. We omit this analytic detail in this thesis, as the core of these proofs has to do with structural properties of scalings.*

Lemma 3.2.10 shows that for small $t \approx 0$, $\|\nabla_{A_t}\|_\infty \approx \|\nabla_A\|_\infty$. Note that this lemma did not use strong convexity at all. Below, we show that for strongly convex inputs, when t is large, we can use the exponential convergence of $\|\nabla_{A_t}\|_{\mathfrak{t}}$ derived in Proposition 3.2.2.

Corollary 3.2.12. *If A_t is α-strongly convex for all $t \in [0,T]$, then*

$$\|\nabla_{A_T}^L\|_\infty \leq \sqrt{d}\|\nabla_A\|_{\mathfrak{t}} \cdot e^{-\alpha T}.$$

Proof. We simply translate the exponential convergence to the infinity norm:

$$\|\nabla^L_{A_T}\|_\infty \leq \sqrt{d}\|\nabla^L_{A_T}\|_{\mathfrak{t}} \leq \sqrt{d}\|\nabla_{A_T}\|_{\mathfrak{t}} \leq \sqrt{d}\|\nabla_A\|_{\mathfrak{t}} e^{-\alpha T},$$

where the first step was by Lemma 3.2.7 since the Y part is 0, and the third step is by Proposition 3.2.2. □

Note that for $T \gtrsim \frac{\log d}{\alpha}$, the above Corollary 3.2.12 gives a bound on $\|\nabla^L_{A_T}\|_\infty \lesssim \|\nabla^L_A\|_\infty$ as the leading d factor is canceled by the exponential convergence for T time. By combining the above two bounds, we can give an improved analysis of the left scaling under the strong convexity assumption. Specifically, we improve the bound on $\|(X_t, Y_t)\|_\infty$ by directly using the bound on $\|\nabla_{A_t}\|_\infty$ for all time instead of resorting to the inequality $\|\nabla_{A_t}\|_\infty \leq \sqrt{d+n}\|\nabla_{A_t}\|_{\mathfrak{t}}$. This allows us to replace the leading $\sqrt{d+n}$ factor by $\log d$ for the left scaling. Following this proposition, we prove a similar bound on the right scaling by a slightly technical comparison argument.

Proposition 3.2.13. *Consider ε-doubly balanced matrix tuple $A \in \mathrm{Mat}(d,n)^K$ of size $s(A) = 1$, and assume A_t is α-strongly convex for all $t \in [0,T]$. Then*

$$\|X_T\|_\infty \leq \frac{\varepsilon \log d}{2\alpha} + \frac{\varepsilon^2 \log d}{2\alpha^2} + \frac{\sqrt{2}\varepsilon}{\alpha}.$$

Proof. We break the evolution into two stages using cutoff $\kappa := \frac{\log d}{2\alpha}$. In the first stage, we use the slow growth shown in Lemma 3.2.10 to bound $\|\nabla^L_{A_t}\|_\infty \approx \varepsilon$, and in the second stage we resort to the exponential convergence of Proposition 3.2.2. The cutoff is chosen so as to balance the bounds coming from both stages. In the rest of this proof, we will use the shorthand $\nabla_t := \nabla_{A_t}$ in order to avoid too many subscripts.

By the fundamental theorem of calculus, we have

$$\|X_T\|_\infty = \left\| \int_0^T -\nabla^L_t \right\|_\infty \leq \int_0^T \|\nabla^L_t\|_\infty = \int_0^\kappa \|\nabla^L_t\|_\infty + \int_\kappa^T \|\nabla^L_t\|_\infty,$$

where in the first step we used $X_0 = 0$ and $\partial_t X_t = -\nabla^L_{A_t}$ by Definition 3.1.14 of gradient flow, and the second was by triangle inequality on $\|\cdot\|_\infty$. To bound the first stage, we use monotonicity:

$$\begin{aligned}\int_0^\kappa \|\nabla^L_t\|_\infty &\leq \int_0^\kappa \max\{\|\nabla^L_t\|_\infty, \|\nabla^R_t\|_\infty\} \leq \int_0^\kappa \left(\max\{\|\nabla^L_A\|_\infty, \|\nabla^R_A\|_\infty\} + \int_0^t \|\nabla_{\tilde{t}}\|^2_{\mathfrak{t}}\right) \\ &\leq \int_0^\kappa \left(\varepsilon + \|\nabla_A\|^2_{\mathfrak{t}} \int_0^t e^{-2\alpha\tilde{t}}\right) \leq \kappa\left(\varepsilon + \frac{2\varepsilon^2}{2\alpha}\right),\end{aligned} \tag{3.1}$$

where the second step was by the fundamental theorem of calculus applied to $\|\nabla_t^L\|_\infty$ and $\|\nabla_t^R\|_\infty$ along with the bound from Lemma 3.2.10, in the third step we used the ε-doubly balanced condition on A to bound the first term and Proposition 3.2.2 to bound the second, and in the final step we used Fact 3.1.13 to bound $\|\nabla_A\|_{\mathsf{t}}^2 \leq 2\varepsilon^2$. Note that this actually give a bound for both the left and right errors.

For the second stage, we use the fast convergence in Corollary 3.2.12.

$$\int_\kappa^T \|\nabla_{A_t}^L\|_\infty \leq \int_\kappa^T \sqrt{d}e^{-\alpha t}\|\nabla_A\|_{\mathsf{t}} = \sqrt{d}e^{-\alpha\kappa}\|\nabla_A\|_{\mathsf{t}} \int_0^{T-\kappa} e^{-\alpha t} \leq \frac{\sqrt{2}\varepsilon}{\alpha},$$

where the first step was by Corollary 3.2.12, the second step was by a simple change of variable in the integral $t \to t - \kappa$, and in the final step we used the definition of $\kappa = \frac{\log d}{2\alpha}$ to cancel the leading term and Fact 3.1.13 to bound $\|\nabla_A\|_{\mathsf{t}}^2 \leq 2\varepsilon^2$. This argument explains our choice of κ, as this cancels the leading $\sqrt{d}$ factor coming from the translation $\|\cdot\|_\infty \to \|\cdot\|_{\mathsf{t}}$ and makes the bounds on the two stages comparable.

Combining the bounds for both stages, we conclude that

$$\|X_T\|_\infty \leq \kappa\left(\varepsilon + \frac{\varepsilon^2}{\alpha}\right) + \frac{\sqrt{2}\varepsilon}{\alpha} = \frac{\varepsilon \log d}{2\alpha}\left(1 + \frac{\varepsilon}{\alpha}\right) + \frac{\sqrt{2}\varepsilon}{\alpha}.$$

□

Repeating the proof of Proposition 3.2.13 for the right error Y_t would give the same bound with $\log n$ instead of $\log d$. For the Paulsen problem in Chapter 4, we are interested in the case $n \gg d$, so we would like to remove the dependency on n. To do so, we compare $\|\nabla_{A_t}^L\|_\infty$ and $\|\nabla_{A_t}^R\|_\infty$ at any given time, and use the fact that $\|\nabla_{A_t}^L\|_\infty$ is converging exponentially to argue that $\|\nabla_{A_t}^R\|_\infty$ cannot be large for too long.

Definition 3.2.14. *For matrix gradient flow $A_t \in \mathrm{Mat}(d,n)^K$, let δ_t satisfy*

$$\|\nabla_{A_t}^R\|_\infty = (1+\delta_t)\|\nabla_{A_t}^L\|_\infty.$$

A simple argument shows that large imbalance δ_t implies decrease of the right error.

Lemma 3.2.15. *For A_t following gradient flow and δ_t according to Definition 3.2.14,*

$$-\partial_t \log\|\nabla_{A_t}^R\|_\infty \geq \min\left\{\frac{s(A_t)\delta_t - \|\nabla_{A_t}\|_\infty}{1+\delta_t}, \frac{(s(A_t) - \|\nabla_{A_t}^R\|_\infty)\delta_t}{1+\delta_t}\right\}.$$

Proof. We prove the statement at time $t = 0$, from which the lemma follows by considering $A = A_t$. Let $\delta := \delta_0$ for shorthand, and let $j \in \arg\max_{j' \in [n]} |n \cdot c_{j'} - s|$. We will separate into two cases depending on sign, and show the lower bound in the first and second term, respectively. First consider the case $\|\nabla_A^R\|_\infty = n \cdot c_j - s$, so that c_j is larger than average and we want to show that this it is decreasing.

$$\begin{aligned}
-\partial_{t=0}(n \cdot c_j(A_t) - s(A_t)) &= \sum_{i=1}^{d} \sum_{k=1}^{K} n|(A_k)_{ij}|^2 \Big((n \cdot c_j - s) + (d \cdot r_i - s) \Big) - \|\nabla_A\|_{\mathfrak{t}}^2 \\
&\geq n \cdot c_j \Big(\|\nabla_A^R\|_\infty - \|\nabla_A^L\|_\infty \Big) - \|\nabla_A\|_{\mathfrak{t}}^2 \\
&\geq (s(A) + \|\nabla_A^R\|_\infty) \Big(\|\nabla_A^R\|_\infty - \|\nabla_A^L\|_\infty \Big) - \Big(\|\nabla_A^R\|_\infty^2 + \|\nabla_A^L\|_\infty^2 \Big) \\
&= s(A)\|\nabla_A^R\|_\infty - \Big(s(A) + \|\nabla_A^R\|_\infty + \|\nabla_A^L\|_\infty \Big) \|\nabla_A^L\|_\infty,
\end{aligned}$$

where in the first step we calculated $\partial_t c_j(A_t)$ using Fact 3.2.9 and $\partial_t s(A_t)$ by Proposition 3.1.15, the second step was by the case assumption $\|\nabla_A^R\| = n \cdot c_j - s$, and in the third step we again used the case assumption that c_j is the largest column sum, as well as Lemma 3.2.7 to bound $\|\nabla\|_{\mathfrak{t}}^2$. For the log derivative (see Remark 3.2.11 for differentiability considerations), we continue

$$-\partial_{t=0} \log \|\nabla_{A_t}^R\|_\infty = \frac{-\partial_{t=0}(n \cdot c_j(A_t) - s(A_t))}{\|\nabla_A^R\|_\infty} \geq s(A) - \frac{s(A) + \|\nabla_A\|_\infty}{1+\delta} = \frac{s(A)\delta - \|\nabla_A\|_\infty}{1+\delta},$$

where we used that $\delta := \delta_0$ satisfies $\|\nabla_A^R\|_\infty = (1+\delta)\|\nabla_A^L\|_\infty$ by Definition 3.2.14. This gives the lower bound in the first term.

Now we consider the case $\|\nabla_A^R\| = s - n \cdot c_j$, i.e. c_j is smaller than average and we want to show that this it is increasing. The calculations are almost the same in this case, so we skip some steps.

$$\begin{aligned}
\partial_{t=0}(n \cdot c_j(A_t) - s(A_t)) &= \sum_{i=1}^{d} \sum_{k=1}^{K} n|(A_k)_{ij}|^2 \Big((s - n \cdot c_j) + (s - d \cdot r_i) \Big) + \|\nabla_A\|_{\mathfrak{t}}^2 \\
&\geq n \cdot c_j (\|\nabla_A^R\|_\infty - \|\nabla_A^L\|_\infty) + \|\nabla_A\|_{\mathfrak{t}}^2 \\
&\geq (s(A) - \|\nabla_A^R\|_\infty)(\|\nabla_A^R\|_\infty - \|\nabla_A^L\|_\infty),
\end{aligned}$$

where in the first step we calculated $\partial_t c_j(A_t)$ using Fact 3.2.9 and $\partial_t s(A_t)$ by Proposition 3.1.15, the second step was by the case assumption $\|\nabla_A^R\| = s - n \cdot c_j$, and in the third step we again used the case assumption that c_j is the smallest column sum, as well as the simple fact $\|\nabla\|_t^2 \geq 0$. To bound the log derivative, we continue

$$-\partial_{t=0} \log \|\nabla^R_{A_t}\|_\infty = \frac{\partial_{t=0}(n \cdot c_j(A_t) - s(A_t))}{\|\nabla^R_A\|_\infty} \geq (s(A) - \|\nabla^R_A\|_\infty) \left(1 - \frac{1}{1+\delta}\right),$$

where we used the definition $\|\nabla^R_A\|_\infty = (1+\delta)\|\nabla^L_A\|_\infty$. This proves the bound in the second term, and show the lemma by considering the minimum of both cases. □

Note that this lemma also does not use strong convexity. We state the following simple corollary which gives fast convergence when δ is large. This will be used in conjunction with Corollary 3.2.12 for strongly convex inputs to show fast convergence for all time.

Corollary 3.2.16. *Let $A_t \in \mathrm{Mat}(d,n)^K$ be the solution to gradient flow, and assume $s(A_t) \geq 0.95$. Then for any $\frac{1}{5} \geq \alpha \geq 3\|\nabla_{A_t}\|_\infty$,*

$$\delta_t \geq 2\alpha \implies -\partial_t \log \|\nabla^R_{A_t}\|_\infty \geq \alpha.$$

Proof. By Lemma 3.2.15, we have the lower bound

$$-\partial_t \log \|\nabla^R_{A_t}\|_\infty \geq \min\left\{\frac{s(A_t)\delta_t - \|\nabla_{A_t}\|_\infty}{1+\delta_t}, \frac{(s(A_t) - \|\nabla^R_{A_t}\|_\infty)\delta_t}{1+\delta_t}\right\} \geq 0.75\frac{\delta_t}{1+\delta_t},$$

where we used the conditions $s \geq 0.95, \delta_t \geq 2\alpha \geq 6\|\nabla_A\|_{\infty_t}$. The statement follows by monotonicity of $\frac{x}{1+x}$ for $x \geq 0$:

$$-\partial_t \log \|\nabla^R_{A_t}\|_\infty \geq 0.75\frac{\delta_t}{1+\delta_t} \geq 1.5\frac{\alpha}{1+2\alpha} \geq \alpha,$$

where the second step was by our assumption $\delta_t \geq 2\alpha$, and the last step was by the assumption $\alpha \leq \frac{1}{5}$. □

For the other case, when δ_t is small, Corollary 3.2.12 gives fast convergence of $\|\nabla^L_{A_t}\|_\infty$ after time $t \geq \kappa = \frac{\log d}{2\alpha}$. In the following lemma, we combine the above two arguments to show exponential convergence of $\|\nabla^R_{A_t}\|_\infty$.

Lemma 3.2.17. *Consider ε-doubly balanced matrix tuple $A \in \mathrm{Mat}(d,n)^K$ of size $s(A) = 1$, and assume A_t is α-strongly convex for all $t \in [0,T]$ with $\frac{1}{5} \geq \alpha \geq 7\varepsilon$. Then, for any $T \geq \kappa = \frac{\log d}{2\alpha}$,*

$$\|\nabla^R_{A_T}\|_\infty \leq \sqrt{2}(1+2\alpha)\varepsilon e^{-\alpha(T-\kappa)}.$$

Proof. Our plan is to leverage the exponential convergence of $\|\nabla^L_{A_t}\|_\infty$ from Corollary 3.2.12 to show that either $\|\nabla^R_{A_t}\|_\infty$ is also decreasing exponentially, or δ_t will become large and we can apply fast convergence from Corollary 3.2.16. To this end, we verify that both the conditions of Corollary 3.2.16 (size and error) are satisfied for all $t \in [\kappa, T]$. First note that A_t is strongly convex for $t \in [0, T]$, so the size can be lower bounded by

$$s(A_T) = s(A_0) + \int_0^T \partial_t s(A_t) = 1 - \int_0^T \|\nabla_{A_t}\|_{\mathfrak{t}}^2 \geq 1 - \|\nabla_A\|_{\mathfrak{t}}^2 \int_0^T e^{-2\alpha t} \geq 1 - \frac{2\varepsilon^2}{2\alpha},$$

where the first step was by the fundamental theorem of calculus, in the second step we used Proposition 3.1.15 to calculate $\partial_t s(A_t)$, in the third step we used the exponential convergence of $\|\nabla_{A_t}\|_{\mathfrak{t}}^2$ derived in Proposition 3.2.2, and the final step was by the bound $\|\nabla_A\|_{\mathfrak{t}}^2 \leq 2\varepsilon^2$ given in Fact 3.1.13 for ε-doubly balanced input. By the assumptions $\varepsilon \leq \frac{\alpha}{7} \leq \frac{1}{35}$, this gives lower bound $s(A_T) \geq 0.95$. Since $\partial_t s(A_t) \leq 0$ always by Proposition 3.1.15, the lower bound $s(A_t) \geq s(A_T) \geq 0.95$ follows for all $t \in [0, T]$.

In order to show that α is large as compared to the error for all time, we use the monotonicity lemma to bound

$$\max\{\|\nabla^L_{A_T}\|_\infty, \|\nabla^R_{A_T}\|_\infty\} \leq \varepsilon - \int_0^T \partial_t s(A_t) \leq \varepsilon + s(A_0) - s(A_T) \leq \varepsilon + \frac{\varepsilon^2}{\alpha}, \tag{3.2}$$

where the first step was by Lemma 3.2.10, the second step was by the fundamental theorem of calculus, and the final step was by the lower bound $s(A_T) \geq s(A_0) - \frac{\varepsilon^2}{\alpha}$ calculated above. In fact the above bound holds for any $t \in [0, T]$, so we have

$$\|\nabla_{A_t}\|_\infty \leq 2\Big(\varepsilon + \frac{\varepsilon^2}{\alpha}\Big) \leq \frac{16\varepsilon}{7} \leq \frac{\alpha}{3},$$

where the first step was by Definition 3.2.3 and the bounds on the left and right part calculated above, and in the final two steps we use the assumption $\frac{\varepsilon}{\alpha} \leq \frac{1}{7}$. This verifies both conditions of Corollary 3.2.16 for all $t \in [0, T]$.

To show fast convergence of $\|\nabla_{A_t}\|_\infty$ for all time, we partition $[\kappa, T]$ into two pieces depending on which of ∇^L, ∇^R are converging quickly.

$$T_R := \{t \in [\kappa, T] \mid \delta_t \geq 2\alpha\}, \qquad \text{and} \qquad T_L := \{t \in [\kappa, T] \mid \delta_t \leq 2\alpha\}.$$

Since all quantities are continuous, this gives the decomposition

$$[\kappa, T] = \{[t_0 = \kappa, t_1], [t_1, t_2], ...\},$$

where $[t_m, t_{m+1}]$ are maximal intervals fully contained in either T_L or T_R. For the remainder of this proof, we will use $\nabla_t := \nabla_{A_t}$ as shorthand to avoid triple subscripts. We show exponential convergence of $\|\nabla^R_{A_t}\|_\infty$ for $t \in [\kappa, t_m]$, and induct on m. For the base case, we have already shown in Eq. (3.2) that

$$\|\nabla^R_\kappa\|_\infty \leq \varepsilon + \frac{\varepsilon^2}{\alpha} \leq \frac{8}{7}\varepsilon \leq \varepsilon(\sqrt{2}(1+2\alpha)e^0),$$

where in the second step we used the assumption $\varepsilon \leq \frac{\alpha}{7}$, and the third step only uses $\alpha \geq 0$. This satisfies the requirement at $t_0 = \kappa$, so we assume by induction that the lemma is shown for $t \in [\kappa, t_m]$.

To show the induction step, first consider the case $[t_m, t_{m+1}] \subseteq T_R$. Then for any $t \in [t_m, t_{m+1}]$ we have

$$\log\|\nabla^R_t\|_\infty - \log\|\nabla^R_{t_m}\|_\infty = \int_{t_m}^t \partial_{\tilde{t}} \log\|\nabla^R_{\tilde{t}}\|_\infty \leq -\alpha(t_{m+1} - t),$$

where the last step was by Corollary 3.2.16. Therefore

$$\|\nabla^R_t\|_\infty \leq \|\nabla^R_{t_m}\|_\infty e^{-\alpha(t-t_m)} \leq \sqrt{2}(1+2\alpha)\varepsilon e^{-\alpha(t-\kappa)},$$

where we used the induction hypothesis on t_m for the last step.

In the other case $[t_m, t_{m+1}] \subseteq T_L$, for any $t \in [t_m, t_{m+1}]$ we have

$$\|\nabla^R_t\|_\infty = (1+\delta_t)\|\nabla^L_t\|_\infty \leq (1+\delta_t)\sqrt{d}e^{-\alpha t}\|\nabla_A\|_{\mathfrak{t}} \leq (1+\delta_t)e^{-\alpha(t-\kappa)}\sqrt{2\varepsilon^2} \leq \varepsilon(1+2\alpha)\sqrt{2}e^{-\alpha(t-\kappa)},$$

where the first step was by Definition 3.2.14 of δ_t, the second step was by Corollary 3.2.12, in the third step we canceled the $\sqrt{d}$ term by our choice of $\kappa = \frac{\log d}{2\alpha}$ and used Fact 3.1.13 to bound $\|\nabla_A\|^2_{\mathfrak{t}} \leq 2\varepsilon^2$ for ε-doubly balanced input, and the final step was by the case assumption $\delta_t \leq 2\alpha$ as $t \in T_L$. Since $t \in [t_m, t_{m+1}]$ was arbitrary, the induction is shown and the lemma follows for all $t \in [\kappa, T]$. □

As discussed earlier, applying the same argument as Proposition 3.2.13 to the right scaling Y_t would give a $\log n$ term. We can avoid this by applying Lemma 3.2.17, which shows $\|\nabla^R_{A_t}\|_\infty$ converges quickly after $\kappa = \frac{\log d}{2\alpha}$ time. Therefore, we can give the following bound for the right scaling.

Proposition 3.2.18. *Consider ε-doubly balanced $A \in \mathrm{Mat}(d,n)^K$ of size $s(A) = 1$, and assume A_t is α-strongly convex for all $t \in [0,T]$ with $\frac{1}{5} \geq \alpha \geq 7\varepsilon$. Then*

$$\|Y_T\|_\infty \leq \frac{\varepsilon \log d}{2\alpha} + \frac{\varepsilon^2 \log d}{2\alpha^2} + \frac{2\varepsilon}{\alpha}.$$

Proof. We can follow the proof of Proposition 3.2.13 by choosing cutoff $\kappa = \frac{\log d}{2\alpha}$ and bounding the first stage by the calculation in Eq. (3.1):

$$\|Y_T\|_\infty \leq \int_0^\kappa \|\nabla^R_{A_t}\|_\infty + \int_\kappa^T \|\nabla^R_{A_t}\|_\infty \leq \frac{\varepsilon \log d}{2\alpha}\left(1 + \frac{\varepsilon}{\alpha}\right) + \int_\kappa^T \|\nabla^R_{A_t}\|_\infty .$$

For the second stage, we apply Lemma 3.2.17 for all $t \geq \kappa$ to bound

$$\int_\kappa^T \|\nabla^R_{A_t}\|_\infty \leq \sqrt{2}(1+2\alpha)\varepsilon \int_0^{T-\kappa} e^{-\alpha t} \leq \frac{\sqrt{2}(1+2\alpha)\varepsilon}{\alpha} \leq \frac{2\varepsilon}{\alpha},$$

where the last step was by the assumption $\alpha \leq \frac{1}{5}$. Therefore, we can combine the two stages to bound

$$\|Y_T\|_\infty \leq \int_0^\kappa \|\nabla^R_{A_t}\|_\infty + \int_\kappa^T \|\nabla^R_{A_t}\|_\infty \leq \frac{\varepsilon \log d}{2\alpha}\left(1 + \frac{\varepsilon}{\alpha}\right) + \frac{2\varepsilon}{\alpha}.$$

□

We now follow the proof strategy of Theorem 3.2.8 with improved control on $\|(X_t, Y_t)\|_\infty$. The theorem below weakens the assumption of fast convergence to $\frac{\alpha}{\varepsilon} \gtrsim \log d$ instead of $\frac{\alpha}{\varepsilon} \gtrsim \sqrt{d+n}$ as in Theorem 3.2.8.

Theorem 3.2.19. *If matrix tuple $A \in \mathrm{Mat}(d,n)^K$ of size $s(A) = 1$ is ε-doubly balanced and α-strongly convex with $\frac{1}{5} \geq \alpha \geq \varepsilon(4\log d + 20)$, then*

1. *For all time $t \geq 0$, the scaling solution satisfies*

$$\|(X_t, Y_t)\|_{\mathfrak{t}} \leq \frac{4\varepsilon}{\alpha} \qquad \textit{and} \qquad \max\{\|X_t\|_\infty, \|Y_t\|_\infty\} \leq \frac{\varepsilon(2\log d + 6)}{\alpha};$$

2. *The limit $(X_\infty, Y_\infty) := \lim_{t\to\infty}(X_t, Y_T)$ exists and $A_\infty := e^{X_\infty/2} A e^{Y_\infty/2}$ gives a solution to the matrix scaling problem in Definition 3.1.3 on input A;*

3. *The size of the solution can be lower bounded by*

$$s(A_\infty) = f_A(X_\infty, Y_\infty) \geq 1 - \frac{e \cdot \varepsilon^2}{\alpha}.$$

Proof. We claim that A_t is $\frac{\alpha}{e}$-strongly convex for all time. For contradiction, let T be the first time A_T is $\leq \frac{\alpha}{e}$-strongly convex. Since A is α-strongly convex, by the robustness property of Lemma 3.2.4 in the contrapositive, we must have $\|(X_T, Y_T)\|_\infty \geq 1$. To show a contradiction, we will use Proposition 3.2.13 and Proposition 3.2.18 to upper bound $\|(X_T, Y_T)\|_\infty$. First verify the conditions: A_t is $\frac{\alpha}{e}$-strongly convex for all $t \in [0, T]$, $s(A) = 1$, and A is ε-doubly balanced with

$$\frac{\alpha}{e} \geq \frac{\varepsilon(4 \log d + 20)}{e} \geq 7\varepsilon,$$

where we used our assumption $\varepsilon(4 \log d + 20) \leq \alpha$. Therefore we can bound

$$\|(X_T, Y_T)\|_\infty \leq 2 \max\{\|X_T\|_\infty, \|Y_T\|_\infty\} \leq \frac{2\varepsilon \log d}{2\alpha/e}\left(1 + \frac{\varepsilon}{\alpha/e}\right) + \frac{4\varepsilon}{\alpha/e} < \frac{\varepsilon(4 \log d + 11)}{\alpha} < 1,$$

where we use $\frac{\alpha}{e}$-strong convexity to bound $\|X_T\|_\infty, \|Y_T\|_\infty$ by Proposition 3.2.13 and Proposition 3.2.18 respectively, and the next two steps use are by our assumption $\alpha \geq \varepsilon(4 \log d + 20)$. This is the desired contradiction, so A_t must be $\frac{\alpha}{e}$-strongly convex for all time and the above calculation shows the infinity norm bound in item (1) for all time $t \geq 0$. The $\mathfrak{t}$-norm bound follows simply as

$$\|(X_t, Y_t)\|_{\mathfrak{t}} \leq \frac{\|\nabla_A\|_{\mathfrak{t}}}{\alpha/e} \leq \frac{e\sqrt{2} \cdot \varepsilon}{\alpha},$$

where the first step was by applying Proposition 3.2.2 with $\frac{\alpha}{e}$-strong convexity, and in the second step we use Fact 3.1.13 to bound $\|\nabla_A\|_{\mathfrak{t}}^2 \leq 2\varepsilon^2$.

The proof of item (2) now follows the same steps as in the proof of Theorem 3.2.8, as we can show

$$\lim_{T\to\infty} \int_{t \geq T} \|\partial_t(X_t, Y_t)\|_{\mathfrak{t}} = \lim_{T\to\infty} \int_{t \geq T} \|\nabla_{A_t}\|_{\mathfrak{t}} \leq \lim_{T\to\infty} \|\nabla_A\|_{\mathfrak{t}} \int_{t \geq T} e^{-\alpha t/e} = 0,$$

where the first step was by Definition 3.1.14 of gradient flow, and the last step was by the fast convergence in Proposition 3.2.2. Therefore the limit (X_∞, Y_∞) exists, and further $\nabla_{A_\infty} = 0$ so $A_\infty = e^{X_\infty/2} A e^{Y_\infty/2}$ is doubly balanced by Proposition 3.1.10(2).

To show (3), consider the univariate restriction $h(\eta) := f_A(\eta X_\infty, \eta Y_\infty)$. We want to apply Lemma 2.3.7 to show the lower bound on $f_A(X_\infty, Y_\infty) = s(A_\infty)$. First, we bound the derivative.

$$|h'(0)| = |\langle \nabla_A, (X_\infty, Y_\infty)\rangle| \leq \|\nabla_A\|_{\mathfrak{t}} \|(X_\infty, Y_\infty)\|_{\mathfrak{t}},$$

where the first step was by Definition 2.3.12 of the gradient, and the final step was by Cauchy-Schwarz. Now we show h is sufficiently strongly convex. We know the optimizer $\eta_* := \arg\min_{\eta \in \mathbb{R}} h(\eta)$ is at $\eta_* = 1$, since h is a restriction of f_A, which has $(X_\infty, Y_\infty) = \arg\inf_{(X,Y)\in \mathfrak{t}} f_A(X,Y)$ by Proposition 3.1.10(3). Further, by Lemma 3.2.4 and the bound $\|(X_\infty, Y_\infty)\|_\infty < 1$ shown in item (1), we have f_A is $\frac{\alpha}{e}$-strongly convex at $(\eta X_\infty, \eta Y_\infty)$ for all $|\eta| \leq 1$. Therefore, the restriction h is α'-strongly convex with $\alpha' \geq \frac{\alpha}{e}\|(X_\infty, Y_\infty)\|_{\mathfrak{t}}^2$ for all $\eta \in [0, \eta_*]$. This verifies the requirement in Lemma 2.3.7, so we can bound the size by

$$s(A_\infty) = \inf_{\eta \in \mathbb{R}} h(\eta) \geq h(0) - \frac{|h'(0)|^2}{2\alpha'} \geq f_A(0,0) - \frac{\|\nabla_A\|_{\mathfrak{t}}^2 \|(X_\infty, Y_\infty)\|_{\mathfrak{t}}^2}{2\alpha \|(X_\infty, Y_\infty)\|_{\mathfrak{t}}^2 / e} \geq 1 - \frac{e \cdot \varepsilon^2}{\alpha},$$

where the second step was by Lemma 2.3.7, the third step were by our two calculations above showing $|h'(0)| \leq \|\nabla_A\|_{\mathfrak{t}} \|(X_\infty, Y_\infty)\|_{\mathfrak{t}}$ and that h is $\frac{\alpha}{e}\|(X_\infty, Y_\infty)\|_{\mathfrak{t}}^2$-strongly convex, and in the final step we used the assumption $f_A(0,0) = s(A) = 1$ as well as Fact 3.1.13 for ε-doubly balanced A to bound the gradient $\|\nabla_A\|_{\mathfrak{t}}^2 \leq 2\varepsilon^2$. □

Note that the strong convexity assumption in Theorem 3.2.19 is weaker by a factor of $O(d/\log d)$ as compared to Theorem 3.2.8. In the following Section 3.3, we will further weaken this assumption by analyzing gradient flow on inputs satisfying a combinatorial pseudorandom condition.

This result should be compared to Theorem 1.5 of [63], which showed the same convergence for the more general operator scaling problem with a stronger assumption $\alpha^2 \gtrsim \varepsilon \log d$ (the result of [63] actually used a related spectral gap condition, and we discuss this distinction in more detail in Section 7.1). Subsequent to its publication, Franks and Moitra [35] applied the fast convergence result of [63] for frame scaling to give near-optimal sample complexity results for an important estimation problem in statistics. Our pseudorandom analysis in Section 3.3 can be used to improve the sample complexity bound from [35], as we show in Section 4.4 and Section 8.5. We discuss the α vs α^2 requirement in more detail in Section 4.2.2 as well as in Section 7.3, as it is a make-or-break distinction for our application to the Paulsen problem in Chapter 4.

3.3 Pseudorandom Setting

The main improvement between Theorem 3.2.19 and Theorem 3.2.8 came from our improved analysis of $\|(X_t, Y_t)\|_\infty$ under gradient flow. Specifically, our two-stage analysis of $\|\nabla_{A_t}\|_{\mathfrak{t}}^L$ in Proposition 3.2.13 allowed us to effectively use the fact that the input was nearly doubly balanced. But there is still a $\frac{\log d}{\alpha}$ loss from the first stage while we wait for the exponential convergence of $\|\nabla_{A_t}\|_t$ to kick in from Proposition 3.2.2. It would be interesting to pin down whether or not this loss is necessary for any analysis of strongly convex inputs.

To give optimal bounds in our application to the Paulsen problem in Chapter 4, we would like to further decrease the requirement on $\frac{\alpha}{\varepsilon}$, and in particular to make it dimension independent. We therefore revisit the combinatorial pseudorandom condition that was first defined in Kwok et al. [62]. This new condition allows us to immediately show convergence of $\|\nabla_{A_t}\|_\infty$ for all $t \geq 0$, which in turn allows us to improve the bound on $\|(X_\infty, Y_\infty)\|_\infty$ and decrease the ratio $\frac{\alpha}{\varepsilon}$ for pseudorandom inputs.

We first define the pseudorandom condition and then show how it implies fast convergence of error. The rest of the analysis follows a similar strategy to Proposition 3.2.18.

Definition 3.3.1. *Matrix tuple $A \in \mathrm{Mat}(d,n)^K$ is (α,β)-pseudorandom if for every $S \subseteq [d]$ and $T \subseteq [n]$ with $|T| \geq \beta n$:*

$$\sum_{i \in S} \sum_{j \in T} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \geq \alpha \frac{|S|}{d} \frac{|T|}{n}.$$

We note that the pseudorandom condition, like Definition 3.2.1 of strong convexity, is not homogeneous. Therefore, α in this condition should be compared to the size, with $\frac{1}{dn}J$ again being an extremal example (see Appendix A.2). From a graph-theoretic perspective, the pseudorandom property is reminiscent of the expander mixing lemma for bipartite graph $w_{ij} := \sum_{k=1}^{K} |A_{ij}|^2$. We will use the connection to graphs in Section 3.4, where we compare strong convexity and pseudorandomness.

Remark 3.3.2. *Definition 3.3.1 is slightly different from the pseudorandom condition in Definition 4.3.2 in [62]. We believe the definition in this thesis is slightly more natural, both its implication for fast convergence, as well as for our smoothed analysis result in Chapter 5 showing random inputs are pseudorandom. For more details on the explicit difference, see Remark 3.4.4.*

The next lemma reduces the pseudorandom condition to a much fewer number of sets. This will not be used in this chapter to analyze pseudorandom matrices, but will be helpful in Chapter 5 where we will use a union bound to show pseudorandomness for random inputs.

Lemma 3.3.3. *Matrix tuple $A \in \mathrm{Mat}(d,n)^K$ is (α,β)-pseudorandom iff*

$$\min_{i \in [d]} \min_{T \in \binom{[n]}{\beta n}} \sum_{j \in T} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \geq \alpha \frac{\beta}{d}.$$

Proof. We will prove the following stronger statement: given arbitrary weight function $w : [d] \times [n] \to \mathbb{R}_+$, if $w(S,T) := \sum_{i \in S} \sum_{j \in T} w_{ij} \geq |S||T|$ for every $S \in \binom{[d]}{a}$ and $T \in \binom{[n]}{b}$, then in fact we have $w(S,T) \geq |S||T|$ for all larger sets $|S| \geq a, |T| \geq b$. The lemma follows by considering $w_{ij} := \frac{1}{\alpha} \sum_{k=1}^{K} |(A_k)_{ij}|^2$ and choosing $a = 1$, $b = \beta n$.

The proof is by a simple induction, so assume we have $w(S,T) \geq |S||T|$ for all $|S| = k, |T| = \ell$. Then for any $|S| = k, |T| = \ell + 1$,

$$\ell \cdot w(S,T) = \sum_{j \in T} w(S, T - j) \geq \sum_{j \in T} |S||T - j| = (\ell + 1)|S|\ell,$$

where in first step, every entry of $w(S,T)$ is counted exactly ℓ times in $\sum_{j \in T} w(S, T - j)$, and the second step is by the inductive hypothesis. Canceling ℓ from both sides and using $|T| = \ell + 1$ by definition, we have $w(S,T) \geq |S||T|$. By a symmetric argument, we can show the required lower bound for $|S| = k + 1, |T| = \ell$, so the lemma follows by induction. □

Next we show that the pseudorandom condition is preserved under scalings. This robustness result is similar to Lemma 3.2.4 and will be important in proving our convergence result in Theorem 3.3.10.

Lemma 3.3.4 (Robustness)**.** *If $A \in \mathrm{Mat}(d,n)$ is (α, β)-pseudorandom, then for any $(X', Y') \in \mathfrak{t}$, the scaling $B = e^{X'/2} A e^{Y'/2}$ is at least $(\alpha \cdot e^{-\|(X',Y')\|_\infty}, \beta)$-pseudorandom.*

Proof. The proof of Lemma 3.2.4 showed $|(B_k)_{ij}|^2 \geq e^{-\|(X',Y')\|_\infty} |(A_k)_{ij}|^2$ for each entry, which implies this robustness lemma by Definition 3.3.1 of pseudorandomness. □

Recall that in Section 3.2 we gave two different arguments for exponential convergence depending on whether the left or right error was larger. In this section we will use the pseudorandom condition to improve the argument in the case when δ_t small, i.e. when $\|\nabla^R_{A_t}\|_\infty \approx \|\nabla^L_{A_t}\|_\infty$ according to Definition 3.2.14. We handle the other case, when the $\|\nabla^R_{A_t}\|_\infty$ is larger, similarly to Corollary 3.2.16, and then combine the two to show exponential convergence for all time (instead of just after time $\kappa = \frac{\log d}{2\alpha}$) using a similar argument to Proposition 3.2.18 on strongly convex inputs.

Lemma 3.3.5. *Consider matrix tuple $A \in \mathrm{Mat}(d,n)^K$ with $s(A) \leq 1$ that is (α, β)-pseudorandom for $\beta \leq \frac{1}{2}$. If $\frac{1}{5} \geq \alpha \geq 14 \|\nabla^L_A\|_\infty$, then*

$$\delta \leq \frac{\alpha}{2} \implies -\partial_{t=0} \log \|\nabla^L_{A_t}\|_\infty \geq \frac{\alpha}{3},$$

where $\delta := \delta_0$ is given in Definition 3.2.14.

Proof. We will use pseudorandomness to show that the worst row error is pushed towards the average due to gradient flow. So let $i \in \arg\max_{i' \in [d]} |d \cdot r_{i'}(A) - s(A)|$ and recall by Fact 3.2.9 that

$$\partial_{t=0} r_i(A_t) = \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \Big((s - d \cdot r_i) + (s - n \cdot c_j) \Big). \tag{3.3}$$

Our plan is to show that the row term always pushes r_i towards the average, by bounding the contribution from the column terms using pseudorandomness and the condition $\delta \leq \frac{\alpha}{2}$. First, for n even, we bound

$$\begin{aligned} \left| \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 (n \cdot c_j - s) \right| &\leq \|\nabla_A^R\|_\infty \sup_{y \in H} \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 y_j \\ &= \|\nabla_A^R\|_\infty \left(r_i - 2 \min_{T \in \binom{n}{n/2}} \sum_{j \in T} \sum_{k=1}^{K} |(A_k)_{ij}|^2 \right) \\ &\leq \|\nabla_A^R\|_\infty \left(r_i - 2\alpha \frac{|T|}{dn} \right) = \|\nabla_A^R\|_\infty \left(r_i - \frac{\alpha}{d} \right), \end{aligned} \tag{3.4}$$

where the first step was by $y := \{n \cdot c_j - s\}_{j \in [n]} \in \|\nabla_A^R\|_\infty \cdot H$ where $H := \{y \in \mathbb{R}^n \mid \langle y, 1_n \rangle = 0, \|y\|_\infty \leq 1\}$, the second step used Fact 2.6.4 which shows the maximizers of any linear function over H are of the form $1_n - 21_T$ for some $T \in \binom{[n]}{n/2}$, and in the third step we used the pseudorandom property for $|T| \geq \beta n$ and $\beta \leq \frac{1}{2}$. The calculation for odd n is similar using $\beta n \leq \lfloor \frac{n}{2} \rfloor$ pseudorandomness (see the following Remark 3.3.6).

Now that we have bounded the contribution from the columns, we separate into two cases depending on the sign of the row error. So first consider the case $\|\nabla_A^L\| = d \cdot r_i - s$, meaning we want to show r_i is decreasing.

$$-\partial_{t=0} r_i(A_t) \geq r_i \|\nabla_A^L\|_\infty - \left| \sum_{j=1}^{n} \sum_{k=1}^{K} |(A_k)_{ij}|^2 (n \cdot c_j - s) \right| \geq r_i \|\nabla_A^L\|_\infty - \|\nabla_A^R\|_\infty \left(r_i - \frac{\alpha}{d} \right),$$

where the first step follows from Eq. (3.3) and the case assumption $\|\nabla_A^L\| = d \cdot r_i - s$, and the second step was by our bound in Eq. (3.4).

To show exponential convergence, we combine with the change in s.

$$
\begin{aligned}
-\partial_{t=0}\log\|\nabla^L_{A_t}\|_\infty &= \frac{-\partial_{t=0}(d\cdot r_i - s)}{\|\nabla^L_{A_t}\|_\infty} \geq \frac{\|\nabla^L_A\|_\infty(d\cdot r_i) - \|\nabla^R_A\|_\infty(d\cdot r_i - \alpha) - \|\nabla_A\|_{\mathfrak{t}}^2}{\|\nabla^L_A\|_\infty} \\
&\geq \alpha\frac{\|\nabla^R_A\|_\infty}{\|\nabla^L_A\|_\infty} + (d\cdot r_i)\frac{\|\nabla^L_A\|_\infty - \|\nabla^R_A\|_\infty}{\|\nabla^L_A\|_\infty} - \frac{\|\nabla^L_A\|^2_\infty + \|\nabla^R_A\|^2_\infty}{\|\nabla^L_A\|_\infty} \\
&= \alpha(1+\delta) - \delta\Big(s + \|\nabla^L_A\|_\infty\Big) - (1 + (1+\delta)^2)\|\nabla^L_A\|_\infty \\
&\geq \alpha - \delta\Big(s + \|\nabla^L_A\|_\infty - \alpha\Big) - 2.5\|\nabla^L_A\|_\infty \geq \frac{\alpha}{3},
\end{aligned}
$$

where the first step was by our case assumption $\|\nabla^L_A\| = s - d\cdot r_i$ (for the question of differentiability, see Remark 3.2.11), in the second step we substituted the lower bound for $-\partial_{t=0} r_i(A_t)$ just calculated along with Proposition 3.1.15 to calculate $\partial_t s(A_t)$, in the third step we rearranged terms and used Lemma 3.2.7 to bound $\|\nabla_A\|^2_{\mathfrak{t}} \leq \|\nabla^L_A\|^2_\infty + \|\nabla^R_A\|^2_\infty$, in the fourth step we used $\|\nabla^R_A\|_\infty = (1+\delta)\|\nabla^L_A\|_\infty$ by Definition 3.2.14 and our case assumption $d\cdot r_i = s + \|\nabla^L_A\|$, and the final steps were by our assumptions $s(A) \leq 1$, $\delta \leq \frac{\alpha}{2} \leq \frac{1}{10}$, and $\|\nabla^L_A\|_\infty \leq \frac{\alpha}{14}$. This verifies the lower bound for this case

In the other case $\|\nabla^L_A\|_\infty = s - d\cdot r_i$, we want to show r_i is increasing:

$$
\partial_{t=0} r_i(A_t) \geq r_i\|\nabla^L_A\|_\infty - \left|\sum_{j=1}^n\sum_{k=1}^K |(A_k)_{ij}|^2(n\cdot c_j - s)\right| = r_i\|\nabla^L_A\|_\infty - \|\nabla^R_A\|_\infty\left(r_i - \frac{\alpha}{d}\right),
$$

where the first step was by Eq. (3.3) and the case assumption $\|\nabla^L_A\| = s - d\cdot r_i$, and the second step was by our bound in Eq. (3.4).

Once again, we combine with the change in s to show exponential convergence:

$$
\begin{aligned}
-\partial_{t=0}\log\|\nabla^L_{A_t}\|_\infty &\geq \frac{d\cdot r_i\|\nabla^L_A\|_\infty - \|\nabla^R_A\|_\infty(d\cdot r_i - \alpha) + \|\nabla_A\|^2_{\mathfrak{t}}}{\|\nabla^L_A\|_\infty} \\
&\geq \alpha(1+\delta) - \delta(d\cdot r_i) = \alpha - \delta(s - \|\nabla^L_A\|_\infty - \alpha) \geq \frac{\alpha}{2},
\end{aligned}
$$

where we follow the same steps as the other case and use $\|\nabla_A\|_{\mathfrak{t}} \geq 0$ and the case assumption $d\cdot r_i = s - \|\nabla^L_A\|_\infty$. Therefore the lemma holds in both cases. □

Remark 3.3.6. *Note that Fact 2.6.4 for odd n actually requires $\beta n \leq \lfloor\frac{n}{2}\rfloor < \frac{n}{2}$. We will ignore detail and always require $\beta \leq \frac{1}{2}$ for simplicity, as the difference is negligible ($O(n^{-1})$) for all our results.*

Remark 3.3.7. *Equation 3.4 is the main consequence of pseudorandomness that we use in our analysis, and we could as well have used it as our sufficient condition for fast convergence, instead of pseudorandomness. But pseudorandomness enjoy strong multiplicative robustness as shown in Lemma 3.3.4, whereas the robustness we can establish for Eq.* (3.4) *is weaker and more difficult to prove. In Section 7.2.3, we generalize this analysis to the tensor scaling setting.*

To show that the error is decreasing exponentially when δ is large, we apply Lemma 3.2.15. This is exactly the same argument as Corollary 3.2.16, just with different parameters, so we omit the proof.

Corollary 3.3.8. *Let $A_t \in \mathrm{Mat}(d,n)^K$ be the solution to gradient flow, and assume $s(A_t) \geq 0.95$. Then for any $\frac{1}{5} \geq \alpha \geq 3\|\nabla_{A_t}\|_\infty$,*

$$\delta_t \geq \frac{\alpha}{2} \implies -\partial_t \log \|\nabla^R_{A_t}\|_\infty \geq \frac{\alpha}{3}.$$

Note that Corollary 3.3.8 requires the lower bound $s \geq 0.95$, whereas Lemma 3.3.5 requires an upper bound $s \leq 1$. We will eventually use properties of fast convergence to show these are both satisfied for all time.

We emphasize that the two convergence arguments in Lemma 3.3.5 and Corollary 3.3.8 only require $\frac{\alpha}{\varepsilon} \gtrsim 1$, which is the reason for our improvement over the requirement $\frac{\alpha}{\varepsilon} \gtrsim \log d$ in Theorem 3.2.19. This will be used in Chapter 4 to give optimal bounds for the Paulsen problem.

We can combine the above two lemmas to show that the error converges exponentially for all time. The following argument is elementary, but slightly more technical than the proof of Proposition 3.2.18.

Proposition 3.3.9. *Consider ε-doubly balanced $A \in \mathrm{Mat}(d,n)$, and let A_t be the solution to gradient flow according to Definition 3.1.14. If for all $t \in [0,T]$, the following assumptions are satisfied: (1) $0.95 \leq s(A_t) \leq 1$; (2) A_t is (α,β)-pseudorandom with $\frac{1}{5} \geq \alpha$ and $\beta \leq \frac{1}{2}$; (3) $\alpha \geq 16\varepsilon$; (4) $\max\{\|\nabla^L_{A_t}\|_\infty, \|\nabla^R_{A_t}\|_\infty\} \leq 1.1\varepsilon$; then*

$$\max\{\|\nabla^R_T\|_\infty, \|\nabla^L_T\|_\infty\} \leq \varepsilon(1+\alpha/2)e^{-\alpha T/3} \quad \textit{and} \quad \max\{\|X_T\|_\infty, \|Y_T\|_\infty\} \leq \frac{3(1+\alpha/2)\varepsilon}{\alpha}.$$

Proof. The conditions (1)-(4) are defined so that we can use the previous fast convergence analysis. Explicitly, the conditions of Lemma 3.3.5 are satisfied for all $t \in [0,T]$ as (1) implies $s(A) \leq 1$, (2) implies $\beta \leq \frac{1}{2}$, and (3), (4) together imply $\frac{1}{5} \geq \alpha \geq 16\varepsilon \geq 14\|\nabla^L_{A_t}\|_\infty$.

Similarly, the conditions of Corollary 3.3.8 are satisfied for all $t \in [0,T]$ as (1) implies $s(A) \geq 0.95$, (2) implies $\beta \leq \frac{1}{2}$, and (3), (4) together imply $\frac{1}{5} \geq \alpha \geq 16\varepsilon \geq 3 \cdot 2.2\varepsilon \geq 3\|\nabla_{A_t}\|_\infty$. Therefore, in the sequel, we apply Lemma 3.3.5 and Corollary 3.3.8 freely without checking conditions.

The proof plan is similar to Proposition 3.2.18, so partition $[0,T]$ into two pieces depending on which of ∇^L, ∇^R is converging quickly.

$$T_R := \{t \in [0,T] \mid \delta_t \geq \alpha/2\}, \qquad \text{and} \qquad T_L := \{t \in [0,T] \mid \delta_t \leq \alpha/2\}.$$

Since all quantities are continuous, this gives the decomposition

$$[0,T] = \{[0,t_1],[t_1,t_2],...\},$$

where $[t_m, t_{m+1}]$ are maximal intervals fully contained in either T_L or T_R. In the remainder of this proof, we assume A is fixed and use $\nabla_t := \nabla_{A_t}$ as shorthand to avoid triple subscripts. We show exponential convergence for both $\|\nabla_t^L\|_\infty$ and $\|\nabla_t^R\|_\infty$ for $t \in [0,t_m]$ and induct on m. The argument for the base case is slightly different as we use the fact that A is ε-doubly balanced.

First consider the case $[0,t_1] \subseteq T_R$, and let $t \in [0,t_1]$ be arbitrary. Then we can apply Corollary 3.3.8 to bound

$$\log\|\nabla_t^R\|_\infty = \log\|\nabla_A^R\|_\infty + \int_0^t \partial_{\tilde{t}} \log\|\nabla_{\tilde{t}}^R\|_\infty \leq \log\varepsilon - \frac{\alpha t}{3},$$

where the first step was by the fundamental theorem of calculus, and the last step was by Corollary 3.3.8 applied with $\delta \geq \frac{\alpha}{2}$. Therefore, the statement is true at time t for ∇^R. To transfer this to the left side, we continue to use Definition 3.2.14 of δ to show

$$\|\nabla_t^L\|_\infty = \frac{\|\nabla_t^R\|_\infty}{1+\delta_t} \leq \frac{\varepsilon e^{-\alpha t/3}}{1+\alpha/2},$$

where the first step was by Definition 3.2.14 of δ, and in the last step we used the case assumption $\delta_t \geq \alpha/2$ as $t \in T_R$. Since the choice of $t \in [0,t_1]$ was arbitrary, both the left and right error satisfy the convergence bound for $[0,t_1] \subseteq T_R$.

For the other case $[0,t_1] \subseteq T_L$, consider any $t \in [0,t_1]$. Then we can apply Lemma 3.3.5 to bound

$$\log\|\nabla_t^L\|_\infty = \log\|\nabla_A^L\|_\infty + \int_0^t \partial_{\tilde{t}} \log\|\nabla_{\tilde{t}}^L\|_\infty \leq \log\varepsilon - \frac{\alpha t}{3},$$

where again the first step was by the fundamental theorem of calculus, and in the last step we substituted $\|\nabla_A^L\|_\infty \leq \varepsilon$ by the ε-doubly balanced condition and applied Lemma 3.3.5 with $\delta \leq \frac{\alpha}{2}$. To transfer this to the right side, we continue

$$\|\nabla_t^R\|_\infty = (1+\delta_t)\|\nabla_t^L\|_\infty \leq (1+\alpha/2)\varepsilon e^{-\alpha t/3},$$

where in the last step we used $\delta_t \leq \alpha/2$ by definition of T_L. Since the choice of $t \in [0, t_1]$ was arbitrary, the statement holds for the first segment.

For the induction, assume we have proved the convergence for $[0, t_m]$ so

$$\max\{\|\nabla_{A_t}^L\|_\infty, \|\nabla_{A_t}^R\|_\infty\} \leq \varepsilon(1+\alpha/2)e^{-\alpha t/3}$$

for all $t \in [0, t_m]$ with $m \geq 1$. Consider the case $[t_m, t_{m+1}] \subseteq T_R$ and let $t \in [t_m, t_{m+1}]$ be arbitrary. We will show $\|\nabla_{A_t}^R\|_\infty$ is small enough by Corollary 3.3.8, and then use Definition 3.2.14 of δ_t to transfer this conclusion to $\|\nabla_{A_t}^L\|_\infty$. So we calculate

$$\log\|\nabla_t^R\|_\infty - \log\|\nabla_{t_m}^R\|_\infty = \int_{t_m}^t \partial_{\tilde{t}} \log\|\nabla_{\tilde{t}}^R\|_\infty \leq -\frac{\alpha}{3}(t-t_m),$$

where the first step was by the fundamental theorem of calculus, and the inequality was by Corollary 3.3.8. Exponentiating both sides and using the induction hypothesis gives

$$\|\nabla_t^R\|_\infty \leq \|\nabla_{t_m}^R\|_\infty e^{-\alpha(t-t_m)/3} \leq (1+\alpha/2)\varepsilon e^{-\alpha t/3}.$$

Now we transfer this bound to ∇^L. We will use that $t \in [t_m, t_{m+1}] \subseteq T_R$ and t_m is the endpoint of the interval, so $\delta_{t_m} = \frac{\alpha}{2} \leq \delta_t$.

$$\|\nabla_t^L\|_\infty = \frac{1}{1+\delta_t}\|\nabla_t^R\|_\infty \leq \frac{1+\alpha/2}{1+\delta_t}\varepsilon e^{-\alpha t/3} \leq \varepsilon e^{-\alpha t/3},$$

where we used Definition 3.2.14 of δ_t in the first step, the second step was by the bound on $\|\nabla_t^R\|_\infty$ calculated above, and the final inequality uses the fact that $0 \leq \frac{\alpha}{2} \leq \delta_t$ since $t \in [t_m, t_{m+1}] \subseteq T_R$.

In the other case, if $[t_m, t_{m+1}] \subseteq T_L$, then Lemma 3.3.5 shows for any $t \in [t_m, t_{m+1}]$

$$\log\|\nabla_t^L\|_\infty - \log\|\nabla_{t_m}^L\|_\infty = \int_{t_m}^t \partial_{\tilde{t}} \log\|\nabla_{\tilde{t}}^L\|_\infty \leq -\frac{\alpha}{3}(t-t_m).$$

Exponentiating both sides and using the induction hypothesis gives

$$\|\nabla_t^L\|_\infty \leq \|\nabla_{t_m}^L\|_\infty e^{-\alpha(t-t_m)/3} \leq (1+\alpha/2)\varepsilon e^{-\alpha t/3}.$$

Now we can transfer this convergence to ∇^R. We again want to use that $t \in [t_m, t_{m+1}] \subseteq T_L$ and t_m is the endpoint of the interval, so $\delta_t \leq \frac{\alpha}{2} = \delta_{t_m}$. This allows us to bound

$$\begin{aligned}\|\nabla_t^R\|_\infty &= (1+\delta_t)\|\nabla_t^L\|_\infty \leq (1+\delta_t)\|\nabla_{t_m}^L\|_\infty e^{-\alpha(t-t_m)/3} \\ &= \frac{1+\delta_t}{1+\delta_{t_m}}\|\nabla_{t_m}^R\|_\infty e^{-\alpha(t-t_m)/3} \leq (1+\alpha/2)\varepsilon e^{-\alpha t/3}\end{aligned}$$

where we used Definition 3.2.14 of δ_t in the first and third step, in the second step we used the convergence of ∇_t^L derived above, and in the final step we used the induction hypothesis on $\nabla_{t_m}^R$ as well as the fact that $\delta_t \leq \frac{\alpha}{2} = \delta_{t_m}$ since $t \in [t_m, t_{m+1}] \subseteq T_L$.

Therefore, we have shown the exponential convergence on $\max\{\|\nabla_{A_t}^L\|_\infty, \|\nabla_{A_t}^R\|_\infty\} \leq \varepsilon(1+\alpha/2)e^{-\alpha t/3}$ for all $t \in [0, T]$.

To bound the scalings, we calculate

$$\|X_T\|_\infty = \left\|\int_0^T -\nabla_t^L\right\|_\infty \leq \int_0^T \|\nabla_t^L\|_\infty \leq \varepsilon(1+\alpha/2)\int_0^T e^{-\alpha t/3} \leq \frac{3(1+\alpha/2)\varepsilon}{\alpha},$$

where the first step was by the fundamental theorem of calculus as $X_0 = 0$ and $\partial_t X_t = -\nabla_{A_t}^L$ by Definition 3.1.14 of gradient flow, the second step was by triangle inequality of $\|\cdot\|_\infty$, and the third step was due to bound on ∇^L derived above. The calculation for $\|Y_T\|_\infty$ is exactly the same except that we use the bound on $\|\nabla^R\|_\infty$. □

Finally, we can use the robustness from Lemma 3.3.4 to show that fast convergence follows when the input is sufficiently pseudorandom.

Theorem 3.3.10. *If matrix tuple $A \in \mathrm{Mat}(d,n)^K$ of size $s(A) = 1$ is ε-doubly balanced and (α, β)-pseudorandom for $\frac{1}{5} \geq \alpha \geq 16e \cdot \varepsilon$ and $\beta \leq \frac{1}{2}$, then*

1. *For all time $t \geq 0$,*

$$\max\{\|\nabla_{A_t}^R\|_\infty, \|\nabla_{A_t}^L\|_\infty\} \leq (1+\alpha/2)\varepsilon e^{-\frac{\alpha t}{3e}} \qquad \text{and} \qquad \max\{\|X_t\|_\infty, \|Y_t\|_\infty\} \leq \frac{9\varepsilon}{\alpha}.$$

2. *The limit $(X_\infty, Y_\infty) := \lim_{t\to\infty}(X_t, Y_T)$ exists and $A_\infty := e^{X_\infty/2}Ae^{Y_\infty/2}$ gives a solution to the matrix scaling problem in Definition 3.1.3 on input A.*

3. *The size of the scaling solution can be lower bounded*

$$s(A_\infty) = f_A(X_\infty, Y_\infty) \geq 1 - \frac{10\varepsilon^2}{\alpha}.$$

Proof. We claim that the assumptions of Proposition 3.3.9 always hold, and in fact we can impose the stronger requirement A_t is $(\frac{\alpha}{e}, \beta)$-pseudorandom for all time. This satisfies assumption (3) of Proposition 3.3.9 as $\frac{\alpha}{e} \geq \frac{16e}{e} \cdot \varepsilon \geq 16\varepsilon$ by our assumption $\alpha \geq 16e \cdot \varepsilon$. From this claim, the conclusions of this theorem will follow. So for contradiction, let T be the first time one of the conditions fail. Then the assumptions (1)-(4) hold simultaneously for all $t \in [0, T]$ and we can apply Proposition 3.3.9 freely up till this time. We will show that in fact all the assumptions are strictly satisfied at time T, which will give the desired contradiction.

Note that (4) cannot fail first, as Proposition 3.3.9 shows

$$\max\{\|\nabla^L_{A_T}\|_\infty, \|\nabla^R_{A_T}\|_\infty\} \leq \varepsilon(1 + \alpha/2e)e^{-\alpha T/3e} \leq 1.1\varepsilon,$$

where in the last step we substituted $T \geq 0$ and $\alpha \leq \frac{1}{5}$.

Next, we show that (1) cannot fail first. The upper bound is clear as $s(A) = 1$ and $\partial_t s(A_t) \leq 0$ by Proposition 3.1.15. For the lower bound, we calculate

$$1 - s(A_T) = \int_0^T \|\nabla_{A_t}\|_{\mathfrak{t}}^2 \leq \int_0^T (\|\nabla^L_{A_t}\|^2_\infty + \|\nabla^R_{A_t}\|^2_\infty) \leq \int_0^T 2\varepsilon^2(1.1)^2 e^{-2\alpha t/3e} \leq \frac{10\varepsilon^2}{\alpha} < \frac{1}{100},$$

where the first step was by the fundamental theorem of calculus as well as Proposition 3.1.15 with the assumption $s(A_0) = 1$, in the second step we used Lemma 3.2.7, the third step was by the convergence derived above, and the final steps were by the assumptions $\alpha \leq \frac{1}{5}$ and $\varepsilon \leq \frac{\alpha}{16e} < \frac{1}{100}$. Therefore the size condition (1) could not have failed at time T.

Clearly (2) cannot fail as we can always decrease α to satisfy the upper bound.

Finally, assume that our stronger pseudorandom requirement fails at time T so A_T is not $(\frac{\alpha}{e}, \beta)$-pseudorandom. Since A is (α, β)-pseudorandom, Lemma 3.3.4 in contrapositive implies that $\|(X_T, Y_T)\|_\infty \geq 1$. But then we simply apply Proposition 3.3.9 with $(\frac{\alpha}{e}, \beta)$-pseudorandomness until time T to bound

$$\max\{\|X_T\|_\infty, \|Y_T\|_\infty\} \leq 3\frac{(1 + \alpha/2e)\varepsilon}{\alpha/e} \leq \frac{9\varepsilon}{\alpha} \leq \frac{9}{16e} < \frac{1}{2},$$

where the first step was by the bounds on $\max\{\|X_T\|_\infty, \|Y_T\|_\infty\}$ given in Proposition 3.3.9, the second was by the assumption $\alpha \leq \frac{1}{5}$, and the third is by our assumption $\alpha \geq 16e \cdot \varepsilon$.

This is the desired contradiction, so Proposition 3.3.9 applies for all time. Therefore, items (1) and (3) in this theorem follows by the calculations above with $T \in [0, \infty]$.

To show item (2), note that we already have shown

$$\lim_{t \to \infty} \|\partial_t (X_t, Y_t)\|_\infty = \lim_{t \to \infty} \|\nabla_{A_t}\|_\infty = 0,$$

where the first step was by Definition 3.1.14 of gradient flow and the last step is by item (1). Therefore the limit (X_∞, Y_∞) exists. Further, $\nabla_{A_\infty} = 0$, so A_∞ is doubly balanced by Proposition 3.1.10(2). □

This dimension-independent requirement on $\frac{\alpha}{\varepsilon}$ will be the key to our optimal distance bound for the Paulsen problem in Chapter 4. We will also see a generalization of this argument to tensor scaling in Chapter 7.

3.4 Pseudorandom property and Convexity

In this section, we will study the relation between the pseudorandom property given in Definition 3.3.1 and the strong convexity property given in Definition 3.2.1. The main technical result of this section is given in Theorem 3.4.7, where we show that if A is ε-doubly balanced and (α, β)-pseudorandom for small enough constants ε and β, then A is $\Omega(\alpha)$-strongly convex. This will be useful in Section 8.5, where we will use strong convexity to give algorithmic guarantees for pseudorandom inputs.

We first associate a bipartite graph to each matrix tuple. This will allow us to use ideas from spectral graph theory.

Definition 3.4.1 (Associated Graph). *Let $H = (V, E, w)$ be an undirected graph with vertex set V, edge set $E \subseteq \binom{V}{2}$, and edge weights $w : E \to \mathbb{R}_+$. The adjacency matrix $W \in \mathbb{R}^{V \times V}$ is defined $W_{uv} := w(u, v)$ for vertices $u, v \in V$; The degree of vertex v is $d(v) := \sum_{u \in V} w_{uv}$ and the diagonal degree matrix is $(D)_{vv} := d(v)$; the Laplacian of H is the matrix $L \in \mathbb{R}^{V \times V}$ defined as*

$$L := \sum_{(u,v) \in E} w_{uv}(e_u - e_v)(e_u - e_v)^*$$

where $\{e_v\}_{v \in V}$ is the standard basis in $\mathbb{R}^V$. Note that this can be written $L = D - A$.

For matrix tuple $A \in \mathrm{Mat}(d, n)^K$, we associate bipartite graph $H_A = ([d] \cup [n], w)$ with

$$w_{ij} := \sum_{k=1}^{K} |(A_k)_{ij}|^2$$

for $i \in [d], j \in [n]$. We will often use $w(S, T) := \sum_{i \in S} \sum_{j \in T} w_{ij}$ as shorthand for $S \subseteq [d], T \subseteq [n]$. The Laplacian of $A \in \mathrm{Mat}(d, n)^K$ is the graph Laplacian of H_A:

$$L_A := \sum_{i=1}^{d} \sum_{j=1}^{n} w_{ij}(e_i - e_j)(e_i - e_j)^*,$$

where $\{e_i\}_{i\in[d]}$ is the standard basis for $\mathbb{R}^d \subseteq \mathbb{R}^d \oplus \mathbb{R}^n$ and $\{e_j\}_{j\in[n]}$ is the standard basis for $\mathbb{R}^n \subseteq \mathbb{R}^d \oplus \mathbb{R}^n$. We may drop the subscript $L = L_A$ if the input tuple is understood.

This allows us to extend the definitions of the previous sections to graphs; in particular the definitions of balanced, pseudorandom and strongly convex inputs.

Definition 3.4.2. *Bipartite graph $H = ([d]\cup[n], w)$ is normalized if $w([d],[n]) = w(E) = 1$. It is called ε-bi-regular if*

$$\forall i \in [d] : \sum_{j\in[n]} w_{ij} \in \frac{1 \pm \varepsilon}{d}, \qquad \forall j \in [n] : \sum_{i\in[d]} w_{ij} \in \frac{1 \pm \varepsilon}{n},$$

and it is called bi-regular if $\varepsilon = 0$.

Note if $H = H_A$ for matrix tuple $A \in \mathrm{Mat}(d,n)^K$, then the normalization condition exactly corresponds to the size condition $s(A) = 1$, and ε-bi-regularity exactly corresponds to A being ε-doubly balanced according to Definition 3.1.2.

We also translate the definitions of pseudorandom and strong convexity to graphs.

Definition 3.4.3. *Weighted bipartite graph $H = ([d] \cup [n], w)$ is (α,β)-pseudorandom if for every $S \subseteq [d]$ and $T \subseteq [n]$ such that $|T| \geq \beta n$ we have*

$$w(S,T) = \sum_{i\in S}\sum_{j\in T} w_{ij} \geq \alpha \frac{|S|}{d}\frac{|T|}{n} = \alpha st,$$

where we used shorthand $\frac{|S|}{d} = s$, $\frac{|T|}{n} = t$.

If H_A is the associated graph for matrix tuple $A \in \mathrm{Mat}(d,n)^K$, then this corresponds to (α,β)-pseudorandomness of A according to Definition 3.3.1.

Remark 3.4.4. *[62] gave a slightly different condition named pseudorandomness in Definition 4.3.2: $\forall i \in [d]$ denote the bad elements $B_i := \{j \in [n] \mid w_{ij} < \frac{\alpha}{dn}\}$; then W is pseudorandom if $\forall i : |B_i| \leq \beta n$. This is a strictly stronger condition:*

$$\sum_{i\in S, j\in T} w_{ij} \geq \sum_{i\in S}\sum_{j\in T-B_i} w_{ij} \geq \frac{\alpha}{dn}\sum_{i\in S}|T - B_i| \geq \alpha\frac{|S|}{d}\frac{|T| - \beta n}{n}.$$

For all $|T| \geq 2\beta n$ the last term is at least $|T|/2n$ so a graph satisfying their pseudorandom condition is $(\alpha/2, 2\beta)$-pseudorandom in our definition.

In our proof of Theorem 3.4.7, we show improved strong convexity results using the weaker notion of pseudorandomness given in Definition 3.4.3.

Definition 3.4.5. *Weighted bipartite graph $H = ([d] \cup [n], w)$ is α-strongly convex iff for all $(x, y) \in \mathbb{R}^d \oplus \mathbb{R}^n$ such that $\sum_{i=1}^d x_i = \sum_{j=1}^n y_j = 0$:*

$$(x,y)^* L_H(x,y) = \sum_{i\in[d]} \sum_{j\in[n]} w_{ij}(x_i - y_j)^2 \geq \alpha \left(\frac{1}{d} \sum_{i=1}^d x_i^2 + \frac{1}{n} \sum_{j=1}^n y_j^2 \right).$$

The fact below shows the connection between the Laplacian of H_A and convexity of A.

Lemma 3.4.6. *Let $H = H_A$ be the associated graph for matrix tuple $A \in \operatorname{Mat}(d,n)^K$ according to Definition 3.4.1. For any $(x,y) \in \mathbb{R}^d \oplus \mathbb{R}^n$ such that $\sum_{i=1}^d x_i = \sum_{j=1}^n y_j = 0$, letting $X := \operatorname{diag}(x) \in \operatorname{diag}(d), Y := \operatorname{diag}(y) \in \operatorname{diag}(n)$ gives the relation*

$$\partial^2_{\eta=0} f_A(\eta X, -\eta Y) = \sum_{i=1}^d \sum_{j=1}^n \sum_{k=1}^K |(A_k)_{ij}|^2 (X_i - Y_j)^2 = \sum_{i=1}^d \sum_{j=1}^n w_{ij}(x_i - y_j)^2 = (x,y)^* L_A(x,y).$$

As a consequence, H_A is an α-strongly convex graph according to Definition 3.4.5 iff A is an α-strongly convex matrix tuple according to Definition 3.2.1.

Proof. The first statement follows by Lemma 3.1.9 and Definition 3.4.1 of the Laplacian. Strong convexity follows by considering the infimum over all $(X,Y) \in \mathfrak{t}$ as described in Definition 3.2.1. □

We now present the main technical result of this section, showing (α,β)-pseudorandomness implies $\Omega(\alpha)$-strong convexity. For this purpose we will need some mild conditions on regularity and that β is small enough.

Theorem 3.4.7. *Consider matrix tuple $A \in \operatorname{Mat}(d,n)^K$ of size $s(A) = 1$ that is ε-doubly balanced with $\varepsilon \leq \frac{1}{16}$. If A is (α,β)-pseudorandom according to Definition 3.3.1 for $\alpha \leq \frac{1}{16}$ and $\beta \leq \frac{1}{16}$, then A is $e^{-11}\alpha$-strongly convex according to Definition 3.2.1.*

Equivalently, for a normalized ε-bi-regular bipartite graph H that satisfies the (α,β)-pseudorandom condition according to Definition 3.4.3, if $\varepsilon, \alpha, \beta$ are all at most $\frac{1}{16}$, then H is $e^{-12}\alpha$-strongly convex according to Definition 3.4.5.

Remark 3.4.8. *A weaker version of this theorem was given in Prop 4.3.3 of [62], specifically showing (α,β)-pseudorandomness implies $\Omega(\frac{\alpha}{d})$-strong convexity. Though the language of that work was different, the purpose was to show pseudorandomness for perturbed frames and then use fast convergence derived from strong convexity to show a nearby doubly balanced frame.*

The improvement we make in our proof requires the assumption that A is nearly doubly balanced. This allows us to slightly simplify the case analysis as well as give a quantitatively stronger result in one of the cases.

Recall that our goal is to show, given (α, β)-pseudorandom graph $(H, [d] \cup [n], w)$ that for any $x \in \mathbb{R}^d, y \in \mathbb{R}^n$ such that $\langle x, \mathbf{1}_d \rangle = \langle y, \mathbf{1}_n \rangle = 0$. We want to show

$$(x, y)^* L_H(x, y) = \sum_{i=1}^{d} \sum_{j=1}^{n} w_{ij}(x_i - y_j)^2$$

is large as compared to $\frac{1}{d} \sum_{i=1}^{d} x_i^2 + \frac{1}{n} \sum_{j=1}^{n} y_j^2$.

From this point on, fix $x \in \mathbb{R}^d, y \in \mathbb{R}^n$ with $\sum_{i \in [d]} x_i = \sum_{j \in [n]} y_j = 0$. Therefore we are more interested in lower bounding the terms $w(i, j)(x_i - y_j)^2$ for those rows and columns with large $\frac{x_i^2}{d}, \frac{y_j^2}{n}$ respectively.

Informally, we want to show H concentrates a significant amount of weight in edges (i, j) for which $|x_i - y_j|$ is large. Our plan is to separate the rows and columns into buckets depending on the value x_i, y_j. Then we will find a large subset of rows and columns from different buckets, and use pseudorandomness to show H has large weight in these edges.

Definition 3.4.9. *For fixed $(x, y) \in \mathbb{R}^d \oplus \mathbb{R}^n$, parameter $\gamma \in (0, 1)$ chosen later, and $k, \ell \in \mathbb{N}$ we define:*

$$R_{k^\pm} := \{i \in [d] \mid \gamma^{2k+1} \le x_i^2 \le \gamma^{2k-1}, sgn(x_i) = \pm\},$$

$$C_{\ell^\pm} := \{j \in [n] \mid \gamma^{2\ell+1} \le y_j^2 \le \gamma^{2\ell-1}, sgn(y_j) = \pm\}.$$

We will often use the shorthand $R_{k\pm 1} := R_{k-1} \cup R_k \cup R_{k+1}$ or $C_{\ell \pm 1} = C_{\ell-1} \cup C_\ell \cup C_{\ell+1}$.

Note that smaller γ implies better lower bounds for different buckets by Fact 3.4.10, but larger γ means that the values in a single bucket are closer together. Note also that (α, β)-pseudorandomness implies (α, β')-pseudorandom for every $\beta' \ge \beta$. Therefore, we will choose both β and γ at the end of the proof to optimize the lower bound.

Fact 3.4.10. *For buckets of the same sign, consider $i \in R_k, j \in C_\ell$ with $|k - \ell| > 1$. Then*

$$(x_i - y_j)^2 \ge (1 - \gamma)^2 \max\{x_i^2, y_j^2\}.$$

Proof. Assume without loss that $i \in R_k, j \in C_\ell$ for some $\ell \geq k+2$ and the signs are both positive. Then

$$y_j^2 \leq \gamma^{2\ell-1} = \gamma^{2(\ell-k-1)}\gamma^{2k+1} \leq \gamma^2 x_i^2,$$

where we used Definition 3.4.9 of buckets and the fact that $\ell \geq k+2$. This implies

$$(x_i - y_j)^2 \geq (1-\gamma)^2 x_i^2 = (1-\gamma)^2 \max\{x_i^2, y_j^2\},$$

as $i \in R_k$ and $j \in C_\ell$ for $\ell > k$. The other case $y_j^2 \geq x_i^2$ follows from a similar calculation. □

The above fact is our main tool for lower bounding contributions in the quadratic form

$$(x,y)^* L_H(x,y) = \sum_{i\in[d]} \sum_{j\in[n]} w_{ij}(x_i - y_j)^2.$$

We will mostly refer to pairs of buckets R_k, C_ℓ which have the same sign and drop the superscript. If we can show the weights $w(R_k, \overline{C_{k\pm1}})$ and $w(\overline{R_{\ell\pm1}}, C_\ell)$ are large, these buckets provides a strong lower bound. This may not be possible for every row/column bucket, so our plan is to use pseudorandomness and regularity to find a large subset with significant contribution. Our proof will divide into two cases, depending on which of $\frac{\|x\|_2^2}{d}$ or $\frac{\|y\|_2^2}{n}$ is larger, and lower bound the quadratic form by rows/columns, respectively.

To this end, we first define the set of good rows as those whose contribution can be lower bounded using pseudorandomness.

Definition 3.4.11. *For fixed x, y and constant β chosen later, we define*

$$K_B := \{k \mid |C_{k\pm1}| > (1-\beta)n\}$$

to be the indices of bad row buckets, and we define good and bad rows as

$$R_B := \cup_{k\in K_B} R_k, \qquad R_G := [d] - R_B.$$

Recall that Definition 3.4.3 of pseudorandomness implies that every row has significant weight going to every large enough subset of columns. Therefore, the above definition of good rows simply implies the following lower bound on the quadratic form.

Lemma 3.4.12. *Let H be a normalized ε-bi-regular graph that is (α,β)-pseudorandom. For fixed $(x,y) \in \mathbb{R}^d \oplus \mathbb{R}^n$, the Laplacian quadratic form can be lower bounded by*

$$(x,y)^* L(x,y) \geq \sum_{i\in R_G} \sum_{j=1}^{n} w_{ij}(x_i + y_j)^2 \geq \alpha\beta(1-\gamma)^2 \sum_{i\in R_G} \frac{x_i^2}{d},$$

where R_G are the good row buckets with respect to (x,y) given in Definition 3.4.11.

Proof. Consider any good row $i \in R_k$ with $k \in K_G$ according to Definition 3.4.11. This means $|\overline{C_{k\pm1}}| \geq \beta n$, so we can apply the pseudorandom condition to lower bound $w(i, \overline{C_{k\pm1}}) \geq \alpha \frac{|\overline{C_{k\pm1}}|}{dn} \geq \alpha \frac{\beta}{d}$. For any such $i \in R_k \subseteq R_G$, this gives the lower bound

$$\sum_{j=1}^{n} w_{ij}(x_i - y_j)^2 \geq \sum_{j \notin C_{k\pm1}} w_{ij}(x_i - y_j)^2 \geq (1-\gamma)^2 x_i^2 w(i, \overline{C_{k\pm1}}) \geq \alpha(1-\gamma)^2 \beta \frac{x_i^2}{d},$$

where the second step was by Fact 3.4.10 applied to $i \in R_k, j \notin C_{k\pm1}$, and the third was by the bound $w(i, \overline{C_{k\pm1}}) \geq \alpha \frac{\beta}{d}$ derived above using pseudorandomness. The lemma then follows by combining the contributions of these good rows. □

At this point, we would be done if R_G was a large portion of the total, i.e.

$$\sum_{i \in R_G} \frac{x_i^2}{d} \gtrsim \frac{\|x\|_2^2}{d} + \frac{\|y\|_2^2}{n}.$$

We will show below that for small enough β, the bad rows R_B are concentrated in a specific way and have all their weight concentrated in a few columns. With this structure, in the case when $\frac{\|x\|_2^2}{d} \gg \frac{\|y\|_2^2}{n}$, we will be able to show

$$\sum_{i \in R_B} \frac{x_i^2}{d} \lesssim \frac{\|y\|_2^2}{n} \ll \frac{\|x\|_2^2}{d},$$

and we will be done by Lemma 3.4.12.

To upper bound the bad row buckets, we will also use the pseudorandom condition. We first show that all the bad rows are close together and that they have weight concentrated in a specific subset of columns.

Lemma 3.4.13. *If $R_B \neq \emptyset$, there is a subset of columns $C \subseteq [n]$ such that $|C| > (1-\beta)n$, and further for every $i \in R_B, j \in C$, it holds that*

$$\gamma^6 y_j^2 \leq x_i^2 \leq \gamma^{-6} y_j^2.$$

Proof. For $k \in K_B$, by Definition 3.4.11 we have $|C_{k\pm1}| > (1-\beta)n$. We first show that all bad row buckets are close together. If $k, k' \in K_B$ such that $|k - k'| \geq 3$, then $C_{k\pm1} \cap C_{k'\pm1} = \emptyset$ and

$$n \geq |C_{k\pm1}| + |C_{k'\pm1}| > 2(1-\beta)n,$$

which is a contradiction for $\beta < \frac{1}{2}$. Similarly if R_B contains buckets of different signs, this will produce the same contradiction.

Therefore, without loss of generality let k be such that $R_k \subseteq R_B \subseteq R_{k\pm1}$. Then we choose $C := C_{k\pm1}$ and note $|C| = |C_{k\pm1}| > (1-\beta)n$ by Definition 3.4.11. Further, $i \in R_B, j \in C$ implies

$$x_i^2 \leq \gamma^{2(k-1)-1} = \gamma^{-6}\gamma^{2(k+1)+1} \leq \gamma^{-6} y_j^2,$$

where the first and last steps were by Definition 3.4.9 of bucketing. The lower bound follows by an analogous calculation, so the lemma follows. □

We will use the structure on R_B and C shown in Lemma 3.4.13 to upper bound the contribution of the bad rows in terms of the columns. Specifically, we will use the fact that $\sum_{j\in[n]} y_j = 0$ along with the small size $|\overline{C}| < \beta n$ to show

$$\sum_{i\in R_B} \frac{x_i^2}{d} \leq c \sum_{j\notin C} \frac{y_j^2}{n} \leq c \frac{\|y\|_2^2}{n}.$$

Importantly, we will be able to tune the constant c to be as small as desired by requiring smaller β. We emphasize that this is the only place the assumption $\sum_{j\in[n]} y_j = 0$ is used.

Lemma 3.4.14. *For R_B, C defined according to Definition 3.4.11 and 3.4.13 respectively:*

$$\sum_{i\in R_B} \frac{x_i^2}{d} \leq \frac{\beta}{\gamma^6(1-\beta)^2} \frac{\|y\|_2^2}{n}.$$

Proof. We will prove the following sequence of inequalities to show the lemma:

$$\sum_{i\in R_B} \frac{x_i^2}{d} \leq \sum_{i\in R_B} \frac{x_i^2}{|R_B|} \leq \frac{\beta}{\gamma^6(1-\beta)^2} \sum_{j\in\overline{C}} \frac{y_j^2}{n} \leq \frac{\beta}{\gamma^6(1-\beta)^2} \frac{\|y\|_2^2}{n}.$$

The first and last inequalities are clear as $R_B \subseteq [d]$ and $C \subseteq [n]$. To show the middle inequality, we will use the fact that all rows of R_B are close together, and charge them to the columns of C.

$$\sum_{i\in R_B} \frac{x_i^2}{|R_B|} \leq \min_{j\in C} \left(\frac{y_j}{\gamma^3}\right)^2 \leq \left(\frac{\sum_{j\in C} y_j}{\gamma^3 |C|}\right)^2 = \frac{\left(\sum_{j\in\overline{C}} y_j\right)^2}{\gamma^6 |C|^2} \leq \frac{|\overline{C}| \sum_{j\in\overline{C}} y_j^2}{\gamma^6 |C|^2} \leq \frac{\beta}{\gamma^6(1-\beta)^2} \sum_{j\in\overline{C}} \frac{y_j^2}{n},$$

where the first step was by the approximation ratio $x^2_{i\in R_B} \leq \gamma^{-6} y^2_{j\in C}$ shown in Lemma 3.4.13, in the second step we simply bounded the min by the average, in the third step we used the assumption $\sum_{j\in[n]} y_j = 0$, the fourth step was by Cauchy-Schwarz, and the final step was by Lemma 3.4.13 giving $|C| > (1-\beta)n$. □

Remark 3.4.15. *While Definition 3.4.5 includes the condition $\sum_{i\in[d]} x_i = 0$, we do not use this condition anywhere in our proof. We can use this condition to bound $|R_B|$ in a manner similar to the proof of Lemma 3.4.14. But this argument does not give an improvement in the final constant in the proof of Theorem 3.4.7, so we choose to omit it.*

Remark 3.4.16. *Up till this point, the argument is very similar to section 4.4 of [62], and in that work we were also able to give a strong lower bound ($\Omega(\alpha)$ vs $\Omega(\alpha/d)$) for the case $\frac{\|x\|_2^2}{d} \gg \frac{\|y\|_2^2}{n}$ discussed above. Our improvement comes from a different case analysis and the use of regularity to bound the other case.*

We emphasize that the ratio $\frac{\beta}{(1-\beta)^2}$ can be made arbitrarily small by decreasing β. Since the pseudorandom property of Definition 3.4.3 is an asymmetric condition, we will not be able to use the same argument to lower bound $(x,y)^*L(x,y)$ in terms of the columns. The fact that we can tune β will allow us to effectively join these two cases and give an $\Omega(\alpha)$ lower bound on strong convexity.

It remains to handle the other case, when $\frac{\|y\|_2^2}{n} \geq \frac{1}{c}\frac{\|x\|_2^2}{d}$ for appropriate chosen (small) constant c. In this case, we use more elementary arguments based on regularity. We first define the notion of good and bad columns.

Definition 3.4.17. *For fixed x, y, $\ell \in \mathbb{N}$ is a bad column bucket index if*

$$\sum_{j\in C_\ell}\sum_{i\in[d]} w(i,j)(x_i - y_j)^2 < \alpha\gamma^2(1-\gamma)^2 \sum_{j\in C_\ell} \frac{y_j^2}{n}.$$

We define bad columns to be $C_B = \cup_\ell C_\ell$ where the union is over bad column bucket indices, and good columns to be $C_G = [n] - C_B$.

As an immediate consequence, the quadratic form can be lower bounded by

$$(x,y)^*L(x,y) \geq \alpha\gamma^2(1-\gamma)^2 \sum_{j\in C_G} \frac{y_j^2}{n}.$$

The good columns are defined just so that our proof plan goes through. Specifically, we will be able to use regularity to show that bad columns have their weight concentrated in nearby rows. This will then allow a similar charging argument to Lemma 3.4.14 in order to bound the bad rows.

Lemma 3.4.18. *For normalized ε-bi-regular graph H, if $C_\ell \subseteq C_B$ is a bad column bucket according to Definition 3.4.17, then*

$$(1-\varepsilon-\alpha)\sum_{j\in C_\ell}\frac{y_j^2}{n} \le \frac{1+\varepsilon}{\gamma^4}\sum_{i\in R_{\ell\pm1}}\frac{x_i^2}{d}.$$

As a corollary,

$$\sum_{j\in C_B}\frac{y_j^2}{n} \le \frac{3}{\gamma^4}\frac{1+\varepsilon}{1-\varepsilon-\alpha}\frac{\|x\|_2^2}{d}.$$

Proof. The second statement follows from the first by summing over all bad column indices $C_\ell \subseteq C_B$ and noting each R_k appears at most three times on the right hand side (for each $\ell \in \{k-1,k,k+1\}$).

We first show that for $C_\ell \subseteq C_B$, the weight from columns $j \in C_\ell$ are concentrated in

$$w(R_{\ell\pm1}, C_\ell) > \frac{|C_\ell|}{n}(1-\varepsilon-\alpha).$$

By Definition 3.4.17, we can upper bound the contribution of bad column bucket C_ℓ by

$$\sum_{j\in C_\ell}\sum_{i\in[d]} w_{ij}(x_i-y_j)^2 < \alpha\gamma^2(1-\gamma)^2\sum_{j\in C_\ell}\frac{y_j^2}{n} \le \alpha\gamma^2(1-\gamma)^2\frac{|C_\ell|}{n}\cdot\gamma^{2\ell-1},$$

where in the last step we used $\max_{j\in C_\ell} y_j^2 \le \gamma^{2\ell-1}$ by Definition 3.4.9. We can also lower bound the contribution of C_ℓ as

$$\sum_{j\in C_\ell}\sum_{i\in[d]} w_{ij}(x_i-y_j)^2 \ge \sum_{j\in C_\ell}\sum_{i\notin R_{\ell\pm1}} w_{ij}(1-\gamma)^2 y_j^2 \ge (1-\gamma)^2 w(\overline{R_{\ell\pm1}}, C_\ell)\gamma^{2\ell+1},$$

where in the first step we considered just the terms from $\overline{R_{\ell\pm1}}$ and used Fact 3.4.10 to lower bound $(x_i-y_j)^2 \ge (1-\gamma)^2 y_j^2$ for $i \notin R_{\ell\pm1}, j \in C_\ell$, and in the last step we used $\min_{j\in C_\ell} y_j^2 \ge \gamma^{2\ell+1}$ by Definition 3.4.9.

These two inequalities can be rearranged to show

$$w(\overline{R_{\ell\pm1}}, C_\ell) < \alpha\frac{\gamma^2(1-\gamma)^2\gamma^{2\ell-1}}{(1-\gamma)^2\gamma^{2\ell+1}}\frac{|C_\ell|}{n} \le \alpha\frac{|C_\ell|}{n}.$$

Finally, we can use column regularity to show

$$w(R_{\ell\pm1}, C_\ell) = w([d], C_\ell) - w(\overline{R_{\ell\pm1}}, C_\ell) > (1-\varepsilon)\frac{|C_\ell|}{n} - \alpha\frac{|C_\ell|}{n},$$

where in the last step we used that each column satisfies $w([d], j) \geq \frac{1-\varepsilon}{n}$ by ε-bi-regularity.

We can use the above and row-regularity to show that $R_{\ell\pm1}$ cannot be too small as compared to C_ℓ:

$$(1+\varepsilon)\frac{|R_{\ell\pm1}|}{d} \geq w(R_{\ell\pm1}, [n]) \geq w(R_{\ell\pm1}, C_\ell) > (1-\varepsilon-\alpha)\frac{|C_\ell|}{n},$$

where the first step was because each row satisfies $w(i, [n]) \leq \frac{1+\varepsilon}{d}$ by ε-bi-regularity, and the last step was by the lower bound derived above.

The first statement in the lemma now follows as

$$(1-\varepsilon-\alpha)\sum_{j\in C_\ell}\frac{y_j^2}{n} \leq (1-\varepsilon-\alpha)\frac{|C_\ell|}{n}\cdot\gamma^{2\ell-1} \leq (1+\varepsilon)\frac{|R_{\ell\pm1}|}{d}\cdot\gamma^{-4}\gamma^{2(\ell+1)+1} \leq \frac{1+\varepsilon}{\gamma^4}\sum_{i\in R_{\ell\pm1}}\frac{x_i^2}{d},$$

where the first step ($y^2_{j\in C_\ell} \leq \gamma^{2\ell-1}$) and third step ($x^2_{i\in R_{\ell\pm1}} \geq \gamma^{2(\ell+1)+1}$) were both due to Definition 3.4.9, and the second step was by the bound $(1+\varepsilon)\frac{|R_{\ell\pm1}|}{d} > (1-\varepsilon-\alpha)\frac{|C_\ell|}{n}$ derived above. □

The purpose of Lemma 3.4.18 is to show that the good columns outweigh the bad columns. In order for this to be true, we need $\frac{\|x\|_2^2}{d} \leq \frac{4}{\gamma^4}\frac{\|y\|_2^2}{n}$. Note that this constant $\frac{4}{\gamma^4}$ is strictly bounded away from 1 as $\gamma < 1$. This is why we need the other case in Lemma 3.4.14 to be able to handle arbitrarily large constant, so that we can combine with Lemma 3.4.18.

We are now ready to prove strong convexity. We will separate into two cases, depending on whether we can bound the contribution from bad rows or columns respectively.

Proof of Theorem 3.4.7. By Definition 3.4.11 and Definition 3.4.17 of good rows and columns, we can lower bound

$$(x,y)^* L(x,y) \geq \max\left\{\alpha\beta(1-\gamma)^2\sum_{i\in R_G}\frac{x_i^2}{d}, \qquad \alpha\gamma^2(1-\gamma)^2\sum_{j\in C_G}\frac{y_j^2}{n}\right\},$$

where the first term is by Lemma 3.4.12 and the second is by the corollary at the end of Definition 3.4.17. For the remainder of the proof, we separate into two cases, depending on whether we want to lower bound the quadratic form in terms of the first term (good rows) or the second term (good columns).

Case 1: $\frac{\|x\|_2^2}{d} \leq \frac{\gamma^4}{4}\frac{\|y\|_2^2}{n}$. We use the lower bound on columns and show that the bad columns contribute only a constant fraction of the total:

$$\sum_{j \in C_B} \frac{y_j^2}{n} \leq \frac{3(1+\varepsilon)}{\gamma^4(1-\varepsilon-\alpha)}\frac{\|x\|_2^2}{d} \leq \frac{3(1+\varepsilon)}{4(1-\varepsilon-\alpha)}\frac{\|y\|_2^2}{n},$$

where the first step is by Lemma 3.4.18, and the second step was by the case assumption $\frac{\|x\|_2^2}{d} \leq \frac{\gamma^4}{4}\frac{\|y\|_2^2}{n}$. By the assumptions $\varepsilon \leq \frac{1}{16}, \alpha < \frac{1}{16}$, the above constant is at most $\frac{3}{4}\frac{17/16}{14/16} \leq \frac{11}{12}$ and we can finish the lower bound on the first term

$$\begin{aligned}(x,y)^* L(x,y) &\geq \alpha\gamma^2(1-\gamma)^2 \sum_{j \in C_G} \frac{y_j^2}{n} \geq \Big(1 - \frac{11}{12}\Big)\alpha\gamma^2(1-\gamma)^2\frac{\|y\|^2}{n} \\ &\geq \frac{\alpha\gamma^2(1-\gamma)^2}{12}\left(\frac{4}{4+\gamma^4}\right)\left(\frac{\|x\|_2^2}{d} + \frac{\|y\|_2^2}{n}\right),\end{aligned}$$

where again the last step was by the case assumption $\frac{\|x\|_2^2}{d} \leq \frac{\gamma^4}{4}\frac{\|y\|_2^2}{n}$.

Case 2: $\frac{\|y\|_2^2}{n} \leq \frac{4}{\gamma^4}\frac{\|x\|_2^2}{d}$. We use the lower bound on rows and show the bad rows contribute only a constant fraction.

$$\sum_{i \in R_B} \frac{x_i^2}{d} \leq \frac{\beta}{\gamma^6(1-\beta)^2}\frac{\|y\|_2^2}{n} \leq \frac{4\beta}{\gamma^{10}(1-\beta)^2}\frac{\|x\|_2^2}{d},$$

where the first step was by Lemma 3.4.14, and the last step is by the case assumption $\frac{\|y\|_2^2}{n} \leq \frac{4}{\gamma^4}\frac{\|x\|_2^2}{d}$. For $\beta \leq \frac{1}{16}$ we choose $\gamma^{10} = \frac{8\beta}{(1-\beta)^2} < 1$ so that $\frac{4\beta}{\gamma^{10}(1-\beta)^2} \leq \frac{1}{2}$. So we can finish the lower bound on the second term in the max:

$$\begin{aligned}(x,y)^* L(x,y) &\geq \alpha\beta(1-\gamma)^2 \sum_{i \in R_G} \frac{x_i^2}{d} \geq \Big(1 - \frac{1}{2}\Big)\alpha\beta(1-\gamma)^2\frac{\|x\|_2^2}{d} \\ &\geq \frac{\alpha\beta(1-\gamma)^2}{2}\left(\frac{\gamma^4}{4+\gamma^4}\right)\left(\frac{\|x\|_2^2}{d} + \frac{\|y\|_2^2}{n}\right),\end{aligned}$$

where the last step was by the case assumption $\frac{\|y\|_2^2}{n} \leq \frac{4}{\gamma^4}\frac{\|x\|_2^2}{d}$.

Combining both cases gives the lower bound

$$(x,y)^* L(x,y) \geq \min\left\{\frac{\gamma^2(1-\gamma)^2}{15}, \frac{\beta\gamma^4(1-\gamma)^2}{10}\right\} \cdot \alpha\left(\frac{\|x\|_2^2}{d} + \frac{\|y\|_2^2}{n}\right).$$

Substituting $\gamma^{10} = \frac{8\beta}{(1-\beta)^2}$ and choosing $\beta = \frac{1}{16}$ gives leading constant at least e^{-11}. □

Remark 3.4.19. *We never used the assumption $\sum_{i\in[d]} x_i = 0$. This would give a separate argument bounding bad columns in terms of rows, analogous to Lemma 3.4.14. But this does not improve the final constant, and only complicates the case anaysis, so we omit it.*

3.5 Lift to Frame and Operator Scaling

In this short section, we briefly discuss how the matrix scaling problem studied in this chapter can be generalized to the frame setting considered in Chapter 4. This will be covered more formally in Chapter 6, where we will discuss scaling problems from the perspective of Lie group optimization.

In all of the previous sections of this chapter, we assumed that we are explicitly given a tuple of matrices $A = \{A_1, ..., A_K\} \in \text{Mat}(d, n)^K$ with a specified standard basis for $\mathbb{F}^d$ and $\mathbb{F}^n$. But none of the above ideas depended on the choice of standard basis. Therefore, all the results carry over verbatim to an arbitrary choice of orthonormal bases $\Xi = \{\xi_1, ..., \xi_d\} \subseteq \mathbb{F}^d, \Psi = \{\psi_1, ..., \psi_n\} \subseteq \mathbb{F}^n$.

Specifically, we can perform the change of basis operation $M_k = \Xi^* A_k \Psi$ as in Eq. (2.1) to find the matrix representation of A with respect to (Ξ, Ψ). This gives a family of matrix scaling problems on A, one for each choice of bases (Ξ, Ψ), which are just the standard matrix scaling problem in Definition 3.1.3 applied to $M := \Xi^* A \Psi$.

We can ask what operation corresponds to performing matrix scaling on $M := \Xi^* A \Psi$. Consider $L \in \text{diag}(d)$ which gives scaling $LM = L(\Xi^* A \Psi)$. Then, we can invert the change of basis to find A is scaled by

$$\Xi(LM)\Psi^* = (\Xi L \Xi^*)A, \tag{3.5}$$

where we used the fact that Ξ and Ψ are orthonormal bases (i.e. $\Xi \in \text{U}(d)$ or $\Xi \in \text{O}(d)$ depending on the field).

Now consider the set of matrix inputs $\{M_\Xi := \Xi^* A\}$ where Ξ is an arbitrary orthonormal basis, and restrict the scalings to the form $(X, Y) \in \mathfrak{t}$ as discussed in Definition 3.1.5. Then by considering all the induced scalings together, we find

$$\begin{aligned}\cup_\Xi \{e^X M_\Xi e^Y \mid (X, Y) \in \mathfrak{t}\} &\to \cup_\Xi \{(\Xi e^X \Xi^*) A e^Y \mid (X, Y) \in \mathfrak{t}\} \\ &= \{e^X A e^Y \mid X \in \text{H}(d), Y \in \text{diag}(n), \text{Tr}[X] = \text{Tr}[Y] = 0\},\end{aligned}$$

where the first step was by the induced scaling as described in Eq. (3.5), and the final step was by the discussion in Theorem 2.1.13 showing that the Spectral Theorem gives a decomposition of Hermitian matrices $\text{H}(d) = \cup_\Xi \Xi\, \text{diag}(d) \Xi^*$.

It turns out that this set above is exactly the domain of frame scaling. In fact, frame scaling can be viewed as a simultaneous matrix scaling problem on all the inputs $\{M_\Xi := \Xi^* A \mid \Xi \in \mathrm{U}(d)\}$. We will discuss this relation more formally in Section 4.2. Similarly, if we consider all possible right bases as well, then this gives the domain of operator scaling. This problem was first defined by Gurvits [45] in the context of the polynomial identity testing problem in algebraic complexity. Recently, Gurvits, Garg, Oliveira, and Wigderson [38] showed that a simple alternating scaling algorithm for operator scaling converges in polynomial time, implying polynomial time algorithms for a variety of problems in algebraic complexity. This work is in fact what drew our attention to the scaling framework, and parts of their results were key to our work in [62] on the Paulsen problem. In Section 6.3, we will show how the results in this chapter on strongly convex matrix scaling can be lifted to the operator setting. This will be applied in Chapter 9 in order to give near-optimal sample complexity results for the matrix normal model in statistics.

The main take-away for this section is that the results of this chapter can be applied in a basis independent way. In fact, in Chapter 6, we will show that these results hold even in the setting of abstract inner product spaces U, V with tuples of abstract linear operators $A \in L(U, V)^K$ as input. By the discussion above, we see that any choice of orthonormal bases (Ξ, Ψ) induces a concrete matrix scaling problem on the matrix tuple $\Xi^* A \Psi$. We will not need this level of generality for our applications in Chapter 4 or Chapter 9, but we note that scaling problems are well-defined in this abstract setting as well.

Chapter 4

Paulsen Problem Revisited

This chapter is devoted to the Paulsen problem in frame theory.

Question 4.0.1. *Let $U = \{u_1, ..., u_n\} \subseteq \mathbb{C}^d$ be a spanning set of vectors satisfying*

$$\frac{1-\varepsilon}{d} I_d \preceq \sum_{j=1}^{n} u_j u_j^* \preceq \frac{1+\varepsilon}{d} I_d, \qquad \forall j \in [n] : \frac{1-\varepsilon}{n} \leq \|u_j\|_2^2 \leq \frac{1+\varepsilon}{n}. \tag{4.1}$$

What is the minimum distance $\sum_{j=1}^{n} \|v_j - u_j\|_2^2$ over all $V = \{v_1, ..., v_n\}$ satisfying these conditions exactly:

$$\sum_{j=1}^{n} v_j v_j^* = \frac{1}{d} I_d, \qquad \forall j \in [n] : \|v_j\|_2^2 = \frac{1}{n}?$$

This was listed as a major open problem in frame theory ([24], [22]), for which little was known despite considerable effort. Frames satisfying the conditions of Eq. (4.1) with $\varepsilon = 0$ are called unit norm tight frames, and these give optimal constructions for certain applications in signal processing and coding theory [49]. In this chapter, we are able to give optimal distance bounds for Question 4.0.1 in both the average case and the worst case, improving on the polynomial distance bounds of Kwok et al. [62], [63] and the previously best known bound from Hamilton and Moitra [46]. We accomplish this by a similar approach to that of [62], while simultaneously simplifying the procedure and quantitatively improving many parts of the proof. Our key tool is the optimization framework for scaling, discussed in more detail in Chapter 6, which allows us to leverage the work of Chapter 3 on fast convergence for matrix scaling. Our main technical contribution for the Paulsen problem is the smoothed analysis performed in Chapter 5. In this chapter, we mainly combine the matrix scaling analysis of Chapter 3 and the probabilistic analysis of Chapter 5 to prove a distance bound for the worst-case and average-case settings of the Paulsen problem.

Overview: In Section 4.1 we give the relevant background from frame theory and previous approaches to the Paulsen problem. We then discuss the approach of Kwok et al. [62], which combined ideas from frame scaling with smoothed analysis. In Section 4.2, we explicitly connect the work of [62] to the optimization framework for scaling. Then, we present a reduction to the matrix scaling problem. The proof of this reduction relies on the concept of geodesic convexity and is deferred to Chapter 6. In Section 4.3, we leverage the results of Chapter 3 to prove strong distance bounds for matrices satisfying certain fast convergence conditions (strong convexity and pseudorandomness). By the reduction presented in Section 4.2, this implies strong distance bounds for special classes of frames. In Section 4.4, we show that random frames satisfy the sufficient conditions for fast convergence, implying optimal distance bounds for the average case and improving on the work of [63] and [35]. Finally, in Section 4.5, we combine these fast convergence results with a perturbation argument to give our optimal distance bound for the Paulsen problem. This settles the question (up to constant factors) for nearly all parameter regimes (see Remark 4.5.4 for the remaining cases). We defer the smoothed analysis component of the proof to Chapter 5, where we use techniques from random matrix theory [94] to show that the random frames and perturbations discussed above satisfy conditions of fast convergence.

4.1 Introduction

In this section, we will formally define the Paulsen problem. Section 4.1.1 also contains some background from frame theory which motivated the question. Then, in Section 4.1.2 we will discuss the known partial results given by previous approaches to the Paulsen problem. The final Section 4.1.3 contains a description of the dynamical system and smoothed analysis approach of [62]. We follow a similar approach, with some refinements, in order to give our optimal distance bounds.

4.1.1 Frame Theory Background

The Paulsen problem had been listed as a central problem in frame theory [25], and prior to the work of Kwok et al. [62], had been open for over fifteen years despite receiving quite a bit of attention [22], [16], [23]. It concerns frames, which are natural generalizations of orthonormal bases. Finite frames can be thought of as overcomplete bases, and they are used in applications that may require flexibility or robustness in representations of vector data. In this subsection, we will discuss some desirable properties of frames, especially those investigated by Holmes and Paulsen [49] in the context of coding theory. For a much more comprehensive look at the history and applications of finite frames, see the book [24].

Below, we define properties of frames which are relevant to the Paulsen problem.

Definition 4.1.1. *A frame is a spanning set $U := \{u_1, ..., u_n\} \in \mathrm{Mat}(d, n)$ for $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. The size of frame U is defined as*

$$s(U) := \|U\|_F^2 = \sum_{j=1}^{n} \|u_j\|_2^2.$$

Doubly-balanced and Grassmannian frames are two well-studied classes of structured frames that we study in this chapter.

Definition 4.1.2. *Frame $U \in \mathrm{Mat}(d, n)$ is called ε-doubly balanced if it satisfies*

$$(1-\varepsilon)\frac{s(U)}{d} I_d \preceq \sum_j u_j u_j^* \preceq (1+\varepsilon)\frac{s(U)}{d} I_d, \qquad \forall j \in [n] : (1-\varepsilon)\frac{s(U)}{n} \le \|u_j\|_2^2 \le (1+\varepsilon)\frac{s(U)}{n}.$$

In the literature, frames satisfying the first condition are called ε-Parseval, and frames satisfying the second are called ε-equal norm. U is called doubly balanced if $\varepsilon = 0$.

The above condition on frames can be viewed as a basis-independent version of the doubly balanced matrix condition in Definition 3.1.2, and we discuss this connection further in Section 4.2.

Another well-studied property of a frame is its pairwise correlation.

Definition 4.1.3. *The correlation of frame $U \in \mathrm{Mat}(d, n)$ is measured by*

$$\Theta(U) := \max_{j \neq j' \in [n]} |\langle u_j, u_{j'} \rangle|^2.$$

Doubly-balanced frames with minimal $\Theta(U)$ are called Grassmannian frames. If in addition, $|\langle u_j, u_{j'} \rangle|^2$ are equal for all $j \neq j'$, then these are called Equiangular frames.

In [49], various classes of frames were studied for their robustness properties in the context of linear algebraic codes. Specifically, it was shown that doubly-balanced frames are optimally robust with respect to a single erasure, and Grassmannian frames are optimally robust with respect to two erasures. Therefore, an understanding of the optimal behavior of the correlation Θ would be of interest to many communities.

Equiangular frames in particular have an extensive literature and have been studied from a variety of perspectives. Still, it is not known whether equiangular tight frames exist for all choices of d, n (see Chapter 5 of [24]). In the quantum information theory literature, equiangular frames are known as SICPOVM's, and are information-theoretically optimal quantum measurements.

Doubly balanced frames are known to exist for any $d \leq n$, and there are even elementary algorithms that can explicitly construct them (Chapter 2 of [24]). But often, frames are used in applications that require additional properties, e.g. small pairwise angles or sparsity, which are not always satisfied by these generic constructions. In some of these instances, the required frames can be produced via complicated algebraic constructions (see e.g. [5]). These may not be robust to numerical error, and are often expensive to produce algorithmically. It is known that the set of doubly balanced frames contains a manifold of nontrivial dimension [32], and recently Needham and Shonkwiler [74] have shown that this set is even (topologically) connected. Therefore the known algorithmic constructions of doubly balanced frames are far from comprehensive.

On the other hand, there are many simple procedures which can construct ε-doubly balanced frames, as a set of random equal-norm vectors is nearly Parseval with high probability. Tropp et al. [91] proposed alternating projection algorithms to construct doubly balanced frames from these nearly doubly balanced ones. They show positive experimental results and some partial convergence analysis. Holmes and Paulsen [49] studied the optimal parameters for Grassmannian frames which are even harder to construct. They construct some nearly equal norm Parseval frames with small maximal inner product, and ask whether they are close to the optimal parameters for Grassmannian frames. This work led Paulsen to ask a number of people whether a nearly doubly balanced frame is always close to a doubly balanced one, and eventually this became known as the Paulsen problem first formally stated in [16].

Conjecture 4.1.4 (Paulsen Problem)**.** *Let $p(d, n, \varepsilon)$ be the minimum value such that for every ε-doubly balanced frame $U \in \mathrm{Mat}(d, n)$ with $s(U) = 1$, there exists a doubly balanced $V \in \mathrm{Mat}(d, n)$ with $s(V) = 1$ such that*

$$\|V - U\|_F^2 \leq p(d, n, \varepsilon).$$

Then p can be bounded by a polynomial function in d and ε. In particular, this function can be taken to be independent of n.

Proving a good upper bound for the Paulsen problem would give us a firm foundation to work with nearly doubly balanced frames, both in theory and in applications. Indeed, our method can be seen as a continuous version of the alternating projection algorithm of Tropp et al. [91], and our results can be viewed as a rigorous justification of the numerical approach for constructing equal norm Parseval frames. We are also able to give an improved randomized construction of nearly optimal Grassmannian frames in Theorem 4.4.5. We hope that our techniques will be useful to the difficult open question of constructing equiangular frames, as Tropp et al. [91] also proposed an alternating projection algorithm for constructing them.

In [62], we were able to resolve the conjecture affirmatively and show the bound $p(d, n, \varepsilon) \lesssim d^{11/2}\varepsilon$. That work relied on techniques from the scaling framework, and we discuss the approach in more detail at the end of this section.

Even in the case of random inputs, very little was known about the distance bound in Conjecture 4.1.4. In fact, many of the known constructions of nearly doubly balanced frames are random frames. The numerical approach suggested in [49] was to generate random frames and then fix the doubly balanced constraints. As random frames have small Θ with high probability, this approach was also suggested in [49] as a procedure to construct Grassmannian frames. In [63], we were able to use the scaling framework to prove beyond worst-case distance bounds for random inputs (on the order of ε^2 instead of ε), and we were able to use this approach to give simple constructions of near-optimal Grassmannian frames. These results will be discussed further in Section 4.1.3.

In this thesis, we refine the scaling approach of [62] and [63] in order to give optimal distance bounds for the Paulsen problem in both worst-case (Section 4.5) and average case (Section 4.4) settings.

4.1.2 Previous Work

In this subsection we discuss the historical developement of both the upper and lower bounds on the distance function $p(d, n, \varepsilon)$ in Conjecture 4.1.4.

A first compactness argument of Hadwin (see [16]) showed that the $p(d, n, \varepsilon)$ is finite for every $\varepsilon > 0$, but this argument did not give any quantitative result. On the other hand, it is easy to correct the Parseval or equal-norm condition individually, as shown by the following transformations.

Fact 4.1.5. *For any input frame $U \in \mathrm{Mat}(d,n)$ with $s(U) = 1$, the two transformations*

$$u_j^L := \frac{u_j}{\sqrt{n}\|u_j\|_2}, \qquad u_j^R := \left(d \sum_{j=1}^n u_j u_j^*\right)^{-\frac{1}{2}} u_j \tag{4.2}$$

produce equal-norm and Parseval frames, respectively. U^L is the nearest Parseval frame to U, and U^R is the nearest equal-norm frame to U.

If U is ε-doubly balanced for $\varepsilon \leq \frac{1}{3}$, then both U^L and U^R are 3ε-doubly balanced and satisfy the distance bound

$$\|U^L - U\|_F^2 \leq \varepsilon^2 \qquad \text{and} \qquad \|U^R - U\|_F^2 \leq \varepsilon^2.$$

The above results, as well as the simple examples showing this distance bound is tight, are well-known in the literature (see [24], [16], [23]). We reproduce the proof in Fact A.3.1 for completeness.

This suggests the following natural algorithm to find a nearby doubly balanced frame: alternate the transformations in Eq. (4.2) until both conditions are satisfied simultaneously. This alternating procedure is a natural generalization of Sinkhorn's algorithm [83] for matrix scaling, and in Chapter 8 we will formally discuss this algorithm from the perspective of tensor scaling.

If this sequence of transformations $\{U(t)\}_{t\geq 0}$ eventually converges to some doubly balanced $U(T)$, then we can attempt to give a distance bound for the Paulsen problem by bounding each step individually

$$\|U(T) - U\|_F \leq \sum_{t=1}^{T} \|U(t) - U(t-1)\|_F.$$

Unfortunately, this alternate scaling procedure does not always reach a doubly balanced frame. Further, Example 11.1 in [24] shows a frame for which this procedure does not even converge to a fixed point. We believe this is the reason the alternating scaling approach to the Paulsen problem was not pursued further in the literature.

According to [24], the main difficulty of the Paulsen problem was the lack of tools available to control both the Parseval and equal-norm conditions simultaneously. Therefore the work of Bodmann and Casazza [16] and Casazza, Fickus, and Mixon [23] used certain constrained procedures to give partial results for the distance function. Specifically, [23] used a gradient descent algorithm which maintained the equal norm property and iteratively improved the nearly-Parseval condition. Similarly, [16] defined a dynamical system on Parseval frames that improved the nearly-equal norm condition. Both of these works were able to show $p(d,n,\varepsilon) \lesssim \text{poly}(d,n) \cdot \varepsilon^2$ in the case when d, n are relatively prime and ε is small enough.

The best known lower bound $p(d,n,\varepsilon) \gtrsim \varepsilon$ comes from from simple examples presented in [16]. We repeat the example and give a detailed proof in Appendix A.1 for completeness. Note importantly that the dependence on ε is linear, and so the $\text{poly}(d,n) \cdot \varepsilon^2$ results in [16], [23] cannot hold in general for arbitrary d, n.

In fact, both of these approaches can be fruitfully analyzed from the perspective of scaling, and we will elaborate on this in a future work (see Chapter 10). We emphasize that our dynamical system approach in [62] was not the first, though there are a few key differences to [16] and [23] which allow us to rely on tools from convex analysis.

In the following subsection, we discuss the approach of [62], which gave the first proof that the distance function p could be bounded by a function that is independent of n. A key component of this approach will be the framework of operator scaling, which provides tools to design and analyze a procedure that improves both the Parseval and equal-norm condition simultaneously. Another key component is to apply smoothed analysis to the input to derive better distance bounds. We will follow almost the same approach with some refinements in order to prove our optimal results.

Subsequently, this result was improved by Hamilton and Moitra [46] to $p \lesssim d\varepsilon$. We discuss this result in more detail in the following Section 4.1.3. Therefore, prior to the work in this thesis, the best known lower and upper bounds for the Paulsen problem differed by a single $O(d)$ factor.

4.1.3 The Dynamical System Approach

In our first attempt, we tried to control the distance of each step in the alternate scaling algorithm described in Eq. (4.2). As discussed, this does not give meaningful results for a variety of reasons. In [62], our intuition was that the alternate scaling algorithm may be taking large steps but moving the frame very little. So we defined a dynamical system on frames that can be viewed as an infinitesimal and continuous version of the discrete alternate scaling algorithm.

Definition 4.1.6. *For frame $U \in \text{Mat}(d,n)$, the dynamical system $U_t = \{u_1(t), ..., u_n(t)\}$*

has initial condition $U_0 = U$ and is the solution to differential equation

$$\partial_t u_j(t) = \Big(d\sum_{j=1}^{n} u_j(t)u_j(t)^* - s(U_t)I_d\Big)u_j(t) + u_j\Big(n\|u_j(t)\|_2^2 - s(U_t)\Big). \tag{4.3}$$

Note that if U is doubly balanced, then it is a fixed point of Definition 4.1.6. Further, the two terms in the differential equation correspond to the error in the Parseval and equal-norm conditions respectively, and serve to pull the frame to satisfy these conditions. To bound the distance of input U to a doubly balanced frame, our plan is to bound the path length of Definition 4.1.6.

Unfortunately, this dynamical system does not always converge to a doubly balanced frame. In [62], our solution was to use smoothed analysis: by adding noise to the input $V := U + E$, we were able to show that the dynamical system converges quickly on this perturbed input. This led to the following result.

Theorem 4.1.7 (Theorem 1.3.1 in [62])**.** *For any ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$ of size $s(U) = 1$, there is a doubly balanced $V \in \mathrm{Mat}(d,n)$ of size $s(V) = 1$ such that*

$$\|V - U\|_F^2 \lesssim d^{11/2}\varepsilon.$$

In particular, the function in Conjecture 4.1.4 can be taken to be independent of n.

The proof of [62] involved a combination of tools from the operator scaling framework of [38] as well as an involved probabilistic analysis of the perturbation argument.

Subsequently, this was improved to $p \lesssim d\varepsilon$ by the work of Hamilton and Moitra [46]. Their proof is by frame scaling, under the name of radial isotropic position [47], [10], along with a cleverly defined distance function related to the Wasserstein metric. It should be noted that this result is requires no assumptions on d, n, ε, whereas both the argument of [62] and the refinement in this thesis have various corner cases that are solved with different arguments. Also the proof is dramatically shorter and simpler due to the slick distance analysis that argues directly about the frame scaling solution.

In this thesis, we also use the two-stage dynamical system and smoothed analysis approach of [62]. Key to our results will be a refined use of the scaling framework shown in Section 4.2. These improvements imply optimal distance bounds for Conjecture 4.1.4 in Section 4.5, as well as an optimal analysis of the average case in Section 4.4.

4.2 Improved Scaling Approach

The key result in this section is a reduction from frame scaling to the simpler matrix scaling setting. In Section 4.2.1, we formally define the frame scaling problem and discuss some applications in theoretical computer science. Then in Section 4.2.2, we show the connection between the dynamical system approach in Definition 4.1.6 and frame scaling. This was a key contribution of [62] which allowed us to use tools from the scaling framework to prove a distance bound for the Paulsen problem. At the end of this subsection, we will briefly discuss why the approach of [62], as well as the spectral analysis in [63], is not sufficient to prove optimal distance bounds for the Paulsen problem. Then in Section 4.2.3, we present a reduction from frame scaling to matrix scaling. This is the new key ingredient which allows us to use the fast convergence properties of matrix scaling studied in Chapter 3 to prove optimal distance bounds for the Paulsen problem.

4.2.1 The Frame Scaling Problem

We will show in the next subsection that both the alternating algorithm of Eq. (4.2) and the dynamical system in Definition 4.1.6 output frames of the form LUR for some $L \in \text{Mat}(d), R \in \text{diag}(n)$ for input $U \in \text{Mat}(d, n)$. As discussed in Section 4.1, both of these algorithms do not always converge to a doubly balanced frame, and therefore they cannot immediately be used to solve the Paulsen problem. To understand convergence of these algorithms, we are led naturally to the following problem.

Definition 4.2.1 (Frame Scaling Problem)**.** *Given frame $U \in \text{Mat}(d, n)$, find scalings $L \in \text{Mat}(d), R \in \text{diag}(n)$ such that $V := LUR$ is a doubly balanced frame according to Definition 4.1.2.*

Remark 4.2.2. *The above definitions hold in the more abstract setting where $u_1, ..., u_n$ are elements of a d-dimensional inner product space $\mathcal{U}$ over either $\mathbb{R}$ or $\mathbb{C}$. In this chapter, we assume without loss that $\mathcal{U} \simeq \mathbb{F}^d$ where $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and the frame $U \in \text{Mat}(d, n)$ is given according to the standard basis. The generalization to abstract vector spaces is discussed further in Chapter 6.*

Note the similarity to the matrix scaling problem in Definition 3.1.3. In fact, the frame scaling problem can be viewed as performing matrix scaling simultaneously in all bases. We formalize this by defining the appropriate generalizations of row and column sums.

Definition 4.2.3. *Given frame $U = \{u_1, ..., u_n\} \in \mathrm{Mat}(d, n)$, the row and column sums are given with respect to $\xi \in S^{d-1}$ and $j \in [n]$, respectively, and are defined as*

$$r_\xi(U) := \langle \xi\xi^*, UU^* \rangle = \sum_{j=1}^{n} |\langle \xi, u_j \rangle|^2, \qquad \textit{and} \qquad c_j(U) := \langle E_{jj}, U^*U \rangle = \|u_j\|_2^2.$$

The left and right error of a frame are $\nabla_U := (\nabla_U^L \in \mathrm{Mat}(d), \nabla_U^R \in \mathrm{diag}(n))$, and are defined by

$$\nabla_U^L := d \cdot UU^* - s(U)I_d, \qquad \nabla_U^R = \mathrm{diag}\{n \cdot c_j(U) - s(U)\}_{j=1}^n.$$

We show that the above definitions generalize matrix row and column sums according to Definition 3.1.1, as well as the gradient in Proposition 3.1.12. The column sums $c_j(U)$ are defined in exactly the same way as in Definition 3.1.1 for the matrix setting, and therefore the right error ∇_U^R is exactly the same as the right part of the matrix gradient defined in Proposition 3.1.12. For the row sums, consider $e_i \in \mathbb{F}^d$ an element of the standard basis. Then the frame row sum $r_{e_i}(U)$ is exactly the same as the i-th row sum of matrix U as given in Definition 3.1.1. Similarly, for arbitrary $\xi \in S^{d-1}$, let Ξ be an orthonormal basis of $\mathbb{F}^d$ with $\xi_1 = \xi$; then the frame row sum $r_\xi(U)$ is exactly the 1-st row sum $r_1(M)$ of the matrix representation $M := \Xi^* U$, where we performed a change of basis. The fact below formally connects the doubly balanced condition for frames and matrices.

Lemma 4.2.4. *Frame $U \in \mathrm{Mat}(d, n)$ is ε-doubly balanced iff the matrix representation $(\Xi^* U)_{ij} := \langle \xi_i, u_j \rangle$ is an ε-doubly balanced matrix according to Definition 3.1.2 for every orthonormal basis $\Xi \in \mathrm{Mat}(d)$.*

Proof. As both definitions are homogenous, we can assume without loss that $s(U) = 1$. Definition 4.1.2 of the ε-doubly balanced frame condition can be written equivalently as

$$\|dUU^* - I_d\|_{\mathrm{op}} \leq \varepsilon, \qquad \text{and} \qquad \max_{j \in [n]} |n \cdot c_j(U) - 1| \leq \varepsilon,$$

where we used the condition $s(U) = 1$. The condition on columns is clearly the same as the matrix version. We can show the row condition by the following sequence of equalities:

$$\|dUU^* - I_d\|_{\mathrm{op}} = \sup_{\xi \in S^{d-1}} |\langle \xi\xi^*, dUU^* - I_d \rangle| = \sup_{\Xi} \sup_{i \in [d]} |d\|\xi_* U\|_2^2 - 1| = \sup_{\Xi} \sup_{i \in [d]} |d \cdot r_i(\Xi^* U) - 1|,$$

where the first step was by the dual description of $\|\cdot\|_{\mathrm{op}}$ given in Eq. (2.5) for Hermitian input, in the second step we used the fact that $\cup_\Xi \{\xi_1, ..., \xi_d\} = S^{d-1}$ where the union is over all orthonormal bases, and the final step was by Definition 3.1.1 of the row sum. Note that by the discussion above

$$r_i(\Xi^* U) = \sum_{j=1}^{n} |\langle \xi_i, u_j \rangle|^2 = r_{\xi_i}(U),$$

so the right hand side characterizes the balance condition for each matrix $\Xi^* U$. □

This lemma suggests that the matrix scaling results in Chapter 3 can be applied to analyze frame scaling problem if we consider all bases simultaneously. This is in fact the approach we will take, and in Definition 4.2.11 we define the frame versions of strong convexity and pseudorandomness in order to apply the analyses from Chapter 3.

We now place the frame scaling problem in a wider context by discussing some past work and related problems. The frame scaling problem has appeared previously in many fields in computer science, including radial isotropic positions in machine learning [47], [50], and geometric conditions in Brascamp-Lieb inequalities [39], [10]. An early application was discovered by Forster [34], who solved a special case of the frame scaling problem in order to derive a lower bound on the sign rank of the Hadamard matrix. This was applied to a breakthrough result in communication complexity. We note that Forster's scaling result was proved earlier in a more general setting by Gurvits and Samorodnitsky [44] in their work on mixed discriminants, and is also implicit in the work of Barthe [10] on Brascamp-Lieb inequalities. Two recent applications of frame scaling in machine learning are found in the work of Hardt and Moitra [47] in robust subspace discovery, as well as Hopkins et al. [50] in point location. The notion of radial isotropic position was key to the work of Hamilton and Moitra [46] which gave the previously best known distance bound for the Paulsen problem. We discuss their approach in more detail in Remark 4.2.6.

Operator scaling is a generalization of frame scaling that was introduced by Gurvits [45] in an attempt to design a deterministic polynomial time algorithm for the important polynomial identity testing problem in algebraic complexity. Continuing this approach, Garg, Gurvits, Oliveira, and Wigderson [38] improved Gurvits' analysis to prove that the alternating algorithm for operator scaling can be used to compute the non-commutative rank of a symbolic matrix in polynomial time. This line of work was what drew our attention to the scaling framework in [62]. We discuss this problem in the context of the general scaling framework in Chapter 7 and Chapter 8.

Many of the scaling and distance results established in this chapter apply to the operator scaling setting as well. But currently, our strongest smoothed analysis results are restricted to the frame setting. We believe that they can be extended to more general problems. For preliminary results for distance bounds in the operator setting, see Theorem 3.7.3 in [62].

4.2.2 Previous Approach by Frame Scaling

A key component of [62] was to connect the Paulsen problem to the scaling framework of [38]. In this subsection, we will discuss how frame scaling arises naturally from the dynamical system in Definition 4.1.6. At the end, we point out some key differences between frame and matrix gradient flow and especially the difference between the convergence analysis of [62] and the results in this thesis.

Since the alternate scaling algorithm always performs steps of the form $U \leftarrow LU$ or $U \leftarrow UR$ (see Eq. (4.2)), it is clear that for $U \in \mathrm{Mat}(d,n)$, the output will always be of the form LUR for some scaling matrices $L \in \mathrm{Mat}(d), R \in \mathrm{diag}(n)$.

It turns out that this is also the case for the dynamical system in Definition 4.1.6. To prove this, we rewrite the dynamical system of [62] in terms of scalings.

Proposition 4.2.5. *For frame $U \in \mathrm{Mat}(d,n)$, the dynamical system in Definition 4.1.6 can be equivalently written as a dynamical system on scalings: $U_t := L_t U R_t$ for $L_t \in \mathrm{SL}(d), R_t \in \mathrm{SL}(n) \cap \mathrm{diag}(n)$ satisfying the differential equation*

$$(L_0, R_0) = (I_d, I_n), \qquad \partial_t L_t = -\nabla^L_{U_t} \cdot L_t, \qquad \partial_t R_t = -R_t \cdot \nabla^R_{U_t},$$

where the gradient $\nabla^{L/R}_U$ is given in Definition 4.2.3.

Proof. Using the notation in Definition 4.2.3, the differential equation in Definition 4.1.6 can be rewritten as

$$U_0 = U, \qquad \partial_t U_t = -(\nabla^L_{U_t} \cdot U_t + U_t \cdot \nabla^R_{U_t}).$$

We verify that this is equivalent to the differential equation on scalings given in this proposition. Clearly the initial conditions are the same as $U_0 = U = L_0 U R_0$. To show that the solutions are the same, it is enough to show that for all frames, the differential equations are equivalent at time $t = 0$.

$$\partial_{t=0}(L_t U R_t) = (\partial_{t=0} L_t) U I_n + I_d U (\partial_{t=0} R_t) = -(\nabla^L_U \cdot U + U \cdot \nabla^R_U),$$

where we used the initial conditions $L_0 = I_d, R_0 = I_n$. Comparing this to the previous equation at $t = 0$, we see that the two differential equation are equivalent for every input frame, and therefore they induces the same dynamical system.

Now we prove the determinant condition, i.e. $L_t \in \mathrm{SL}(d), R_t \in \mathrm{SL}(n)$. Note that $\det(L_0) = \det(I_d) = 1$ by the initial conditions. We show that it is invariant:

$$\partial_t \log\det(L_t) = \mathrm{Tr}[L_t^{-1} \partial_t L_t] = -\mathrm{Tr}[\nabla^L_{U_t}] = s(U_t)\,\mathrm{Tr}[I_d] - d \cdot \mathrm{Tr}[U_t U_t^*] = 0,$$

where the first step is by standard matrix calculus (see e.g. [90]), and the last step was by Definition 4.1.1 of size $s(V) := \|V\|_F^2$. By a similar calculation, $\det(R_0) = \det(I_n) = 1$ and $\partial_t \log\det(R_t) = -\operatorname{Tr}[\nabla_{U_t}^R] = 0$. Therefore, for all time $\det(L_t) = \det(R_t) = 1$. □

Recall that in Definition 3.1.14, we gave a similar dynamical system for matrix scaling as the gradient flow of a particular convex function. This convexity in fact comes from a general phenomenon that captures the frame dynamical system as well (see Section 6.1.3). In Definition 7.1.5, we discuss the geodesic convex formulation for tensor scaling which will allow us to formally define the geodesic gradient flow for tensor scaling, of which this frame scaling dynamical system is a special case. We will also heavily use this connection to gradient flows in order to give our distance analysis in Section 4.3.1.

To close out this subsection, we elaborate on some technical details of [62] and [63] in order to motivate the reduction to matrix scaling in Section 4.2.3. Note that a crucial part of the analysis in Chapter 3 relied on the multiplicative robustness properties of strong convexity (Lemma 3.2.4) and pseudorandomness (Lemma 3.3.4). It turns out that this multiplicative robustness is not true for strongly convex frames (as shown in Appendix A.4). The work of [63] was able to give a sufficient condition for fast convergence in the more general operator scaling case, but with a requirement that was quadratically worse (compare $\alpha^2 \gtrsim \varepsilon \log d$ in Theorem 1.5 of [63] to $\alpha \gtrsim \varepsilon \log d$ in Theorem 3.2.19). This means that in order to apply this fast convergence analysis to an ε-doubly balanced frame, we would need to perturb the frame by distance $\Omega(\sqrt{\varepsilon})$, which is the wrong order for the Paulsen conjecture in Conjecture 4.1.4. Similarly, in [62], the analysis proceeded by a slightly different pseudorandom condition on matrices. In that case, we were able to prove multiplicative robustness, but only for the matrix dynamical system in Definition 3.1.14. This meant that in order to guarantee that fast convergence was maintained throughout the frame dynamical system (Definition 4.1.6), we had to follow a much more convoluted path by repeatedly perturbing the frame. In the following subsection, we describe a reduction to matrix scaling which allows us to bypass both of these issues. This will allow us to use a simpler perturbation argument in order to guarantee fast convergence for the perturbed frame.

Remark 4.2.6. *In [46], Hamilton and Moitra avoided all of these convergence issues of the dynamical system by a far simpler argument. They were able to use the simple fact that the set of non-scalable frames, those inputs for which there is no non-zero doubly balanced frame scaling, is of measure* 0*. Therefore, even an infinitesimal perturbation suffices for their result. The core of their proof is then a very clever way to bound the distance to the doubly balanced scaling by the initial error of the frame. The smoothed analysis strategy of [62] can be thought of as a robust version of this argument, as there the input frame* *is perturbed by some finite amount, and this guarantees fast convergence to the scaling solution and a strong bound on the path length of the dynamical system.*

4.2.3 Reduction to Matrix Scaling

In this subsection, we will present a reduction from the distance analysis for the Paulsen problem to the much simpler matrix scaling setting. We defer its proof to Chapter 6, where we will formally define and use the notion of geodesic convexity. This sets up the following Section 4.3, in which we will be able to use fast convergence properties of matrix scaling derived in Chapter 3 in order to prove strong distance bounds for these matrix inputs. This will immediately imply the same distance bounds for the frame setting by the reduction.

We first show some invariance properties of frames. This will allow us to simplify the domain of frame scalings, paralleling the development in Section 3.1.2. First note that Definition 4.1.2 is homogenous so we can normalize scalings to have unit determinant without loss. Below, we show a much larger set of transformations which does not affect the doubly balanced conditions for frames (Definition 4.1.2).

Fact 4.2.7. *The ε-doubly balanced condition for frames given in Definition 4.1.2 is invariant under the group $\mathrm{U}(d) \times (S^1)^n$ if $\mathbb{F} = \mathbb{C}$ and $\mathrm{O}(d) \times \{\pm 1\}^n$ if $\mathbb{F} = \mathbb{R}$. Here $S^1 = \{\lambda \in \mathbb{C} \mid |\lambda| = 1\}$ is the unit circle in the complex plane.*

Proof. Explicitly, we will show that $U \in \mathrm{Mat}(d,n)$ is an ε-doubly balanced frame iff $V := \Xi U e^Y$ is ε-doubly balanced for any $(\Xi, e^Y) \in \mathrm{U}(d) \times (S^1)^n$ or $(\Xi, e^Y) \in \mathrm{O}(d) \times \{\pm 1\}^n$ depending on the field. We will focus on the case $\mathbb{F} = \mathbb{C}$. The simpler $\mathbb{F} = \mathbb{R}$ case follows by an analogous calculation.

First note that the size does not change as

$$s(V) = \|V\|_F^2 = \|\Xi U e^Y\|_F^2 = \|U\|_F^2 = s(U),$$

where we used invariance of $\|\cdot\|_F$ under unitaries $\Xi \in \mathrm{U}(d), e^Y \in \mathrm{U}(n)$. We can similarly bound the left error

$$\|dVV^* - s(V)I_d\|_{\mathrm{op}} = \|d\Xi U e^Y e^{-Y} U^* \Xi^* - s(U)I_d\|_{\mathrm{op}} = \|dUU^* - s(U)I_d\|_{\mathrm{op}},$$

where in the first step we used $s(V) = s(U)$ as calculated above and $\overline{e^Y} = e^{-Y}$ as $e^{Y_{jj}} \in S^1$, and in the last step we used unitary invariance of $\|\cdot\|_{\mathrm{op}}$ as $\Xi \in \mathrm{U}(d)$. For the right error, we similarly calculate

$$|n \cdot c_j(V) - s(V)| = |n\|\Xi u_j e^{Y_{jj}}\|_2^2 - s(U)| = |n\|u_j\|_2^2 - s(U)| = |n \cdot c_j(U) - s(U)|,$$

where in the first and the last steps we used Definition 4.2.3 of the column sum, and in the second step we used unitary invariance of $\|\cdot\|_2$ to remove $\Xi \in \mathrm{U}(d)$ and $e^{Y_{jj}} \in S^1$.

The above calculations verify Definition 4.1.2 showing V is ε-doubly balanced. Since $(\Xi, e^Y) \in G_i$ were arbitrary, the statement is shown. □

Remark 4.2.8. *This unitary invariance is a key component of the Kempf-Ness theory described in Section 6.1.3 for the general scaling framework. In fact, one of the main reasons that many of these scaling problems are tractable is due to the underlying convexity which is revealed after we reduce from general scalings to positive definite ones using this unitary invariance.*

With this fact in hand, we can restrict the domain of the frame scaling problem to the following subset. This is another instance of the polar decomposition in Theorem 2.1.13, which will be used extensively in Chapter 6 in order to give an optimization problem for all tensor scaling problems.

Definition 4.2.9. *For the purpose of the frame scaling problem on input $U \in \mathrm{Mat}_{\mathbb{C}}(d,n)$ as given in Definition 4.2.1, we can restrict to scalings of the form $e^X U e^Y$ where (X, Y) are elements of the following vector space:*

$$\mathfrak{p} := \{(X \in \mathrm{H}(d), Y \in \mathrm{diag}_{\mathbb{R}}(n)) \mid \mathrm{Tr}[X] = \mathrm{Tr}[Y] = 0\}.$$

If $U \in \mathrm{Mat}_{\mathbb{R}}(d,n)$ is a real frame, then we replace the Hermitian matrices $\mathrm{H}(d)$ by the symmetric matrices $\mathrm{S}(d)$ (as defined in Section 2.1.3).

Recalling the discussion in Section 3.5, this is exactly the set of scalings induced by the family of matrix scaling problem applied to the various representations $\Xi^* U$ where Ξ runs over all orthonormal bases. Explicitly, by the discussion in Theorem 2.1.13, we can rewrite

$$\mathfrak{p} = \cup_{\Xi} \{(\Xi X \Xi^*, Y) \mid (X, Y) \in \mathfrak{t}\},$$

where the index of the union runs over all orthonormal bases. This decomposition will be an important component in our proof of the reduction from frame to matrix scaling.

We can also lift the norms $\|\cdot\|_{\mathfrak{t}}$ and $\|\cdot\|_{\infty}$ to the frame scaling setting.

Definition 4.2.10. *For vector space $\mathfrak{p}$ given in Definition 4.2.9 and element $(X, Y) \in \mathfrak{p}$, the $\mathfrak{p}$-norm and operator norm are defined as*

$$\|(X,Y)\|_{\mathfrak{p}}^2 := \frac{\|X\|_F^2}{d} + \frac{\|Y\|_F^2}{n}, \qquad \|(X,Y)\|_{\mathrm{op}} := \|X\|_{\mathrm{op}} + \|Y\|_{\mathrm{op}}.$$

We emphasize the difference between $\|\cdot\|_{\mathfrak{p}}$ on the vector space $\mathfrak{p}$ and the L_p-norm $\|\cdot\|_p$ given in Definition 2.1.14.

Note that $Y \in \operatorname{diag}(n)$, so $\|Y\|_{\text{op}} = \max_{j\in[n]} |Y_{jj}|$, matching the matrix case. Further, note that $\|\cdot\|_{\mathfrak{p}}$ reduces to $\|\cdot\|_{\mathfrak{t}}$ and $\|\cdot\|_{\text{op}}$ reduces to $\|\cdot\|_\infty$ when the domain is restricted to $\mathfrak{t} \subseteq \mathfrak{p}$. Since Y is always diagonal, we will use $\|Y\|_{\text{op}} = \|Y\|_\infty$ interchangeably.

As shown in Lemma 4.2.4, the doubly balanced frame condition can be seen as a basis-independent version of the doubly balanced matrix condition in Definition 3.1.2. In the next definition, we give frame versions of strong convexity and pseudorandomness. These will be used as sufficient conditions for fast convergence in an analogous way to the analyses in Chapter 3.

Definition 4.2.11 (Frame Strong Convexity and Pseudorandomness)**.** *For frame $U = \{u_1, ..., u_n\} \in \operatorname{Mat}_{\mathbb{F}}(d,n)$, U is an α-strongly convex frame iff the matrix representation $\Xi^* U$ is α-strongly convex according to Definition 3.2.1 for every choice of orthonormal basis $\Xi \in \mathrm{U}(d)$ if $\mathbb{F} = \mathbb{C}$ and $\Xi \in \mathrm{O}(d)$ if $\mathbb{F} = \mathbb{R}$.*

Similarly, U is an (α,β)-pseudorandom frame iff $M := \Xi^* U$ is an (α,β)-pseudorandom matrix according to Definition 3.3.1 for every orthonormal basis Ξ.

This definition already allows us to simply lift the result of Theorem 3.4.7 to frames.

Corollary 4.2.12. *Consider frame $V \in \operatorname{Mat}(d,n)$ of size $s(V) = 1$ that is $\varepsilon \leq \frac{1}{16}$-doubly balanced according to Definition 4.1.2. If V is an (α,β)-pseudorandom frame for $\alpha \leq \frac{1}{16}$ and $\beta \leq \frac{1}{16}$, then V is $e^{-11}\alpha$-strongly convex according to Definition 4.2.11.*

Proof. According to Definition 4.2.11, V is an (α,β)-pseudorandom frame iff the matrix representation $M_\Xi := \Xi^* V$ is an (α,β)-pseudorandom matrix according to Definition 3.3.1 for every orthonormal basis $\Xi \subseteq \mathbb{F}^d$. Therefore, since $\max\{\alpha,\beta,\varepsilon\} \leq \frac{1}{16}$, we can apply Theorem 3.4.7 to show that each M_Ξ is $e^{-11}\alpha$-strongly convex as a matrix according to Definition 3.2.1. Since this applies to every matrix representation M_Ξ, this is equivalent to $e^{-11}\alpha$-strong convexity of the frame V according to Definition 4.2.11. □

We have shown that frame scaling can be seen as a simultaneous version of matrix scaling for all basis representations. There are simple examples, like the all-ones matrix $J \in \operatorname{Mat}(d,n)$, which show that the doubly balanced frame condition is much more restrictive than the doubly balanced matrix condition. But intuitively, if we could find the basis in which the solution to the frame scaling problem lies, then we could just analyze matrix scaling in this basis. This is formalized in our reduction below.

Theorem 4.2.13. *Consider frame $U = \{u_1, ..., u_n\} \in \mathrm{Mat}(d,n)$. Assume that for every choice of orthonormal basis Ξ, the matrix scaling problem in Definition 3.1.3 on input $M_\Xi := \Xi^* U$ has a doubly balanced solution $e^{X_\Xi/2} M_\Xi e^{Y_\Xi/2}$ with $(X_\Xi, Y_\Xi) \in \mathfrak{t}$ and $\|(X_\Xi, Y_\Xi)\|_{\mathfrak{t}} \leq R$ for some $R < \infty$. Then there exists a choice of orthonormal basis Ξ such that the frame scaling*

$$U_* := (\Xi e^{X_\Xi/2} \Xi^*) U e^{Y_\Xi/2}$$

is a doubly balanced frame.

We defer the proof of this theorem to Chapter 6 after we have defined the notion of geodesic convexity. As a consequence, we get the following strong convergence analysis of frame scaling.

Theorem 4.2.14. *If frame $U \in \mathrm{Mat}(d,n)$ of size $s(U) = 1$ is ε-doubly balanced according to Definition 4.1.2 and (α, β)-pseudorandom according to Definition 4.2.11 for $\frac{1}{5} \geq \alpha \geq 16e \cdot \varepsilon$ and $\beta \leq \frac{1}{2}$, then there is a scaling $U_* = e^{X_*/2} U e^{Y_*/2}$ with $(X_*, Y_*) \in \mathfrak{p}$ satisfying:*

1. *$U_* := e^{X_*/2} U e^{Y_*/2}$ is a doubly balanced frame;*
2. *$\max\{\|X_*\|_{\mathrm{op}}, \|Y_*\|_{\mathrm{op}}\} \leq \frac{9\varepsilon}{\alpha}$;*
3. *The size of the scaling solution is lower bounded by $s(U_*) \geq 1 - \frac{10\varepsilon^2}{\alpha}$.*
4. *The distance to the scaling solution is bounded by $\|U_* - U\|_F^2 \leq \frac{8\varepsilon^2}{\alpha}$.*

Proof. For every orthonormal basis Ξ, the matrix representation $M^\Xi := \Xi^* U$ is an ε-doubly balanced matrix by Lemma 4.2.4, and is an (α, β)-pseudorandom matrix by Definition 4.2.11. Therefore, we can apply Theorem 3.3.10 to each such matrix M^Ξ to find scalings $(X_\Xi, Y_\Xi) \in \mathfrak{t}$ such that $e^{X_\Xi/2} M^\Xi e^{Y_\Xi/2}$ is a doubly balanced matrix.

Since this holds for every basis, Theorem 4.2.13 implies that there is some choice of orthonormal basis Ξ such that the induced frame scaling $(X_*, Y_*) = (\Xi X_\Xi \Xi^*, Y_\Xi)$ produces doubly balanced frame

$$U_* := e^{X_*/2} U e^{Y_*/2}.$$

The conclusions of (2) and (3) now follow exactly from the definitions $M^\Xi := \Xi^* U$ and $(X_*, Y_*) = (\Xi X_\Xi \Xi^*, Y_\Xi)$ by the corresponding conclusions of Theorem 3.3.10, as

$$s(U_*) = \|e^{X_*/2} U e^{Y_*/2}\|_F^2 = \|\Xi e^{X_\Xi/2} \Xi^* U e^{Y_\Xi/2}\|_F^2 = \|e^{X_\Xi/2} M_\Xi e^{Y_\Xi/2}\|_F^2,$$

$$\|X_*\|_{\mathrm{op}} = \|\Xi X_\Xi \Xi^*\|_{\mathrm{op}} = \|X_\Xi\|_\infty, \qquad \|Y_*\|_{\mathrm{op}} = \|Y_\Xi\|_\infty,$$

where we used unitary invariance of $\|\cdot\|_F, \|\cdot\|_{\text{op}}$. The distance bound in item (4) follows similarly from the distance bound for the matrix Paulsen problem given in Proposition 4.3.1 in the following section, as

$$\|U_* - U\|_F = \|(\Xi e^{X_\Xi/2}\Xi^*)Ue^{Y_\Xi/2} - U\|_F = \|e^{X_\Xi/2}(\Xi^* U)e^{Y_\Xi/2} - \Xi^* U\|_F = \|e^{X_\Xi/2}M^\Xi e^{Y_\Xi/2} - M^\Xi\|_F,$$

where we repeatedly used invariance of $\|\cdot\|_F$ under orthonormal Ξ. □

The same proof strategy can be used to give frame and operator versions of the strong convexity analysis in Theorem 3.2.19. We give this argument in Section 7.3.3 after we have defined geodesic convexity, as the remainder of this chapter only uses the pseudorandom distance analysis for frames.

The distance bound in Theorem 4.2.14 will be applied to give optimal bounds for the Paulsen problem Conjecture 4.1.4. Specifically, we will show optimal average case bounds in Section 4.4, and we will use a perturbation argument to show optimal worst case bounds in Section 4.5. In Section 8.5, we are able to use the scaling bound in Theorem 4.2.14(2) to give a tight sample complexity and algorithmic convergence guarantee for Tyler's M-estimator in statistics, improving the results in [35].

4.3 Distance Analysis for Matrix Scaling

The goal of this section is to prove the following strong distance bounds on matrix instances satisfying the pseudorandom condition as in Theorem 3.3.10.

Proposition 4.3.1. *Let $A \in \text{Mat}(d,n)$ be a matrix of size $s(A) = 1$ that is ε-doubly balanced and (α,β)-pseudorandom for $\frac{1}{5} \geq \alpha \geq 16e \cdot \varepsilon$ and $\beta \leq \frac{1}{2}$, and consider the scaling $A_\infty := \lim_{t\to\infty} A_t$ where A_t is defined according to gradient flow in Definition 3.1.14. Then this limit exists and A_∞ satisfies the distance bound*

$$\|A_\infty - A\|_F^2 \leq \frac{64\varepsilon^2}{\alpha}.$$

We also give a distance bound for strongly convex inputs in Proposition 4.3.6. This will be applied, in Chapter 8, in order to make the results in this chapter algorithmic. The remainder of this chapter only uses the pseudorandom distance analysis as this gives better (dimension independent) bounds for the Paulsen problem.

As discussed in the previous Section 4.2.3, the above two propositions imply the same distance bounds for frames satisfying strong convexity or pseudorandom conditions via the reduction given in Theorem 4.2.13. We will use the pseudorandom analysis to give optimal average case bounds for the Paulsen problem in Section 4.4 and optimal worst case bounds in Section 4.5.

In Section 4.3.1, we present the Kempf-Ness equivalence [58] from algebraic geometry, specialized to the matrix scaling setting. This allows us to bound the path length of the gradient flow given in Definition 3.1.14 using a natural potential function analysis. Then in Section 4.3.2 and Section 4.3.3, we use the strong convergence properties derived in Section 3.2 and Section 3.3, respectively, in order to give distance bounds for the scaling solutions of strongly convex and pseudorandom inputs.

4.3.1 Kempf-Ness Equivalence

The dynamical system on matrices that is induced by Definition 3.1.14 can be written as

$$\partial_t A_t = -(\nabla^L_{A_t} A_t + A_t \nabla^R_{A_t}),$$

where the left and right errors ∇ given in Proposition 3.1.12 act by diagonal scaling $(\nabla^L_A A \nabla^R_A)_{ij} = (\nabla^L_A)_{ii} A_{ij} (\nabla^R_A)_{jj}$. Our approach will be to bound the magnitude of this infinitesimal movement in terms of the following potential function.

Definition 4.3.2. *For $A \in \mathrm{Mat}(d,n)$, we measure the error to doubly balanced by*

$$\Delta(A) := \|\nabla_A\|_{\mathfrak{t}}^2 = \frac{1}{d}\sum_{i=1}^{d} (d \cdot r_i(A) - s(A))^2 + \frac{1}{n}\sum_{j=1}^{n} (n \cdot c_j(A) - s(A))^2 .$$

We give this quantity a new name to emphasize that it is a function on $\mathrm{Mat}(d,n)$, whereas the Kempf-Ness function f_A in Definition 3.1.6 is a function on $\mathfrak{t}$. In the rest of this subsection, we are essentially following the distance analysis of [62]. We reproduce these results as we believe the Kempf-Ness theory and scaling perspective give a principled approach to the various useful equalities proved in Section 3.4 of [62].

By Definition 3.1.11 of $\|\cdot\|_{\mathfrak{t}}$, we can simply bound this error measure for nearly doubly balanced matrices.

Fact 4.3.3. *For $A \in \mathrm{Mat}(d,n)$ that is ε-doubly balanced according to Definition 3.1.2, the error can bounded by*

$$\Delta(A) = \|\nabla_A\|_{\mathfrak{t}}^2 \leq \|\nabla^L_A\|_\infty^2 + \|\nabla^R_A\|_\infty^2 \leq 2s(A)^2\varepsilon^2.$$

Proof. The first step is by Definition 4.3.2 of Δ, and the rest is exactly Fact 3.1.13. □

A simple consequence is that $\Delta(A) = 0$ iff A is doubly balanced. Therefore, we can follow the approach laid out in [62] to solve the Paulsen problem: follow the gradient flow of Δ and bound the distance using Δ as a potential function. The following shows that this approach is exactly equivalent to the dynamical system on scalings given in Definition 3.1.14. This is a special case of a general phenomenon coming from the work of Kempf and Ness [58] in geometric invariant theory.

Theorem 4.3.4 (Kempf-Ness Equivalence)**.** *For any matrix $A \in \mathrm{Mat}(d,n)$, the Euclidean gradient flow of Δ is (up to constants) equivalent to the dynamical system on matrices induced by Definition 3.1.14. Explicitly, the gradient flow on scalings $\partial_t(X_t, Y_t) = -\nabla f_A(X_t, Y_t)$ induces the matrix dynamical system $A_t = e^{X_t/2} A e^{Y_t/2}$ which is equivalently the solution to the following differential equation:*

$$A_0 = A, \quad \partial_t(A_t)_{ij} = -\frac{1}{8}\partial_{ij}\Delta(A_t).$$

Proof. The initial conditions are clearly equivalent as $A_0 = A = e^{X_0/2} A e^{Y_0/2}$. We show that for all $A \in \mathrm{Mat}(d,n)$, the direction induced by the gradient of Δ is the same as the direction induced by gradient flow of scalings in Definition 3.1.14 at time $t = 0$.

We check the statement for each entry $a \in [d], b \in [n]$. We can write out $\partial_{ab}\Delta$ according to Definition 4.3.2 as

$$\partial_{ab}\Delta(A) = d\sum_{i=1}^{d} \partial_{ab}\left(\langle E_{ii}, AA^*\rangle - \frac{s(A)}{d}\right)^2 + n\sum_{j=1}^{n} \partial_{ab}\left(\langle E_{jj}, A^*A\rangle - \frac{s(A)}{n}\right)^2.$$

Note that the row sum $\langle E_{ii}, AA^*\rangle = \sum_{j=1}^{n} |A_{ij}|^2$ depends on A_{ab} iff $a = i$, and similarly the column sum $\langle E_{jj}, A^*A\rangle = \sum_{i=1}^{d} |A_{ij}|^2$ depends on A_{ab} iff $b = j$. Also, the size depends on every element as $\partial_{ab}s(A) = \partial_{ab}\sum_{ij}|A_{ij}|^2 = 2A_{ab}$. We calculate the first term as

$$\begin{aligned} d\sum_{i=1}^{d} \partial_{ab}\left(\langle E_{ii}, AA^*\rangle - \frac{s}{d}\right)^2 &= d\sum_{i=1}^{d} 2\left(\langle E_{ii}, AA^*\rangle - \frac{s}{d}\right)\left(\partial_{ab}\langle E_{ii}, AA^*\rangle - \partial_{ab}\frac{s}{d}\right) \\ &= 2d\left(\langle E_{aa}, AA^*\rangle - \frac{s}{d}\right)(2A_{ab}) - 2d\sum_{i=1}^{d}\left(\langle E_{ii}, AA^*\rangle - \frac{s}{d}\right)(2A_{ab}). \end{aligned}$$

Note that the second term in the equation above vanishes because $\sum_{i=1}^{d} \langle E_{ii}, AA^* \rangle = \|A\|_F^2 = s(A)$. The other term involving column error from Δ can be calculated similarly as

$$\begin{aligned} n \sum_{j=1}^{n} \partial_{ab} \left(\langle E_{jj}, A^*A \rangle - \frac{s}{n} \right)^2 &= n \sum_{j=1}^{n} 2 \left(\langle E_{jj}, A^*A \rangle - \frac{s}{n} \right) \left(\partial_{ab} \langle E_{jj}, A^*A \rangle - \partial_{ab} \frac{s}{n} \right) \\ &= 2n \left(\langle E_{bb}, A^*A \rangle - \frac{s}{n} \right) (2A_{ab}) - 2n \sum_{j=1}^{n} \left(\langle E_{jj}, A^*A \rangle - \frac{s}{n} \right) (2A_{ab}), \end{aligned}$$

and again the second term vanishes because $\sum_{j=1}^{n} \langle E_{jj}, A^*A \rangle = \|A\|_F^2 = s$. Therefore, we can combine the two terms to show

$$\begin{aligned} \frac{1}{4} \partial_{ab} \Delta(A) &= \Big(d \langle E_{aa}, AA^* \rangle - s \Big) A_{ab} + A_{ab} \Big(n \langle E_{bb}, A^*A \rangle - s \Big) \\ &= \Big(d \cdot r_a(A) - s(A) \Big) A_{ab} + A_{ab} \Big(n \cdot c_b(A) - s(A) \Big), \end{aligned}$$

where in the last line we used Definition 3.1.1 on rows and columns.

We next calculate the direction induced by Definition 3.1.14 of matrix gradient flow (note the leading factor 2):

$$2\partial_{t=0}(e^{X_t/2} A e^{Y_t/2}) = (\partial_{t=0} X_t) A + A(\partial_{t=0} Y_t) = -(\nabla_A^L A + A \nabla_A^R),$$

where the first step was by the product rule and initial conditions $(X_0, Y_0) = (0, 0)$, and the final step was by Definition 3.1.14. We can write the (a, b) entry of this equation as

$$\Big[2\partial_{t=0}(e^{X_t/2} A e^{Y_t/2}) \Big]_{ab} = \Big(s(A) - d \cdot r_a(A) \Big) A_{ab} + A_{ab} \Big(s(A) - n \cdot c_b(A) \Big),$$

where the final equality follows by Proposition 3.1.12. Since this is equivalent to the equation derived for $-\frac{1}{4}\partial_{ab}\Delta$, the theorem is shown. □

This equivalence allows us to more directly analyze the path length of the dynamical system, as it is defined in terms of the frame instead of the scalings.

Lemma 4.3.5. *For $A \in \mathrm{Mat}(d, n)$ and dynamical system A_t according to Definition 3.1.14,*

$$\partial_t \Delta(A_t) = -8 \|\partial_t A_t\|_F^2.$$

Further, the distance traveled can be bounded by

$$\|A_T - A_0\|_F \leq \frac{1}{\sqrt{8}} \int_0^T \sqrt{-\partial_t \Delta(A_t)}.$$

Proof. The first statement can be shown by the following calculation:

$$\partial_t \Delta(A_t) = \sum_{i=1}^{d} \sum_{j=1}^{n} \overline{(\partial_{ij}\Delta(A_t))}(\partial_t A_t)_{ij} = -8 \sum_{ij} |(\partial_t A_t)_{ij}|^2 = -8\|\partial_t A_t\|_F^2,$$

where the first step was by the chain rule, and in the final step we used Theorem 4.3.4 to relate $\partial_t (A_t)_{ij} = -\frac{1}{8}\partial_{ij}\Delta(A_t)$, and the last step is by definition of the Frobenius norm.

Next, we show the distance bound:

$$\|A_T - A_0\|_F = \left\| \int_0^T \partial_t A_t \right\|_F \leq \int_0^T \|\partial_t A_t\|_F = \frac{1}{\sqrt{8}} \int_0^T \sqrt{-\partial_t \Delta(A_t)},$$

where the first step is the fundamental theorem of calculus, the second is by the triangle inequality, and the final step is by the equality just derived. □

Following this approach, we have reduced the Paulsen problem for matrices to a convergence analysis on $\Delta(A) = \|\nabla_A\|_{\mathfrak{t}}^2$. In the next section, we will use the fast convergence properties of strongly convex and pseudorandom inputs to give strong distance bounds.

4.3.2 Strong Convexity Analysis

Assuming strong convexity according to Definition 3.2.1, the distance bound now follows by a simple calculation on the change in Δ. This is the same analysis as in Lemma 4.16 of [63] which was used to bound the distance for operator scaling.

Proposition 4.3.6. *Consider $A \in \mathrm{Mat}(d, n)$ with A_t the solution of gradient flow according to Definition 3.1.14 (or equivalently Theorem 4.3.4). If A_t is α-strongly convex for all $t \in [0, T]$, then the distance of the dynamical system is bounded by*

$$\|A_T - A\|_F^2 \leq \frac{\Delta(A)}{4\alpha}.$$

Proof. Lemma 4.3.5 bounds the distance travelled by

$$\|A_T - A_0\|_F \leq \frac{1}{\sqrt{8}} \int_0^T \sqrt{-\partial_t \Delta(A_t)}.$$

Below, we will show that α-strong convexity implies $\sqrt{-\partial_t \Delta(A_t)} \leq -\partial_t \sqrt{\frac{2\Delta(A_t)}{\alpha}}$, from which the proposition will follow by integration.

The assumption that A_t is α-strongly convex is equivalent to

$$-\partial_t \Delta(A_t) = -\partial_t \|\nabla_{A_t}\|_{\mathfrak{t}}^2 \geq 2\alpha \|\nabla_{A_t}\|_{\mathfrak{t}}^2 = 2\alpha \cdot \Delta(A_t),$$

where the first and last equalities are by Definition 4.3.2 of Δ, and the inequality is due to the exponential convergence shown in Proposition 3.2.2. Therefore, we can continue

$$\sqrt{2\alpha} \leq \sqrt{\frac{-\partial_t \Delta(A_t)}{\Delta(A_t)}} \iff \sqrt{-\partial_t \Delta(A_t)} \leq \frac{-\partial_t \Delta(A_t)}{\sqrt{2\alpha \Delta(A_t)}} = -\sqrt{\frac{2}{\alpha}} \cdot \partial_t \sqrt{\Delta(A_t)}, \tag{4.4}$$

where the last step was by the chain rule. Therefore, we can bound the distance by

$$\|A_T - A_0\|_F \leq \frac{1}{\sqrt{8}} \int_0^T \sqrt{-\partial_t \Delta(A_t)} \leq \frac{-1}{\sqrt{4\alpha}} \int_0^T \partial_t \sqrt{\Delta(A_t)} = \frac{\sqrt{\Delta(A_0)} - \sqrt{\Delta(A_T)}}{\sqrt{4\alpha}},$$

where the first step is by Lemma 4.3.5, and the second is by Eq. (4.4). Recalling that $A = A_0$ and $\Delta(A_T) \geq 0$, the proposition follows by squaring both sides. □

Remark 4.3.7. *Eq. (4.4) is an instance of the Lojaseiwicz gradient inequality from analysis (see Prop 6.8 in [14]). This inequality is used in the result of Lerman [65] to show properties of this same gradient flow in the much more general setting of Hamiltonian manifolds.*

This result can be combined with the fast convergence analysis in Theorem 3.2.19 to replace the strong convexity assumption throughout gradient flow by sufficient strong convexity at the initial input. We omit this proof, as well as its lift to the frame setting, as will only use the pseudorandom analysis of the following subsection for our results on the Paulsen problem. This strongly convex distance bound will be applied in Section 8.5 to show algorithmic convergence of the distance to a doubly balanced frame.

4.3.3 Pseudorandom Analysis

When the input matrix satisfies the pseudorandom condition given in Definition 3.3.1, Theorem 3.3.10 shows exponential convergence of $\|\nabla\|_\infty$ instead of $\|\nabla\|_{\mathfrak{t}}^2 = \Delta$. In this section, we use this condition to directly prove a strong distance bound.

We first show an elementary result on the distance traveled that requires no conditions.

Lemma 4.3.8. *For A_t the solution to gradient flow according to Definition 3.1.14 and interval $0 \leq t_1 \leq t_2 \leq \infty$:*

$$\|A_{t_2} - A_{t_1}\|_F \leq \sqrt{\frac{\Delta(A_{t_1})(t_2 - t_1)}{8}}.$$

Proof. The proof is by a simple Cauchy-Schwarz as

$$\sqrt{8}\|A_{t_2}-A_{t_1}\|_F \leq \int_{t_1}^{t_2} \sqrt{-\partial_t \Delta(A_t)} \leq \sqrt{\int_{t_1}^{t_2} -\partial_t \Delta(A_t)}\sqrt{\int_{t_1}^{t_2} 1} = \sqrt{(\Delta(A_{t_1}) - \Delta(A_{t_2}))(t_2 - t_1)},$$

where the first step is by Lemma 4.3.5, the second is by the Cauchy-Schwarz inequality, and the last step is by the fundamental theorem of calculus. Note that $\Delta(A_{t_2}) \leq \Delta(A_{t_1})$ by Theorem 4.3.4 as we are following the gradient flow of Δ, so the term in the square root is non-negative. The lemma follows as $\Delta(A_{t_2}) \geq 0$. □

Combining this with the exponential convergence of $\|\nabla_{A_t}\|_\infty$ for pseudorandom inputs gives the strong distance bound.

Proof of Proposition 4.3.1. Since A satisfies the conditions of Theorem 3.3.10, we can bound Δ for all time:

$$\Delta(A_t) = \|\nabla_{A_t}\|_{\mathfrak{t}}^2 \leq \|\nabla_{A_t}^L\|_\infty^2 + \|\nabla_{A_t}^R\|_\infty^2 \leq 2(1+\alpha/2)^2\varepsilon^2 e^{-\frac{2\alpha t}{3e}},$$

where the first step was by Definition 4.3.2 of Δ, the second was by Lemma 3.2.7, and the final step was by the convergence guarantee of Theorem 3.3.10(1).

To bound the distance travelled, we break the convergence into intervals $t_k := k\frac{3e}{2\alpha}$ and bound the distance of each interval:

$$\|A_\infty - A\|_F \leq \sum_{k\geq 0} \|A_{t_{k+1}} - A_{t_k}\|_F \leq \sum_{k\geq 0} \sqrt{\frac{\Delta(A_{t_k})}{8} \cdot \frac{3e}{2\alpha}} \leq \sum_{k\geq 0} \sqrt{\frac{3e(1+\alpha/2)^2\varepsilon^2 e^{-k}}{8\alpha}} \leq \sqrt{\frac{8\varepsilon^2}{\alpha}},$$

where the second step was by Lemma 4.3.8 and our choice of $t_{k+1} - t_k = \frac{3e}{2\alpha}$, the third was by substituting $t_k = k\frac{3e}{2\alpha}$ into the bound on $\Delta(A_t)$ derived above, and the final step was by a geometric sum and the assumption $\alpha \leq \frac{1}{5}$. □

This result implies item (4) of Theorem 4.2.14, which will be used in the remainder of this chapter to give optimal bounds for the Paulsen problem. Specifically, we show optimal distance bounds for average case inputs in Section 4.4, and for worst case inputs in Section 4.5.

4.4 Average Case Analysis

In this section, we show that random frames satisfy the conditions of Theorem 4.2.14. As a consequence we get an optimal average case analysis for the Paulsen problem. As another consequence of this strong convergence result on random frames, we are also able to give very simple constructions of near-optimal Grassmannian frames, which was one of the original motivations of [49] for the Paulsen problem.

In [63], we used fast convergence properties of Definition 4.1.6 to give strong distance bounds for the Paulsen problem when the input satisfied a certain "spectral gap" condition. In fact, [63] was already able to give beyond worst-case bounds by showing random frames satisfied this condition with high probability. Franks and Moitra subsequently improved this result in [35] and used it to show near-optimal sample complexity bounds for an important statistical estimation problem. In Section 8.5, we use the pseudorandom analysis in Theorem 4.2.14 to improve this result and give tight sample complexity results for the statistical problem studied in [35].

The next theorem is a standard result in random matrix theory that was used in [35] to show that random frames are nearly doubly balanced with high probability. Note that each vector in the frame is a random unit vector, so the frame is equal-norm by definition. Otherwise, if we had used random Gaussians, then the frame becomes less and less balanced as n grows, which would make it difficult to apply any of our convergence results, and also would defeat the purpose of generating these random frames.

Theorem 4.4.1 (Theorem 5.39 in [94], Theorem 5.14 in [35])**.** *Let $U \in \mathrm{Mat}(d,n)$ be a random matrix where the columns are independent and uniformly distributed as $u_j \sim n^{-1/2}S^{d-1}$. Then there exists a universal constant C such that, for any $\varepsilon \leq \frac{1}{C}$ with $n \geq \frac{d}{\varepsilon^2}$,*

$$\left\| \sum_{j=1}^{n} u_j u_j^* - \frac{1}{d} I_d \right\|_{\mathrm{op}} \lesssim \frac{\varepsilon}{d},$$

with probability at least $1 - 2\exp(-\Omega(\varepsilon^2 n))$.

The next theorem states that random frames are pseudorandom with high probability. This allows us to use Theorem 4.2.14 in order to give tight bounds for the Paulsen problem in the average case.

Theorem 4.4.2. *Let $U \in \mathrm{Mat}(d,n)$ be a random matrix where the columns are independent and uniformly distributed as $u_j \sim n^{-1/2}S^{d-1}$. For any $20e^{-d/9} \leq \beta \leq \frac{1}{2}$, if $n \geq 15\frac{d}{\beta}$, then U is an $(e^{-(6-3\log_2 \beta)}, \beta)$-pseudorandom frame according to Definition 4.2.11 with probability at least $1 - 2\exp(-\frac{\beta n}{10})$.*

Remark 4.4.3. *In both [63] and [35], the authors used a "spectral gap" strategy in order to show fast convergence of random frames. We discuss the relationship between this spectral condition and our strong convexity condition in Section 7.1.3.*

This is the main result of the probabilistic analysis in Section 5.1. This probabilistic analysis combined with the fast convergence from Theorem 4.2.14 implies a near-optimal distance bound for the Paulsen problem on random inputs.

Theorem 4.4.4. *Let $U \in \operatorname{Mat}(d,n)$ be a random frame where the columns are independent and uniformly distributed as $u_j \sim n^{-1/2}S^{d-1}$. Then there exists a universal constant C such that if $n \geq Cd$, the following results holds simultaneously with probability at least $1 - \exp(-\Omega(n))$:*

1. *U has size $s(U) = 1$ and is an ε-doubly balanced frame with $\varepsilon^2 \lesssim \frac{d}{n}$;*
2. *U is an $(\alpha \geq \Omega(1), \frac{1}{2})$-pseudorandom frame;*
3. *The solution to the frame scaling problem in Definition 4.2.1 on input U is a doubly balanced frame $V \in \operatorname{Mat}(d,n)$ with*

$$\|U - V\|_F^2 \lesssim \varepsilon^2.$$

Proof. Item (1) is by Theorem 4.4.1 and item (2) is by Theorem 4.4.2 applied for $\beta = \frac{1}{2}$. By the union bound, both hold simultaneously with probability at least $1 - \exp(-\Omega(n))$. Therefore we have

$$\alpha \geq \Omega(1) \geq 16e \cdot \varepsilon,$$

where the last step was by our assumption $\varepsilon \leq \frac{1}{C}$, so we can apply Theorem 4.2.14 to show the frame scaling solution on input U satisfies the distance bound in item (3). □

Theorem 5.1 in [63] gave the same conclusion as Theorem 4.4.4 but required the stronger conditions $n \gtrsim d^{4/3}$. This was improved to the condition $n \gtrsim d \log^2 d$ in [35] though the authors did not require a distance bound in that work. Both of these works used a variant of the strong convexity analysis in Section 3.2 in order to show fast convergence. The key to our improvement is the pseudorandom analysis in Section 3.3, which has an optimal requirement $\frac{\alpha}{\varepsilon} \gtrsim 1$, instead of $\frac{\alpha}{\varepsilon} \gtrsim \log d$ for strong convexity.

Note that by Example A.1.1, the distance function for the Paulsen problem must in general grow linearly with ε. Therefore, the above theorem provides a beyond-worst case analysis for random inputs. Further, it can be shown that these random frames are typically

ε-Parseval for $\varepsilon^2 \gtrsim \frac{d}{n}$, so by Fact 4.1.5 this gives a lower bound for the distance that matches Theorem 1.1.5 up to constants.

Random frames can also be shown to satisfy strong pairwise correlation properties according to Definition 4.1.3. Therefore, this scaling approach gives a simple procedure to construct near-optimal Grassmannian frames.

Theorem 4.4.5. *Let $U \in \operatorname{Mat}(d,n)$ be a random matrix where the columns are independent and uniformly distributed as $u_j \sim n^{-1/2} S^{d-1}$. For any $n \gtrsim d$ large enough, with probability at least $1 - \frac{1}{\operatorname{poly}(n)}$, if U_* is the solution to the frame scaling problem on input U, then $V := \frac{U_*}{\|U_*\|_F}$ is doubly balanced and*

$$\Theta(V) \lesssim \frac{1}{n^2}\left(\frac{\log n}{d} + \frac{d}{n}\right)$$

where Θ is the correlation parameter given in Definition 4.1.3.

Proof. We follow the proof of Theorem 4.20 in [63] but supply our own pseudorandom convergence analysis using Theorem 4.2.14. The result is trivial for $\log n \gtrsim d$ as

$$\Theta(V) = \max_{j \neq j' \in [n]} \langle v_j, v_{j'} \rangle^2 \leq \max_j \|v_j\|_2^4 = \frac{1}{n^2},$$

where the first step was by Definition 4.1.3 of Θ, the second step was by the Cauchy-Schwarz inequality, and the final step was by the fact that V is equal-norm of size $s(V) = 1$. Therefore, for the following, we assume $\log n \lesssim d$.

The random frame U has size $s(U) = 1$ by construction, and the condition $n \gtrsim d$ implies that U is $\varepsilon \lesssim \sqrt{\frac{d}{n}}$-doubly balanced by Theorem 4.4.1, and is $(\alpha = \Omega(1), \frac{1}{2})$-pseudorandom frame by Theorem 4.4.2 simultaneously with probability at least $1 - \exp(-\Omega(n))$. Our plan is to bound $\Theta(U)$ with high probability, and then show that $\Theta(V) \approx \Theta(U)$ by the strong scaling bounds of Theorem 4.2.14(2).

First note that by independence and orthogonal invariance, the random variable $X_{jj'} := n^2 \langle u_j, u_{j'} \rangle^2$ for $j \neq j'$ has the same distribution as $X := \langle v, e_1 \rangle^2$ with $\mathbb{E}X = \frac{1}{d}$. Further, Lemma 5.9 in [35] shows that $X - \frac{1}{d}$ is $(O(d^{-2}), O(d^{-1}))$-subexponential according to Definition 2.5.3. Therefore we can apply the Bernstein bound in Lemma 2.5.4 to show

$$Pr[X - \mathbb{E}X \geq \theta] \leq \exp\left(-\min\left\{\frac{\theta^2}{O(d^{-2})}, \frac{\theta}{O(d^{-1})}\right\}\right).$$

Since $X_{jj'}$ has the same distribution as X for each pair $j \neq j'$, we can apply the union bound with $\theta = C\frac{logn}{d}$ for some large enough constant C to show

$$Pr\Big[\Theta(U) \gtrsim \frac{\log n}{dn^2}\Big] \leq n^2 Pr[X - \mathbb{E}X \geq \theta] \leq n^2 \exp\Big(-\Omega(1)\min\Big\{\log^2 n, \log n\Big\}\Big) \leq n^{-\Omega(1)},$$

where the first step was by our choice of $\theta = C\frac{\log n}{d}$ and the union bound over all pair $j \neq j' \in [n]$, the second step was by the concentration bound on X derived above, and the final step was again by our choice of $\theta = C\frac{\log n}{d}$ for large enough leading constant C.

Now we apply fast convergence via scaling to show $\Theta(V) \approx \Theta(U)$. For $n \gtrsim d$ large enough, U satisfies the conditions of Theorem 4.2.14, so we can lower bound the size by

$$s(U_*) \geq s(U) - O\Big(\frac{\varepsilon^2}{\alpha}\Big) \geq 1 - O(\varepsilon^2) = 1 - O\Big(\frac{d}{n}\Big),$$

where the first step was by Theorem 4.2.14(3), the second step was because U is $(\alpha = \Omega(1), \frac{1}{2})$-pseudorandom, and in the final step substituted $\varepsilon^2 \lesssim \frac{d}{n}$ by the doubly balanced condition on U. Similarly, the scaling $U_* := e^{X_*/2}Ue^{Y_*/2}$ satisfies

$$\max\{\|X_*\|_{\text{op}}, \|Y_*\|_{\text{op}}\} \lesssim \frac{\varepsilon}{\alpha} \lesssim \sqrt{\frac{d}{n}},$$

where the first step was by the bound in Theorem 4.2.14(2) and the final step was by the bounds shown above: $\alpha \gtrsim 1$ by pseudorandomness and $\varepsilon^2 \lesssim \frac{d}{n}$ by the doubly balanced condition on U.

We can use the above facts to bound the change in the correlation of $V := \frac{U_*}{\|U_*\|_F} = \frac{U_*}{\sqrt{s(U_*)}}$. We first rewrite the correlations of V in terms of U:

$$\langle v_j, v_{j'}\rangle = \frac{1}{s(U_*)}\langle e^{X_*/2}u_j e^{(Y_*)_{jj}/2}, e^{X_*/2}u_j e^{(Y_*)_{j'j'}/2}\rangle = \frac{e^{(Y_*)_{jj}/2}e^{(Y_*)_{j'j'}/2}}{s(U_*)}\langle u_j, e^{X_*}u_{j'}\rangle.$$

Now we use the fact that (X_*, Y_*) are close to the origin, so we can bound

$$\begin{aligned}|\langle v_j, v_{j'}\rangle| &\leq \frac{e^{((Y_*)_{jj}+(Y_*)_{j'j'})/2}}{s(U_*)}\Big(|\langle u_j, u_{j'}\rangle| + |\langle u_j, (e^{X_*} - I_d)u_{j'}\rangle|\Big) \\ &\lesssim |\langle u_j, u_{j'}\rangle| + |\langle u_j, (e^{X_*} - I_d)u_{j'}\rangle| \lesssim |\langle u_j, u_{j'}\rangle| + \|u_j\|_2\|u_{j'}\|_2\|X_*\|_{\text{op}},\end{aligned}$$

where in the first step we separated $e^{X_*} = I_d + (e^{X_*} - I_d)$, in the second step we used $\max_{j\in[n]}|Y_{jj}| = \|Y_*\|_{\text{op}} \lesssim \frac{\varepsilon}{\alpha} \lesssim 1$ and the lower bound $s(U_*) \geq 1 - O(\varepsilon^2)$ to bound the first term by a constant, and in the third step we used the definition of the operator norm and the Taylor approximation $|e^x - 1| \lesssim |x|$ for $|x| \lesssim 1$ to bound the second term.

By the correlation bound on $\Theta(U)$ derived above, this implies

$$\Theta(V) = \max_{j \neq j' \in [n]} |\langle v_j, v_{j'} \rangle|^2 \lesssim \max_{j \neq j'} \Big(|\langle u_j, u_{j'} \rangle| + \|u_j\|_2 \|u_{j'}\|_2 \|X_*\|_{\mathrm{op}} \Big)^2$$
$$\leq \left(\sqrt{\Theta(U)} + \frac{O(\varepsilon)}{\alpha n} \right)^2 \lesssim \frac{1}{n^2} \left(\frac{\log n}{d} + \frac{d}{n} \right),$$

where the first step was by Definition 4.1.3 of correlation, in the second step we applied the perturbation bound derived above to rewrite $|\langle v_j, v_{j'} \rangle|$ in terms of U and the scaling X_*, in the third step we used $u_j \in n^{1/2} S^{d-1}$ by definition as well as the bound $\|X_*\|_{\mathrm{op}} \lesssim \frac{\varepsilon}{\alpha} \lesssim \varepsilon$ derived from Theorem 4.2.14(2), and in the final step we used the simple fact $(a+b)^2 \leq 2(a^2 + b^2)$ and substituted in the value $\Theta(U) \lesssim \frac{\log n}{dn^2}$, $\alpha \gtrsim 1$, and $\varepsilon^2 \lesssim \frac{d}{n}$ derived above. □

4.5 Solution to the Paulsen Problem

This section collects together the results necessary to show our optimal distance bound for the Paulsen problem. The perturbation argument and its proof are given in Section 5.2 and Section 5.3.

We follow the same strategy as [62], by showing that for any ε-doubly balanced input $U \in \mathrm{Mat}(d, n)$ to the Paulsen problem, there is a nearby perturbation $V := U + E$ such that V satisfies the conditions of Theorem 4.2.14. Combining these two steps gives a strong distance bound for Conjecture 4.1.4. The next two theorems contain the main perturbation step requirements.

We separate the perturbation argument into two cases for the following technical reasons. To show existence of the frame satisfying the properties above, we will add random noise to the input; in the asymptotic case of $n \to \infty$, with high probability there will exist some $j \in [n]$ such that the random noise in this column will make the perturbation V have large error ∇_V^R. To combat this, we re-normalize all the columns in this large n case, and so we must argue that this different process still satisfies our requirements. These arguments are proven in full detail in Section 5.2 and Section 5.3 respectively.

Theorem 4.5.1. *Let frame $U \in \mathrm{Mat}(d, n)$ have size $s(U) = 1$ and be ε-doubly balanced. Then, for any $\beta \leq \frac{1}{2}$, there is a universal constant $C = C_\beta$ depending only on β such that if $d \geq C, Cd \leq n \leq e^{d/C}, \varepsilon \leq 1/C$, then there exists a frame V satisfying*

1. *V has size $s(V) = 1$ and is 2ε-doubly balanced;*

2. *V is (α, β)-pseudorandom as a frame with $\alpha \geq 16e(2\varepsilon)$;*

3. *$\|V - U\|_F^2 \leq O_\beta(\varepsilon)$.*

Theorem 4.5.2. *Let frame $U \in \operatorname{Mat}(d, n)$ have size $s(U) = 1$ and be ε-doubly balanced. Then, for any $\beta \leq \frac{1}{2}$, there is a universal constant $C = C_\beta$ depending only on β such that if $d \geq C, n \geq Cd, \varepsilon \leq \frac{1}{Cd}$, then there exists a frame V satisfying*

1. *V has size $s(V) = 1$ and is 4ε-doubly balanced;*

2. *V is (α, β)-pseudorandom as a frame with $\alpha \geq 16e(4\varepsilon)$;*

3. *$\|V - U\|_F^2 \leq O_\beta(\varepsilon)$.*

Note that the first theorem has a two-sided condition $\Omega(d) \leq n \leq e^{O(d)}$, whereas the goal of the Paulsen problem is to develop bounds that are independent of n. Therefore, we supplement with the second theorem, which only has a lower bound condition $n \gtrsim d$, but requires much smaller initial error $\varepsilon \lesssim \frac{1}{d}$.

Given these perturbation theorems, we complete the distance analysis for the Paulsen problem in almost the entire parameter regime.

Theorem 4.5.3. *There exists a universal constant C such if $n \geq Cd$ and*

$$n \leq e^{d/C} \qquad \text{or} \qquad \varepsilon \leq \frac{1}{Cd},$$

then for any ε-doubly balanced frame $U \in \operatorname{Mat}(d, n)$ of size $s(U) = 1$, there exists a doubly balanced $V \in \operatorname{Mat}(d, n)$ with size $s(V) = 1$ such that

$$\|U - V\|_F^2 \lesssim \varepsilon.$$

In the language of Conjecture 4.1.4, $p(d, n, \varepsilon) \lesssim \varepsilon$ when the above conditions are satisfied.

Proof. We first deal with some corner cases. If $d \leq C$, then the main results of [46] (or even its weaker form in [62]) give

$$p(d, n, \varepsilon) \lesssim d\varepsilon \lesssim \varepsilon.$$

Similarly, if $\varepsilon \geq \frac{1}{C}$ then choosing V as an arbitrary doubly balanced frame gives

$$\sum_{j=1}^{n} \|v_j - u_j\|_2^2 \leq \sum_{j=1}^{n} (\|v_j\|_2 + \|u_j\|_2)^2 \lesssim 1 \lesssim \varepsilon,$$

where the first step was by triangle inequality, the second was because $\|u_j\|_2 \approx \|v_j\|_2 = \frac{1}{n}$, and in the last step we used the assumption $\varepsilon \geq \frac{1}{C}$.

In the remaining cases, we apply Theorem 4.5.1 for $d \geq C, Cd \leq n \leq e^{cd}, \varepsilon \leq \frac{1}{C}$, and Theorem 4.5.2 for $d \geq C, n \geq Cd, \varepsilon \leq \frac{1}{Cd}$, both for $\beta = \frac{1}{2}$ and $C = C_\beta$ large enough. So let V be the output of those theorems. Then V satisfies the pseudorandom condition in Theorem 4.2.14, which implies V_∞ is a doubly balanced frame. To satisfy the size condition, we normalize $\tilde{V} := \frac{V_\infty}{\sqrt{s(V_\infty)}}$ and bound the distance:

$$\begin{aligned}\|\tilde{V} - U\|_F &\leq \|\tilde{V} - V_\infty\|_F + \|V_\infty - V\|_F + \|V - U\|_F \\ &\lesssim |(1 - O(\varepsilon))^{-1/2} - 1| + \sqrt{\frac{\varepsilon^2}{\varepsilon}} + \sqrt{\varepsilon} \lesssim \sqrt{\varepsilon},\end{aligned}$$

where the first step was by triangle inequality, in the second step we bounded the first term using the definition $\tilde{V} := \frac{V_\infty}{\sqrt{s(V_\infty)}}$ and the bound $s(V_\infty) \geq s(V) - O(\varepsilon^2/\alpha) = 1 - O(\varepsilon)$ from item (3) of Theorem 4.2.14 since $\alpha \gtrsim \varepsilon$, the second term by the distance bound in item (4) of Theorem 4.2.14, and the third term by the perturbation bound in Theorem 4.5.1(3) and Theorem 4.5.2(3) respectively, and the final inequality is by Taylor approximation $|\sqrt{1-x} - 1| \lesssim |x|$ for $|x| \lesssim 1$. □

Remark 4.5.4. *In view of the lower bound in Example A.1.1, the above theorem gives a tight distance bound for all cases except: (1) $d \geq C, n \leq Cd, \varepsilon \leq \frac{1}{C}$, and (2) $d \geq C, n \geq e^{cd}, \frac{1}{C} \geq \varepsilon \geq \frac{1}{Cd}$. For these remaining cases, we resort to the $O(d\varepsilon)$ bound of Hamilton and Moitra [46]. We once again remark that their result is entirely unconditional and the procedure is quite a bit simpler. We believe that our perturbation approach can be improved to cover all cases, but some new ideas are required to give a unified argument for all parameter settings.*

In the following Chapter 5, we will prove that Theorem 5.1.6 showing random frames are pseudorandom, as well as the perturbation statements in Theorem 4.5.1 and Theorem 4.5.2.

Chapter 5

Smoothed Analysis of the Paulsen Problem

In this chapter we prove strong error and pseudorandom properties for certain random distributions of frames. In Section 5.1 we consider the setting of random unit vectors, which will be used to prove optimal average-case bounds for the Paulsen problem in Section 4.4. Then in Section 5.2 and Section 5.3 we show that a random perturbation of a nearly doubly balanced frames satisfies the pseudorandom property of Theorem 4.2.14 with high probability. This smoothed analysis approach allows us to prove an optimal distance bound for the Paulsen problem in Section 4.5.

All of the results in this chapter will deal with real vectors and vector spaces. This is not without loss of generality, but the real case is easier to understand (for the author). We will mention how each definition and result can be simply lifted to the complex case as we go along.

5.1 Random Frames

In this section, we will study the random frame $V = \{v_1, ..., v_n\}$, where each v_j is drawn independently and uniformly from $\frac{1}{\sqrt{n}}S^{d-1}$. Note that V is equal norm and of size $s(V) = 1$ by construction. It is also nearly Parseval for n large enough with high probability according to Theorem 4.4.1. Here we will show that for every $\beta \leq \frac{1}{2}$ and n large enough (as a function of d, β), V is an $(\Omega(1), \beta)$-pseudorandom frame with high probability.

We first show that Definition 4.2.11 of $(\cdot, \beta)$-pseudorandom frames is equivalent to a spectral lower bound on all subsets $\{v_j\}_{j\in T}$ for $|T| = \beta n$.

Lemma 5.1.1. *Frame $V \in \mathrm{Mat}(d,n)$ is (α,β)-pseudorandom according to Definition 4.2.11 iff*

$$\min_{T\in\binom{[n]}{\beta n}} \lambda_d\left(\sum_{j\in T} v_j v_j^*\right) = \min_{T\in\binom{[n]}{\beta n}} \inf_{\xi\in S^{d-1}} \sum_{j\in T} |\langle \xi, v_j\rangle|^2 \geq \alpha\frac{\beta}{d},$$

where λ_d denotes the d-th largest eigenvalue.

Proof. Recall that frame V is pseudorandom according to Definition 4.2.11 iff $\Xi^* V$ is a pseudorandom matrix according to Definition 3.3.1 for every orthonormal basis Ξ. By Lemma 3.3.3, the pseudorandom matrix condition reduces to a lower bound on each row $i \in [d]$ and column subsets $T \in \binom{[n]}{\beta n}$. We lift this to the frame setting by taking an infimum over all orthonormal bases ($\mathrm{U}(d)$ if $\mathbb{F} = \mathbb{C}$ and $\mathrm{O}(d)$ if $\mathbb{F} = \mathbb{R}$). So for fixed $T \in \binom{[n]}{\beta n}$,

$$\inf_{\Xi} \min_{i\in[d]} \sum_{j\in T} |\langle \xi_i, v_j\rangle|^2 = \inf_{\xi\in S^{d-1}} \sum_{j\in T} |\langle \xi, v_j\rangle|^2,$$

where we used $(\Xi^* V)_{ij} = \langle \xi_i, v_j\rangle$ by Eq. (2.1), and the equality follows as $\cup_\Xi\{\xi_1, ..., \xi_d\} = S^{d-1}$ where the union is over all orthonormal bases (by a similar reasoning to the proof of Lemma 4.2.4). The lemma follows by requiring this condition for every $T \in \binom{[n]}{\beta n}$ as in Definition 4.2.11 of pseudorandomness. □

Our plan is to show this spectral lower bound for a single subset with high probability, and then show the pseudorandom condition by a union bound over subsets. Unfortunately, the standard concentration properties of unit vectors, i.e. the ones used in Theorem 4.4.1, can only give $\exp(-\beta n/4)$ as an upper bound on the failure probability for each subset, which is slightly weaker than necessary for a union bound over $\binom{n}{\beta n} \approx 2^{\beta n(1-\log_2 \beta)}$ subsets.

Therefore, in the next lemma, we show that the normalization $v_j = \frac{g_j}{\sqrt{n}\|g_j\|_2}$ for random Gaussian vectors $\{g_j \sim N(0, \frac{1}{n} I_d)\}_{j\in[n]}$ only decreases the pseudorandom condition by a small amount with high probability. This allows us to use the stronger lower tail bounds in Corollary 2.5.16 for Gaussian random vectors to give a tighter failure probability for the sets. We emphasize that this reduction is only for the analysis of pseudorandomness, and the frame we construct still comprises of random unit vectors.

Lemma 5.1.2. *Let $U \in \mathrm{Mat}(d,n)$ be an (α,β)-pseudorandom frame according to Definition 4.2.11, and $V = UR$ a right-scaling with $R \in \mathrm{diag}(n)$. If the subset*

$$T_B := \{j \in [n] \mid |R_j|^2 \leq \tau\}$$

satisfies $|T_B| \leq \beta' n$, *then* V *is a* $(\tau\alpha\frac{\beta}{\beta+\beta'}, \beta+\beta')$*-pseudorandom frame.*

Proof. We expand the condition of Lemma 5.1.1 for V: for $T \in \binom{[n]}{(\beta+\beta')n}$ and $\xi \in S^{d-1}$,

$$\sum_{j\in T} |\langle \xi, v_j\rangle|^2 = \sum_{j\in T} |R_j|^2 |\langle \xi, u_j\rangle|^2 \geq \sum_{j\in T-T_B} \tau |\langle \xi, u_j\rangle|^2 \geq \tau\alpha \frac{|T-T_B|}{dn} \geq \tau\alpha \frac{\beta}{\beta+\beta'} \frac{|T|}{dn},$$

where in the second step we used $|R_j|^2 \geq \tau$ for $j \in T_B$ by definition, the third step is by the pseudorandom property of U applied to $|T| - |T_B| \geq \beta n$, and the final step is again by $|T| - |T_B| \geq \beta n$ and $|T| = (\beta+\beta')n$. □

Below, we show that the normalization $v_j = \frac{g_j}{\sqrt{n}\|g_j\|_2}$ does not affect the vectors by too much, so that we can use the lemma above to reduce our analysis to showing the random Gaussian frame is pseudorandom.

Lemma 5.1.3. *Let* $G \in \mathrm{Mat}(d,n)$ *be a random frame where each column is generated independently as* $g_j \sim N(0, \frac{1}{dn} I_d)$*, and let* V *be the normalization, defined as* $v_j := \frac{g_j}{\sqrt{n}\|g_j\|_2}$*. Then for any* $\beta' \geq 20e^{-d/9}$*, the random variable* $T_B := \{j \in [n] \mid n\|g_j\|_2^2 \geq 2\}$ *satisfies*

$$Pr[|T_B| \geq \beta' n] \leq \exp(-\Omega(\beta' n)).$$

Proof. Note $nd\|g_j\|_2^2 \sim \chi(d)$ is distributed as a chi-squared variable with d degrees of freedom by Definition 2.5.9, so letting X_j be the indicator variable for the event $j \in T_B$, we can bound the mean by

$$\mathbb{E}X_j = Pr[n\|g_j\|_2^2 \geq 2] \leq Pr\Big[\chi(d) \geq d(1 + \frac{2}{\sqrt{9}} + \frac{2}{9})\Big] \leq \exp\Big(-\frac{d}{9}\Big),$$

where we used the bound in Theorem 2.5.11 for $\chi(d)$ with $\theta = \sqrt{\frac{d}{9}}$.

Therefore $\mathbb{E}|T_B| = \sum_{j\in[n]} \mathbb{E}X_j \leq ne^{-d/9}$. To show a high probability bound on $|T_B|$, we can use the fact that $\{X_j\}$ are mutually independent since the vectors $\{g_j\}$ are mutually independent. Now we can apply the Chernoff bound to show

$$Pr\Big[|T_B| \geq \beta' n\Big] = Pr\Big[\sum_{j=1}^n X_j \geq \beta' n\Big] \leq e^{-\mathbb{E}|T_B|}\Big(\frac{e\mathbb{E}|T_B|}{\beta' n}\Big)^{\beta' n} \leq \exp(-\Omega(\beta' n)),$$

where the first step was by definition $|T_B| = \sum_{j\in[n]} \mathbb{E}X_j$, the second step was by Theorem 2.5.2 applied to Bernoulli random variables $\{X_j\}_{j\in[n]}$, and in the final step we used the bound $\beta' n \geq 20\mathbb{E}|T_B|$. □

At this point, we can show pseudorandomness of V by lower bounding eigenvalues of random Gaussian matrices. In the next two lemmas, we use standard Gaussian concentration and net arguments to prove stronger lower tail bounds for eigenvalues of Gaussian random matrices. We first show that for each ξ the quantity we want to lower bound is very well concentrated.

Lemma 5.1.4. *For fixed $\xi \in S^{d-1}, T \in \binom{[n]}{\beta n}$ and independent random Gaussian vectors $g_1, ..., g_N \sim N(0, I_d)$, the random variable*

$$r_\xi := \langle \xi\xi^*, \sum_{j=1}^{N} g_j g_j^* \rangle = \sum_{j=1}^{N} |\langle \xi, g_j \rangle|^2$$

is distributed as $\chi(N)$, a chi-squared variable with N degrees of freedom. This implies that $\mathbb{E}[r_\xi] = N$ and for any $c \geq 5$,

$$Pr\Big[r_\xi \geq N(1+c)\Big] \leq \exp\left(-\frac{cN}{4}\right), \qquad \textit{and} \qquad Pr\Big[r_\xi \leq e^{-c}N\Big] \leq \exp\left(-\frac{2}{5}cN\right).$$

Proof. The distribution and mean statements follow from Definition 2.5.9 of chi-squared variables. The concentration of the upper bound follows from Theorem 2.5.11 by choosing $\theta^2 = \frac{c}{4} \geq 1$, and the concentration for the lower bound is exactly Corollary 2.5.16. □

Note that we only need a lower bound for $\sum_{j \in T} |\langle \xi, g_j \rangle|^2$ by Lemma 5.1.1. We use the upper bound from Lemma 5.1.4 in the following proof so that we can bound the operator norm and give a more precise net argument for the lower bound using Lemma 2.6.6.

Lemma 5.1.5. *For random Gaussian matrix $G = [g_1, ..., g_N]$ where $g_j \sim N(0, I_d)$ and any $c \geq 5$, if $N \gtrsim 10d$ then*

$$\lambda_{\min}(GG^*) = \inf_{\xi \in S^{d-1}} \|\xi^* G\|_2^2 \geq \frac{4}{9} e^{-c} N$$

with probability at least $1 - 2\exp(-\frac{cN}{3})$.

Proof. Our plan is to use Lemma 5.1.4 to bound each direction ξ in an appropriate net $N_L \subseteq S^{d-1}$. Since random Gaussian matrices are well-conditioned with very high probability, we can decrease the size of the net required for our union bound if we first bound the largest eigenvalue (according to Lemma 2.6.6).

So let $N_U \subseteq S^{d-1}$ be an $\eta_U = \frac{1}{3}$-net according to Definition 2.6.2. Then

$$Pr\Big[\sup_{\xi \in N_U} \|\xi^* G\|_2 \geq \sqrt{N(1+2c)}\Big] \leq \sum_{\xi \in N_U} Pr\Big[\|\xi^* G\|_2 \geq \sqrt{N(1+2c)}\Big]$$
$$\leq \exp\Big(2d - \frac{cN}{2}\Big) \leq \exp\Big(-\frac{cN}{3}\Big),$$

where the first step was by union bound over N_U, in the second step we applied Fact 2.6.3 to bound $|N_U| \leq (1+2\eta_U^{-1})^d \leq 7^d \leq e^{2d}$ and Lemma 5.1.4 with $2c$ to bound the probability for $r_\xi(G) = \|\xi^* G\|_2^2$, and the last step was by the assumption $N \geq 10d, c \geq 5$ so $d \leq \frac{cN}{50}$.

Assuming this event occurs, Lemma 2.6.5 shows

$$\sup_{\xi \in S^{d-1}} \|\xi^* G\|_2 \leq \frac{3}{2} \sup_{\xi \in N_U} \|\xi^* G\|_2 \leq \frac{3}{2}\sqrt{N(1+2c)}.$$

Now we perform a similar argument for the lower bound, so let $N_L \subseteq S^{d-1}$ be an η_L-net with $\eta_L = \frac{1}{3}\sqrt{\frac{e^{-c}}{(\frac{3}{2})^2(1+2c)}}$. Then Fact 2.6.3 gives a bound on the size of the net

$$|N_L| \leq (1 + 2\eta_L^{-1})^d \leq (1 + 9\sqrt{e^c(1+2c)})^d \leq e^{2cd},$$

where the final inequality was by the assumption $c \geq 5$. Therefore, we can lower bound

$$Pr\Big[\inf_{\xi \in N_L} \|\xi^* G\|_2^2 \leq e^{-c} N\Big] \leq \exp\left(2cd - \frac{2}{5}cN\right) \leq \exp\left(-\frac{cN}{3}\right),$$

where we used the union bound, the derived bound on $|N_L|$, and Lemma 5.1.4 to bound the probability, and the final step was by the assumption $c \geq 5$ so that $N \geq 30d$.

Now assume that both events occurred:

$$\inf_{\xi \in N_L} \|\xi^* G\|_2 \geq \sqrt{e^{-c}N}, \qquad \text{and} \qquad \sup_{\xi \in S^{d-1}} \|\xi^* G\|_2 \leq \frac{3}{2}\sqrt{N(1+2c)},$$

which by the union bound happens with probability at least $1 - 2\exp(-\frac{cN}{3})$. Then by Lemma 2.6.6,

$$\inf_{\xi \in S^{d-1}} \|\xi^* G\|_2 \geq \inf_{\xi \in N_L} \|\xi^* G\|_2 - \eta_L \sup_{\xi \in S^{d-1}} \|\xi^* G\|_2 \geq \sqrt{e^{-c}N}\left(1 - \frac{1}{3}\right),$$

where the last step was by our choice of η_L. Squaring both sides gives the result. □

By taking a union bound over all sets of size βn, we can show pseudorandomness of G. And combining this with our reduction gives pseudorandomness of V.

Theorem 5.1.6. *Let $G \in \mathrm{Mat}(d,n)$ be a random frame where each column is generated independently as $g_j \sim N(0, \frac{1}{dn} I_d)$, and let V be the right normalization given by $v_j := \frac{g_j}{\sqrt{n}\|g_j\|_2}$. Then for any $20e^{-d/9} \leq \beta \leq \frac{1}{2}$, if $n \geq 15\frac{d}{\beta}$, then G is $(\frac{4}{9}e^{-(4-3\log_2 \beta)}, \frac{4}{5}\beta)$-pseudorandom with probability at least $1 - \exp(-\beta n/10)$, and V is $(e^{-(6-3\log_2 \beta)}, \beta)$-pseudorandom with probability at least $1 - 2\exp(-\beta n/10)$.*

Note that the second statement in this theorem on random unit vectors is exactly a restatement of Theorem 4.4.2.

Proof. By Lemma 5.1.1, to show pseudorandomness of G it is enough to lower bound the smallest eigenvalue of every $\frac{4}{5}\beta n$ subset of vectors. Note that each subset $G_T = \{g_j\}_{j \in T}$ with $T \in \binom{[n]}{\frac{4}{5}\beta n}$ is a random Gaussian matrix where each entry is from $N(0, \frac{1}{dn})$ (not the standard Gaussian), and further that $|T| = \frac{4}{5}\beta n \geq 10d$ by our assumption $n \geq 15\frac{d}{\beta}$. Therefore, we apply Lemma 5.1.5 to each $T \in \binom{[n]}{\frac{4}{5}\beta n}$ with $c = 3(1 - \log_2(\frac{4}{5}\beta)) \leq 4 - 3\log_2 \beta$ to show

$$Pr\left[\min_{T \in \binom{[n]}{\frac{4}{5}\beta n}} \lambda_{\min}(G_T G_T^*) \leq \frac{4}{9}\frac{e^{-c}}{d} \cdot \frac{4}{5}\beta\right] \leq \left(\frac{2}{e}\right)^{\frac{4}{5}\beta n(1-\log_2(\frac{4}{5}\beta))} \leq \exp(-\beta n/10),$$

where $G_T = \{g_j\}_{j \in T}$, the bound $\binom{n}{\beta' n} \leq 2^{\frac{4}{5}\beta n(1-\log_2(\frac{4}{5}\beta))}$ is by Fact 2.6.1, and the bound on the failure probability of each subset comes from Lemma 5.1.5 with $c = 3(1 - \log_2(\frac{4}{5}\beta))$. This shows G is $(\alpha := \frac{4}{9}e^{-(4-3\log_2 \beta)}, \frac{4}{5}\beta)$-pseudorandom with probability at least $1 - \exp(-\beta n/10)$.

Next we show that this implies pseudorandomness for V. Lemma 5.1.3 shows that if $T_B := \{j \in [n] \mid \|g_j\|_2^2 \geq 2\}$, then

$$Pr\Big[|T_B| \geq \frac{1}{5}\beta n\Big] \leq \exp\left(-\frac{\beta n}{10}\right).$$

Since $v_j = \frac{g_j}{\sqrt{n}\|g_j\|_2}$, this is equivalent to considering $V = GR$ with $R_j = \frac{1}{\sqrt{n}\|g_j\|_2}$ so that $T_B := \{j \in [n] \mid |R_j|^2 \leq \tau = \frac{1}{2}\}$ satisfies $|T_B| \leq \frac{1}{5}\beta n$. Then assuming that we are in the event where G is $(\alpha := \frac{4}{9}e^{-(4-3\log_2 \beta)}, \frac{4}{5}\beta)$-pseudorandom, we can apply Lemma 5.1.2 with

R and τ defined above to show that, with additional failure probability at most $e^{-\beta n/10}$, V is $(\alpha', \frac{4}{5}\beta + \frac{1}{5}\beta)$-pseudorandom for

$$\alpha' \geq \tau\alpha \frac{\frac{4}{5}\beta}{\frac{4}{5}\beta + \frac{1}{5}\beta} \geq \frac{1}{2} \cdot \frac{4}{9} e^{-(4-3\log_2 \beta)} \cdot \frac{4}{5} \geq e^{-(6-3\log_2 \beta)},$$

where we substituted $\tau = \frac{1}{2}$ and the pseudorandomness of G shown above. □

Theorem 5.1.6 is used along with the error bound in Theorem 4.4.1 to give an optimal average-case analysis for the Paulsen problem in Section 4.4. Below we discuss the relationship of this result to the previous works of [63] and [35] on random frames.

In [63], we first studied random frames in the setting of the Paulsen problem. In that work, we used a slightly different spectral gap condition to show fast convergence of the scaling dynamical system of Definition 4.1.6. The relation between the spectral gap condition of [63] and strong convexity will be discussed in more detail in Chapter 7. We were able to show, using a trace method, that $n \gtrsim d^{4/3}$ vectors suffice in order for the random frame to satisfy the spectral gap condition of [63]. This implied strong distance bounds for the Paulsen problem as well as strong bounds on the scaling solution for random frames with $n \gtrsim d^{4/3}$.

In [35], Franks and Moitra improved this analysis by showing $n \gtrsim d \log d$ vectors suffice in order for random frames to satisfy the spectral gap condition. Their result went through an intermediate step of defining the Cheeger constant of a frame and then borrowing from spectral graph theory to show that a large Cheeger constant implied the spectral gap assumption of [63].

In our analysis, we are able to use the pseudorandom analysis of Section 3.3. Therefore, we show in Theorem 5.1.6 that $n \gtrsim d$ vectors suffice for a random frame to satisfy the pseudorandom condition in Definition 4.2.11, which then implies strong distance bounds and strong bounds on the scaling solution of random frames. This gives a strictly stronger result than the random frame analysis of [63]. Further, the main application in [35] was to show that bounds on the scaling solution of a random frame implied strong accuracy bounds on Tyler's M-estimator for a problem in high-dimensional statistics. They were also able to prove that when $n \gtrsim d \log^2 d$, this M-estimator could be computed by fast algorithm. In Section 8.5, we will be able to improve their result to show $n \gtrsim d$ vectors suffice for both of these results.

5.2 Perturbation Argument for small n

In this section, we will use smoothed analysis to prove the following generalization of Theorem 4.5.1.

Theorem 5.2.1. *For ε-doubly balanced frame $U \in \mathrm{Mat}_{\mathbb{R}}(d,n)$ of size $s(U) = 1$ and $\varepsilon \le \frac{1}{4}$, let $\frac{1}{4} \ge \delta > 0$ be the magnitude of noise added in the perturbation process $V = U + \delta G$ in Definition 5.2.2. Then for any $\beta \le \frac{1}{2}, \theta \le \frac{1}{4}$, if $d \ge \frac{100}{\min\{\beta,\theta^2\}}$ and $100\frac{d}{\min\{\beta,\theta^2\}} \le n \le e^{\frac{\theta^2 d}{100}}$, then the output of the perturbation process $V = U + \delta G$ satisfies:*

1. *(Distance): $\|V-U\|_F^2 = \delta^2 s(G)$ and $s(V) = 1 + \delta^2 s(G)$ with $s(G) \in 1 \pm \theta$;*
2. *(Error): $\max\{\|\nabla_V^L\|_{\mathrm{op}}, \|\nabla_V^R\|_{\mathrm{op}}\} \le (1+\delta^2)\varepsilon + 6\theta\delta^2$;*
3. *(Pseudorandom): V is an $(e^{-(4-3\log_2\beta)}\delta^2, \beta)$-pseudorandom frame;*

simultaneously with probability at least $1 - 6\exp(-\frac{\theta^2 d}{10})$.

Before we give the formal details of the perturbation argument, let us see how this smoothed analysis argument implies the existence of perturbation given in Theorem 4.5.1 in the $\mathbb{F} = \mathbb{R}$ case. This can be simply lifted to $\mathbb{C}$ by the discussion in Remark 5.2.4.

Proof of Theorem 4.5.1. We will show that there is an appropriate choice of parameters such that, for V the output of Theorem 5.2.1, the normalization $V' := \frac{V}{\|V\|_F}$ satisfies the three requirements of Theorem 4.5.1 with non-zero probability.

To this end, let $\delta^2 = e^{9-3\log_2\beta}\varepsilon$ and $\theta = e^{-(11-3\log_2\beta)}$. Clearly $\theta \le \frac{1}{2}$ by the assumption that $\beta \le \frac{1}{2}$, and we can take C_β large enough in Theorem 4.5.1 so that

$$d \ge \frac{100}{\min\{\theta^2,\beta\}} \qquad \text{and} \qquad \frac{100d}{\min\{\beta,\theta^2\}} \le n \le e^{\theta^2 d/40},$$

and we can apply Theorem 5.2.1. Therefore, its three conclusions hold simultaneously with failure probability at most

$$6\exp\Big(-\frac{\theta^2 d}{10}\Big) < 1,$$

where the last inequality was by the assumption that $d \ge \frac{100}{\theta^2}$. Below, we verify the distance, error, and pseudorandom conditions of V'.

1. (Distance): By item (1) of Theorem 5.2.1, $s(V) = s(U) + \delta^2 s(G) > s(U) = s(V')$. Note that V' is the projection of V onto the Euclidean ball in $\mathrm{Mat}(d,n)$, so by Lemma 2.3.13, this projection only shrinks the distance to U:

$$\|V' - U\|_F^2 \leq \|V - U\|_F^2 = \delta^2 s(G) \leq (1+\theta)\delta^2 \leq 1.1 \cdot e^{9-3\log_2\beta}\varepsilon,$$

where the last inequality was by the distance bound in Theorem 5.2.1(1) and the choice of parameters $\delta^2 = e^{9-3\log_2\beta}\varepsilon$ and $\theta = e^{-(11-3\log_2\beta)}$. This proves item (3) of Theorem 4.5.1.

2. (Error): We can apply item (2) of Theorem 5.2.1 to show

$$\max\{\|\nabla_V^L\|_{\mathrm{op}}, \|\nabla_V^R\|_{\mathrm{op}}\} \leq \varepsilon(1+\delta^2) + 6\theta\delta^2 \leq \varepsilon\Big(1 + \frac{1}{16} + \frac{6}{e^2}\Big) \leq 2\varepsilon,$$

where the second step was by our choice of parameters $\theta\delta^2 = e^{-(11-3\log_2\beta)}e^{9-3\log_2\beta}\varepsilon = e^{-2}\varepsilon$. Since $s(V) > s(U) = 1$ by the calculation above, the normalization V' is 2ε-doubly balanced. This proves item (1) of Theorem 4.5.1.

3. (Pseudorandom): By item (3) of Theorem 5.2.1, V' is (α,β)-pseudorandom with

$$\alpha \geq \frac{e^{-(4-3\log_2\beta)}\delta^2}{s(V)} \geq \frac{e^5\varepsilon}{1+\delta^2(1+\theta)} \geq 16e(2\varepsilon),$$

where in the second step we used our choice of $\delta^2 = e^{9-3\log_2\beta}\varepsilon$ for the numerator and the size bound $s(V) \leq 1 + \delta^2(1+\theta)$ from item (1) for the denominator, and the final inequality was by $\delta^2(1+\theta) \leq \frac{1}{8}$ by our choice of $\theta = e^{-(11-3\log_2\beta)} \leq 1$ and the assumption $\delta^2 \leq \frac{1}{16}$. This proves item (2) of Theorem 4.5.1.

□

Therefore the goal of this section will be to establish the more general Theorem 5.2.1. We will describe the exact construction of noise G in Section 5.2.1, and then prove the error and pseudorandom properties in Section 5.2.2 and Section 5.2.3 respectively.

5.2.1 Perturbation Process

The work of Section 5.1 intuitively shows that adding random Gaussian noise $V = U + \delta G$ will improve the pseudorandom property of V by $\Omega(\delta^2)$. But if U is nearly doubly balanced, this noise may cause the error of V to blow up. Specifically, consider for any $\xi \in S^{d-1}$:

$$r_\xi(V) - r_\xi(U) = \langle \xi\xi^*, (U+\delta G)(U+\delta G)^*\rangle - \langle \xi\xi^*, UU^*\rangle = 2\delta\langle U\xi, G\xi\rangle + \delta^2\langle \xi\xi^*, GG^*\rangle,$$

so that $\|\nabla_V^L\|_{\text{op}} = \sup_{\xi \in S^{d-1}} |d \cdot r_\xi(V) - s(V)|$ could grow by $\Omega(\delta)$ due to the cross-term $2\delta\langle U\xi, G\xi\rangle$. For small $\delta \ll 1$, this means that adding δG random noise will cause error to grow much faster than pseudorandomness, which only grows at rate $O(\delta^2)$, and so we will not be able to apply Theorem 4.2.14 for fast convergence.

To maintain small error of the perturbation and satisfy the $\alpha \gtrsim \varepsilon$ condition required for our pseudorandom analysis, we add linear constraints to the noise so that the first order term vanishes for both marginals. Explicitly, for fixed input $U \in \text{Mat}(d, n)$, we add two sets of constraints to the noise $G \in \text{Mat}(d, n)$: $UG^* = GU^* = 0$ so that the row error does not blow up; and $\text{diag}(G^*U) = \{\langle Ge_j, Ue_j\rangle\}_{j=1}^n = 0$ so that the column error does not blow up. This is formalized below by choosing the appropriate covariance matrix for our random noise.

Definition 5.2.2 (Perturbation Process)**.** *For input frame $U \in \text{Mat}(d, n)$ we define two subspaces of $\mathbb{R}^d \otimes \mathbb{R}^n$:*

$$\overline{L} := \{\text{vec}(X) \mid UX^* = XU^* = 0\}, \quad \overline{R} := \{\text{vec}(X) \mid \text{diag}(X^*U) = 0\}.$$

Then the orthogonal projections onto these subspaces are denoted $I_{dn} - P_L$ and $I_{dn} - P_R$ respectively. Further, $P_U : \mathbb{R}^d \otimes \mathbb{R}^n \to \mathbb{R}^d \otimes \mathbb{R}^n$ is defined as the orthogonal projection onto the intersection $\overline{L} \cap \overline{R}$. To reduce clutter, we may drop the subscript $P = P_U$ if the input frame U is understood.

For any $\delta \geq 0$, the input is δ-perturbed to $V := U + \delta G$ where $\text{vec}(G) \sim N(0, \frac{1}{dn}P_U)$.

We can show that adding this random noise will increase the error by $O(\delta^2)$ and increase the pseudorandom property by $\Omega(\delta^2)$ with high probability. We need the conditions on d, n, β, θ to apply Gaussian concentration, as we require various random variables to have sufficient degrees of freedom. In the next proposition, we collect properties which will be helpful to analyze the noise, since Definition 5.2.2 does not give an explicit formula for the projection P_U.

Proposition 5.2.3. *For input frame $U \in \text{Mat}(d, n)$ and P_L, P_R, P_U from Definition 5.2.2:*

1. *P_L, P_R can be explicitly defined as*

$$P_L := I_d \otimes U^*(UU^*)^{-1}U, \quad P_R := \sum_{j=1}^n \frac{u_j u_j^*}{\|u_j\|_2^2} \otimes E_{jj},$$

and we can implicitly define P_U such that $\ker(P_U) = Im(P_L) + Im(P_R)$.

2. *The following relations hold for any* $\mathrm{vec}(X) \in \overline{L} \cap \overline{R}$ *and for any* $\xi \in \mathbb{R}^d, j \in [n]$*:*

$$r_\xi(U+X) = \langle \xi\xi^*, (U+X)(U+X)^* \rangle = \langle \xi\xi^*, UU^* + XX^* \rangle = r_\xi(U) + r_\xi(X),$$

$$c_j(U+X) = \|(U+X)e_j\|_2^2 = \|Ue_j\|_2^2 + \|Xe_j\|_2^2 = c_j(U) + c_j(X).$$

As a corollary $s(U+X) = s(U) + s(X)$*.*

3. P_U *satisfies the spectral bounds*

$$\max\{0, I_{dn} - P_L - P_R\} \preceq P_U \preceq I_{dn} - \max\{P_L, P_R\}.$$

Here max *is used to denote that* $P_U \preceq I_{dn} - P_L$ *and* $P_U \preceq I_{dn} - P_R$*.*

4. *If* $\mathrm{vec}(G) \sim N(0, \frac{1}{dn} P_U)$ *then the size* $s(G)$ *is distributed as* $\frac{1}{dn}\chi(k)$ *where* $\chi(k)$ *has* $k = \mathrm{Tr}[P_U] = \mathrm{rk}(P_U)$ *degrees of freedom and* $nd - d^2 - n \leq k \leq \min\{(d-1)n, d(n-d)\}$*.*

Proof. 1. By Definition 5.2.2, we have that for any $\mathrm{vec}(X) \in \overline{R}$, each column of X is orthogonal to the corresponding column of U. This means that the image of P_R is exactly the span of the vectors $\{u_j \otimes e_j\}_{j \in [n]}$. Note that these are all orthogonal (since $\langle e_j, e_{j'} \rangle = 0$ for $j \neq j'$) so we can explicitly define

$$P_R = \sum_{j=1}^{n} \frac{u_j u_j^*}{\|u_j\|_2^2} \otimes E_{jj}.$$

Similarly, Definition 5.2.2 shows that for any $\mathrm{vec}(X) \in \overline{L}$, each row of X is orthogonal to the row span of U. The projection onto the row span of U is given by Eq. (2.3) as $U^*(UU^*)^{-1}U$. Since the constraint on every row is the same, item (1) follows.

2. Assume $\mathrm{vec}(X) \in Im(P_U)$ and calculate

$$(U+X)(U+X)^* = UU^* + UX^* + XU^* + XX^* = UU^* + XX^*,$$

$$\|(U+X)e_j\|_2^2 = \|Ue_j\|_2^2 + 2\langle Ue_j, Xe_j \rangle + \|Xe_j\|_2^2 = \|Ue_j\|_2^2 + \|Xe_j\|_2^2,$$

where the cross terms in both lines vanish by definition of subspace $\mathrm{vec}(X) \in \overline{L}$ and $\mathrm{vec}(X) \in \overline{R}$ respectively. The corollary follows simply as

$$s(U+X) = \mathrm{Tr}[(U+X)(U+X)^*] = \mathrm{Tr}[UU^* + XX^*] = s(U) + s(X),$$

where the first step and the last step were by Definition 4.1.1 of size, and the second step was by the previous calculation to showing the cross terms vanish.

3. In this proof, we will use the various properties of orthogonal projections described in Section 2.1.7. By part (1), $Im(P_U) \subseteq \overline{L}$ so $P_U \preceq I - P_L$. Similarly $Im(P_U) \subseteq \overline{R} \implies P_U \preceq I - P_R$.

   For the lower bound, we have that $P_U \succeq 0$ since it is an orthogonal projection. Then, note that $I_{dn} - P_L - P_R$ has eigenvalue 1 on $\overline{L} \cap \overline{R}$ and eigenvalue 0 or -1 on the complement. This is dominated by P_U, which has eigenvalue 1 on $\overline{L} \cap \overline{R}$ and eigenvalue 0 on the complement.

4. $\text{rk}(P_U) = \text{Tr}[P_U]$ because P_U is an orthogonal projection. Now we perform dimension counting to bound the ranks:

   $$\dim(L) = \text{rk}(P_L) = \text{rk}(I_d)\,\text{rk}(U^*(UU^*)^{-1}U) = d^2,$$

   where we used the property of tensor products in the second step and the fact that U is full rank in the final step; $\dim(R) = \text{rk}(P_R) = n$ since there are n orthogonal terms. Therefore, the bounds follow by taking the trace of the spectral bounds in item (3).

   Also, $s(G) = \frac{1}{dn}\langle \text{vec}(G), P_U \,\text{vec}(G)\rangle$, and the eigenvalues of P_U are in $\{0,1\}$, so the statement follows by Definition 2.5.9 of chi-square variables with $k = \text{rk}(P_U)$.

□

We will use these properties to bound the error of V in Section 5.2.2 and show V is a pseudorandom frame in Section 5.2.3 with high probability.

Remark 5.2.4. *We could have chosen the weaker constraint $UG^* + GU^* = 0$, which would decrease the number of constraints by a constant factor. This would improve some of the results below as the noise would have more degrees of freedom. We chose this model for simplicity, as it is easier (for this author) to reason about row spaces.*

Also, here we focus on the case $\mathbb{F} = \mathbb{R}$. The extension to $\mathbb{C}$ is quite simple: We can replace the real Gaussian distribution with the complex one, and the orthogonality relations are with respect to the complex inner product on the vector space. Again this could improve the number of degrees of freedom by a constant factor.

We sacrifice these small constant factors for the sake of clarity. Of course, our approach gives a real perturbation if the original vector space is real, and the scaling algorithm also remains within this real vector space, which may be desirable for applications.

5.2.2 Error

In this subsection, we bound the error of $V = U + \delta G$ after the perturbation process in Definition 5.3.2. Recall that according to Definition 4.2.3 and Definition 4.2.10, the error can be written as

$$\|\nabla_V^L\|_{\text{op}} = \sup_{\xi \in S^{d-1}} \Big| d \cdot r_\xi(V) - s(V) \Big| \qquad \text{and} \qquad \|\nabla_V^R\|_\infty = \max_{j \in [n]} \Big| n \cdot c_j(V) - s(V) \Big|.$$

We use the orthogonality properties in Proposition 5.2.3 to decompose this error into two terms, one depending on U and the other on the noise G. This allows us to use the fact that U is ε-doubly balanced to bound the error of perturbation V.

Lemma 5.2.5. *For ε-doubly balanced $U \in \operatorname{Mat}(d,n)$ of size $s(U) = 1$, and perturbation $V = U + \delta G$ according to Definition 5.2.2,*

$$\|\nabla_V^L\|_{\text{op}} \leq \varepsilon + \delta^2 \sup_{\xi \in S^{d-1}} \left| \langle \xi\xi^*, dGG^* - s(G) I_d \rangle \right|, \qquad \|\nabla_V^R\|_\infty \leq \varepsilon + \delta^2 \max_{j \in [n]} \Big| n \|Ge_j\|_2^2 - s(G) \Big|.$$

Proof. Item (2) of Proposition 5.2.3 shows

$$r_\xi(V) = r_\xi(U) + \delta^2 r_\xi(G), \quad c_j(V) = c_j(U) + \delta^2 c_j(G), \quad s(V) = s(U) + \delta^2 s(G).$$

So by the the triangle inequality, we can bound

$$\|\nabla_V^L\|_{\text{op}} = \sup_{\xi \in S^{d-1}} \Big| (d \cdot r_\xi(U) - s(U)) + \delta^2 (d \cdot r_\xi(G) - s(G)) \Big| \leq \varepsilon + \delta^2 \|\nabla_G^L\|_{\text{op}},$$

where the first step is by item (2) of Proposition 5.2.3 and the definition of $\|\cdot\|_{\text{op}}$, and in the last step we used that U is ε-Parseval. By the same calculation, we have

$$\|\nabla_V^R\|_\infty = \max_{j \in [n]} \Big| (n \cdot c_j(U) - s(U)) + \delta^2 (n \cdot c_j(G) - s(G)) \Big| \leq \varepsilon + \delta^2 \|\nabla_G^R\|_\infty,$$

where the first step is by item (2) of Proposition 5.2.3 and the definition of $\|\cdot\|_\infty$, and the final inequality is due to U being ε-nearly equal norm. □

To show these errors are small, we first give mean and concentration inequalities for $s(G), r_\xi(G), c_j(G)$. Then we apply a net argument to control the left error of G and a union bound over columns to control the right error of G. We will repeatedly use the spectral bounds of item (3) in Proposition 5.2.3 to control these quantities.

Lemma 5.2.6. *For ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2, the size can be bounded by*

$$1 - \frac{1}{d} - \frac{d}{n} \leq \mathbb{E}[s(G)] \leq 1 - \max\left\{\frac{1}{d}, \frac{d}{n}\right\}.$$

Further, if $d \geq 100, n \geq 100d$, then for any $\theta \leq \frac{1}{4}$ the size concentrates as

$$Pr[|s(G) - \mathbb{E}[s(G)]| \geq 3\theta] \leq 2\exp(-\theta^2 dn).$$

Proof. Item (4) of Proposition 5.2.3 shows that $s(G) = \|G\|_F^2$ is distributed as $\frac{1}{dn}\chi(k)$ where $dn - n - d^2 \leq k \leq dn - \max\{n, d^2\}$. So $\mathbb{E}[s(G)] = \frac{k}{dn}$ and the bounds on the mean follow.

To show concentration, we can use Theorem 2.5.11 to show

$$Pr\Big[|s(G) - \mathbb{E}[s(G)]| \geq 3\theta\Big] \leq 2\exp\left(-\frac{(\frac{6}{5}\theta dn)^2}{k}\right) \leq 2\exp\left(-\theta^2 dn\right),$$

where in the first step we used the assumption $\theta \leq \frac{1}{4}$ to bound $3\theta \geq \frac{6}{5}\frac{5}{2}\theta \geq \frac{6}{5}(2\theta + 2\theta^2)$, and the last step was by $k \geq nd - n - d^2 \geq \frac{8}{9}dn$ by the assumptions $d \geq 100, n \geq 100d$. □

Next, we show that $r_{\xi \in S^{d-1}}$ also concentrates around this value.

Lemma 5.2.7. *For ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2, the row sum for any $\xi \in S^{d-1}$ can be bounded by*

$$1 - \frac{d}{n} - \frac{1+3\varepsilon}{d} \leq \mathbb{E}[d \cdot r_\xi(G)] \leq 1 - \max\left\{\frac{d}{n}, \frac{1-3\varepsilon}{d}\right\}.$$

Further, for any $0 \leq \theta \leq \frac{1}{2}$,

$$Pr\Big[\Big|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\Big| \geq \theta\Big] \leq 2\exp\left(-\frac{\theta^2 n}{8}\right).$$

Proof. For fixed $\xi \in S^{d-1}$, we use Definition 4.2.3 to rewrite $r_\xi(G)$ as the quadratic form of standard Gaussian $g \sim N(0, I_{dn})$:

$$r_\xi(G) = \sum_{j=1}^{n} \langle \xi, Ge_j \rangle^2 = \Big\langle \xi\xi^* \otimes I_n, \mathrm{vec}(G)\,\mathrm{vec}(G)^* \Big\rangle = \frac{1}{dn}\Big\langle \xi\xi^* \otimes I_n, Pgg^*P \Big\rangle, \qquad (5.1)$$

where the last step is by $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P)$. Therefore, the mean can be calculated as

$$\mathbb{E}[d \cdot r_\xi(G)] = \frac{1}{n}\mathbb{E}\Big\langle \xi\xi^* \otimes I_n, Pgg^*P \Big\rangle = \frac{1}{n}\Big\langle \xi\xi^* \otimes I_n, P \Big\rangle,$$

where we used $\mathbb{E}[gg^*] = I_{dn}$ and the fact that $P^2 = P$ for projection P according to Definition 2.1.12. Since we don't have an explicit formula for P, we apply spectral bounds from item (3) of Proposition 5.2.3 to bound

$$\frac{1}{n}\Big\langle \xi\xi^* \otimes I_n, I_{dn} - P_L - P_R \Big\rangle \leq \frac{1}{n}\Big\langle \xi\xi^* \otimes I_n, P \Big\rangle \leq \frac{1}{n}\Big\langle \xi\xi^* \otimes I_n, I_{dn} - \max\{P_L, P_R\} \Big\rangle. \tag{5.2}$$

To control the mean, we bound the inner products with P_L, P_R.

$$\langle \xi\xi^* \otimes I_n, P_L \rangle = \langle \xi\xi^* \otimes I_n, I_d \otimes U^*(UU^*)^{-1}U \rangle = \langle \xi\xi^*, I_d \rangle \cdot \mathrm{Tr}[U^*(UU^*)^{-1}U] = d,$$

where in the first step we substituted the explicit form of P_L given in item (1) of Proposition 5.2.3, and in the final step we used the fact that $\xi \in S^{d-1}$ as well as the cyclic property of trace $\mathrm{Tr}[U^*(UU^*)^{-1}U] = \mathrm{Tr}[(UU^*)(UU^*)^{-1}]$.

We bound the inner product with P_R similarly as

$$\langle \xi\xi^* \otimes I_n, P_R \rangle = \left\langle \xi\xi^* \otimes I_n, \sum_{j=1}^n \frac{u_j u_j^*}{\|u_j\|_2^2} \otimes E_{jj} \right\rangle = \sum_{j=1}^n \frac{\langle \xi, u_j \rangle^2}{\|u_j\|_2^2} \langle I_n, E_{jj} \rangle \in \frac{n}{1 \pm \varepsilon} r_\xi(U),$$

where in the first step we substituted the explicit form of P_R given in item (1) of Proposition 5.2.3, and in the final step we used the fact that U is ε-equal norm so $n\|u_j\|_2^2 \in 1 \pm \varepsilon$ and then substituted r_ξ by Definition 4.2.3. We can now bound the mean using the bounds in Eq. (5.2):

$$1 - \frac{d}{n} - \frac{r_\xi(U)}{1-\varepsilon} \leq \mathbb{E}[d \cdot r_\xi(G)] = \frac{1}{n}\langle \xi\xi^* \otimes I_n, P \rangle \leq 1 - \max\left\{\frac{d}{n}, \frac{r_\xi(U)}{1+\varepsilon}\right\}.$$

Since U is assumed to be ε-Parseval, the statement in the lemma follows by the bounds $d \cdot r_\xi(U) \in 1 \pm \varepsilon$ and the Taylor approximation $\frac{1+x}{1-x} \in 1 \pm 3x$ for $|x| \leq \frac{1}{3}$.

To prove concentration on $r_\xi(G)$, we recall by Eq. (5.1) that we can write

$$d \cdot r_\xi(G) = \frac{1}{n}\Big\langle P(\xi\xi^* \otimes I_n)P, gg^* \Big\rangle,$$

where $g \sim N(0, I_{dn})$. Then, we apply Corollary 2.5.14 to this quadratic form to show

$$Pr\left[\left|d \cdot r_\xi(G) - \mathbb{E}d \cdot r_\xi(G)\right| \geq \theta\right] = Pr\left[\left|\langle P(\xi\xi^* \otimes I_n)P, gg^*\rangle - \mathrm{Tr}[P(\xi\xi^* \otimes I_n)P]\right| \geq \theta n\right]$$
$$\leq 2\exp\left(-\frac{\theta n}{8\|P(\xi\xi^* \otimes I_n)P\|_{\mathrm{op}}} \min\left\{\frac{\theta n}{\mathrm{Tr}[P(\xi\xi^* \otimes I_n)P]}, 1\right\}\right),$$

where in the first step we used $\mathbb{E}gg^* = I_{dn}$ to calculate the mean. To finish the lemma, we can plug in bounds $\mathrm{Tr}[P(\xi\xi^* \otimes I_n)P] \leq n$ using $P^2 = P \preceq I_{dn}$, and $\|P(\xi\xi^* \otimes I_n)P\|_{\mathrm{op}} \leq \|P^2\|_{\mathrm{op}} \leq 1$ to conclude that

$$Pr\left[\left|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\right| \geq \theta\right] \leq \exp\left(-\frac{\theta^2 n}{8}\right).$$

□

At the end of this subsection, we will use the concentration of r_ξ to control the left error of V by a standard net argument. The next lemma show that each column concentrates around a similar value to $s(G), r_\xi$, which we will use to control the right error.

Lemma 5.2.8. *For ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2, the column sum for any $j \in [n]$ can be bounded by*

$$1 - \frac{(1+3\varepsilon)d}{n} - \frac{1}{d} \leq \mathbb{E}[n \cdot c_j(G)] \leq 1 - \max\left\{\frac{(1-3\varepsilon)d}{n}, \frac{1}{d}\right\}.$$

Further for any $\theta \geq 0$,

$$Pr\left[\left|n \cdot c_j(G) - \mathbb{E}[n \cdot c_j(G)]\right| \geq \theta\right] \leq 2\exp\left(-\frac{\min\{\theta^2, \theta\}d}{8}\right).$$

Proof. For fixed $j \in [n]$, we use Definition 4.2.3 to rewrite $n \cdot c_j(G)$ as a quadratic form with standard Gaussian $g \sim N(0, I_{dn})$ as

$$n \cdot c_j(G) = n\left\langle I_d \otimes E_{jj}, \mathrm{vec}(G)\,\mathrm{vec}(G)^*\right\rangle = \frac{1}{d}\left\langle I_d \otimes E_{jj}, Pgg^*P\right\rangle, \tag{5.3}$$

where the last step is by $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P)$. The mean can then be calculated as

$$\mathbb{E}[n \cdot c_j(G)] = \frac{1}{d}\mathbb{E}\left\langle I_d \otimes E_{jj}, Pgg^*P\right\rangle = \frac{1}{d}\left\langle I_d \otimes E_{jj}, P\right\rangle,$$

where we used $\mathbb{E}[gg^*] = I_{dn}$ and the fact that $P^2 = P$ as P is a projection. Since we don't have an explicit formula for P, we apply spectral bounds from item (3) of Proposition 5.2.3 to show

$$\frac{1}{d}\langle I_d \otimes E_{jj}, I_{dn} - P_L - P_R\rangle \le \frac{1}{d}\langle I_d \otimes E_{jj}, P\rangle \le \frac{1}{d}\langle I_d \otimes E_{jj}, I_{dn} - \max\{P_L, P_R\}\rangle. \quad (5.4)$$

So to control the mean, we bound the inner products with P_L, P_R.

$$\begin{aligned}\Big\langle I_d \otimes E_{jj}, P_L\Big\rangle &= \Big\langle I_d \otimes E_{jj}, I_d \otimes U^*(UU^*)^{-1}U\Big\rangle = d\langle E_{jj}, U^*(UU^*)^{-1}U\rangle \\ &= d\langle u_j u_j^*, (UU^*)^{-1}\rangle \in d\left(\frac{d}{1 \pm \varepsilon}\right)\|u_j\|_2^2 \in (1 \pm 3\varepsilon)\frac{d^2}{n},\end{aligned}$$

where in the first step we substituted the explicit form of P_L given in item (1) of Proposition 5.2.3, in the fourth step we used that U is ε-Parseval so $\|dUU^* - I_d\|_{\text{op}} \le \varepsilon$, and in the final step we used that U is ε-equal norm, as well the Taylor approximation $\frac{1+x}{1-x} \in 1 \pm 3x$ for $|x| \le \frac{1}{3}$.

We similarly calculate the inner product with P_R as

$$\langle I_d \otimes E_{jj}, P_R\rangle = \left\langle I_d \otimes E_{jj}, \sum_{j'=1}^{n} \frac{u_{j'}u_{j'}^*}{\|u_{j'}\|_2^2} \otimes E_{j'j'}\right\rangle = \frac{\langle I_d, u_j u_j^*\rangle}{\|u_j\|_2^2} = 1,$$

where in the first step we substituted the explicit form of P_R given in item (1) of Proposition 5.2.3, and in the second step the cross terms vanished by $\langle E_{jj}, E_{j'j'}\rangle = 1[j = j']$. We can now bound the mean using the bounds in Eq. (5.4):

$$1 - \frac{(1+3\varepsilon)d}{n} - \frac{1}{d} \le \mathbb{E}[n \cdot c_j(G)] = \frac{1}{d}\langle I_d \otimes E_{jj}, P\rangle \le 1 - \max\left\{\frac{(1-3\varepsilon)d}{n}, \frac{1}{d}\right\}.$$

To prove concentration, we recall the expression that by Eq. (5.3) we can write

$$n \cdot c_j(G) = \frac{1}{d}\langle P(I_d \otimes E_{jj})P, gg^*\rangle,$$

where $g \sim N(0, I_{dn})$. So we can apply Corollary 2.5.14 to get

$$\begin{aligned}Pr\Big[\Big|n \cdot c_j(G) - \mathbb{E}[n \cdot c_j(G)]\Big| \ge \theta\Big] &= Pr\Big[\Big|\langle P(I_d \otimes E_{jj})P, gg^*\rangle - \text{Tr}[P(I_d \otimes E_{jj})P]\rangle\Big| \ge \theta d\Big] \\ &\le 2\exp\left(-\frac{\theta d}{8\|P(I_d \otimes E_{jj})P\|_{\text{op}}}\min\left\{\frac{\theta d}{\text{Tr}[P(I_d \otimes E_{jj})P]}, 1\right\}\right).\end{aligned}$$

To finish the lemma, we plug in the bounds $\text{Tr}[P(I_d \otimes E_{jj})P] \leq d$ using $P^2 = P \preceq I_{dn}$, and $\|P(I_d \otimes E_{jj})P\|_{\text{op}} \leq \|P^2\|_{\text{op}} \leq 1$ to conclude that

$$Pr\left[\left|n \cdot c_j(G) - \mathbb{E}[n \cdot c_j(G)]\right| \geq \theta\right] \leq 2\exp\left(-\min\{\theta^2, \theta\}, \frac{d}{8}\right).$$

□

The above lemmas have shown that $dr_i(G) \approx s(G)$ and $nc_j(G) \approx s(G)$ for each row and column. To bound the left error of G, we use a net argument over S^{d-1}, and to bound the right error of G we use a simple union bound over columns $j \in [n]$.

Proposition 5.2.9. *For ε-doubly balanced frame $U \in \text{Mat}(d, n)$ and any $\frac{d}{n} + \frac{1}{d} \leq \theta \leq \frac{1}{4}$ such that $d \geq 100$ and $100\frac{d}{\theta^2} \leq n \leq \exp(\frac{\theta^2 d}{100})$, the output $V = U + \delta G$ of the perturbation process in Definition 5.2.2 satisfies the following simultaneously with probability at least $1 - 6\exp(-\frac{\theta^2 d}{2})$:*

1. $|s(G) - \mathbb{E}[s(G)]| \leq \frac{\theta}{2}$;
2. $\|\nabla_G^L\|_{\text{op}} \leq \varepsilon + 6\theta$;
3. $\|\nabla_G^R\|_\infty \leq \varepsilon + 5\theta$.

Proof. Item (1) follows exactly from Lemma 5.2.6 with failure probability at most $2\exp(-\theta^2 dn/9) \leq 2\exp(-\theta^2 d/2)$ as $n \geq 100$ by assumption.

To show the remaining items, we use a net argument and union bound to control the errors of G, and then apply Lemma 5.2.5 to control the error of $V = U + \delta G$.

To show (2), we bound the left error of G by

$$\|\nabla_G^L\|_{\text{op}} \leq \sup_{\xi \in S^{d-1}} \left|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\right| + \left|\mathbb{E}[d \cdot r_\xi(G)] - \mathbb{E}[s(G)]\right| + \left|s(G) - \mathbb{E}[s(G)]\right|, \tag{5.5}$$

using the triangle inequality. We first apply the bounds on expectations given in Lemma 5.2.6 and Lemma 5.2.7 to show:

$$\left|\mathbb{E}[s(G)] - \mathbb{E}[d \cdot r_\xi(G)]\right| \leq 3\varepsilon \max\left\{\frac{d}{n}, \frac{1}{d}\right\} + \min\left\{\frac{d}{n}, \frac{1}{d}\right\} \leq \frac{d}{n} + \frac{1}{d},$$

where we used the assumption $\varepsilon \leq \frac{1}{3}$ in the final inequality.

Then we can use the concentration bound in Lemma 5.2.7 to show for any fixed $\xi \in S^{d-1}$

$$Pr\Big[\Big|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\Big| \geq 3\theta\Big] \leq \exp(-\theta^2 n).$$

To show the final bound on ∇_G^L, we use a standard net argument. So let $N \subseteq S^{d-1}$ be an $\eta = \frac{1}{9}$-net according to Definition 2.6.2. By Fact 2.6.3 we can assume $|N| \leq (1 + 2\eta^{-1})^d = 19^d \leq e^{3d}$. So we can bound

$$Pr\Big[\sup_{\xi \in N}\Big|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\Big| \geq 3\theta\Big] \leq 2\exp(3d - \theta^2 n) \leq 2\exp\Big(-\frac{\theta^2 n}{2}\Big),$$

where we used the union bound over N and Lemma 5.2.7, and the last step was by the assumption $n \geq 100\frac{d}{\theta^2}$.

Then we can apply Lemma 2.6.5 for quadratic forms on the matrix $\langle \xi\xi^*, dGG^* - \mathbb{E}[dGG^*]\rangle$ to show

$$\sup_{\xi \in S^{d-1}}\Big|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\Big| \leq (1 - 2\eta - \eta^2)^{-1}\sup_{\xi \in N}\Big|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\Big| \leq \frac{3}{2} \cdot 3\theta. \tag{5.6}$$

where the last step was by our choice of $\eta = \frac{1}{9}$.

The concentration of s and r_ξ all hold simultaneously with failure probability at most $2\exp(-\theta^2 dn/9) + 2\exp(4d - \theta^2 n/10) \leq 4\exp(-\theta^2 n/20)$. So we can bound the error

$$\|\nabla_G^L\|_{\text{op}} \leq \frac{9}{2}\theta + \left(\frac{d}{n} + \frac{1}{d}\right) + \theta \leq 6\theta,$$

where the first step was by bounding each of the three terms in the decomposition in Eq. (5.5), and the last inequality was by the condition $d \geq \frac{100}{\theta^2}, n \geq 100\frac{d}{\theta^2}$ and $\theta \leq \frac{1}{4}$.

Similarly, we can bound the column error

$$\|\nabla_G^R\|_\infty \leq \max_{j \in [n]}\Big|n \cdot c_j(G) - \mathbb{E}[n \cdot c_j(G)]\Big| + \Big|\mathbb{E}[n \cdot c_j(G)] - \mathbb{E}[s(G)]\Big| + \Big|s(G) - \mathbb{E}[s(G)]\Big|. \tag{5.7}$$

Applying the union bound over all columns gives

$$Pr\Big[\max_{j \in [n]}\Big|n \cdot c_j(G) - \mathbb{E}[n \cdot c_j(G)]\Big| \geq 3\theta\Big] \leq 2n\exp(-\theta^2 d) \leq \exp\left(-\frac{\theta^2 d}{2}\right),$$

where the first step follows from the bound in Lemma 5.2.8 on each individual column, and the last step was by our assumption $n \leq \exp(\frac{\theta^2 d}{100})$. Note that this is a crucial assumption to bound the column error.

Therefore, we can collect terms to show

$$\|\nabla_G^R\|_\infty \leq \left(\frac{d}{n} + \frac{1}{d}\right) + 3\theta + \theta \leq 5\theta,$$

where the first step was by bounding each of the three terms in the decomposition in Eq. (5.7), and the last inequality was by the conditions $d \geq \frac{100}{\theta^2}, n \geq 100\frac{d}{\theta^2}$ with $\theta \leq \frac{1}{4}$.

Finally, when these error bounds of ∇_G hold, we can apply Lemma 5.2.5 to show

$$\|\nabla_V^L\|_{\text{op}} \leq \varepsilon + \delta^2\|\nabla_G^L\|_{\text{op}} \leq \varepsilon + 6\theta\delta^2, \qquad \|\nabla_V^R\|_\infty \leq \varepsilon + \delta^2\|\nabla_G^R\|_\infty \leq \varepsilon + 5\theta\delta^2$$

□

Note that the only place we used the upper bound $n \leq \exp(\frac{\theta^2 d}{100})$ was to bound the column error of G in Proposition 5.2.9. In the next section, we will deal with the case of larger n by right-normalizing our perturbation.

5.2.3 Pseudorandom Condition

In this subsection we will show that the perturbed frame $V = U + \delta G$ in Definition 5.2.2 is pseudorandom with high probability. The argument will be somewhat similar to Section 5.1, but more complicated due to dependencies in the entries of noise G. We will show that the noise G has $\Omega(\delta^2)$ contribution in each term of Lemma 5.1.1.

Recall that Lemma 5.1.1 gives a spectral characterization of the frame pseudorandom condition which requires a lower bound on

$$\inf_{\xi \in S^{d-1}} \sum_{j \in T} |\langle \xi, v_j \rangle|^2 = \inf_{\xi \in S^{d-1}} \|\xi^* V P_T\|_2^2$$

for every $T \in \binom{[n]}{\beta n}$ where we use $P_T = \text{diag}(1_T) \in \text{Mat}(n)$ to denote the orthogonal coordinate projection onto T.

Substituting in $V = U + \delta G$ according to Definition 5.2.2 of the perturbation process, we can rewrite $\|\xi^* V P_T\|_2 = \|\xi^* U P_T - \delta \xi^* G P_T\|_2$. Now if we had orthogonality conditions for these two terms (like in Proposition 5.2.3(2) for rows and columns), then we could separate into two terms, one depending only on U and other depending only on G, and focus on lower bounding the random part. The first issue with this strategy is that the U term is totally arbitrary, so it does not necessarily help with the pseudorandom condition, unlike in Section 5.3.2 where we were able to use the ε-doubly balanced condition of U to bound the error of V. The second issue is that, even though we only require a lower bound, the U and G parts are not necessarily orthogonal, so we cannot ignore the U term.

What we can do is show that the random matrix G part is mostly uncorrelated with the deterministic matrix U. Explicitly, we will show that for every ξ, T, there is a large component of $\xi^* G P_T$ that is orthogonal to $\xi^* U P_T$. For this purpose, we define the following projections.

Definition 5.2.10. *For ε-doubly balanced frame $U \in \text{Mat}(d,n)$, let $P_T := \text{diag}(1_T) \in \text{Mat}(n)$ be the orthogonal projection onto column subset $T \subseteq [n]$. Let $P_{U,T} \in \text{Mat}(n)$ be the orthogonal projection onto the row space of UP_T, which according to Eq. (2.3), can be written as*

$$P_{U,T} = (UP_T)^*(UP_T P_T U^*)^{-1}(UP_T) = (UP_T)^* \left(\sum_{j \in T} u_j u_j^* \right)^{-1} (UP_T).$$

Note that we have defined $P_{U,T}$ such that its image contains the entire row space of UP_T (i.e. $\xi^* U P_T$ for any $\xi \in \mathbb{R}^d$). So our plan is to show $\xi^* G P_T$ has a large component in the kernel of P_{UT}. This may seem like overkill, as we really only need orthogonality for each fixed pair $\xi^* U P_T$ and $\xi^* G P_T$. But this property is less robust for changes in ξ, so it will be difficult to perform the approximation portion of our net argument below with only pairwise orthogonality.

Therefore, we can use these projections to isolate the random contribution in order to lower bound $\|\xi^* V P_T\|_F$. We will use the fact that $Im(P_{U,T}) \subseteq Im(P_T)$, and in particular that these projections commute.

Lemma 5.2.11. *For ε-doubly balanced frame $U \in \text{Mat}(d,n)$ and perturbation $V = U + \delta G$ according to Definition 5.2.2,*

$$\|\xi^* V P_T\|_F^2 \geq \delta^2 \|\xi^* G (P_T - P_{U,T})\|_2^2$$

for any $\xi \in S^{d-1}$ and $T \in \binom{[n]}{\beta n}$ where $P_T, P_{U,T}$ are given according to Definition 5.2.10.

Proof. Note that $Im(P_{UT}) \subseteq Im(P_T)$ and therefore $P_T = (P_T - P_{UT}) + P_{UT}$ gives a decomposition of $Im(P_T)$ into orthogonal subspaces. This will separate the components of random vector $\xi^* G P_T$ into two parts, depending on whether or not they are orthogonal to the deterministic part $\xi^* U P_T$. So for any $\xi \in S^{d-1}$ we have

$$\begin{aligned}\|\xi^* V P_T\|_2^2 &= \|\xi^*(UP_T + \delta G P_{UT}) + \delta \xi^* G(P_T - P_{UT})\|_2^2 \\ &= \|\xi^* U P_T + \delta \xi^* G P_{UT}\|_2^2 + \delta^2 \|\xi^* G(P_T - P_{UT})\|_2^2,\end{aligned}$$

where the first step was by the decomposition $P_T = (P_T - P_{UT}) + P_{UT}$, and the last step was by the fact that $\xi^* G P_{UT}$ and $\xi^* U P_T$ are both always contained in $Im(P_{UT})$, whereas $\xi^* G(P_T - P_{UT})$ is in the orthogonal component $\ker(P_{UT})$. Both terms are non-negative, so lemma follows by ignoring the first. □

This lemma allows us to show that V is pseudorandom by lower bounding the smallest singular value of the set of random matrices $G(P_T - P_{U,T})$ for $T \in \binom{[n]}{\beta n}$. There are two problems with using standard two-sided concentration results on Gaussian random matrices: first our Gaussian noise has complicated dependencies; second, similar to the issue in Section 5.1, standard Gaussian concentration can give an upper bound of at most $\exp(-\beta n/4)$ on the failure probability, which is slightly too weak for a union bound over $\binom{n}{\beta n} \approx 2^{\beta n(1-\log_2 \beta)}$ many subsets. So we will rely on Lemma 2.5.15 which gives stronger probability bounds for the lower tail.

We first show mean and concentration bounds for an individual $\xi \in S^{d-1}$.

Lemma 5.2.12. *For any fixed $\xi \in S^{d-1}, T \in \binom{[n]}{\beta n}$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn} P_U)$ according to Definition 5.2.2:*

$$\frac{1}{d}\left(\beta - \frac{d}{n} - \frac{(1+\varepsilon)^2}{d}\right) \le \mathbb{E}\|\xi^* G(P_T - P_{U,T})\|_2^2 \le \frac{\beta}{d}.$$

Further, if $n \ge 100\frac{d}{\beta}$ and $d \ge \frac{100}{\beta}$, then for any $c \ge 5$

$$Pr\left[\|\xi^* G(P_T - P_{U,T})\|_2^2 \ge (1+c)\frac{\beta}{d}\right] \le \exp\left(-\frac{c\beta n}{9}\right) \quad \textit{and}$$
$$Pr\left[\|\xi^* G(P_T - P_{U,T})\|_2^2 \le 0.95 e^{-c}\frac{\beta}{d}\right] \le \exp\left(-\frac{2}{5}c(0.95\beta n)\right)$$

give high probability bounds for the upper and lower tail, respectively.

Proof. We first rewrite the term as a quadratic form with standard Gaussian $g \sim N(0, I_{dn})$:

$$\left\langle \xi\xi^*, G(P_T - P_{UT})^2 G^* \right\rangle = \left\langle \xi\xi^* \otimes (P_T - P_{UT})^2, \mathrm{vec}(G)\,\mathrm{vec}(G)^* \right\rangle = \frac{1}{dn}\left\langle \xi\xi^* \otimes (P_T - P_{UT}), P g g^* P \right\rangle,$$

where the first step is a straightforward calculation on tensor products, and in the last step we used that $\mathrm{vec}(G) \sim N(0, \frac{1}{dn} P_U)$ along with the fact that $(P_T - P_{UT})^2 = (P_T - P_{UT})$ as it is an orthogonal projection. Now we can calculate the mean:

$$\mathbb{E}\|\xi^* G(P_T - P_{UT})\|_2^2 = \frac{1}{dn}\mathbb{E}\Big\langle \xi\xi^* \otimes (P_T - P_{UT}), Pgg^*P \Big\rangle = \frac{1}{dn}\langle \xi\xi^* \otimes (P_T - P_{UT}), P\rangle,$$

where the last step was by $P^2 = P$ as it is an orthogonal projection. Since we don't have an explicit formula for P, we bound this expectation using the spectral bounds from item (3) Proposition 5.2.3 as

$$\begin{aligned}\langle \xi\xi^* \otimes (P_T - P_{UT}), I_{dn} - P_L - P_R\rangle &\leq \langle \xi\xi^* \otimes (P_T - P_{UT}), P\rangle \\ &\leq \langle \xi\xi^* \otimes (P_T - P_{UT}), I_{dn} - \max\{P_L, P_R\}\rangle. \quad (5.8)\end{aligned}$$

So we calculate inner products with I_{dn}, P_L, P_R to bound the mean. First,

$$\langle \xi\xi^* \otimes (P_T - P_{UT}), I_{dn}\rangle = \langle \xi\xi^*, I_d\rangle \operatorname{Tr}[P_T - P_{UT}] = \mathrm{rk}(P_T - P_{UT}),$$

where the last step was because $P_T - P_{UT}$ is an orthogonal projection. By comparing dimensions, we get

$$|T| - d \leq \mathrm{rk}(P_T) - \mathrm{rk}(P_{UT}) \leq \langle \xi\xi^* \otimes (P_T - P_{UT}), I_{dn}\rangle \leq \mathrm{rk}(P_T) = |T|,$$

where the lower bound was because $\mathrm{rk}(P_{UT}) \leq \mathrm{rk}(U) = d$. Similarly we can bound the inner product with P_L as

$$\begin{aligned}\langle \xi\xi^* \otimes (P_T - P_{UT}), P_L\rangle &= \langle \xi\xi^* \otimes (P_T - P_{UT}), I_d \otimes U^*(UU^*)^{-1}U\rangle \\ &= \langle \xi\xi^*, I_d\rangle\langle P_T - P_{UT}, U^*(UU^*)^{-1}U\rangle = \langle P_T - P_{UT}, U^*(UU^*)^{-1}UP_T\rangle = 0, \quad (5.9)\end{aligned}$$

where in the first step we substituted the explicit form for P_L given in item (1) of Proposition 5.2.3, in the third step we used $P_T - P_{UT} = P_T(P_T - P_{UT})$ since all of these are commuting projections, and the last step was because $P_T - P_{UT}$ is the projection to the orthogonal subspace of the row span of UP_T, so $UP_T \in \ker(P_T - P_{UT})$. We also calculate the inner product with P_R as

$$\langle \xi\xi^* \otimes (P_T - P_{UT}), P_R\rangle = \left\langle \xi\xi^* \otimes (P_T - P_{UT}), \sum_{j=1}^n \frac{u_j u_j^*}{\|u_j\|_2^2} \otimes E_{jj} \right\rangle \in \sum_{j=1}^n \frac{n\langle \xi, u_j\rangle^2}{1 \pm \varepsilon}\langle P_T - P_{UT}, E_{jj}\rangle,$$

where the last step was because U is ε-equal norm. Since $P_T - P_{UT}$ is an orthogonal projection, we can bound the inner product $0 \leq \langle P_T - P_{UT}, E_{jj} \rangle \leq 1$. So in total, we can bound the P_R term

$$0 \leq \langle \xi\xi^* \otimes (P_T - P_{UT}), P_R \rangle \leq \frac{n}{1-\varepsilon} r_\xi(U) \leq (1+3\varepsilon)\frac{n}{d},$$

where in the last step we used that U is ε-Parseval and the Taylor approximation $\frac{1+x}{1-x} \in 1 \pm 3x$ for $|x| \leq \frac{1}{3}$.

Plugging the above calculations into Eq. (5.8) gives upper and lower bounds

$$\mathbb{E}\|\xi^* G(P_T - P_{UT})\|_2^2 \leq \frac{1}{dn}\langle \xi\xi^* \otimes (P_T - P_{UT}), I_{dn} \rangle \leq \frac{\beta}{d} \qquad \text{and} \tag{5.10}$$

$$\begin{aligned} \mathbb{E}\|\xi^* G(P_T - P_{U,T})\|_2^2 &\geq \frac{1}{dn}\langle \xi\xi^* \otimes (P_T - P_{UT}), I_{dn} - P_L - P_R \rangle \\ &\geq \frac{|T| - d}{dn} - 0 - \frac{1+3\varepsilon}{dn}\frac{n}{d} = \frac{1}{d}\left(\beta - \frac{d}{n} - \frac{1+3\varepsilon}{d}\right) \geq \frac{0.95\beta}{d}, \end{aligned} \tag{5.11}$$

where the first step was by Eq. (5.8), in the second step we plugged in the bounds for I_{dn}, P_L, and P_R, in the third step we used $|T| = \beta n$, and the last step was by our assumptions $d \geq \frac{100}{\beta}, n \geq 100\frac{d}{\beta}$ so that $\frac{d}{n} \leq \frac{\beta}{100}$ and $\frac{1}{d} \leq \frac{\beta}{100}$.

To show concentration for the upper bound, we can apply Corollary 2.5.14 to the quadratic form $\langle P(\xi\xi^* \otimes (P_T - P_{UT}))P, gg^* \rangle$ along with the bounds

$$\|P(\xi\xi^* \otimes (P_T - P_{UT}))P\|_{\text{op}} \leq \|P^2\|_{\text{op}} \leq 1,$$

since $P_T - P_{UT} \preceq I_n$ as it is an orthogonal projection, and $\text{Tr}[P(\xi\xi^* \otimes (P_T - P_{UT}))P] \leq \langle I_{dn}, \xi\xi^* \otimes (P_T - P_{UT}) \rangle \leq \beta n$ from Eq. (5.10) to show, for $c \geq 1$:

$$\begin{aligned} Pr\Big[\|\xi^* G(P_T - P_{UT})\|_2^2 \geq (1+c)\frac{\beta}{d}\Big] &= Pr\Big[\langle P(\xi\xi^* \otimes (P_T - P_{UT}))P, gg^* \rangle \geq (1+c)\beta n\Big] \\ &\leq \exp\Big(-\min\Big\{\frac{(c\beta n)^2}{8(\beta n)(1)}, \frac{c\beta n}{8(1)}\Big\}\Big) = \exp\Big(-\frac{\beta n}{8}\Big). \end{aligned}$$

Similarly, to show a high probability bound for the lower tails, we apply Lemma 2.5.15 with $c \geq 5$ to show

$$Pr\Big[\|\xi^* G(P_T - P_{UT})\|_2^2 \leq e^{-c} \cdot 0.95\frac{\beta}{d}\Big] \leq \exp\left(-\frac{2}{5}c(0.95\beta n)\right),$$

where we used the lower bound on the mean given in Eq. (5.11). □

We emphasize that this lemma allows us to tune the failure probability to be arbitrarily high at the cost of worse upper and lower bounds. Note that the cost of the lower tail is especially high, as the bound grows exponentially with respect to c.

To show pseudorandomness of V, we continue with a standard net argument to give high probability lower tail bounds for the random matrix in Lemma 5.2.11.

Lemma 5.2.13. *For ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$, consider the perturbation $V := U + \delta G$ with $\delta \leq \frac{1}{4}$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2. If $\beta \leq \frac{1}{2}$, $d \geq \frac{100}{\beta}$ and $n \geq \frac{100d}{\beta}$, then for any $T \in \binom{[n]}{\beta n}, c \geq 5$*

$$Pr\left[\inf_{\xi \in S^{d-1}} \|\xi^* G(P_T - P_{U,T})\|_2^2 \leq e^{-(c+1)}\frac{\beta}{d}\right] \leq 2\exp\left(-\frac{c\beta n}{3}\right).$$

Proof. Our plan is to bound each direction $\xi \in N_L \subseteq S^{d-1}$ for an appropriate net $N_L \subseteq S^{d-1}$, and then use Lemma 2.6.6 to bound the infimum over S^{d-1}. Intuitively, the random matrix $G(P_T - P_{UT})$ will be well-conditioned with high probability, so we can decrease the cardinality of the net N_L by first proving high probability upper bounds according to Lemma 2.6.5.

So we choose $\eta_U = \frac{1}{9}$ and let $N_U \subseteq S^{d-1}$ be an η_U-net. By Fact 2.6.3 we have

$$|N_U| \leq (1 + 2\eta_U^{-1})^d = 19^d \leq e^{3d}.$$

For $c \geq 5$, this gives the upper bound

$$Pr\Big[\sup_{\xi \in N_U} \|\xi^* G(P_T - P_{UT})\|_2 \geq \sqrt{(1+3c)\beta/d}\Big] \leq \exp\left(3d - \frac{3}{8}c\beta n\right) \leq \exp\Big(-\frac{c\beta n}{3}\Big),$$

where the first step was by the union bound over N_U as well as the concentration shown in Lemma 5.2.12, and the final step was by our assumption $n \geq 100\frac{d}{\beta}$.

Assuming that these bad events do not occur, we can bound every $\xi \in S^{d-1}$ by the multiplicative upper bound in Lemma 2.6.5:

$$\sup_{\xi \in S^{d-1}} \|\xi^* G(P_T - P_{UT})\|_2 \leq \frac{9}{8}\sup_{\xi \in N_U} \|\xi^* G(P_T - P_{UT})\|_2 \leq \frac{9}{8}\sqrt{(1+3c)\frac{\beta}{d}}.$$

Now let $N_L \subseteq S^{d-1}$ be an $\eta_L = \frac{1}{3}\sqrt{\frac{0.95e^{-c}}{(\frac{9}{8})^2(1+3c)}}$. By Lemma 2.6.6, we can bound the cardinality of the net

$$|N_L| \leq (1 + 2\eta_L^{-1})^d \leq \left(1 + 2 \cdot 3 \cdot \frac{9}{8}\sqrt{\frac{e^c(1+3c)}{0.95}}\right)^d \leq e^{2cd},$$

where the last inequality was due to the assumption $c \geq 5$. This allows us to simultaneously lower bound every $\xi \in N_L$ by

$$Pr\Big[\inf_{\xi \in N_L} \|\xi^* G(P_T - P_{UT})\|_2 \leq \sqrt{0.95e^{-c}\beta/d}\Big] \leq \exp\Big(2cd - \frac{2}{5}c\beta n\Big) \leq \exp\Big(-\frac{c\beta n}{3}\Big),$$

where the first step was by the union bound, the cardinality bound on N_L, and the concentration shown in Lemma 5.2.12, and the last step was by the assumption $n \geq 100\frac{d}{\beta}$. Now assume both events occur:

$$\inf_{\xi \in N_L} \|\xi^* G(P_T - P_{UT})\|_2 \geq \sqrt{0.95e^{-c}\beta/d} \quad \text{and} \quad \sup_{\xi \in S^{d-1}} \|\xi^* G(P_T - P_{UT})\|_2 \leq \frac{9}{8}\sqrt{(1+3c)\frac{\beta}{d}},$$

which by the union bound occurs with probability at least $1 - 2\exp(-c\beta n/3)$. Then we can simultaneously lower bound all $\xi \in S^{d-1}$:

$$\begin{aligned}\inf_{\xi \in S^{d-1}} \|\xi^* G(P_T - P_{UT})\|_2 &\geq \inf_{\xi \in N_L} \|\xi^* G(P_T - P_{UT})\|_2 - \eta_L \sup_{\xi \in S^{d-1}} \|\xi^* G(P_T - P_{UT})\|_2 \\ &\geq \sqrt{0.95e^{-c}\frac{\beta}{d}} - \frac{1}{3}\sqrt{\frac{0.95e^{-c}}{(\frac{9}{8})^2(1+3c)}}\sqrt{(\frac{9}{8})^2(1+3c)\frac{\beta}{d}} \geq \frac{2}{3}\sqrt{0.95e^{-c}\frac{\beta}{d}},\end{aligned}$$

where the first step was by Lemma 2.6.6, the second step was by choice of η_L and the upper bound assumption. The result follows by squaring both sides and noting $0.95\frac{4}{9} \geq e^{-1}$. □

We can finally show pseudorandomness of the perturbation $V = U + \delta G$ by a union bound over sets.

Proposition 5.2.14. *For ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$ let $V := U + \delta G$ be the perturbation according to Definition 5.2.2. If $\delta \leq \frac{1}{4}, \beta \leq \frac{1}{2}, d \geq \frac{100}{\beta}, n \geq \frac{100d}{\beta}$, then V is $(\exp(-(4 - 3\log_2 \beta))\delta^2, \beta)$-pseudorandom with probability at least $1 - 2\exp(-\beta n/5)$.*

Proof. We apply Lemma 5.2.13 with $c = 3(1 - \log_2 \beta)$ simultaneously to every $T \in \binom{[n]}{\beta n}$. With failure probability at most $2^{\beta n(1-\log_2 \beta)}\exp(-c\beta n/3) \leq \exp(-c\beta n/5)$, this gives the lower bound

$$\min_{T \in \binom{[n]}{\beta n}} \inf_{\xi \in S^{d-1}} \|\xi^* V P_T\|_2^2 \geq \min_{T \in \binom{[n]}{\beta n}} \inf_{\xi \in S^{d-1}} \delta^2 \|\xi^* G(P_T - P_{UT})\|_2^2 \geq \delta^2 e^{-(c+1)}\frac{\beta}{d} \geq e^{-(4-3\log_2 \beta)}\delta^2\frac{\beta}{d},$$

where the first step is by Lemma 5.2.11, the second is the conclusion of Lemma 5.2.13, and the final step is by our choice of c. This is exactly the sufficient condition for pseudorandomness given by Lemma 5.1.1. □

5.2.4 Putting it Together

Proof of Theorem 5.2.1. Let V be the output of Definition 5.2.2. Then the first part of item (1) follows as $V - U = \delta G$ so $\|V - U\|_F^2 = \delta^2 \|G\|_F^2 = \delta^2 s(G)$. The second follows by the orthogonality condition in Proposition 5.2.3(2). To show the bound on $s(G)$, we use the mean and concentration bounds to show

$$|s(G) - 1| \leq |s(G) - \mathbb{E}[s(G)]| + |\mathbb{E}[s(G)] - 1| \leq \frac{\theta}{2} + \frac{1}{d} + \frac{d}{n} \leq \theta,$$

where we used the bound on the mean given in Lemma 5.2.6 and the concentration given in Proposition 5.2.9(1) in the first step, and the final inequality follows from our assumptions $d \geq \frac{100}{\theta^2}$ and $n\frac{100d}{\theta^2}$.

Item (2) is exactly Proposition 5.2.9(2) and (3).

Item (3) is exactly Proposition 5.3.17.

So by the union bound these occur simultaneously with probability at least

$$1 - 2\exp\Big(-\frac{\theta^2 n}{10}\Big) - 2\exp\Big(-\frac{\theta^2 d}{10}\Big) - 2\exp\Big(-\frac{\beta n}{10}\Big) 1 - 6\exp\Big(-\frac{\theta^2 d}{10}\Big) > 0,$$

where the first inequality is by our assumptions $n \geq \beta n \geq 100d$, and the last inequality is by our assumptions $d \geq \frac{100}{\theta^2}$. □

5.3 Perturbation Argument for Large n

The goal of this section is to prove Theorem 4.5.2. We will in fact use a modified perturbation process for the smoothed analysis argument to prove the following generalization.

Theorem 5.3.1. *Let $U \in \mathrm{Mat}(d,n)$ be an equal-norm frame with size $s(U) = 1$ that is ε-Parseval for $\varepsilon \leq \frac{1}{10}$. If $\delta \leq \frac{1}{4}, \beta \leq \frac{1}{2}$, then for any choice of $\frac{1}{d} + \frac{d}{n} \leq \theta \leq \frac{1}{4}, C \geq 10$ such that $d \geq \frac{100}{\beta}, n \geq \frac{100d}{\min\{\beta,\theta^2\}}$, and $\beta \geq 20Ce^{-d/9}$, the output V' of the perturbation process given in Definition 5.3.2 satisfies the following properties simultaneously:*

1. *(Distance): $\|V' - U\|_F^2 \leq (1+\theta)\delta^2$;*
2. *(Error): $\|\nabla_{V'}^L\|_{\mathrm{op}} \leq (1+\delta^2)\varepsilon + \delta^2(8\theta + 2\sqrt{\delta^2 \cdot (1+3\theta)7Cd} + \delta^2 \cdot 21Cd + \delta^4 \cdot 20Cd)$ and $\nabla_{V'}^R = 0$;*

3. *(Pseudorandom): V' is an $(e^{-(6-3\log_2\beta)}\delta^2, \beta)$-pseudorandom frame;*

with probability at least $1 - \frac{4}{C} - 4\exp(-\theta^2 n/10) - 4\exp(-\beta n/10)$.

Before we give the formal details of the smoothed analysis argument, let us see how this implies the existence of the perturbation given in Theorem 4.5.2 in the $\mathbb{F} = \mathbb{R}$ case. This can be simply lifted to $\mathbb{C}$ according to Remark 5.2.4.

Proof of Theorem 4.5.2. Our input U is ε-doubly balanced, so we first column-normalize $u'_j := \frac{u_j}{\sqrt{n}\|u_j\|_2}$ so that we can apply Theorem 5.3.1. This produces U' that has size $s(U') = 1$ and $\nabla^R_{U'} = 0$ by construction. Further by Fact 4.1.5, U' is 3ε-doubly balanced and

$$\|U' - U\|_F^2 \leq \varepsilon^2.$$

Now we can apply Theorem 5.3.1 with the appropriate choice of parameters to get output V' which will satisfy the three properties with non-zero probability. To this end, let $\delta^2 = e^{11-3\log_2\beta}\varepsilon$ and $\theta = e^{-(14-3\log_2\beta)}$ and $C = 10$. Clearly $\theta \leq \frac{1}{2}$, and we can choose C' large enough in Theorem 4.5.2 such that

$$d \geq \frac{100}{\min\{\beta, \theta^2\}}, \qquad \text{and} \qquad n \geq 100\frac{d}{\min\{\beta, \theta^2\}},$$

so that $\beta \geq 10Ce^{-d/9}$. Therefore, the three conclusions Theorem 5.3.1 hold simultaneously with failure probability at most

$$\frac{4}{C} + 4\exp(-\beta n/10) + 4\exp(-\theta^2 n/10) \leq \frac{8}{10} < 1,$$

where the first inequality was by the assumption that $n \geq 100\frac{d}{\theta^2}$. Below we verify the distance, error, and pseudorandom conditions of V'.

1. (Distance): We calculate

$$\|V' - U\|_F^2 \leq (\|V' - U'\|_F + \|U' - U\|_F)^2 \leq (\sqrt{(1+\theta)\delta^2} + \sqrt{\varepsilon^2})^2 \leq e^{12-3\log_2\beta}\varepsilon,$$

   where the first step was by triangle inequality, in the second step we bounded the first term by item (1) of Theorem 5.3.1 and the second term by the calculation above using Fact 4.1.5, and the third step was by our choice of $\delta^2 = e^{11-3\log_2\beta}\varepsilon$ and $\theta = e^{-(14-3\log_2\beta)}$ and the assumption $\varepsilon \leq \frac{1}{e^{40-9\log_2\beta}d}$.

2. (Error): $\nabla^R_{V'} = 0$ by construction, so we can bound the left error by item (2) of Theorem 5.3.1 to show

$$\begin{aligned}
\|\nabla^L_{V'}\|_{\mathrm{op}} &\leq 3\varepsilon + \delta^2(3\varepsilon + 8\theta) + 2\delta^3\sqrt{(1+3\theta)7Cd} + \delta^4(21Cd) + \delta^6(20Cd) \\
&\leq \Big(3\varepsilon + e^{11-3\log_2\beta}\varepsilon(3\varepsilon + 8e^{-(14-3\log_2\beta)}) \\
&\quad + 2e^{16.5-4.5\log_2\beta}\varepsilon\sqrt{(1+3e^{-(14-3\log_2\beta)})\varepsilon\cdot 70d} \\
&\quad + e^{22-6\log_2\beta}\varepsilon^2(210d) + e^{33-9\log_2\beta}\varepsilon^3(200d)\Big) \\
&\leq \varepsilon\Big(3 + 3e^{-10} + 8e^{-3} + e^{-2} + e^{-10} + e^{-40}d^{-1}\Big) \leq 4\varepsilon,
\end{aligned}$$

where we used Fact 4.1.5 to bound $\|\nabla^L_{U'}\|_{\mathrm{op}} \leq 3\varepsilon$, in the second step we substituted our choice of $C = 10, \delta^2 = e^{11-3\log_2\beta}\varepsilon$, and $\theta = e^{-(14-3\log_2\beta)}$, and in the final step we used our the assumption that $\varepsilon \leq \frac{1}{e^{40-9\log_2\beta}d}$. Note that the higher order terms are the only place we use the assumption $\varepsilon \lesssim \frac{1}{d}$, and we believe this part of the analysis can be improved. This will be discussed in more detail in Section 5.3.2.

3. (Pseudorandom): By item (3) of Theorem 5.2.1, V' is (α, β)-pseudorandom with

$$\alpha \geq e^{-(6-3\log_2\beta)}\delta^2 = e^5\varepsilon \geq 16e(4\varepsilon),$$

where we plugged $\delta^2 = e^{11-3\log_2\beta}\varepsilon$ in the second step.

□

The formal description of the perturbation will be given in Section 5.3.1, and the proof of the error and pseudorandom properties will be given in Section 5.3.2 and Section 5.3.3 respectively.

5.3.1 Perturbation Process

For the previous perturbation argument, we used the assumption $n \leq e^{d/C}$ in Lemma 5.2.8 to show that the error of every column $|c_j(G) - 1|$ remains bounded. As $n \to \infty$, the union bound over these n columns will fail because some column $\|g_j\|_2^2$ will be large with probability approaching 1. To get around this, we perform a right-normalization after our perturbation and show that we still have enough randomness (degrees of freedom) to maintain the error and pseudorandom properties.

Definition 5.3.2 (Perturbation Process)**.** *Let frame $U \in \text{Mat}(d,n)$ of size $s(U) = 1$ be ε-Parseval ($\|\nabla_U\|_{\text{op}} \le \varepsilon$) and equal-norm ($\|\nabla_U^R\|_\infty = 0$), and consider some $\delta > 0$. First, add noise $V := U + \delta G$ where $\text{vec}(G) \sim N(0, \frac{1}{dn} P_U)$ according to Definition 5.2.2, then output the column normalization $V' := \{v_1', ..., v_n'\} \subseteq \mathbb{C}^d$ defined as*

$$v_j' := \frac{v_j}{\sqrt{n}\|v_j\|_2} = \frac{u_j + \delta g_j}{\sqrt{n}\|u_j + \delta g_j\|_2}.$$

Note that the output V' has size $s(V') = 1$ and $\nabla_{V'}^R = 0$ by construction.

We emphasize that in the sequel we will use $g \sim N(0, I_{dn})$ for the standard Gaussian, which is not to be confused by $\{g_j = Ge_j\}$ the columns of the random matrix $\text{vec}(G) \sim N(0, \frac{1}{dn} P_U)$.

In the following two section, we will show that for nearly doubly balanced frame U, the normalized output V' still has small error and satisfies the pseudorandom property. To simplify calculations, we have assumed that input U is also equal-norm. As shown in the proof of Theorem 4.5.2, this assumption can be satisfied with only a small loss.

5.3.2 Error

In this subsection, we bound the error of V' after the perturbation process in Definition 5.3.2. By definition, $\nabla_{V'}^R = 0$, so the main result of this subsection is the following bound on the left error.

Proposition 5.3.3. *For frame $U \in \text{Mat}(d,n)$ with size $s(U) = 1$ which is equal-norm ($\nabla_U^R = 0$) and ε-Parseval ($\|\nabla_U^L\|_{\text{op}} \le \varepsilon$), let V' be the output of the perturbation process in Definition 5.3.2. Then, for any $10(\frac{1}{d} + \frac{d}{n}) \le \theta \le \frac{1}{4}$ and $C \ge 10$ with $d \ge 100, n \ge 100\frac{d}{\theta^2}$,*

$$\|\nabla_{V'}^L\|_{\text{op}} \le \varepsilon + \delta^2(\varepsilon + 8\theta + \sqrt{\delta^2 \cdot 100Cd} + \delta^2 \cdot 21Cd + \delta^4 \cdot 20Cd),$$

with probability at least $1 - 10\exp(-\theta^2 n/10) - \frac{3}{C}$.

Note that $\nabla_{V'}^R = 0$ by construction. By Definition 4.2.3, we can write the left error as

$$\|\nabla_{V'}^L\|_{\text{op}} = \|d \cdot V'V'^* - s(V')I_d\|_{\text{op}} = \sup_{\xi \in S^{d-1}} |\langle \xi\xi^*, d \cdot V'V'^* - I_d \rangle|, \tag{5.12}$$

where V' is normalized so $s(V') = 1$ by construction. In Section 4.4 and Section 5.2.2, we were able to control the error using Gaussian concentration. But the normalization step in Definition 5.3.2 makes the distribution much more complicated, and we cannot simply apply Gaussian concentration e.g. Corollary 2.5.14. Therefore we will use some Taylor approximations and simple moment bounds to control the error of V', which will only give constant success probability instead of high probability results. The proof is quite long as there are many terms to bound, but the calculations are elementary.

We begin by decomposing the error of V' into various terms that we control separately.

Lemma 5.3.4. *For perturbation V' given according to Definition 5.3.2, the left error can be bounded by the following:*

$$\|\nabla^L_{V'}\|_{\mathrm{op}} \le \underbrace{\|\nabla^L_U\|_{\mathrm{op}}}_{(0)} + \delta^2 \underbrace{\left\| d\sum_{j=1}^n g_j g_j^* - d\sum_{j=1}^n (n\|g_j\|_2^2) u_j u_j^* \right\|_{\mathrm{op}}}_{(2)}$$

$$+ \delta^3 \underbrace{\sqrt{4 \left\| d\sum_{j=1}^n (n\|g_j\|_2^2) u_j u_j^* \right\|_{\mathrm{op}} \left\| d\sum_{j=1}^n (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}}}}_{(3)}$$

$$+ \delta^4 \underbrace{\left\| d\sum_{j=1}^n (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}}}_{(4)} \underbrace{+ 2\delta^4 \left\| d\sum_{j=1}^n (n\|g_j\|_2^2)^2 u_j u_j^* \right\|_{\mathrm{op}} + 2\delta^6 \left\| d\sum_{j=1}^n (n\|g_j\|_2^2)^2 g_j g_j^* \right\|_{\mathrm{op}}}_{(H)},$$

where the terms are named based on the order of δ, and (H) stands for "higher-order".

Proof. Note that item (2) of Proposition 5.2.3 shows that

$$n\|v_j\|_2^2 = n\|u_j\|_2^2 + \delta^2 n\|g_j\|_2^2 = 1 + \delta^2 n\|g_j\|_2^2,$$

where the last step was by the equal norm property of U. Below, we use the Taylor expansion $\frac{1}{1+x} = 1 - x + \frac{x^2}{1+x}$ to write out the left marginal as

$$V'V'^* = \sum_{j=1}^n \frac{v_j v_j^*}{n\|v_j\|_2^2} = \sum_{j=1}^n \frac{v_j v_j^*}{1 + \delta^2 n\|g_j\|_2^2} = \sum_{j=1}^n \left(1 - \delta^2 n\|g_j\|_2^2 + \frac{(\delta^2 n\|g_j\|_2^2)^2}{1 + \delta^2 n\|g_j\|_2^2} \right) v_j v_j^*. \tag{5.13}$$

According to Eq. (5.12), we want to bound the difference $\|dV'V'^* - I_d\|_{\mathrm{op}}$. We show the lemma by substituting $v_j := u_j + \delta g_j$ into Eq. (5.13) and grouping terms by the exponent of δ:

$$\begin{aligned}
\|dV'V'^* - I_d\|_{\mathrm{op}} \leq \underbrace{\|dUU^* - I_d\|_{\mathrm{op}}}_{(0)} &+ \delta d \underbrace{\|UG^* + GU^*\|_{\mathrm{op}}}_{(1)} \\
+ \delta^2 \underbrace{\left\| d\sum_{j=1}^{n} g_j g_j^* - d\sum_{j=1}^{n}(n\|g_j\|_2^2) u_j u_j^* \right\|_{\mathrm{op}}}_{(2)} &+ \delta^3 \underbrace{\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)\Big(u_j g_j^* + g_j u_j^*\Big) \right\|_{\mathrm{op}}}_{(3)} \\
+ \delta^4 \underbrace{\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}}}_{(4)} &+ \delta^4 \underbrace{\left\| d\sum_{j=1}^{n} \frac{(n\|g_j\|_2^2)^2}{1+\delta^2 n\|g_j\|_2^2}\Big(u_j + \delta g_j\Big)\Big(u_j + \delta g_j\Big)^* \right\|_{\mathrm{op}}}_{(H)} ,
\end{aligned} \tag{5.14}$$

where we applied triangle inequality on $\|\cdot\|_{\mathrm{op}}$, and names are based on the exponent of δ ((H) stands for "higher order").

Now we bound term-by-term. First note U that term (0) matches exactly as $\|dUU^* - I_d\|_{\mathrm{op}} = \|\nabla_U^L\|_{\mathrm{op}}$ by Definition 4.2.3, term (1) vanishes due to the orthogonality condition of Proposition 5.2.3(2), and the terms (2) and (4) in Eq. (5.14) match exactly with the same terms in the statement of Lemma 5.3.4. The following two claims will match (3) to (3) and (H) to (H), from which the lemma follows.

Claim 5.3.5. *Term (3) in Eq. (5.14) can be bounded by*

$$\left\| \sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)\delta(u_j g_j^* + g_j u_j^*) \right\|_{\mathrm{op}}^2 \leq 4 \left\| \sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2) u_j u_j^* \right\|_{\mathrm{op}} \left\| \sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)\delta^2 g_j g_j^* \right\|_{\mathrm{op}} .$$

Proof. For any $\xi \in S^{d-1}$, we consider the quadratic form

$$\begin{aligned}
\left| \langle \xi\xi^*, \sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)\delta(u_j g_j^* + g_j u_j^*)\rangle \right|^2 &= \left| 2\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)\langle \xi, u_j\rangle\langle \delta g_j, \xi\rangle \right|^2 \\
&\leq 4\left(\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)\langle \xi, u_j\rangle^2\right)\left(\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)\delta^2\langle \xi, g_j\rangle^2\right),
\end{aligned}$$

where the final step was by Cauchy-Schwarz. The claim then follows by taking supremum over $\xi \in S^{d-1}$ according to the definition of $\|\cdot\|_{\mathrm{op}}$. □

Next we deal with the higher-order terms in (H).

Claim 5.3.6. *(H) in Eq. (5.14) can be bounded by*

$$\left\|\sum_{j=1}^{n}\frac{(\delta^2 n\|g_j\|_2^2)^2}{1+\delta^2 n\|g_j\|_2^2}v_jv_j^*\right\|_{\mathrm{op}}\leq 2\left\|\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)^2 u_ju_j^*\right\|_{\mathrm{op}}+2\delta^2\left\|\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)^2 g_jg_j^*\right\|_{\mathrm{op}}.$$

Proof. Since the term (H) on the left hand side is positive semidefinite, we only increase the operator norm by removing the normalization $1+\delta^2 n\|g_j\|_2^2\geq 1$. Therefore we have

$$\sum_{j=1}^{n}\frac{(\delta^2 n\|g_j\|_2^2)^2}{1+\delta^2 n\|g_j\|_2^2}v_jv_j^*\preceq\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)^2\Big(u_ju_j^*+u_j(\delta g_j)^*+(\delta g_j)u_j^*+\delta^2 g_jg_j^*\Big).$$

To deal with the cross-term, note that for any $\xi\in S^{d-1}$

$$\begin{aligned}\left\langle\xi\xi^*,\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)^2(\delta u_jg_j^*+\delta g_ju_j^*)\right\rangle&=\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)^2 2\big(\langle\xi,u_j\rangle\big)\big(\delta\langle\xi,g_j\rangle\big)\\&\leq\sum_{j=1}^{n}(\delta^2 n\|g_j\|_2^2)^2\Big(\langle\xi,u_j\rangle^2+\delta^2\langle\xi,g_j\rangle^2\Big),\end{aligned}$$

where the last step was by $2ab\leq a^2+b^2$. The claim then follows by taking supremum over $\xi\in S^{d-1}$ according to the definition of $\|\cdot\|_{\mathrm{op}}$. □

The right hand sides of Claim 5.3.5 and Claim 5.3.6 are exactly the terms (3) and (H) in Lemma 5.3.4, so the lemma is shown. □

Note that the term (3) in Lemma 5.3.4 combines terms from (2) and (4). We will control both of these terms separately, which will imply a bound on (3) by Claim 5.3.5.

We first bound (2) in Lemma 5.3.4 by Gaussian concentration and a net argument, similar to how we bounded $\|\nabla_V^L\|_{\mathrm{op}}$ in the previous section. We first define a shorthand for the quadratic term that comes from the Taylor approximation for the normalization.

Definition 5.3.7. *Consider equal-norm frame* $U\in\mathrm{Mat}(d,n)$ *with* $\mathrm{vec}(G)\sim N(0,\frac{1}{dn}P_U)$ *according to Definition 5.2.2. Define* $Y:=\mathrm{diag}\{\sqrt{n}\|g_j\|_2\}_{j=1}^{n}$ *so that, for any* $\xi\in S^{d-1}$,

$$r_\xi(UY)=\sum_{j=1}^{n}\langle\xi,u_j\rangle^2(n\|g_j\|_2^2).$$

With this definition, we can rewrite quadratic term (2) in Lemma 5.3.4 as

$$\left\| d\sum_{j=1}^{n} g_j g_j^* - d\sum_{j=1}^{n}(n\|g_j\|_2^2)u_j u_j^* \right\|_{\mathrm{op}} = \sup_{\xi\in S^{d-1}} d\left|r_\xi(G) - r_\xi(UY)\right|.$$

In Eq. (5.6) in the proof of item Proposition 5.2.9(2), we have already shown high probability bounds on $r_\xi(G)$. We will use a similar argument to show that $r_\xi(UY)$ also concentrates around the same value, so the quadratic term can be bounded.

Lemma 5.3.8. *If frame $U \in \mathrm{Mat}(d,n)$ of size $s(U) = 1$ is $\varepsilon \leq \frac{1}{8}$-Parseval and exactly equal norm (according to Definition 4.1.2), then for every $\xi \in S^{d-1}$,*

$$1 - \varepsilon - \frac{(1+3\varepsilon)d}{n} - \frac{1-\varepsilon}{d} \leq \mathbb{E}[d\cdot r_\xi(UY)] \leq 1 + \varepsilon - \max\left\{\frac{(1-3\varepsilon)d}{n}, \frac{1+\varepsilon}{d}\right\}.$$

Further, for any $0 \leq \theta \leq \frac{1}{4}$,

$$Pr\left[\left|d\cdot r_\xi(UY) - \mathbb{E}[d\cdot r_\xi(UY)]\right| \geq \theta\right] \leq 2\exp\Big(-\frac{\theta^2 n}{9}\Big).$$

Proof. We rewrite $d\cdot r_\xi(UY)$ as a quadratic form with standard Gaussian $g \sim N(0, I_{dn})$. We emphasize that $g_j \in \mathbb{R}^d$ is the j-th column of random matrix $g_j = Ge_j$ where $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P)$, so g_j is not a subset of the coordinates of $g \sim N(0, I_{dn})$.

$$d\cdot r_\xi(UY) = dn\sum_{j=1}^{n}\langle\xi, u_j\rangle^2\|g_j\|_2^2 = \sum_{j=1}^{n}\langle\xi, u_j\rangle^2\langle I_d\otimes E_{jj}, Pgg^*P\rangle, \tag{5.15}$$

where the last step is by $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P)$ and $g_j = Ge_j$. Now we can control the mean:

$$\mathbb{E}[d\cdot r_\xi(UY)] = \mathbb{E}\sum_{j=1}^{n}\langle\xi, u_j\rangle^2\langle I_d\otimes E_{jj}, Pgg^*P\rangle = \Big\langle \sum_{j=1}^{n} I_d\otimes(\langle\xi, u_j\rangle^2 E_{jj}), P\Big\rangle,$$

where the last step was by $P^2 = P$ as P is an orthogonal projection and $\mathbb{E}gg^* = I_{dn}$. Since we don't have an explicit formula for P, we control the expectation using the spectral bounds in Proposition 5.2.3(3):

$$\begin{aligned}\sum_{j=1}^{n}\langle\xi, u_j\rangle^2\langle I_d\otimes E_{jj}, I_{dn} - P_L - P_R\rangle &\leq \sum_{j=1}^{n}\langle\xi, u_j\rangle^2\langle I_d\otimes E_{jj}, P\rangle \\ &\leq \sum_{j=1}^{n}\langle\xi, u_j\rangle^2\langle I_d\otimes E_{jj}, I_{dn} - \max\{P_L, P_R\}\rangle,\end{aligned} \tag{5.16}$$

where again we use max as shorthand to denote that the inequality is satisfied with both terms separately. Now, we bound the inner products with I_{dn}, P_L, P_R. First,

$$\sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle I_d \otimes E_{jj}, I_d \otimes I_n \rangle = d \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 = d \cdot r_\xi(U).$$

Then, we calculate the inner product with P_L as

$$\begin{aligned}\sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle I_d \otimes E_{jj}, P_L \rangle &= \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle I_d \otimes E_{jj}, I_d \otimes U^*(UU^*)^{-1}U \rangle \\ &= \langle I_d, I_d \rangle \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle E_{jj}, U^*(UU^*)^{-1}U \rangle = d \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle u_j u_j^*, (UU^*)^{-1} \rangle \\ &\in d \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \left(\frac{d}{1 \pm \varepsilon} \right) \|u_j\|_2^2 \in d \cdot r_\xi(U) \left(\frac{d}{n(1 \pm \varepsilon)} \right),\end{aligned}$$

where the fourth step was because U is ε-Parseval with $s(U) = 1$ so $dUU^* \in (1 \pm \varepsilon) I_d$, and the final step was because U is equal norm so $n\|u_j\|_2^2 = s(U) = 1$. Finally, we calculate the inner product with P_R as

$$\sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle I_d \otimes E_{jj}, P_R \rangle = \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 \langle I_d \otimes E_{jj}, \sum_{j'=1}^{n} \frac{u_{j'} u_{j'}^*}{\|u_{j'}\|_2^2} \otimes E_{j'j'} \rangle = \sum_{j=1}^{n} \langle \xi, u_j \rangle^2 = r_\xi(U),$$

where we used $\langle E_{j'j'}, E_{jj} \rangle$ vanishes unless $j = j'$, and that $\mathrm{Tr}[u_j u_j^*] = \|u_j\|_2^2$. Plugging these calculations into Eq. (5.16), we bound the mean

$$dr_\xi(U) \left(1 - \frac{d}{(1-\varepsilon)n} - \frac{1}{d} \right) \leq \mathbb{E}[d \cdot r_\xi(UY)] \leq dr_\xi(U) \left(1 - \max\left\{ \frac{d}{(1+\varepsilon)n}, \frac{1}{d} \right\} \right).$$

where we used Taylor approxamation $|(1+x)^{-1} - 1| \leq 2|x|$ for $|x| \leq \frac{1}{3}$ to bound the ε terms. The bound in the lemma now follows by the fact that U is ε-Parseval of size $s(U) = 1$ so $d \cdot r_\xi(U) \in 1 \pm \varepsilon$ for all $\xi \in S^{d-1}$. Therefore we can use the Taylor approximation $|\frac{1+x}{1-x} - 1| \leq 3|x|$ for $|x| \leq \frac{1}{3}$ to bound the error terms.

To prove concentration, we recall Eq. (5.15) where we wrote $d \cdot r_\xi(UY)$ in terms of standard Gaussian $g \sim N(0, I_{dn})$:

$$d \cdot r_\xi(UY) = \left\langle P \left(\sum_{j=1}^{n} I_d \otimes (\langle \xi, u_j \rangle^2 E_{jj}) \right) P, gg^* \right\rangle.$$

To get concentration, we will plug the following bounds into Corollary 2.5.14:

$$Tr\left[P\left(\sum_{j=1}^{n} I_d \otimes (\langle \xi, u_j\rangle^2 E_{jj})\right)P\right] \leq \sum_{j=1}^{n} \langle \xi, u_j\rangle^2 \langle I_d \otimes E_{jj}, I_{dn}\rangle \leq d \cdot r_\xi(U) \leq 1+\varepsilon,$$

$$\left\|P\left(\sum_{j=1}^{n} I_d \otimes (\langle \xi, u_j\rangle^2 E_{jj})\right)P\right\|_{\text{op}} \leq \max_{j\in[n]} \langle \xi, u_j\rangle^2 \|P^2\|_{\text{op}} \leq \max_{j\in[n]} \|u_j\|_2^2 = \frac{1}{n},$$

where in the first line we used the spectral bound $P \preceq I_{dn}$ from Proposition 5.2.3(3) and the fact that U is ε-Parseval of size $s(U) = 1$, and in the second line we used Cauchy-Schwarz for the second step and the fact that U is exactly equal-norm with $s(U) = 1$ for the final step. Therefore, applying Corollary 2.5.14 shows

$$Pr\left[\left|d \cdot r_\xi(UY) - \mathbb{E}[d \cdot r_\xi(UY)]\right| \geq \theta\right] \leq 2\exp\left(-\min\{\theta^2, \theta\}\frac{n}{8(1+\varepsilon)}\right) \leq 2\exp\Big(-\frac{\theta^2 n}{9}\Big),$$

where the last step was by our assumptions $\theta \leq \frac{1}{4}$ and $\varepsilon \leq \frac{1}{8}$. □

Now that we have proved concentration results for both quadratic terms, we can bound term (2) in Lemma 5.3.4.

Lemma 5.3.9. *If $\varepsilon \leq \frac{1}{8}$ and $\frac{1}{4} \geq \theta \geq 10(\frac{d}{n} + \frac{1}{d})$, then the term (2) in Lemma 5.3.4 satisfies*

$$\left\|d\sum_{j=1}^{n} g_j g_j^* - d\sum_{j=1}^{n} (n\|g_j\|_2^2) u_j u_j^*\right\|_{\text{op}} = \sup_{\xi\in S^{d-1}} \left|d \cdot r_\xi(G) - d \cdot r_\xi(UY)\right| \leq \varepsilon + 8\theta$$

with probability at least $1 - 4\exp(-\frac{\theta^2 n}{2})$.

Proof. By the triangle inequality, we can bound

$$\left|r_\xi(G) - r_\xi(UY)\right| \leq \left|r_\xi(G) - \mathbb{E}[r_\xi(G)]\right| + \left|\mathbb{E}[r_\xi(G)] - \mathbb{E}[r_\xi(UY)]\right| + \left|r_\xi(UY) - \mathbb{E}[r_\xi(UY)]\right|.$$

In Eq. (5.6) in the proof of Proposition 5.2.9(2), we have already shown

$$\sup_{\xi\in S^{d-1}} \left|d \cdot r_\xi(G) - \mathbb{E}[d \cdot r_\xi(G)]\right| \leq 4.5\theta$$

with probability at least $1 - 2\exp(-\frac{\theta^2 n}{2})$ as $n \geq 100\frac{d}{\theta^2}$. We can also control the difference in expectations:

$$\sup_{\xi \in S^{d-1}} \Big| \mathbb{E}[d \cdot r_\xi(G)] - \mathbb{E}[d \cdot r_\xi(UY)] \Big| \leq \varepsilon + (1+3\varepsilon)\Big(\frac{1}{d} + \frac{d}{n}\Big),$$

where the first step was by the two-sided bounds in Lemma 5.2.7 and Lemma 5.3.8.

We apply a net argument (similar to Proposition 5.2.9) to give high probability bounds for $r_\xi(UY)$. So let $N \subseteq S^{d-1}$ be an $\eta = \frac{1}{25}$-net so that $1 - 2\eta - \eta^2 \geq 9.9$. Fact 2.6.3 gives us the bound $|N| \leq (1 + 2\eta^{-1})^d \leq e^{4d}$, so we can use the union bound to show

$$Pr\Big[\sup_{\xi \in S^{d-1}} \Big| r_\xi(UY) - \mathbb{E}[r_\xi(UY)] \Big| \geq \frac{3\theta}{d}\Big] \leq Pr\Big[\sup_{\xi \in N} \Big| r_\xi(UY) - \mathbb{E}[r_\xi(UY)] \Big| \geq \frac{(0.9)3\theta}{d}\Big]$$
$$\leq \sum_{\xi \in N} Pr\Big[\Big| r_\xi(UY) - \mathbb{E}[r_\xi(UY)] \Big| \geq \frac{(0.9)3\theta}{d}\Big] \leq 2\exp(4d - (0.9\theta)^2 n) \leq 2\exp\Big(-\frac{\theta^2 n}{2}\Big),$$

where the first step was by Lemma 2.6.5, the second was by the union bound over N, the third was by by the concentration probability shown in Lemma 5.3.8, and the final step was by the assumption $n \geq 100\frac{d}{\theta^2}$.

In total, with probability $1 - 4\exp(-\frac{\theta^2 n}{2})$, the quadratic term can be bounded by

$$\Big\| d\sum_{j=1}^n g_j g_j^* - d\sum_{j=1}^n (n\|g_j\|_2^2) u_j u_j^* \Big\|_{\mathrm{op}} \leq 4.5\theta + \varepsilon + (1+3\varepsilon)\Big(\frac{1}{d} + \frac{d}{n}\Big) + 3\theta \leq \varepsilon + 8\theta,$$

where the last step was by the assumption $\theta \geq 10(\frac{1}{d} + \frac{d}{n})$ and $\varepsilon \leq \frac{1}{8}$. □

To bound the higher order terms (4) and (H) of Lemma 5.3.4, we use a crude triangle inequality and the following bound on higher order moments of Gaussian norm $c_j(G) = \|g_j\|_2^2$.

Lemma 5.3.10. *For every $j \in [n]$, if frame $U \in \mathrm{Mat}(d,n)$ is $\varepsilon \leq \frac{1}{3}$-doubly balanced, then for $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2 and any $p \geq 1$:*

$$\mathbb{E}(dn\|g_j\|_2^2)^p \leq (8p)^p d + (2d)^p.$$

Proof. We rewrite $n \cdot \|g_j\|_2^2$ as a quadratic form with standard Gaussian $g \sim N(0, I_{dn})$.

$$n\|g_j\|_2^2 = n\langle I_d \otimes E_{jj}, \mathrm{vec}(G)\,\mathrm{vec}(G)^* \rangle = \frac{1}{d}\langle I_d \otimes E_{jj}, Pgg^*P \rangle,$$

where the last step is because $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P)$.

We have derived the following bounds in Lemma 5.2.8:

$$\|P(I_d \otimes E_{jj})P\|_{\mathrm{op}} \le \|P^2\|_{\mathrm{op}} \le 1, \qquad \text{and} \qquad \mathrm{Tr}[P(I_d \otimes E_{jj})P] \le \langle I_d \otimes E_{jj}, I_{dn}\rangle \le d,$$

where we used $P \preceq I_{dn}$. Therefore, we can apply Corollary 2.5.17 with the bounds to show, for any $p \ge 1$:

$$\mathbb{E}(dn\|g_j\|_2^2)^p \le (8p)^p \,\mathrm{Tr}[P(I_d \otimes E_{jj})P]\|P(I_d \otimes E_{jj})P\|_{\mathrm{op}}^{p-1} + (2\,\mathrm{Tr}[P(I_d \otimes E_{jj})P])^p \le (8p)^p d + (2d)^p.$$

□

This allows us to bound the expectation of (4) and (H) in Lemma 5.3.4. Unlike the arguments of Lemma 5.3.9, for these terms we bound the whole $\|\cdot\|_{\mathrm{op}}$ instead of each inner product $\langle \xi\xi^*, \cdot\rangle$, and we will not use a net argument. We will also only be able to prove constant probability bounds on these terms using Markov's inequality.

We begin by bounding term (4) in Lemma 5.3.4.

Claim 5.3.11. *For equal-norm frame $U \in \mathrm{Mat}(d,n)$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2,*

$$\mathbb{E}\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}} \le 4d + 256.$$

Proof. We use a crude triangle inequality to bound in terms of $\|g_j\|_2^2$:

$$\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}} \le dn \sum_{j=1}^{n} \|g_j\|_2^4.$$

So we can bound the expectation using the Gaussian moment bounds:

$$\mathbb{E}\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}} \le \frac{1}{dn}\sum_{j=1}^{n} \mathbb{E}(dn\|g_j\|_2^2)^2 \le \frac{1}{dn}\sum_{j=1}^{n}((2d)^2 + (8\cdot 2)^2 \cdot d) = 4d + 256,$$

where the first step was by the triangle inequality, and the second step was by Lemma 5.3.10 with $p = 2$. □

Next, we bound the first term in (H) in Lemma 5.3.4.

Claim 5.3.12. *For equal-norm frame $U \in \mathrm{Mat}(d,n)$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2,*

$$\mathbb{E}2\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)^2 u_j u_j^* \right\|_{\mathrm{op}} \leq 2(4d+256).$$

Proof. We use a crude triangle inequality to bound in terms of $\|g_j\|_2^2$:

$$\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)^2 u_j u_j^* \right\|_{\mathrm{op}} \leq dn^2 \sum_{j=1}^{n} \frac{\|g_j\|_2^4}{n},$$

where we used that U is equal-norm in the last step so $\|u_j\|_2^2 = \frac{1}{n}$. So we can bound the expectation using the Gaussian moment bounds:

$$\mathbb{E}\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)^2 u_j u_j^* \right\|_{\mathrm{op}} \leq \frac{1}{dn}\sum_{j=1}^{n}\mathbb{E}(dn\|g_j\|_2^2)^2 \leq 4d+256,$$

where the first step was by the triangle inequality, and the second step was by Lemma 5.3.10 with $p=2$ (we omit the calculation as it is the same term as Claim 5.3.11). □

Finally, we bound the second term in (H) in Lemma 5.3.4.

Claim 5.3.13. *For equal-norm frame $U \in \mathrm{Mat}(d,n)$ and $\mathrm{vec}(G) \sim N(0, \frac{1}{dn}P_U)$ according to Definition 5.2.2,*

$$\mathbb{E}2\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)^2 g_j g_j^* \right\|_{\mathrm{op}} \leq \left(16d + \frac{2(24)^3}{d}\right).$$

Proof. We use a crude triangle inequality to bound in terms of $\|g_j\|_2^2$:

$$\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)^2 g_j g_j^* \right\|_{\mathrm{op}} \leq dn^2 \sum_{j=1}^{n} \|g_j\|_2^6.$$

So we can bound the expectation using the Gaussian moment bounds:

$$\mathbb{E}2\left\| d\sum_{j=1}^{n}(n\|g_j\|_2^2)^2 g_j g_j^* \right\|_{\mathrm{op}} \leq \frac{2}{d^2 n}\sum_{j=1}^{n}\mathbb{E}(dn\|g_j\|_2^2)^3 \leq \frac{2}{d^2 n}\sum_{j=1}^{n}((2d)^3+(8\cdot 3)^3 d) \leq \left(16d+\frac{2(24)^3}{d}\right),$$

where the first step was by the triangle inequality, and the second step was by Lemma 5.3.10 with $p=3$. □

We can now combine these terms to get a bound on the left marginal.

Proof of Proposition 5.3.3. We use the decomposition in Lemma 5.3.4 and bound term by term. First, $\|\nabla_U\|_{\mathrm{op}} \le \varepsilon$ by the ε-Parseval assumption on U. Then, Lemma 5.3.9 shows the quadratic term is bounded by $\varepsilon + 8\theta$ with probability at least $1 - 4\exp(-\frac{\theta^2 n}{2})$.

For the higher order terms in Lemma 5.3.4, we apply the expectation bounds given in Claims 5.3.11, 5.3.12, and 5.3.13, along with Markov's inequality for $C \ge 1$ to show that simultaneously with probability at least $1 - \frac{3}{C}$

$$\begin{aligned}
\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}} &\le C\mathbb{E}\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2) g_j g_j^* \right\|_{\mathrm{op}} \le C(4d + 256) \le 7Cd; \\
\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2)^2 u_j u_j^* \right\|_{\mathrm{op}} &\le C\mathbb{E}\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2)^2 u_j u_j^* \right\|_{\mathrm{op}} \le 2C(4d + 256) \le 14Cd; \\
\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2)^2 g_j g_j^* \right\|_{\mathrm{op}} &\le C\mathbb{E}\left\| d\sum_{j=1}^{n} (n\|g_j\|_2^2)^2 g_j g_j^* \right\|_{\mathrm{op}} \le C\Big(16d + \frac{2(24)^3}{d}\Big) \le 20Cd;
\end{aligned}$$

where the last step in each line was by the assumption $d \ge 100$.

This also allows us to bound term (3) in Eq. (5.14), as

$$\begin{aligned}
\left\| d\sum_{j=1}^{n} (\delta^2 n\|g_j\|_2^2)\delta(u_j g_j^* + g_j u_j^*) \right\|_{\mathrm{op}}^2 &\le 4\left\| d\sum_{j=1}^{n} (\delta^2 n\|g_j\|_2^2) u_j u_j^* \right\|_{\mathrm{op}} \left\| d\sum_{j=1}^{n} (\delta^2 n\|g_j\|_2^2)\delta^2 g_j g_j^* \right\|_{\mathrm{op}} \\
&\le 4\Big(\delta^2(1 + \varepsilon + 3\theta)\Big)\Big(\delta^4 \cdot 7Cd\Big),
\end{aligned}$$

where the first step was by Claim 5.3.5, we bounded the first term by using Lemma 5.3.8 to bound the expectation and following the proof of Lemma 5.3.9 for concentration, and we used the bound derived for term (4) above to bound the second term.

In total, we can collect all terms in Lemma 5.3.4 and show that with probability $1 - 4\exp(-\frac{\theta^2 n}{10}) - \frac{3}{C}$:

$$\|\nabla_{V'}^L\|_{\mathrm{op}} \le \varepsilon + \delta^2(\varepsilon + 8\theta) + 2\delta^3\sqrt{(1 + \varepsilon + 3\theta)(7Cd)} + \delta^4(21Cd) + \delta^6(20Cd).$$

□

Remark 5.3.14. *We believe the moment bounds in Claims 5.3.11, Claim 5.3.12, Claim 5.3.13 are not necessarily the best way to analyze these terms. These are in fact the bottleneck in our assumption $\varepsilon \lesssim \frac{1}{d}$. Note that Theorem 5.2.1 covers all $n \le e^{d/C}$, so for this case, we could try to use the assumption $n \ge e^{d/C}$ to improve the error bound.*

5.3.3 Pseudorandom Condition

In Section 5.2.3, we were able to use Gaussian concentration to give high probability lower bounds for the spectral pseudorandom condition given in Lemma 5.1.1. In this section, we have right-normalized the perturbed input $V = U + \delta G$, so the distribution of columns becomes more complicated. Therefore, instead of proving the lower bound directly, we follow a strategy similar to Section 5.1 by using Lemma 5.1.2 to bound the effect of normalization on pseudorandomness. For this purpose, we use the following lemma to show the normalization affects most columns very little.

Lemma 5.3.15. *For equal norm frame $U \in \mathrm{Mat}(d,n)$, let $V = U + \delta G$ and V' be the output of the perturbation process according to Definition 5.3.2. Then for any $C \geq 10$, the random variable $T_B := \{j \in [n] \mid n\|g_j\|_2^2 \geq 6\}$ satisfies*

$$Pr[|T_B| \geq 2Ce^{-d/2}n] \leq \frac{1}{C}.$$

Proof. To show the claim, let X_j be the indicator variable for the event $j \in T_B$ and note

$$\mathbb{E}X_j = Pr[n\|g_j\|_2^2 \geq 6] \leq Pr[n \cdot c_j(G) \geq 1 + 5] \leq 2\exp\Big(-\frac{d}{2}\Big),$$

where in the second step we used the definition of c_j, and in the final step we applied concentration from Lemma 5.2.8.

Therefore we use Markov's inequality to show

$$Pr\Big[|T_B| \geq 2Ce^{-d/2}\Big] \leq Pr\Big[|T_B| \geq C\mathbb{E}|T_B|\Big] \leq \frac{1}{C},$$

where the first step was by the above bound on $\mathbb{E}|T_B|$. □

Remark 5.3.16. *This is the part of the argument most ready for improvement. We believe that the different columns will only be weakly dependent and so we should be able to get a high probability bound in the above statement.*

We can now prove pseudorandomness of V' simply by applying Lemma 5.1.2.

Proposition 5.3.17. *For ε-doubly balanced $U \in \mathrm{Mat}(d,n)$ with $\varepsilon \leq \frac{1}{4}$, and V' the output of the perturbation process in Definition 5.3.2, if $\delta \leq \frac{1}{4}, d \geq \frac{100}{\beta}, n \geq \frac{100d}{\beta}, \beta \geq 10Ce^{-d/2}$, then V' is $(e^{-(6-3\log_2 \beta)}\delta^2, \beta)$-pseudorandom with probability at least $1 - \frac{2}{C}$.*

Proof. Applying Proposition 5.2.14 with $\frac{4}{5}\beta$ shows that V is $(e^{-(5-3\log_2\beta)}, \frac{4}{5}\beta)$-pseudorandom. Then we normalize $v'_j := \frac{v_j}{\sqrt{n}\|v_j\|_2}$, so we want to bound the size of $T_B := \{j \in [n] \mid n\|g_j\|_2^2 \geq 6\}$. Applying Lemma 5.3.15 with $C = 10$ gives that $|T_B| \leq Ce^{-d/2}n \leq \frac{1}{5}\beta n$ by our assumption $\beta \geq 10Ce^{-d/2}$. By the union bound, both of these events occur simultaneously with the probability stated.

Our plan is to reduce pseudorandomness of V' to the pseudorandomness of V, so we rewrite the normalization as $V' = VR$ with $R_{jj} = \frac{1}{\sqrt{n}\|v_j\|_2}$. Note $n\|v_j\|_2^2 = n\|u_j\|_2^2 + \delta^2 n\|g_j\|_2^2 = 1 + \delta^2 n\|g_j\|_2^2$ by Proposition 5.2.3(2) and the equal-norm property of U. So $j \in T_B \implies R_j^2 \leq \frac{1}{1+6\delta^2}$. By the assumption $\delta \leq \frac{1}{4}$, we can apply Lemma 5.1.2 with $\tau = \frac{8}{11} \geq \frac{1}{1+6\delta^2}$ to show V' is $(\alpha'\delta^2, \beta')$-pseudorandom for $\beta' \leq \frac{4}{5}\beta + \frac{1}{5}\beta = \beta$ and

$$\alpha' \geq \frac{8}{11} \cdot e^{-(5-3\log_2\beta)} \cdot \frac{4}{5} \geq e^{-(6-3\log_2\beta)},$$

where we used $(e^{-(5-3\log_2\beta)}, \frac{4}{5}\beta)$-pseudorandomness of V and substituted in $\tau = \frac{8}{11}$. □

5.3.4 Putting it Together

Proof of Theorem 5.3.1. Let V' be the output of Definition 5.3.2. Then item (2) follows from Proposition 5.3.3 by the assumption that U is ε-Parseval with $s(U) = 1$ (so $\|\nabla_U^L\|_{\text{op}} \leq \varepsilon$) and item (3) is exactly the content of Proposition 5.3.17, so by the union bound these occur simultaneously with failure probability at most

$$4\exp\Big(-\frac{\theta^2 n}{10}\Big) + 2\exp\Big(-\frac{\beta n}{10}\Big) + \frac{4}{C} < 1,$$

where the last step is by our assumptions $C \geq 10, n \geq 100\frac{d}{\min\{\beta,\theta^2\}}$.

To show item (1), we claim that the distance from $U \to V'$ is less than the distance from $U \to V$. The claim then follows as

$$\|V' - U\|_F^2 \leq \|V - U\|_F^2 \leq (1+\theta)\delta^2,$$

where the last step is by item (1) of Theorem 5.2.1.

To show the claim, we use the fact that both U and V' are equal norm. In particular, for every $j \in [n]$ we have

$$v'_j = \arg\min_w\{\|w - v_j\|_2^2 \mid n\|w\|_2^2 = 1\}.$$

By the orthogonality conditions in Proposition 5.2.3(2), $\|v_j\|_2^2 = \|u_j\|_2^2 + \delta^2 \|g_j\|_2^2 \geq \|u_j\|_2^2 = \frac{1}{n}$, so in fact v_j' is the projection of v_j onto B_2^d. Therefore, the claim follows by Lemma 2.3.13. □

This completes the smoothed analysis section of the thesis. With the results of Section 5.2 and Section 5.3, we can complete the optimal distance bound for the Paulsen problem as shown in Section 4.5.

Chapter 6

Geodesic Convexity and Scaling

The main goal of this chapter is to present a framework for our analysis of the general tensor scaling problem. To this end, we first present the scaling framework from a more abstract perspective in Section 6.1. This background comes from a long line of work in mathematical physics and algebraic geometry, and is quite abstract and technical. This general theory is not required in order to understand our results, as the analysis of tensor scaling mostly relies on tools from basic linear algebra and convexity theory. In Section 6.2, we formally define the tensor scaling problem, which is a special case of the scaling framework and the most general problem we study in this thesis. Then, we use the theory of Kempf-Ness functions presented in Section 6.1 to give a geodesically convex formulation for tensor scaling. This is main result of this chapter, and in Chapter 7 we use it to generalize the matrix scaling analysis of Chapter 3 to the tensor setting. Finally in Section 6.3, we use this geodesically convex formulation to prove the frame-to-matrix reduction given in Theorem 4.2.13. We also present a general reduction theorem from the non-commutative tensor scaling problem to the simpler commutative setting. Using this reduction, we are able to unify the arguments in [63] and [36], as well as to give quantitative improvements in Chapter 7.

6.1 Background on Scaling

In this section, we would like to give some background for the geodesic convex formulation for tensor scaling that we present in Section 6.2. The ideas presented here come from the fields of Hamiltonian geometry and geometric invariant theory. We reiterate to the reader that most of this abstract perspective is only discussed for context, and will not be used in subsequent analyses. Since we will only require a small bit of this theory, we mostly give sketches here without formal definitions.

The following subsections proceed in order from general to specific, ending with the a description of the convex optimization framework that we specialize to our analysis of tensor scaling. Section 6.1.1 contains a very brief sketch of the moment map, using the Schur-Horn Theorem and Horn's problem as illustrating examples. The moment map is used to understand (compact) groups acting on (symplectic) spaces in a specific (Hamiltonian) way. Then in Section 6.1.2, we present some ideas and questions from geometric invariant theory, which studies group actions with additional algebraic structure. For this and the remaining sections, we will use the example of matrix normalization to illustrate concepts. It turns out that this setting can be seen as a special case of the Hamiltonian group actions discussed in Section 6.1.1. The focus of Section 6.1.3 is the Kempf-Ness Theorem [58], which presents an optimization formulation for some of the group orbit questions of Section 6.1.2. This is also the context for the most general form of scaling, which can be viewed as a dual problem to the null cone question from geometric invariant theory. One of the key contributions of the Kempf-Ness theorem is to explain the underlying geodesic convex structure of these group optimization problems. Finally, in Section 6.1.4, we follow the presentation of Bürgisser et al. [20], which makes quantitative the relationship between group optimization and scaling. In particular, this is the place we discuss scaling algorithms and their analysis at a high level. This sets the stage for Section 6.2, where we use this geodesic convex optimization framework to analyze tensor scaling.

6.1.1 The Moment Map

The moment map and moment polytope are central objects in the study of Hamiltonian manifolds. This subject has a long history with many connections across mathematics and physics. We will only need a tiny sliver of the full theory for the work in this thesis. Therefore we invite the reader to consult the monograph by Guillemin and Sjamaar [41] for a much more thorough exposition of the history and main results. In this subsection, we will discuss some concepts via the following illustrating example.

Example 6.1.1. *Given two vectors $\lambda, a \in \mathbb{R}^d$, when does there exist a matrix $A \in \mathrm{H}(d)$ with spectrum λ and diagonals $\mathrm{diag}(A) = a$?*

The answer to the above question is a classical result in matrix analysis and is known as the Schur-Horn Theorem [51]. It states that A exists iff a is in the convex hull of permutations of λ, i.e. $S_d \cdot \lambda := \{(\lambda_{\sigma(1)}, ..., \lambda_{\sigma(d)}) \mid \sigma \in S_d\}$ where S_d are the set of permutations on $[d]$. One elementary proof of this result relies on the theory of majorization and is explained in detail in the wonderful book by Bhatia [12]. Below, we translate this result into a more general language in order to illustrate the concepts of moment maps and moment polytopes.

First note that by the Spectral Theorem (Theorem 2.1.8), $A \in \mathrm{H}(d)$ has spectrum $\lambda \in \mathbb{R}^d$ iff $A = U\Lambda U^*$ for some unitary matrix of eigenvectors $U \in \mathrm{U}(d)$, where $\Lambda = \mathrm{diag}(\lambda)$. Therefore, we can rewrite the set under consideration as the orbit of Λ under conjugation by the group $\mathrm{U}(d)$. We want to understand the image of this orbit under diagonal projection $\mathrm{diag}(B) := \{b_{ii}\}_{i \in [d]}$, and the Schur-Horn theorem tells us that this image is in fact a convex polytope with vertices $S_d \cdot \lambda$.

We can now present the main characters in the general story: the input is a set X (symplectic manifold), a group G (compact Lie group), and a special kind of (Hamiltonian) action $G \cdot X \to X$, and we want to understand the properties of this action using a specific (moment map) $\mu : X \to \mathfrak{g}$, where $\mathfrak{g}$ is a vector space associated with G (its Lie algebra, see the discussions in Section 2.2.3). The role of the moment map is to encode the infinitesimal behavior of the group action at each point. We will not detail how exactly this is done, but will present a more concrete example in Section 6.1.3.

In the context of Example 6.1.1, we have X the subset of $\mathrm{H}(d)$ with spectrum λ, group $G = \mathrm{U}(d)$ acts by conjugation on X, and we want to understand this group action through its diagonal projection $\mathrm{diag} : \mathrm{H}(d) \to \mathbb{R}^d$. The main content of the Schur-Horn Theorem is that the image $\mathrm{diag}(U\Lambda U^*)$ is a convex polytope.

As the culmination of a long line of work in symplectic geometry [6], [42], [43], Kirwan [59] proved the following grand generalization: the image of the moment map for any Hamiltonian action is always a convex polytope! This allows us to unambiguously define this image as the moment polytope. In the next Section 6.1.2, we restrict to the some problems in geometric invariant theory which are illuminated by the theory of moment maps.

6.1.2 Geometric Invariant Theory

In this section, we present some basic ideas from geometric invariant theory. This is a subfield of algebraic geometry where we want to study the action of a group G on a vector space V via the geometry of polynomial functions on V. Once again this is a powerful and deep subject which we will barely sketch, so we point the interested reader to the classical text of Mumford et al. [73] for a more thorough treatment.

In order to avoid too many technical details, we will mostly focus on the following illustrating example.

Example 6.1.2. *Given matrix $A \in \mathrm{Mat}(n)$, when is it diagonalizable according to Definition 2.1.6, i.e. when is there some $V \in \mathrm{GL}(n)$ such that $VAV^{-1} \in \mathrm{diag}(n)$? Further, if it is diagonalizable, what is the condition number defined as*

$$\inf_V \kappa(V) := \inf_V \|V\|_{\mathrm{op}}\|V^{-1}\|_{\mathrm{op}}?$$

Here, the infimum is over all possible bases of eigenvectors $V \in \mathrm{GL}(n)$.

Recall that according to Definition 2.1.6, A is diagonalizable iff it has a linearly independent basis of eigenvectors. On the other hand, if $A^k \equiv 0$ for some $k \in \mathbb{N}$, then A is said to be nilpotent, and A cannot be diagonalized if this is the case. Below, we present some stability concepts from geometric invariant theory which generalize the diagonalizability property in Example 6.1.2. We will revisit the condition number in Section 6.1.4.

Definition 6.1.3. *Let $v \in V$ for some inner product space V over field $\mathbb{C}$, and let $G \subseteq \mathrm{GL}(V)$ be an appropriately nice (complex algebraic reductive) group with a linear action on V. We denote the orbit of v by $G \cdot v$, and the orbit closure with respect to the Euclidean topology by $\overline{G \cdot v}$. We say v is*

1. *unstable: if $0 \in \overline{G \cdot v}$;*
2. *semi-stable: if $0 \notin \overline{G \cdot v}$;*
3. *stable: if $G \cdot v$ is closed;*

The set of unstable points is called the null cone of the group action.

Note that $0 \in V$ is a singleton closed orbit under any linear action, so v stable implies v is semi-stable. In the context of Example 6.1.2, it can be shown that for vector space $\mathrm{Mat}(n)$ and conjugation action VAV^{-1} of $V \in \mathrm{SL}(n)$, the null cone is exactly the set of nilpotent matrices, and that A is diagonalizable iff this orbit is closed, i.e. A is stable. This suggests the following natural problem

Definition 6.1.4 (Null Cone Membership). *Given (V, G) as in Definition 6.1.3, decide whether input $v \in V$ is stable or in the null cone.*

This basic problem contains many natural questions in computational complexity and algebra as special cases. It turns out that in this general setting, classical results by Hilbert and Mumford (see [73]) show that the null cone can always be written as the common zeros of some set of G-invariant polynomials. This motivates the following approach to null cone membership: compute the set of invariant polynomials and test whether they all vanish on the given input. This algebraic approach has been made constructive in some cases (e.g. [72]), but often with prohibitively large runtime.

By another set of classical results [58], it turns out that we can connect the null cone problem to the moment map from Section 6.1.1. For now, we will just state the result for Example 6.1.2. In this setting, $\mu(A) := AA^* - A^*A$ is the moment map. Note that $\mu(A) = 0$ is exactly the set of normal matrices which can be diagonalized by unitaries (see Theorem 2.1.8). Therefore the diagonalization problem can be rephrased as finding an element of the orbit $B \in \mathrm{SL}(d) \cdot A := \{VAV^{-1} \mid V \in \mathrm{SL}(d)\}$ such that $\mu(B) = 0$. Equivalently, A is stable iff A is diagonalizable iff $0 \in \mu(\mathrm{SL}(d) \cdot A)$. This is known as the scaling problem associated to null cone membership problem in Definition 6.1.4.

In Section 6.1.3, we explore this connection between stability and scaling in more detail.

6.1.3 Kempf-Ness Equivalence

In this subsection, we will introduce an optimization formulation for the null cone problem given in Definition 6.1.4. We will follow the work of Kempf and Ness [58], which connects the moment map from Section 6.1.1 to the group optimization setting. We will once again use the example of matrix diagonalization from Example 6.1.2.

First note that deciding the stability properties of $v \in V$ according to Definition 6.1.3 can be rewritten in terms of the following optimization problem:

$$cap(v) := \inf_{g \in G} \|g \cdot v\|_2^2 = 0,$$

where $\|\cdot\|_2$ is the standard Euclidean norm on inner product space V. This is called the capacity of group orbit $G \cdot v$ and has a long history not only in geometric invariant theory, but also in various instances of the scaling framework in theoretical computer science, namely operator scaling [45] and the Brascamp-Lieb inequalities [39]. Indeed, it can be shown that $cap(v) = 0$ iff $0 \in \overline{G \cdot v}$ iff v is in the null cone, i.e. v is unstable. This reduces the null cone membership problem in Definition 6.1.4 to solving this group optimization problem, or even approximating it to sufficient precision.

Going back to our diagonalizability example, consider the standard upper triangular Jordan block $N \in \mathrm{Mat}(2)$ which sends $Ne_2 = e_1$ and $Ne_1 = 0$. Clearly N is nilpotent as $N^2 \equiv 0$. We can simply verify that N is in the null cone of the SL(2) conjugation action: let $V_t := \mathrm{diag}(e^{-t}, e^t) \in SL(2)$ and observe that

$$\lim_{t\to\infty} V_t N V_t^{-1} = \lim_{t\to\infty} \begin{pmatrix} e^{-t} & 0 \\ 0 & e^t \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} = \lim_{t\to\infty} \begin{pmatrix} 0 & e^{-2t} \\ 0 & 0 \end{pmatrix} = 0,$$

so 0 is in the orbit closure of N and N is in the null cone according to Definition 6.1.3.

This answers one direction of the null cone problem, but for the stable case, we would like to certify that *cap* > 0. If the function $g \to \|g \cdot v\|_2^2$ were convex, then we could attempt to find the minimizer of this function, and the vanishing gradient would give the required certificate of optimality (see Lemma 2.3.4). It turns out that this is the right idea given an appropriate geometry on G. In particular, the Kempf-Ness function [58] for group G and vector v is defined as

$$f_v(g \in G) := \|g \cdot v\|_2^2,$$

and the capacity of $G \cdot v$ is the optimum value of f_v over G. Further, by defining the appropriate notion of geodesic convexity and geodesic gradients on G, we can certify optimality of the Kempf-Ness function and show that v is stable according to Definition 6.1.3. Below we explicitly describe these optimality certificates in the matrix conjugation example.

Proposition 6.1.5. *For $A \in \mathrm{Mat}(d)$ and group $\mathrm{SL}(d)$ acting by conjugation, the Kempf-Ness function is defined as*

$$f_A(V) := \|VAV^{-1}\|_F^2.$$

By unitary invariance of $\|\cdot\|_F$, the value of $f_A(g)$ depends only on the polar part V^*V. Further, for any $P \in \mathrm{SPD}(d)$ and $X \in \mathfrak{spd}(d)$ given in Definition 2.1.10, the univariate restriction

$$h(t) := f_A(\sqrt{P} e^{tX} \sqrt{p})$$

is convex on $t \in \mathbb{R}$. Finally, the moment map $\mu(B) := BB^* - B^*B$ encodes the first order derivative of f_A in the following sense: for any $P \in \mathrm{SPD}(d)$ and $X \in \mathfrak{spd}(d)$,

$$\partial_{t=0} h(t) = \partial_{t=0} f_A(\sqrt{P} e^{tX} \sqrt{p}) = \mathrm{Tr}[\mu(\sqrt{P} \cdot A)X].$$

Therefore, V is an optimizer of f_A iff $\mu(VAV^{-1}) = 0$ iff VAV^{-1} is normal.

We omit the proof of these properties, as refer to Proposition 6.2.18 for the explicit calculations in the tensor scaling setting. Note that we have used the Kempf-Ness function to explicitly connect the moment map to the group stability conditions in Definition 6.1.3. In fact, in this group optimization setting, the Kempf-Ness function gives an equivalent way to define the moment map as $\mu(g \cdot v) := \nabla f_v(g)$ where ∇ denotes the appropriate notion of (geodesic) gradient on the group G. We will discuss this more formally later (see Definition 7.1.1). This gives us enough machinery to state the Kempf-Ness theorem [58], which completely characterizes unstable and stable points.

Theorem 6.1.6 (Kempf-Ness Theorem)**.** *Let $v \in V$ for some inner product space V over field $\mathbb{C}$, and let G be a nice (complex reductive) group with a linear action on V. Then v is*

1. *unstable iff $cap(v) = \inf_{g \in G} f_v(g) = 0$.*
2. *semi-stable iff $cap(v) > 0$.*
3. *stable iff $\exists g \in G$ such that $cap(v) = f_v(g)$ iff $\mu(g \cdot v) = \nabla f_v(g) = 0$.*

This tells us that we can decide the stability of v in two ways: either by solving the group optimization problem $cap(v) = \inf_{g \in G} f_v(g) = \inf_{g \in G} \|g \cdot v\|_2^2$, or by solving the so-called scaling problem $g_* := \arg\inf_{g \in G} \|\mu(g \cdot v)\|_{\mathfrak{g}}^2 = \arg\inf_{g \in G} \|\nabla f_v(g)\|_{\mathfrak{g}}^2$, where $\| \cdot \|_{\mathfrak{g}}$ is an appropriately chosen norm for the gradient of f_v. Note that item (3) shows that for the Kempf-Ness function, the local optimality condition given by $\nabla f_v(g) = 0$ is equivalent to the global optimality condition $cap(v) = f_v(g)$. This is reminiscent of convex analysis (Lemma 2.3.4), and indeed an important result of [58] shows that f_v is in fact a convex function when the domain G is given the appropriate geodesic geometry (see Definition 6.2.13). In the next Section 6.1.4, we will discuss how this can be viewed as a duality theory between the null cone problem and scaling. Then we will present a quantitative strengthening of the Kempf-Ness theorem due to [20] which makes this duality effective and allows us to transfer results between the two problems.

6.1.4 Optimization for Scaling

In this subsection, we will discuss the non-commutative duality theory due to [20]. This will allow us to effectively connect the group optimization framework to scaling problems, which are the main focus of this thesis.

One of the main results of [20] is the following quantitative strengthening of the Kempf-Ness Theorem (6.1.6).

Theorem 6.1.7 (Non-commutative Duality Theorem 1.17 of [20]). *Let V be an inner product space with linear action of complex reductive group G. Then there are constants γ (weight margin) and L (weight norm), depending only on (G, V) such that, for any $v \in V$, the capacity $cap(v) := \inf_{g \in G} \|g \cdot v\|_2^2$ and moment map $\mu : V \to \mathfrak{g}$ are related by*

$$1 - \frac{1}{\gamma} \cdot \frac{\|\mu(v)\|_{\mathfrak{g}}}{\|v\|_2} \leq \frac{cap(v)}{\|v\|_2^2} \leq 1 - \frac{1}{2L} \cdot \frac{\|\mu(v)\|_{\mathfrak{g}}^2}{\|v\|_2^2}.$$

Assuming the normalization $\|v\|_2 = 1$ for simplicity, we can rewrite this in terms of the Kempf-Ness function $f_v(g) := \|g \cdot v\|_2^2$ as

$$1 - \frac{\|\nabla f_v(I_G)\|_{\mathfrak{g}}}{\gamma} \leq \inf_{g \in G} f_v(g) \leq 1 - \frac{\|\nabla f_v(I_G)\|_{\mathfrak{g}}^2}{2L},$$

where ∇ denotes the geodesic gradient for f_v.

These inequalities now allow us to transfer results between the group optimization and scaling problems. Explicitly, if we have a scaling algorithm that can output a point $v \neq 0$ with $\|\nabla_v\|_{\mathfrak{g}} < \gamma$, then this implies

$$cap(v) = \inf_{g \in G} f_v(g) \geq \|v\|_2^2 \Big(1 - \frac{\|\nabla_v\|_{\mathfrak{g}}}{\gamma}\Big) > 0,$$

which by Theorem 6.1.6 certifies that v is semi-stable, i.e. v is not in the null cone. Conversely, if we have a group optimization algorithm which can output a δ-optimizer g satisfying $f_v(g) = \|g \cdot v\|_2^2 \geq (1 - \delta)^{-1} cap(v)$, then this gives an approximate solution to the scaling problem as

$$\|\nabla f_v(g)\|_g^2 \leq 2L\Big(1 - \frac{cap(v)}{\|g \cdot v\|_2^2}\Big) \leq 2L\delta.$$

Now that we have this framework, the duality theory of Theorem 6.1.7 along with the geodesic convexity of the Kempf-Ness function suggests that we can borrow ideas from classical convex optimization in order to solve scaling problems. This perspective was a major contribution of [20] and allowed them to give new algorithms for a variety of scaling problems that were previously intractable, as well as to give a principled analysis for many known algorithms. We discuss these algorithmic results in more detail in Chapter 8.

Below we present some concrete instances of the scaling framework and discuss the consequences of Theorem 6.1.7 as they relate to the results in this thesis. We will specifically discuss known bounds on the parameters γ and L from Theorem 6.1.7.

In the work of Linial et al. [66] on the matrix scaling problem, it was shown γ^{-1} is bounded by a polynomial in the dimension, though they did not use this language. This was a key step in their strongly polynomial algorithm for matrix scaling. It turns out that a similar polynomial bound holds for the matrix balancing problem [76], which also has many well-known polynomial time algorithms.

For the non-commutative generalizations of frame and operator scaling, it had been known since the work of [45] that $\gamma^{-1} \lesssim dn$ for inputs in $\mathrm{Mat}(d, n)$, though once again this was not the language used. This bound was a key step in our preliminary bound $p(d, n, \varepsilon) \lesssim dn\varepsilon$ on the Paulsen problem in [62]. The polynomial bound for operator scaling also explains the polynomial time algorithm for operator scaling and its various downstream applications in algebraic complexity given in [38].

On the other hand, almost all other scaling problems are not known to have polynomial bounds for γ^{-1}. In fact, even for the next simplest case of 3-tensor scaling, a result of Kravtsov [60] shows that γ^{-1} must be exponential in the dimensions (see also the extension to higher order tensor scaling in [37]). This can be seen as analogous to the jump in difficulty from graph matching to hypergraph matching.

In Chapter 3, we were able to prove much stronger results for matrix scaling inputs that satisfied the strong convexity or pseudorandom conditions. Similarly, in Chapter 7 we will give very strong results for tensor scaling, but not in the worst case. We generalize the strongly convex and pseudorandom analyses of Chapter 3 to the tesor setting in order to bypass the worst-case convergence results given in Theorem 6.1.7. This is sufficient for our applications to the Paulsen problem (Chapter 4) and the tensor normal model (Chapter 9), as we can show random instances of tensor scaling satisfy these conditions with high probability.

Our subsequent analyses on the tensor scaling problem can be derived in a self-contained manner. The background presented in this section serves to motivate and contextualize the concepts used, but is not required to understand the main results in this thesis. Of course, many of the key observations and ideas for the analysis are heavily inspired by the theory of general scaling problems.

In the subsequent sections, we restrict our attention to the tensor scaling setting, though we can now use language that connects our work to the general setting of [20].

6.2 Tensor Scaling and Geodesic Convex Formulation

In this section, we will introduce the tensor scaling problem, which is a generalization of both matrix and frame scaling. The problem will use the language of classical Lie groups and Lie algebras, so we refer the reader to Section 2.2.3 for the relevant notation and definitions. This section provides the framework for our analyses in Chapter 7, where we provide quantitatively stronger analyses for inputs to the tensor scaling problem satisfying additional conditions.

In Section 6.2.1, we formally define the tensor scaling problem. We also prove some simple properties of scalings that will be useful for our optimization formulation. Then in Section 6.2.2, we explicitly define the Kempf-Ness function for tensor scaling. This function will be crucial to our analysis, as it provides a tractable optimization formulation for the tensor scaling problem. In Section 6.2.3, we define the notion of geodesic convexity on positive definite matrices. This reveals the proper geometry in which the Kempf-Ness function is convex. Finally, in Section 6.2.4, we formally show that the Kempf-Ness function gives a geodesic convex optimization formulation for tensor scaling. This will allow us to use tools from convex optimization in our analysis in Chapter 7.

6.2.1 Tensor Scaling Problem

The tensor scaling problem is a generalization of matrix scaling in two directions: the inputs are higher order tensors instead of matrices, and scalings are general matrices instead of diagonal ones. We eventually want to describe the general tensor scaling problem, which involves finding a scaling of a particular form in order to satisfy certain balance conditions on the input. We begin by giving some basic definitions about tensors.

Definition 6.2.1. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$. Then for tuple $x := \{x_1, ..., x_K\} \in V^K$, its size and associated operator are defined as*

$$s(x) := \sum_{k=1}^{K} \|x_k\|_2^2, \qquad \textit{and} \qquad \rho_x := \sum_{k=1}^{K} x_k x_k^*,$$

where $\|\cdot\|_2$ is the standard Euclidean norm on V. Note that $\rho_x \succeq 0$ is positive semidefinite and $\mathrm{Tr}[\rho_x] = \sum_{k=1}^{K} \|x_k\|_2^2 = s(x)$.

We recall that the input to matrix and frame scaling are (tuples of) elements from $\mathrm{Mat}(d, n)$ which is isomorphic to the tensor product $\mathbb{F}^d \otimes \mathbb{F}^n$ by $A \to \mathrm{vec}(A)$. The size of tensor tuple $\{\mathrm{vec}(A_1), ..., \mathrm{vec}(A_k)\} \in \mathbb{F}^d \otimes \mathbb{F}^n$ as given in Definition 6.2.1 is consistent with the size of matrix tuple $A \in \mathrm{Mat}(d, n)^K$ as given by Definition 3.1.1.

The following notion of tensor marginals gives an analogous generalization of row and column sums of matrix tuples.

Definition 6.2.2. *Consider linear operator $\rho \in L(V)$ on tensor product $V = \otimes_{a\in[m]} V_a$, and any $S \subseteq [m]$ with $V_S := \otimes_{a\in S} V_a$. Then the S-marginal of ρ is defined as $\rho^S := \mathrm{Tr}_{\overline{S}}[\rho]$, where $\mathrm{Tr}_{\overline{S}}$ denotes the partial trace over the complement $V_{\overline{S}}$ as given in Definition 2.4.7. For small subsets e.g. $S = \{a\}$ or $S = \{a, b, c\}$, we use shorthand $\rho^{(a)}$ or $\rho^{(abc)}$.*

The appropriate notion of balance for tensors will be defined using these marginals. Below, we show how this definition generalizes the row and column sums of frames and matrices, which were the quantities of interest in the matrix and frame scaling problems.

Consider tuple $A = \{\mathrm{vec}(A_1), ..., \mathrm{vec}(A_k)\} \in (\mathbb{F}^d \otimes \mathbb{F}^n)^K$. Then the associated operator according to Definition 6.2.1 is

$$\rho_A := \sum_{k=1}^{K} \mathrm{vec}(A_k)\,\mathrm{vec}(A_k)^*.$$

If we use $\rho_A^L \in L(d), \rho_U^R \in L(n)$ to denote the left and right marginals, then these can be explicitly calculated as follows. For the left marginal, we calculate the inner product with $E_{ii} = e_i e_i^* \in L(d)$ for standard basis $\{e_i\}_{i\in[d]} \subseteq \mathbb{F}^d$ as

$$\langle \rho_A, E_{ii} \otimes I_n \rangle = \Big\langle \sum_{k=1}^{K} \mathrm{vec}(A_k)\,\mathrm{vec}(A_k)^*, \sum_{j=1}^{n} (e_i \otimes e_j)(e_i \otimes e_j)^* \Big\rangle = \sum_{k=1}^{n} \sum_{j=1}^{n} |(A_k)_{ij}|^2,$$

where the first step was by the formula above for ρ_A and the decomposition $I_n = \sum_{j=1}^{n} E_{jj}$ for standard basis $\{e_j\}_{j\in[n]} \subseteq \mathbb{F}^n$. Definition 2.4.7 tells us that $\rho_A^L = \mathrm{Tr}_R[\rho_A]$ is the unique operator satisfying $\langle \rho_A^L, X \rangle = \langle \rho_A, X \otimes I_n \rangle$ for all $X \in L(d)$. Matching this to the above, we see that the diagonals of ρ_A^L are exactly the row sums $\mathrm{diag}\{r_i(A)\}_{i=1}^{d}$ according to Definition 3.1.1. The symmetric calculation shows that the diagonals of $\rho_A^R = \mathrm{diag}\{c_j(A)\}_{j=1}^{n}$ are exactly the column sums.

Next, we show how this generalizes the row and column sums of Definition 4.2.3 for frames. Consider frame $\{u_1, ..., u_n\} \in \mathcal{U}^n$ for inner product space $\mathcal{U}$. This can equivalently be viewed as the tuple $U := \{u_1 \otimes e_1, ..., u_n \otimes e_n\} \in (\mathcal{U} \otimes \mathbb{R}^n)^n$ for standard basis $\{e_j\}_{j=1}^{n}$. Then the associated operator according to Definition 6.2.1 is

$$\rho_U := \sum_{j=1}^{n} (u_j \otimes e_j)(u_j \otimes e_j)^* = \sum_{j=1}^{n} (u_j u_j^* \otimes E_{jj}).$$

We use $\rho_U^L \in L(\mathcal{U}), \rho_U^R \in L(n)$ to denote the left and right marginals according to Definition 6.2.2, and compute them explicitly as follows. For the left marginal, we calculate the inner product with arbitrary $X \in L(\mathcal{U})$ as

$$\langle \rho_U, X \otimes I_n \rangle = \sum_{j=1}^{n} \langle u_j u_j^* \otimes E_{jj}, X \otimes I_n \rangle = \sum_{j=1}^{n} \langle u_j u_j^*, X \rangle,$$

where the first step was by the formula above for ρ_U, and the second step was by definition of tensor products. Definition 2.4.7 tells us that $\rho_U^L = \mathrm{Tr}_{\mathbb{R}}[\rho_U]$ is the unique operator satisfying $\langle \rho_U^L, X \rangle = \langle \rho_U, X \otimes I_n \rangle$ for all $X \in L(\mathcal{U})$, so matching this to the above we see

$$\rho_U^L = \sum_{j=1}^{n} u_j u_j^*.$$

A similar calculation shows $\rho_U^R = \mathrm{diag}\{\|u_j\|_2^2\}_{j=1}^n$. These exactly produce the row and column marginals of a frame given in Definition 4.2.3.

Recall that both frames and matrices were defined by elements of $\mathrm{Mat}(d, n)$. The matrix and frame settings were different as the left marginal for frames could be scaled by arbitrary matrices instead of diagonal ones, and we required a stronger balance condition on this left marginal for frames in Definition 4.1.2 instead of just on the diagonals in Definition 3.1.2.

To properly generalize these to the tensor setting, we need to specify a set of scaling operations and balance conditions. These will be defined based on the following notion of a scaling group. The definition is a bit long and technical, and the reader can keep in mind the matrix scaling setting, where we acted on the left and right by diagonal matrices.

Definition 6.2.3 (Tensor Scaling Group). *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces over field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. A tensor scaling group on V is defined as*

- *$G = (G_1, ..., G_m)$ where each G_a is a choice of one of the following groups:*

  $$\text{(non-commutative)} \qquad G_a = \mathrm{SL}_{\mathbb{F}}(V_a), \qquad \text{OR}$$

  $$\text{(commutative)} \qquad G_a = ST_{\mathbb{F}}^{\Xi^a}(V_a) \quad \text{with} \begin{cases} \Xi^a \in \mathrm{SU}(V_a) & \text{if } \mathbb{F} = \mathbb{C} \\ \Xi^a \in \mathrm{SO}(V_a) & \text{if } \mathbb{F} = \mathbb{R} \end{cases}.$$

  G has the natural embedding $G \to \{g_1 \otimes ... \otimes g_m \mid g_a \in G_a\} \subseteq \mathrm{SL}_{\mathbb{F}}(V)$, which is in fact gives an isomorphism of the group structure by component-wise multiplication. We will often use G to refer to this embedding $G \subseteq \mathrm{SL}_{\mathbb{F}}(V)$ by abuse of notation.

The polar part of scaling group G is denoted $(P, \mathfrak{p})$, and is defined as

- $P = (P_1, ..., P_m)$ *where P_a is the polar part of G_a according to Theorem 2.1.13. Explicitly,*

  $$\begin{aligned} &\textit{(non-commutative)} && G_a = \mathrm{SL}_{\mathbb{F}}(V_a) \implies P_a = \mathrm{SPD}_{\mathbb{F}}(V_a), \\ &\textit{(commutative)} && G_a = ST_{\mathbb{F}}^{\Xi^a}(V_a) \implies P_a = \mathrm{ST}_{+}^{\Xi^a}(V_a). \end{aligned}$$

  P has the induced embedding $P \to \{p_1 \otimes ... \otimes p_m \mid p_a \in P_a\} \subseteq \mathrm{SPD}_{\mathbb{F}}(V)$, which is in fact the polar part of the embedding $G \subseteq \mathrm{SL}_{\mathbb{F}}(V)$ by Theorem 2.1.13. We will use P to refer to this embedding $P \subseteq \mathrm{SPD}_{\mathbb{F}}(V)$ by abuse of notation.

- $\mathfrak{p} = \mathfrak{p}_1 \oplus ... \oplus \mathfrak{p}_m$ *where $\mathfrak{p}_a = \log P_a$ is the associated vector space according to the discussion in Section 2.2.3 and Section 2.2.2 for non-commutative and commutative groups respectively. Explicitly,*

  $$\begin{aligned} \textit{(non-commutative)} \quad & P_a = \mathrm{SPD}_{\mathbb{F}}(V_a) \implies \mathfrak{p}_a = \mathfrak{spd}_{\mathbb{F}}(V_a), \\ \textit{(commutative)} \quad & P_a = \mathrm{ST}_{+}^{\Xi^a}(V_a) \implies \mathfrak{p}_a = \mathfrak{st}_{+}^{\Xi^a}(V_a), \end{aligned}$$

  where $\mathfrak{spd}(V) := \log \mathrm{SPD}(V)$ is explicitly given in Eq. (2.7). $\mathfrak{p}$ has the induced embedding

  $$\mathfrak{p} \to \{Z_1 \otimes I_{\overline{1}} + ... + I_{\overline{m}} \otimes Z_m \mid Z_{a\in[m]} \in \mathfrak{p}_{a\in[m]}\},$$

  where $I_{\overline{a}}$ the identity operator on $\otimes_{b\neq a} V_b$. This is the associated vector space of the embedding $P \subseteq \mathrm{SPD}_{\mathbb{F}}(V)$ as discussed in Section 2.2.3, i.e. $\mathfrak{p} = \log P$.

Matrix scaling is specified by choosing $G = (\mathrm{ST}(d), \mathrm{ST}(n))$ in Definition 6.2.3. In Definition 3.1.5, we restricted the set of scalings to the vector space $\mathfrak{t} = \mathfrak{st}(d) \oplus \mathfrak{st}(n)$ by the simple change of variables $x \to e^x$ without loss of generality. Similarly, frame scaling is specified by $G = (\mathrm{SL}(d), \mathrm{ST}(n))$, and in Definition 4.2.9 we reduced the set of scalings to $P = (\mathrm{SPD}(d), \mathrm{ST}_+(n))$ and its Lie algebra $\mathfrak{p} = \mathfrak{spd}(d) \oplus \mathfrak{st}_+(n)$ by the same change of variables. Operator scaling [38] is specified by the choice $G = (\mathrm{SL}(d), \mathrm{SL}(n))$, and it can be shown that the set of scalings can similarly be reduced to the positive definite matrices. In general, the set of scalings for the tensor setting can also be reduced to the polar parts $(P, \mathfrak{p})$. In Section 6.2.3, we can define a natural geometry on P which will allow us to give a tractable optimization formulation for the tensor scaling problem in Definition 6.2.5 below. Similarly, the reduction to $\mathfrak{p} = \log P$ will allow us to use standard convex analysis on vector spaces to analyze this more general group optimization setting.

We can now define the balance condition for tensors. This will depend on the choice of scaling group in Definition 6.2.3.

Definition 6.2.4. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$ with choice of scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then tuple $x = \{x_1, ..., x_K\} \in V^K$ is ε-G-balanced if for every $a \in [m]$:*

$$\frac{1-\varepsilon}{d_a} s(x) I_a \preceq \rho_x^{(a)} \preceq \frac{1+\varepsilon}{d_a} s(x) I_a \quad \textit{if } G_a = \mathrm{SL}(d_a),$$
$$\frac{1-\varepsilon}{d_a} s(x) I_a \preceq \mathrm{diag}^{\Xi^a}(\rho_x^{(a)}) \preceq \frac{1+\varepsilon}{d_a} s(x) I_a \quad \textit{if } G_a = \mathrm{ST}^{\Xi^a}(d_a),$$

where diag^{Ξ} is the diagonal projection in the Ξ basis as given in Section 2.2.2. x is called G-balanced if the above holds with $\varepsilon = 0$.

In our discussion after Definition 6.2.2, we showed that for input $\{\mathrm{vec}(A_1), ..., \mathrm{vec}(A_K)\} \in (\mathbb{F}^d \otimes \mathbb{F}^n)^K$, the diagonal entries of ρ_A^L and ρ_A^R corresponded exactly to the row and column sums of matrix tuple $A \in \mathrm{Mat}(d, n)^K$ given in Definition 3.1.1. Therefore, A is an ε-doubly balanced matrix according to Definition 3.1.2 iff it is ε-G-balanced for $G = (\mathrm{ST}(d), \mathrm{ST}(n))$ according to Definition 6.2.4. Similarly, Definition 4.1.2 of the ε-doubly balanced frame condition corresponds exactly to the ε-G-balance condition for $G = (\mathrm{SL}(d), \mathrm{ST}(n))$. And Definition 2.12 in [63] of an ε-doubly balanced operator corresponds to the ε-G-balance condition for $G = (\mathrm{SL}(d), \mathrm{SL}(n))$.

Now we can collect the above into a formal definition of the tensor scaling problem.

Definition 6.2.5 (Tensor Scaling Problem)**.** *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for input $x = \{x_1, ..., x_K\} \in V^K$, output scaling $g = \otimes_{a\in[m]} g_a \in G$ such that*

$$g \cdot x := \{g \cdot x_k\}_{k=1}^K = \{\otimes_{a\in[m]} g_a \cdot x_k\}_{k=1}^K$$

is G-balanced according to Definition 6.2.4.

Note that from here onwards, we will repeatedly (and implicitly) use the isomorphism $(g_1, ..., g_m) \to g_1 \otimes ... \otimes g_m$ as described in Definition 6.2.3.

This gives a common generalization of matrix, frame, and operator scaling, as well as many other problems in the scaling framework. The first three are known to be solvable in polynomial time. Similarly, if T is a commutative group then this reduces to geometric programming, which can be solved using standard convex optimization techniques (see e.g. [21]). On the other hand, for all non-commutative tensor scaling problems with $m \geq 3$, the framework of [20] requires exponential time even to decide whether there exists a non-trivial $g \neq 0$ for which $g \cdot x$ is G-balanced. This is because the weight margin γ described in Theorem 6.1.7 is exponentially small for tensor scaling, and this parameter controls the algorithmic analysis of [20]. In Chapter 9, we will consider the setting $V = \otimes_{a \in [m]} \mathbb{R}^{d_a}$ with scaling group $G = (\mathrm{SL}(d_1), ..., \mathrm{SL}(d_m))$ for our statistical application. Importantly, our inputs will come from natural random distributions which we show satisfy certain strong convexity and pseudorandom conditions. This allows us to give strong bounds on the solution and algorithms to compute them in this beyond worst-case setting.

To finish this subsection, we give a useful property of the balance conditions which will allow us to restrict the scalings to the polar $(P, \mathfrak{p})$ and eventually will be useful in our optimization formulation. This result generalizes the discussion in Definition 3.1.5 and Fact 4.2.7 showing we could restrict scalings to the polar components for the matrix and frame settings respectively.

Lemma 6.2.6 (Equivariance and Invariance)**.** *The following properties hold for input $x \in V^K$ with $V = \otimes_{a \in [m]} V_a$.*

1. *For scaling $g \in \mathrm{GL}(V)$, the associated operator satisfies $\rho_{g \cdot x} = g \rho_x g^*$.*

2. *For any $S \subseteq [m]$, the S-marginal also satisfies the following equivariance: for any $g_S \in \mathrm{GL}(V_S)$ where $V_S := \otimes_{a \in S} V_a$,*
$$\rho^S_{(g_S \otimes I_{\overline{S}}) \cdot x} = g_S \rho^S_x g^*_S.$$

3. *Let $(G, P, \mathfrak{p})$ be a choice of scaling group according to Definition 6.2.3, and let $U := G \cap \mathrm{SU}(V)$ be the subset of unitary operators. Then x is ε-G-balanced iff $u \cdot x$ is ε-G-balanced for any $u \in U$, i.e. the balance conditions in Definition 6.2.4 is invariant under $G \cap \mathrm{SU}(V)$.*

Proof. The first statement follows by expanding Definition 6.2.1 of ρ:
$$\rho_{g \cdot x} = \sum_{k=1}^{K} (g x_k)(g x_k^*) = g \rho_x g^*.$$

We prove the second statement in the special case when $S = \{a\}$ is a singleton. Note that the general case follows by considering the partition $V = V_S \otimes V_{\overline{S}}$. To do so, we verify that $g_a \rho_x^{(a)} g_a^*$ matches Definition 6.2.2 of the marginal. So consider arbitrary $X \in L(V_a)$ and calculate the inner product
$$\langle \rho_{(g_a \otimes I_{\overline{a}}) \cdot x}, X \otimes I_{\overline{a}} \rangle = \langle \rho_x, g_a^* X g_a \otimes I_{\overline{a}} \rangle = \langle \rho_x^{(a)}, g_a^* X g_a \rangle = \langle g_a \rho_x^{(a)} g_a^*, X \rangle,$$

where in the first step we used $\rho_{g\cdot x} = g\rho_x g^*$ as shown above, and the second step was by Definition 2.4.7 of marginals. Since $\rho^{(a)}_{(g_a\otimes I_{\overline{a}})\cdot x}$ is uniquely defined by the equation

$$\langle \rho^{(a)}_{(g_a\otimes I_{\overline{a}})\cdot x}, X\rangle = \langle \rho_{(g_a\otimes I_{\overline{a}})\cdot x}, X\otimes I_{\overline{a}}\rangle$$

for all $X \in L(V_a)$, this matches with $g_a \rho^{(a)}_x g_a^*$ and the second statement is shown.

Now we prove the third statement. Consider $u_1 \otimes ... \otimes u_m \in U$, which we decompose as

$$u_1 \otimes ... \otimes u_m = \prod_{a=1}^{m} (u_a \otimes I_{\overline{a}}).$$

We will show that each individual transformation maintains the G-balance property.

So consider $u := u_a \otimes I_{\overline{a}}$, and we claim that $\rho_{u\cdot x}$ is ε-G-balanced iff ρ_x is ε-G-balanced. We first show that the size does not change. This follows simply as

$$s(u\cdot x) = \sum_{k=1}^{K} \|u\cdot x_k\|_2^2 = \sum_{k=1}^{K} \|x_k\|_2^2 = s(x),$$

where the first and last steps were by Definition 6.2.1, and the middle step was because $u \in \mathrm{SU}(V)$ is an isometry according to Definition 2.1.11 so Euclidean norm $\|\cdot\|_2$ is invariant by definition.

Now we show that every $b \neq a$ marginal is unchanged. So consider arbitrary $Z_b \in L(V_b)$ and calculate the inner product

$$\langle \rho_{u\cdot x}, Z_b \otimes I_{\overline{b}}\rangle = \langle \rho_x, u_a^* u_a \otimes Z_b \otimes I_{\overline{ab}}\rangle = \langle \rho^{(b)}_x, Z_b\rangle,$$

where the first step was by the equivariance property $\rho_{u\cdot x} = u\rho_x u^*$, and in the last step we used $u_a \in G_a \subseteq \mathrm{SU}(V_a)$ so $u_a^* u_a = I_a$ by Definition 2.1.11. This shows $\rho^{(b)}_{u\cdot x} = \rho^{(b)}_x$ since the marginal is uniquely defined by the equation

$$\langle \rho^{(b)}_{u\cdot x}, Z_b\rangle = \langle \rho_{u\cdot x}, Z_b \otimes I_{\overline{b}}\rangle.$$

Now we show that the a-th marginal $\rho^{(a)}_x$ is ε-balanced iff $\rho^{(a)}_{u\cdot x}$ is. Define $S_a := \{\xi \in V_a \mid \|\xi\|_2^2 = 1\}$ if $G_a = \mathrm{SL}(V_a)$ and $S_a := \{\xi_i \in \Xi\}$ if $G_a = \mathrm{ST}^{\Xi}(V_a)$, and consider arbitrary $\xi \in S_a$. Then for arbitrary $y \in V^K$, we can rewrite the constraint in Definition 6.2.4 of the ε-G-balance condition for the a-th marginal as

$$\frac{1-\varepsilon}{d_a} s(y) I_a \preceq \rho^{(a)}_y \preceq \frac{1+\varepsilon}{d_a} s(y) I_a \quad \text{iff} \ \sup_{\xi\in S_a} |\langle \xi\xi^*, d_a\rho^{(a)}_y - s(y)I_a\rangle| \leq s(y)\varepsilon, \tag{6.1}$$

where we used that S_a is the sphere for $G_a = \mathrm{SL}(V_a)$ and the standard basis for $G_a = \mathrm{ST}^{\Xi}(V_a)$. For $y = (u_a \otimes I_{\overline{a}}) \cdot x$ with $u_a \in G_a$, we have

$$\sup_{\xi \in S_a} |\langle d_a \rho_{u \cdot x}^{(a)} - s(u \cdot x) I_a, \xi \xi^* \rangle| = \sup_{\xi \in S_a} |\langle d_a \rho_x^{(a)} - s(x) I_a, u_a^* \xi \xi^* u_a \rangle| = \sup_{\psi \in S_a} |\langle d_a \rho_x^{(a)} - s(x) I_a, \psi \psi^* \rangle|,$$

where in the first step we used the equivariance $\rho_{u \cdot x}^{(a)} = u_a \rho_x^{(a)} u_a^*$ for the first term and $s(u \cdot x) = s(x)$ shown in the calculation above for the second term, and the second step was by the change of variable $\psi := u_a^* \xi$ as $u_a \in G_a$ preserves the sphere S_a. Matching this to Eq. (6.1), we see that x is ε-balanced in the a-th marginal iff $u \cdot x$ is. Since the other $b \neq a$ marginals were invariant as shown above, this verifies that x is ε-G-balanced according to Definition 6.2.4 iff $u \cdot x$ is ε-G-balanced. Finally, we can apply this iteratively for each part to show the third statement for arbitrary $u \in U$. □

By the polar decomposition in Theorem 2.1.13, we can factor $G = U \cdot P$ for unitary part U and polar part P. The unitary invariance of the balance condition shown in Lemma 6.2.6(3) implies that we do not lose anything by restricting our scalings to P. In Section 6.2.3 we will provide a geometry on the positive definite matrices in P. This will allow us to use the theory of Kempf-Ness functions in algebraic geometry to give a tractable optimization formulation for finding the solution to the tensor scaling problem.

6.2.2 Kempf-Ness Function

In this subsection, we will formally define the Kempf-Ness function [58] for tensor scaling. This will give the optimization formulation we use to analyze the tensor scaling problem. For background on this function in geometric invariant theory, see Section 6.1.2.

Definition 6.2.7. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a \in [m]}$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for tuple $x := \{x_1, ..., x_K\} \in V^K$, the Kempf-Ness function $\tilde{f}_x^G : G \to \mathbb{R}_+$ is defined as*

$$\tilde{f}_x^G(g) := s(g \cdot x) = Tr[\rho_{g \cdot x}] = \langle \rho_x, g^* g \rangle.$$

The goal of the tensor scaling problem in Definition 6.2.5 is to find a G-balanced scaling of the input. In Lemma 6.2.6(3), we showed that if there is a balanced scaling $y \in G \cdot x$, then we can assume without loss that $y \in P \cdot x$. A similar invariance property holds for the Kempf-Ness function.

Fact 6.2.8. *The tensor Kempf-Ness function is unitarily invariant, i.e. the value of $\tilde{f}_x^G$ at g depends only on $g^* g \in P$.*

As a consequence, this function is also well-defined on P, which will be useful for our optimization formulation given in Section 6.2.4

Definition 6.2.9. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for tuple $x := \{x_1, ..., x_K\} \in V^K$, the Kempf-Ness function can be equivalently defined as $f_x^P : P \to \mathbb{R}_+$, where*

$$f_x^P(p) := \langle \rho_x, p \rangle.$$

Note that $\tilde{f}_x^G(g) = f_x^P(g^* g)$ and $f_x^P(p) = \tilde{f}_x^G(p^{1/2})$.

This last line is why we give different names to $\tilde{f}^G, f^P$. Specifically, note that $P \subseteq G$ so the domain of f^P is contained in the domain of $\tilde{f}^G$, but they may have different values $f_x^P(p) \neq \tilde{f}_x^G(p)$ for $p \in P$ on this common domain. Therefore, we will tend to use f^P exclusively for positive definite elements to avoid confusion.

In the next subsection, we will provide a geometry on positive definite matrices P which will allow us to do calculus on the Kempf-Ness function. The geometry is more straightforward when the base point is the identity, so we will repeatedly use the following property to simplify calculations.

Fact 6.2.10 (Equivariance)**.** *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. For tuple $x := \{x_1, ..., x_K\} \in V^K$, the Kempf-Ness functions in Definition 6.2.7 and Definition 6.2.9 satisfy the following relations:*

$$\tilde{f}_x^G(g) = s(g \cdot x) = \tilde{f}_{g\cdot x}^G(I_V), \qquad \text{and} \qquad f_x^P(p) = \langle \rho_x, p \rangle = f_{p^{1/2}\cdot x}^P(I_V)$$

where $I_V = \otimes_{a\in[m]} I_a$ is the identity element of $G \subseteq \mathrm{GL}(V)$.

In the following section, we define the notion of geodesic convexity on positive definite matrices. This will allow us to better understand the Kempf-Ness function f^P in Definition 6.2.9 and eventually show that it gives a tractable optimization perspective for the tensor scaling problem.

6.2.3 Calculus for Positive Definite Operators

This subsection will introduce the geodesic framework that reveals the underlying convexity of the Kempf-Ness function in Definition 6.2.9. This subsection simply lifts the results of Section 2.2.4 to the tensor scaling setting.

The domain of the Kempf-Ness function in Definition 6.2.9 is a tensor product of subsets of positive definite matrices. It turns out that the geodesic curves from Definition 2.2.4 lift naturally to this tensor setting.

Fact 6.2.11. *For tensor product $V = \otimes_{a\in[m]} V_a$, let $(G, P, \mathfrak{p})$ be a scaling group according to Definition 6.2.3. Then P is closed under the geodesics given in Definition 2.2.4. Explicitly, for any $p, q \in P$ and $Z \in \mathfrak{p}$,*

$$\gamma_{p,q}(\eta) = \otimes_{a\in[m]} p_a^{1/2}(p_a^{-1/2} q_a p_a^{-1/2})^{\eta} p_a^{1/2}, \qquad \text{and} \qquad \gamma_p(Z) = \otimes_{a\in[m]} p_a^{1/2} e^{Z_a} p_a^{1/2},$$

where we have used the embeddings given in Definition 6.2.3: $p \to p_1 \otimes, ..., \otimes p_m \in P$, and $Z \to Z_1 \otimes I_{\overline{1}} + ... + Z_m \otimes I_{\overline{m}} \in \mathfrak{p}$.

The symmetry properties of geodesics also lift to this tensor setting by applying Fact 2.2.5 component-wise.

Fact 6.2.12. *Consider tensor product $V = \otimes_{a\in[m]} V_a$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. For any $p, q \in P$, the geodesics satisfy $\gamma_{p,q}(\eta) = \gamma_{q,p}(1-\eta)$ for any $\eta \in [0, 1]$. Further, $\|\log p^{-1/2} q p^{-1/2}\| = \|\log q^{-1/2} p q^{-1/2}\|$ for any unitarily invariant norm $\|\cdot\|$ on $\mathfrak{p}$.*

Since our scaling groups are just direct products of the very simple groups SL(d) and ST(d), the above result follows immediately from our embeddings, so we omit the proofs. This above is a special case of the Cartan decomposition in the general Lie group setting. For details, see the book of Wallach [97].

The convexity of the matrix Kempf-Ness function was a crucial ingredient in our analysis of Chapter 3. The following definition generalizes the notion of a convex function on a vector space to this geodesic setting and will be equally crucial in our analysis of tensor scaling in Chapter 7.

Definition 6.2.13 (Geodesic Convexity)**.** *Let $V = \otimes_{a\in[m]}$ be a tensor product of inner product spaces, and let $(G, P, \mathfrak{p})$ be a scaling group according to Definition 6.2.3. Then function $f : P \to \mathbb{R}$ is geodesically convex if for every $p, q \in P$, the univariate restriction $\eta \to f(\gamma_{p,q}(\eta))$ is convex according to Definition 2.3.1.*

Given norm $\|\cdot\|$ on $\mathfrak{p}$, f is α-geodesically strongly convex at $p \in P$ with respect to $\|\cdot\|$ iff for every $Z \in \mathfrak{p}$,

$$\partial^2_{\eta=0} f(\gamma_p(\eta Z)) \geq \alpha \|Z\|^2_{\mathfrak{p}},$$

where $\gamma_p(\eta Z) = p^{1/2} e^{\eta Z} p^{1/2}$ is the geodesic given in Fact 6.2.11.

In the following subsection, we will show that the Kempf-Ness function for tensor scaling is geodesically convex. This will allow us to use tools from convex optimization to analyze the tensor scaling solution.

As an example, we can relate critical points and optimizers of geodesically convex functions similar to the result of Lemma 2.3.4 for univariate functions.

Definition 6.2.14. *Let $V = \otimes_{a \in [m]}$ be a tensor product of inner product spaces, and let $(G, P, \mathfrak{p})$ be a scaling group according to Definition 6.2.3. Then $p \in P$ is a critical point of function $f : P \to \mathbb{R}$ iff*

$$\forall q \in P : \partial_{\eta=0} f(\gamma_{p,q}(\eta))) = 0.$$

This condition can be equivalently written as

$$\forall Z \in \mathfrak{p} : \partial_{\eta=0} f(\gamma_p(\eta Z)) = \partial_{\eta=0} f(p^{1/2} e^{\eta Z} p^{1/2}) = 0.$$

The following lemma generalizes the natural property of convex functions: that local minimizers are global minimizers.

Lemma 6.2.15. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces, and let $(G, P, \mathfrak{p})$ be a scaling group according to Definition 6.2.3. For geodesically convex function $f : P \to \mathbb{R}$, $p \in P$ is a critical point iff it is a global minimizer of f.*

Proof. If $p \in P$ is the optimizer of f, then in particular for any $Z \in \mathfrak{p}$ we must have

$$\partial_{\eta=0} f(\gamma_p(\eta Z)) \geq 0, \qquad \text{and} \qquad -\partial_{\eta=0} f(\gamma_p(\eta Z)) = \partial_{\eta=0} f(\gamma_p(-\eta Z)) \geq 0.$$

This implies $\partial_{\eta=0} f(\gamma_p(tZ)) \geq 0$ for every $Z \in \mathfrak{p}$ and verifies criticality of p according to Definition 6.2.14.

Conversely, assume $p \in P$ is a critical point of f, and consider arbitrary $q \in P$. By univariate convexity, we have

$$f(q) - f(p) = f(\gamma_{p,q}(1)) - f(\gamma_{p,q}(0)) \geq (1-0)\partial_{\eta=0} f(\gamma_{p,q}(\eta)) = 0,$$

where the first step was by Definition 2.2.4, the second step was by the 1-st order condition of Definition 2.3.2 applied to univariate convex function $t \to f(\gamma_{p,q}(\eta))$, and the final step is because p is a critical point. As $q \in P$ was arbitrary, p is a global minimimizer of f. □

Recall that in Lemma 3.1.8 we were able to show that critical points of the matrix Kempf-Ness function correspond to doubly balanced scalings. This implied that the matrix Kempf-Ness function gives a convex formulation for matrix scaling as shown in Proposition 3.1.10. We will show a similar result for the more general tensor case in the following subsection using Lemma 6.2.15.

6.2.4 Geodesic Convex Formulation for Tensor Scaling

In this subsection we show that the Kempf-Ness function in Definition 6.2.9 gives a geodesically convex optimization formulation for the tensor scaling problem in Definition 6.2.5. These results are well-known in the geometric invariant theory literature (see [58], [40], [73], [11], [20]), and follow by straightforward derivative calculations as shown below.

We first show that G-balanced scalings correspond to critical points of f^P. This generalizes Lemma 3.1.8 for the matrix scaling problem.

Lemma 6.2.16. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. For tuple $x := \{x_1, ..., x_K\} \in V^K$ and $g \in G$, $g \cdot x$ is a G-balanced scaling of x iff $p := g^*g$ is a critical point of $f_x^P(p)$ according to Definition 6.2.14.*

Proof. We will show that $y \in V$ is G-balanced iff the identity I_V is a critical point for f_y^P. This suffices to show the lemma as

$$\partial_{\eta=0} f_x^P(\gamma_p(\eta Z)) = \partial_{\eta=0}\langle \rho_x, p^{1/2} e^{\eta Z} p^{1/2}\rangle = \partial_{\eta=0}\langle \rho_{p^{1/2}\cdot x}, e^{\eta Z}\rangle = \partial_{\eta=0} f_{p^{1/2}\cdot x}(\gamma_{I_V}(\eta Z)),$$

where the first and last steps were by Definition 6.2.9 of the Kempf-Ness function and Fact 6.2.11 of geodesics on P, and the second step was by the equivariance property of ρ shown in Lemma 6.2.6(1). Therefore p is critical for f_x^P iff I_V is critical for $f_{p^{1/2}\cdot x}^P$.

We first calculate the first order derivative of f_y^P as

$$\begin{aligned}\partial_{\eta=0} f_y^P(e^{\eta Z}) &= \partial_{\eta=0}\langle \rho_y, e^{\eta Z}\rangle = \langle \rho_y, Ze^{\eta Z}\rangle|_{\eta=0} \\ &= \Big\langle \rho_y, \sum_{a\in[m]} Z_a \otimes I_{\overline{a}} \Big\rangle = \sum_{a\in[m]} \langle \rho_y^{(a)}, Z_a\rangle = \sum_{a\in[m]} \Big\langle \rho_y^{(a)} - \frac{s(y)}{d_a} I_a, Z_a \Big\rangle, \end{aligned} \tag{6.2}$$

where the first step was by Definition 6.2.9 of the Kempf-Ness function, the second step was by standard matrix calculus $\partial_\eta e^{\eta Z} = Ze^\eta Z$ and the embedding $Z \to \sum_{a\in[m]} Z_a \otimes I_{\overline{a}}$ given in Definition 6.2.3 for $\mathfrak{p}$, in the fourth step we used Definition 6.2.2 of marginals of

ρ, and in the final step we used the fact that for every $a \in [m]$, $Z_a \in \mathfrak{p}_a \subseteq \mathfrak{spd}(V_a)$ so $\langle I_a, Z_a \rangle = \mathrm{Tr}[Z_a] = 0$ by Definition 2.1.10.

Using this formula, we first show that the balance condition implies criticality. So assume y is G-balanced, and first consider the case $G_a = \mathrm{SL}(V_a)$. Then, according to Definition 6.2.4, $\rho_y^{(a)} - \frac{s(y)}{d_a} I_a = 0$ so the a-th term above vanishes. In the other case $G_a = ST^{\Xi}(V_a)$, the G-balance condition in Definition 6.2.4 implies that $\mathrm{diag}^{\Xi}(\rho_y^{(a)} - \frac{s(y)}{d_a} I_a) = 0$. Since $Z_a \in \mathfrak{p}_a$, it is also diagonal in the Ξ basis, so the a-th term vanishes in this case as well. Therefore, for G-balanced y, the entire derivative vanishes in Eq. (6.2). Since $Z \in \mathfrak{p}$ was arbitrary, this verifies Definition 6.2.14 showing I_V is a critical point of f_y^P.

Conversely, assume y is not G-balanced. We will exhibit a $Z \in \mathfrak{p}$ such that the derivative $\partial_{\eta=0} f_y^P(e^{\eta Z}) \neq 0$, which implies that I_V is not a critical point of f_y^P. In the case when $G_a = SL(V_a)$ we choose $Z_a := \rho_y^{(a)} - \frac{s(y)}{d_a} I_a$, and in the case when $G_a = ST^{\Xi}(V_a)$, we choose its diagonal projection $Z_a = \mathrm{diag}^{\Xi}(\rho_y^{(a)} - \frac{s(y)}{d_a} I_a)$. By construction, $Z = \sum_{a \in [m]} Z_a \otimes I_{\overline{a}} \in \mathfrak{p}$, and since y is not G-balanced, $Z \neq 0$. Therefore, we can calculate

$$\partial_{\eta=0} f_y^P(e^{\eta Z}) = \sum_{a \in [m]} \left\langle \rho_y^{(a)} - \frac{s(y)}{d_a} I_d, Z_a \right\rangle = \sum_{a \in [m]} \|Z_a\|_F^2 > 0,$$

where the first step was shown in Eq. (6.2), in the second step we used the definition of Z, and the last inequality is strict as y is not G-balanced so $Z \neq 0$. This shows the identity is not a critical point of f_y^P according to Definition 6.2.14. □

Next we show that the Kempf-Ness function is geodesically convex everywhere.

Lemma 6.2.17. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a \in [m]}$ with choice of scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for any tuple $x := \{x_1, ..., x_K\} \in V^K$, the Kempf-Ness function f_x^P is geodesically convex on P according to Definition 6.2.13.*

Proof. We will show that f_y^P is geodesically convex at the identity for any choice of $y \in V^K$. This will suffice to show the lemma as, for any $p \in P$ and $Z \in \mathfrak{p}$, $\partial^2_{\eta=0} f_x^P(\gamma_p(\eta Z)) = \partial^2_{\eta=0} f_{p^{1/2} \cdot x}(e^{\eta Z})$ by the equivariance property of Fact 6.2.8, so f_x^P is geodesically convex at p iff $f_{p^{1/2} \cdot x}$ is geodesically convex at the identity.

For any $Z \in \mathfrak{p}$, we calculate

$$\partial^2_{\eta=0} f_y^P(e^{\eta Z}) = \partial^2_{\eta=0} \langle \rho_y, e^{\eta Z} \rangle = \langle \rho_y, Z^2 \rangle \geq 0, \tag{6.3}$$

where the first step was by Definition 6.2.9 of the Kempf-Ness function, the second step was by standard matrix calculus $\partial_\eta e^{\eta Z} = Z e^\eta Z$, and the final inequality was because $Z \in \mathfrak{p} \subseteq \mathrm{H}(V)$ so $Z^2 \succeq 0$ and the inner product of two positive definite operators is always non-negative. This verifies that the univariate function $\eta \to f_y^P(e^{\eta Z})$ is convex by Definition 2.3.2, and since $Z \in \mathfrak{p}$ was arbitrary, this verifies geodesic convexity. □

Below we collect the properties we have shown in this subsection.

Proposition 6.2.18. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$ with choice of scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for any tuple $x := \{x_1, ..., x_K\} \in V^K$:*

1. *The Kempf-Ness function f_x^P is geodesically convex on P;*
2. *For $g \in G$, $g \cdot x$ is G-balanced iff $(g^*g)^{1/2}$ is a critical point of f_x^P;*
3. *For $g \in G$, $g \cdot x$ is G-balanced iff g^*g is a global minimizer of f_x^P.*

In Chapter 7, we will prove strong bounds on the tensor scaling solution for special inputs using tools from convex analysis to analyze the geodesically convex formulation.

6.3 General Scaling Reductions

Now that we have properly defined the geodesically convex formulation for tensor scaling, we can give a formal proof of the frame-to-matrix reduction in Theorem 4.2.13. Specifically, we will use Proposition 6.2.18(3) equating the scaling solution to the global minimum of the Kempf-Ness function in Definition 6.2.9. We first present a simple functional statement that will be useful in the proof of Theorem 4.2.13. It will also be used in Chapter 7 to reduce the analysis of the tensor scaling problem to the simpler setting of commutative scaling groups (see Definition 6.2.3 and Definition 7.2.1).

Theorem 6.3.1. *Let $f : P \to \mathbb{R}$ be a continuous function on domain P. Let there be a (not necessarily disjoint) decomposition $P = \cup_{\Xi\in\mathcal{X}} P^\Xi$, and assume for every $\Xi \in \mathcal{X}$ the restriction $f|_{P^\Xi}$ attains its minimum (not necessarily uniquely) at some point $p_\Xi \in P^\Xi$. If there is some compact set $K \subseteq P$ such that $p_\Xi \in K$ for every $\Xi \in \mathcal{X}$, then f attains its global minimum at some point $p_* \in \cup_{\Xi\in\mathcal{X}} p_\Xi$.*

Proof. In order to find the global minimum of f, we can restrict our attention to

$$\inf_{p\in P} f(p) = \inf_{\Xi\in\mathcal{X}} \inf_{q\in P^\Xi} f(q) = \inf_{\Xi\in\mathcal{X}} f(p_\Xi),$$

where the first step was by the decomposition $P = \cup_{\Xi\in\mathcal{X}} P^\Xi$, and the last step was by the assumption $p_\Xi = \arg\inf_{q\in P^\Xi} f(q)$. Therefore, if the global minimum of f is attained, then we can assume that the optimizer is an element of $\cup_{\Xi\in\mathcal{X}} p_\Xi$.

To show that the global minimum is attained, we write the infimum as

$$\inf_{p\in P} f(p) = \inf_{\Xi\in\mathcal{X}} f(p_\Xi) = \inf_{q\in K} f(q),$$

as $\cup_{\Xi\in\mathcal{X}} p_\Xi \subseteq K$ by assumption. The right hand side is the infimum of continuous f over compact set K, so by the extreme value theorem, this infimum is attained at some point p_*. The left hand side shows that p_* is the global minimizer of f, and by the argument above, $p_* \in \cup_{\Xi\in\mathcal{X}} p_\Xi$. □

This simple result is the key to many of our reduction for the analysis of tensor scaling. It also allows us to unify and improve the analyses of [63] and [36] for tensor scaling by choosing different decompositions $P = \cup_\Xi P_\Xi$. We discuss this in more detail at the end of the section.

At this point, our reduction from frame scaling to matrix scaling in Theorem 4.2.13 follows simply by translating the language of Chapter 4 to the geodesic convex formulation of Proposition 6.2.18 and applying the decomposition result of Theorem 6.3.1.

Proof of Theorem 4.2.13. We first rewrite the frame scaling problem in the language of Definition 6.2.5: we are given input $\mathrm{vec}(U) \in \mathbb{F}^d \otimes \mathbb{F}^n$ with scaling group $G = (\mathrm{SL}(d), \mathrm{ST}(n))$ and $P = (\mathrm{SPD}(d), \mathrm{ST}_+(n))$ the associated polar part according to Definition 6.2.3. By Proposition 6.2.18(3), if $p_* = (e^{X_*}, e^{Y_*}) := \arg\inf_{p\in P} f_U^P(p)$ is a global minimum of the Kempf-Ness function in Definition 6.2.9, then the scaling $e^{X_*/2} U e^{Y_*/2}$ produces a doubly balanced frame according to Definition 6.2.4 (or more simply Definition 4.1.2 for frames).

To show the conclusion of the theorem, we will apply Theorem 6.3.1 to f_U^P with the decomposition

$$P = \cup_\Xi T_+^\Xi := \cup_\Xi (\mathrm{ST}_+^\Xi(d), \mathrm{ST}(n)),$$

where the union is over all orthonormal bases $\Xi \subseteq \mathbb{F}^d$ according to the decomposition of $\mathrm{SPD}(d)$ given in Eq. (2.6).

The optimizer of each restriction $f_U^P|_{T_+^\Xi}$ can be found by using the assumption that, for every orthonormal basis Ξ and matrix representation $M^\Xi := \Xi^* U$, there exists a diagonal scaling $(X_\Xi, Y_\Xi) \in \mathfrak{t}$ (where $\mathfrak{t} = \mathfrak{st}_+(d) \oplus \mathfrak{st}_+(n)$ according to Definition 3.1.5) such that $e^{X_\Xi/2} M^\Xi e^{Y_\Xi/2}$ is a doubly balanced matrix according to Definition 3.1.2. Note that for any matrix representation M^Ξ, the diagonal scaling $(X, Y) \in \mathfrak{t}$ induces a frame scaling

$$e^{X/2} M^\Xi e^{Y/2} \to (\Xi e^{X/2} \Xi^*) U e^{Y/2}$$

according to Eq. (3.5).

Therefore, if f_{M^Ξ} is the matrix Kempf-Ness function according to Definition 3.1.6, then it is related to the frame Kempf-Ness function as

$$f_{M^\Xi}((X, Y) \in \mathfrak{t}) = s(e^{X/2} M^\Xi e^{Y/2}) = s((\Xi e^{X/2} \Xi^*) U e^{Y/2}) = f_U^P(\Xi e^X \Xi^*, e^Y), \tag{6.4}$$

where the first step was by Definition 3.1.6 of the matrix Kempf-Ness function, the second was by our calculation above showing $(\Xi e^{X/2} \Xi^*) U e^{Y/2}$ is the frame scaling induced by $(X, Y) \in \mathfrak{t}$, and the final step was by Definition 6.2.9 of the Kempf-Ness function for frame scaling on domain P.

By Proposition 3.1.10(3), $e^{X_\Xi/2} M^\Xi e^{Y_\Xi/2}$ is a doubly balanced matrix scaling iff (X_Ξ, Y_Ξ) is the global minimizer of the matrix Kempf-Ness function f_{M^Ξ} given in Definition 3.1.6. By the equivalence in Eq. (6.4), this means that $p_\Xi := (\Xi e^{X_\Xi} \Xi^*, e^{Y_\Xi}) \in (\mathrm{ST}_+^\Xi(d), \mathrm{ST}(n)) = T_+^\Xi$ is the global minimum of $f_U^P|_{T_+^\Xi}$. Further, by the assumption that $\|(X_\Xi, Y_\Xi)\|_{\mathfrak{t}} \leq R$, we have that these optimizers are contained in a compact set. Therefore, we can apply Theorem 6.3.1 to f_U^P with decomposition $P = \cup_\Xi T_+^\Xi$ to find the global minimizer $p_* \in \cup_\Xi (\Xi e^{X_\Xi} \Xi^*, e^{Y_\Xi})$. By Proposition 6.2.18(3), the induced scaling is a doubly balanced frame.

□

We now discuss how the partition idea in Theorem 6.3.1 allows us to unify the previous analyses of specific tensor scaling groups.

In [62] and [63], our motivation was to analyze the Paulsen problem in Chapter 4, so we defined the dynamical system in Definition 4.1.6 as the simultaneous and continuous version of the alternate scaling algorithm in Eq. (4.2). Because we did not have the perspective of geodesic convex optimization, we directly analyzed the convergence of this dynamical system in terms of the error in the doubly balanced condition. Specifically, we showed that the error $\|\nabla_U\|_{\mathfrak{p}}^2$ defined in Definition 4.2.3 decreased exponentially when the input frame satisfied a natural spectral condition. Now that we have a better understanding of the geodesic convex formulation, we can derive this as the exponential convergence of the gradient under gradient flow for geodesically strongly convex inputs. But as discussed in Section 4.2.3, the reduction in Theorem 4.2.13 is crucial to our application to the Paulsen problem because it allows us to use the stronger robustness properties of matrix scaling (e.g. Lemma 3.3.4) to bound the frame scaling solution.

Similarly, in [36], we analyzed the tensor scaling solution directly using functional arguments and geodesic convexity. Specifically, we bound the optimizer of the Kempf-Ness function by decomposing P into geodesic curves and bounding the optimum for each univariate restriction. We repeat this argument in Theorem 7.1.16. Since each piece of the partition is just a univariate convex function, we can use very simple arguments based on the gradient to bound the optimum. Further, these univariate restrictions enjoy strong robustness properties for tensors similar to matrix scaling. While this argument is straightforward, it loses some information about the error in the balance conditions. Therefore in Section 7.2 we are able to given an improved analysis for the case of commutative tori by analyzing the ε-G-balance condition of Definition 6.2.4 directly throughout gradient flow. Then, we are once again able to lift this to the non-commutative setting using Theorem 6.3.1.

As the above discussion shows, the choice of partition $P = \cup_{\Xi} P_{\Xi}$ affects the analysis of tensor scaling in subtle ways. By Proposition 6.2.18(3), finding the tensor scaling solution for input x with scaling group $(G, P, \mathfrak{p})$ is equivalent to finding the optimizer of the Kempf-Ness function $\min_{p \in P} f_x^P$ given in Definition 6.2.9. By using the partition argument in Theorem 6.3.1, we can reduce this to bounding the optimizer on each piece P_{Ξ}, which may be simpler. But it is also important to consider how much global information about the tensor is preserved when analyzing the restricted optimization problem over P_{Ξ}. The value of Theorem 6.3.1 is that the choice of decomposition is left to the user, and therefore gives flexibility to leverage different partitions and different simpler analyses to approach the full tensor scaling problem. It would be interesting to see whether we could find improved analyses of tensor scaling for other special subsets of P.

At this point, we have described all the general theory required to analyze tensor scaling. In particular, we have shown that Definition 6.2.9 gives a geodesically convex formulation for tensor scaling, and analysis of the non-commutative setting can be reduced to the commutative setting using Theorem 6.3.1. Recall that there were many valuable properties of matrix scaling (e.g. standard convexity, strong robustness in Lemma 3.2.4) which do not necessarily carry over to frame scaling. Similarly, we will be able to give improved results for the tensor scaling problem when the scaling group is commutative, and then use Theorem 6.3.1 to lift these results to the non-commutative setting. As a consequence of these ideas, our work in Chapter 7 will mostly use elementary convex analysis and structural observations about scaling.

Chapter 7

Tensor Scaling

In this chapter, we study the tensor scaling problem described in Definition 6.2.5. This is a common generalization of matrix, frame, and operator scaling. We will generalize the strongly convex and pseudorandom techniques of Chapter 3 in order to analyze the geodesic convex formulation given in Proposition 6.2.18. A key component in our proofs will be the decomposition strategy given by Theorem 6.3.1. This will allow us to reduce to the simpler commutative setting, where we can use arguments from standard convex optimization. Our main application of these results is given in Chapter 9, where we prove strong bounds on sample complexity and error for the tensor normal model from statistics.

Overview: In Section 7.1, we present the necessary definitions relating to our strong convergence results, specifically strong convexity, pseudorandomness, and the spectral condition. This leads to a reasonably simple preliminary analysis of the scaling solution for sufficiently strongly convex inputs. Then, in Section 7.2, we improve this analysis for the commutative tensor scaling problem when the inputs are strongly convex and pseudorandom. This is lifted to the non-commutative setting using the decomposition ideas from Theorem 6.3.1. In Section 7.3, we consider robustness properties of these convergence conditions for non-commutative tensor scaling. These will be helpful in deriving algorithmic guarantees for inputs satisfying these sufficient conditions. Finally, in Section 7.4, we show that the pseudorandom condition implies strong convexity. This is a similar (but incomparable) result to Theorem 3.4.7 on matrix pseudorandomness and strong convexity.

7.1 First Analysis of Strongly Convex Tensor Scaling

In this section, we give our first results on the scaling solution for tensor scaling inputs satisfying a natural strong convexity assumption. This can be accomplished by lifting ideas from convex optimization to analyze the optimizer of the geodesically convex formulation for tensor scaling presented in Proposition 6.2.18 This preliminary analysis can be viewed as a generalization of Theorem 3.2.8, and will be sharpened in two ways in Section 7.2.

In Section 7.1.1 and Section 7.1.2, we extend the notion of gradient and gradient flow to the geodesic setting, specifically for the Kempf-Ness function. In Section 7.1.3, we define the appropriate notion of geodesic strong convexity for tensor scaling. Finally, in Section 7.1.4, we show a bound on the optimizer for inputs satisfying this assumption by lifting arguments for strongly convex functions to the geodesic setting.

7.1.1 Geodesic Gradient

The simplest and most natural approaches to convex optimization are gradient based algorithms (discussed briefly in Section 2.3.2). In this subsection, we formally define the geodesic gradient. We use this to define the geodesic gradient flow algorithm in Section 7.1.2.

Recall, from Definition 2.3.12, that the gradient of function $h : V \to \mathbb{R}$ on vector space V encodes the first order differential information of the function. Explicitly, for any $x \in V, v \in V$ and inner product $\langle \cdot, \cdot \rangle$ on V, the gradient $\nabla h(x) \in V$ is defined to satisfy $\langle \nabla h(x), v \rangle = \partial_{t=0} h(x + tv)$. In our setting, the Kempf-Ness function for tensor scaling is defined on positive definite matrices, so the natural vector gradient is not well-defined. But we can use the geodesics in Fact 6.2.11 to encode first order differential information in the geodesic gradient. Specifically, for scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3, infinitesimal changes at $p \in P$ correspond to geodesic curves $\gamma_p(\eta Z) := p^{1/2} e^{\eta Z} p^{1/2}$ parametrized by $Z \in \mathfrak{p}$. Therefore, we would like the geodesic gradient to capture first order information for these directions.

Definition 7.1.1. *Consider scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3 with inner product $\langle \cdot, \cdot \rangle$ on vector space $\mathfrak{p}$. Then the geodesic gradient of function $F : P \to \mathbb{R}$ at point $p \in P$ satisfies*

$$\forall Z \in \mathfrak{p} : \quad \langle \nabla F(p), Z \rangle_{\mathfrak{p}} = \partial_{\eta=0} F(\gamma_p(\eta Z)) = \partial_{\eta=0} F(p^{1/2} e^{\eta Z} p^{1/2}). \tag{7.1}$$

This definition of geodesic gradient applies in the general setting of Riemannian manifolds. These are spaces which locally look like inner product spaces, and so the geodesic gradient encodes local first order information with respect to these local inner products. As discussed in Section 2.2.3, our domain is a special kind of manifold for which the local structure is invariant under a natural group action. Therefore, in this thesis, we will restrict our definitions to this simpler setting. We will briefly mention extensions to more general geodesic settings in Chapter 10.

Below, we make an appropriate choice of inner product which is adapted to the tensor scaling problem, and then present the explicit form of the geodesic gradient for the tensor Kempf-Ness function.

Definition 7.1.2 ($\mathfrak{p}$ Inner Product)**.** *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces, and let $(G, P, \mathfrak{p})$ be a choice of scaling group on V according to Definition 6.2.3. Then, for elements $Y, Z \in \mathfrak{p}$, the inner product is defined as*

$$\langle Y, Z\rangle_{\mathfrak{p}} := \sum_{a\in[m]} \frac{1}{d_a}\langle Y_a, Z_a\rangle,$$

where the right hand side uses the standard $L(V_a)$ inner product $\langle X, Y\rangle = Tr[X^*Y]$.

This inner product is a natural choice for our tensor setting. As shown in Definition 6.2.3, $\mathfrak{p}$ has a natural embedding into $L(V)$ as the infinitesimal of P:

$$Z \in \mathfrak{p} \quad \to \quad \partial_{\eta=0} \otimes_{a\in[m]} e^{\eta Z_a} = \sum_{a\in[m]} Z_a \otimes I_{\overline{a}},$$

where $I_{\overline{a}}$ is the identity operator on $V_{\overline{a}} := \otimes_{b\neq a\in[m]} V_b$. From this perspective, Definition 7.1.2 is (up to constant factor) the natural Frobenius inner product on this embedding:

$$\begin{aligned}\left\langle \sum_{a\in[m]} Y_a \otimes I_{\overline{a}}, \sum_{b\in[m]} Z_b \otimes I_{\overline{b}} \right\rangle &= \sum_{a\in[m]} \langle Y_a, Z_a\rangle\langle I_{\overline{a}}, I_{\overline{a}}\rangle + \sum_{a\neq b} \langle Y_a, I_a\rangle\langle I_b, Z_b\rangle\langle I_{\overline{ab}}, I_{\overline{ab}}\rangle \\ &= D\sum_{a\in[m]} \frac{1}{d_a}\langle Y_a, Z_a\rangle + 0 = D\cdot\langle Y, Z\rangle_{\mathfrak{p}},\end{aligned}$$

where $D := \prod_{a\in[m]} d_a$ so $\frac{D}{d_a}$ is the dimension of $V_{\overline{a}}$, and the cross-terms vanish because $Y_a, Z_a \in \mathfrak{spd}(V_a)$ so $\langle Y_a, I_a\rangle = \langle I_b, Z_b\rangle = 0$ by Definition 2.1.10. This last expression exactly matches Definition 7.1.2 up to the constant D factor.

This is also a natural generalization of the inner product in Definition 3.1.11 for matrix scaling. In that setting, we are given matrix tuple $A \in \text{Mat}(d, n)^K$ which can be viewed as a tuple of elements in $\mathbb{F}^d \otimes \mathbb{F}^n$ by $A_k \to \mathrm{vec}(A_k)$. The diagonal scaling group is $T := (\mathrm{ST}(d), \mathrm{ST}(n))$, and this gives the associated infinitesimal vector space $\mathfrak{t} := \mathfrak{st}_+(d) \oplus \mathfrak{st}_+(n)$ according to Definition 6.2.3. So for elements $(X, Y), (X', Y') \in \mathfrak{t}$ we have

$$\Big\langle (X, Y), (X', Y') \Big\rangle_{\mathfrak{t}} = \frac{\langle X, X' \rangle}{d} + \frac{\langle Y, Y' \rangle}{n} = \frac{1}{d} \sum_{i=1}^{d} X_{ii} X'_{ii} + \frac{1}{n} \sum_{j=1}^{n} Y_{jj} Y'_{jj},$$

where in the last step we used that operators in $\mathfrak{t}$ are all diagonal in the standard basis. This exactly matches the inner product given in Definition 3.1.11.

Now that we have chosen an inner product on $\mathfrak{p}$, the geodesic structure of P induces a unique geodesic gradient for the Kempf-Ness function.

Proposition 7.1.3. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for input $x = \{x_1, ..., x_K\} \in V^K$, the geodesic gradient of the Kempf-Ness function f_x^P at point $p \in P$ satisfies $\nabla f_x^P(p) = \nabla f_{p^{1/2} \cdot x}^P(I_V)$, and is given by $\nabla f_x^P(p) = \{(\nabla f_x^P(p))^{(a)}\}_{a \in [m]}$ which is defined component-wise as*

$$(\nabla f_x^P(p))^{(a)} = \begin{cases} d_a \cdot \rho_{p^{1/2} \cdot x}^{(a)} - s(p^{1/2} \cdot x) \cdot I_a & \text{if } G_a = \mathrm{SL}(V_a), \\ \mathrm{diag}^{\Xi_a} \left(d_a \cdot \rho_{p^{1/2} \cdot x}^{(a)} - s(p^{1/2} \cdot x) \cdot I_a \right) & \text{if } G_a = \mathrm{ST}^{\Xi_a}(V_a) \end{cases},$$

where diag^{Ξ} is the diagonal projection into basis Ξ. Note that $(\nabla f_x^P(p))^{(a)} \in \mathfrak{p}_a$ for all $a \in [m]$, and $\nabla f_x^P(p) = \{(\nabla f_x^P(p))^{(a)}\}_{a \in [m]} \in \mathfrak{p}$. We will often use shorthand $\nabla_x := \nabla f_x^P(I_V)$, and $\nabla_x = \{\nabla_x^{(a)}\}_{a \in [m]}$ for the marginals.

Proof. We will verify that the above formulas satisfy the requirements for geodesic gradient given in Definition 7.1.1. Recall that

$$f_x^P(\gamma_p(\eta Z)) = \langle \rho_x, p^{1/2} e^{\eta Z} p^{1/2} \rangle = \langle p^{1/2} \rho_x p^{1/2}, e^{\eta Z} \rangle = \langle \rho_{p^{1/2} \cdot x}, e^{\eta Z} \rangle = f_{p^{1/2} \cdot x}^P(\gamma_{I_V}(\eta Z)),$$

where in the first and last steps we used Definition 6.2.9 of the Kempf-Ness function and Fact 6.2.11 for geodesics, and the third step was by the equivariance of ρ as shown in Lemma 6.2.6(1).

Therefore, we can reduce our calculation of the gradient to geodesics from the identity by the change of variable $y := p^{1/2} \cdot x$, as

$$\langle \nabla f_x^P(p), Z \rangle_{\mathfrak{p}} = \partial_{\eta=0} f_x^P(\gamma_p(\eta Z)) = \partial_{\eta=0} f_y^P(\gamma_{I_V}(\eta Z)) = \langle \nabla f_y^P(I_V), Z \rangle_{\mathfrak{p}},$$

where the first and last steps were by Definition 7.1.1 of the geodesic gradient, and the second step was by substituting $y := p^{1/2} \cdot x$ into the equation $f_x^P(\gamma_p(\eta Z)) = f_{p^{1/2}\cdot x}^P(\gamma_{I_V}(\eta Z))$ shown above. We emphasize that $\nabla f_{g\cdot x}^P(I_V) \neq \nabla f_x^P(g^* g)$ unless $g = (g^* g)^{1/2}$ is the unique positive definite square-root.

Now we calculate the first order differential from I_V using Eq. (6.2):

$$\partial_{\eta=0} f_y(e^{\eta Z}) = \sum_{a\in[m]} \left\langle \rho_y^{(a)} - \frac{s(y)}{d_a} I_a, Z_a \right\rangle = \sum_{a\in[m]} \frac{1}{d_a} \langle d_a \cdot \rho_y^{(a)} - s(y) I_a, Z_a \rangle.$$

If $G_a = \mathrm{SL}(V_a)$ we leave the term as is, and if $G_a = \mathrm{ST}^{\Xi}(V_a)$ then $Z_a \in \mathfrak{p}_a = \mathfrak{st}_+^{\Xi}(V_a)$ so the inner product does not change by projecting $d_a \cdot \rho_y^{(a)} - s(y) I_a \to \mathrm{diag}^{\Xi}(d_a \cdot \rho_y^{(a)} - s(y) I_a)$. In either case, the above expression exactly matches the given definition of gradient as

$$\langle \nabla_y, Z \rangle_{\mathfrak{p}} = \sum_{a\in[m]} \frac{1}{d_a} \langle (\nabla f_y(I_V))^{(a)}, Z_a \rangle$$

by Definition 7.1.2 of the inner product $\langle \cdot, \cdot \rangle_{\mathfrak{p}}$. □

The $\mathfrak{p}$-norm of the geodesic gradient gives a natural way to measure error that is compatible with the balance condition of Definition 6.2.4.

Fact 7.1.4. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$ along with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. For ε-G-balanced tensor tuple $x \in V^K$ according to Definition 6.2.4, the gradient $\nabla_x = \nabla f_x^P(I_V)$ satisfies*

$$\|\nabla_x\|_{\mathfrak{p}}^2 \leq m \cdot s(x)^2 \varepsilon^2.$$

Proof. We assume that $G_a = \mathrm{SL}(V_a)$ for every $a \in [m]$. The case of diagonal scaling groups follows simply by applying the calculation below to the diagonal restriction. So let ∇_x be the geodesic gradient according to Proposition 7.1.3, and we calculate

$$\|\nabla_x\|_{\mathfrak{p}}^2 = \sum_{a\in[m]} \frac{\|d_a \rho_x^{(a)} - s(x) I_a\|_F^2}{d_a} \leq \sum_{a\in[m]} \|d_a \rho_x^{(a)} - s(x) I_a\|_{\mathrm{op}}^2 \leq m \cdot s(x)^2 \varepsilon^2,$$

where in the first step we substituted in the expression from Proposition 7.1.3 for the geodesic gradient and Definition 7.1.2 for $\|\cdot\|_{\mathfrak{p}}$, the second step was by the inequality $\|\cdot\|_F^2 \leq d_a \|\cdot\|_{\mathrm{op}}^2$ applied to $\nabla_x^{(a)} \in \mathfrak{p}_a \subseteq L(V_a)$ with $\dim(V_a) = d_a$ for each $a \in [m]$, and the final step was by ε-G-balance condition of x according to Definition 6.2.4, or more precisely, the expression in Eq. (6.1) for ε-G-balanced inputs. □

After defining the appropriate notion of strong geodesic convexity in Section 7.1.3, we will use simple gradient arguments to bound the scaling solution of sufficiently strongly convex inputs in Section 7.1.4. In the following subsection, we will describe the gradient flow dynamical system which is naturally induced by Proposition 7.1.3. This part is not required for our analysis and is only presented for completeness.

7.1.2 Geodesic Gradient Flow

In this subsection, we will present the geodesic gradient flow for tensor scaling. This part can be skipped without loss for any of our results, as the analysis of Section 7.1.4 only uses properties of the geodesic gradient and geodesic strong convexity, and the analyses in Section 7.2 only use gradient flow for the much simpler commutative tensor scaling setting. The goal of this subsection is to discuss the techniques in this thesis in relation to past work on tensor scaling.

We first present the formal definition of the gradient flow dynamical system for non-commutative tensor scaling.

Definition 7.1.5. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then, for input $x = \{x_1, ..., x_K\} \in V^K$, the G-gradient flow is the dynamical system $\{g_t \in G\}_{t\geq 0}$ defined by initial condition $g_0 = I_V$ and differential equation*

$$\partial_t g_t = -\frac{1}{2} \nabla f^P_{g_t \cdot x}(I_V) \cdot g_t.$$

This induces a dynamical system on tensors by $x_t := g_t \cdot x$.

Note that the above definition is with respect to G scalings. For most of the results in this thesis, the properties of interest will be invariant with respect to isometries (e.g. Lemma 6.2.6), and so we will be able to restrict our attention to the polar P and its geodesic geometry. This is not the case for gradient flow, and we will discuss the reason for this at the end of this subsection.

Before this, we prove that the dynamical system in Definition 7.1.5 is natural in the sense that it follows the direction of steepest descent for the Kempf-Ness function $s(g \cdot x) = \tilde{f}^G_x(g) = f^P_x(g^* g)$.

Lemma 7.1.6. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and consider scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then, for input $x = \{x_1, ..., x_K\} \in V^K$*

and $x_t = g_t \cdot x$ the solution to gradient flow given in Definition 7.1.5, the change in size is $\partial_t s(x_t) = -\|\nabla_{x_t}\|_{\mathfrak{p}}^2$. As a consequence

$$s(x_T) - s(x) = -\int_0^T \|\nabla_{x_t}\|_{\mathfrak{p}}^2.$$

Proof. We begin by rewriting the size of $x_t := g_t \cdot x$ in terms of the Kempf-Ness function and the gradient flow:

$$s(x_t) = \operatorname{Tr}[\rho_{g_t \cdot x}] = \langle \rho_x, g_t^* g_t \rangle, \tag{7.2}$$

where the first step was by Definition 6.2.1 for the size of tensor $x_t := g_t \cdot x$, and the second step was by the equivariance property in Fact 6.2.10

Since this only depends on the polar part of g_t, for the purpose of computing the size we write the induced dynamical system as

$$\begin{aligned} -\partial_t(g_t^* g_t) &= -(\partial_t g_t)^* g_t - g_t^*(\partial_t g_t) \\ &= \left(\frac{1}{2}\nabla f^P_{g_t \cdot x}(I_V) \cdot g_t\right)^* g_t + g_t^* \left(\frac{1}{2}\nabla f^P_{g_t \cdot x}(I_V) \cdot g_t\right) = g_t^*(\nabla f^P_{g_t \cdot x}(I_V)) g_t, \end{aligned} \tag{7.3}$$

where the first step was by the product rule, the second step was by Definition 7.1.5 of gradient flow, and in the final step we used that $\nabla f^P_{g_t \cdot x}(I_V) \in \mathfrak{p}$ is self-adjoint.

Now we can simply compute the change in size as

$$\begin{aligned} -\partial_t s(x_t) = -\partial_t \langle \rho_x, g_t^* g_t \rangle = \langle \rho_x, g_t^*(\nabla f^P_{g_t \cdot x}(I_V)) g_t \rangle = \sum_{a \in [m]} \langle \rho_{g_t \cdot x}, (\nabla f^P_{g_t \cdot x}(I_V))^{(a)} \otimes I_{\overline{a}} \rangle \\ = \sum_{a \in [m]} \langle \rho^{(a)}_{x_t}, \nabla^{(a)}_{x_t} \rangle = \sum_{a \in [m]} \frac{1}{d_a} \langle d_a \rho^{(a)}_{x_t} - s(x_t) I_a, \nabla^{(a)}_{x_t} \rangle = \sum_{a \in [m]} \frac{\|\nabla^{(a)}_{x_t}\|_F^2}{d_a} = \|\nabla_{x_t}\|_{\mathfrak{p}}^2, \end{aligned}$$

where the first step was by the calculation in Eq. (7.2), in the second step we used the formula for the derivative of the polar part given in Eq. (7.3), in the third step we used $\rho_{g_t \cdot x} = g_t \rho_x g_t^*$ by the equivariance property in Fact 6.2.10 as well as the embedding $\nabla f^P_{x_t}(I_V) = \sum_{a \in [m]} (\nabla f^P_{x_t}(I_V))^{(a)} \otimes I_{\overline{a}}$ as described in Definition 6.2.3, in the fourth step we substituted $x_t = g_t \cdot x$ as well as Definition 6.2.2 to reduce the inner product to each marginal, in the fifth step we subtracted $s(x_t) I_a$ from each term as $\nabla^{(a)}_{x_t} \in \mathfrak{p}_a \subseteq \mathfrak{spd}(V_a)$ so $\langle \nabla^{(a)}_{x_t}, I_a \rangle = 0$ by Definition 2.1.10, in the sixth step we substitute $\nabla^{(a)}_{x_t} = d_a \rho^{(a)}_{x_t} - s(x_t) I_a$ or its diagonal restriction depending on the scaling group according to Proposition 7.1.3, and the final step was by Definition 7.1.2 of the $\mathfrak{p}$-norm. The second statement follows simply from the first by the fundamental theorem of calculus. □

To complement this lemma, the discussion below gives an intuitive derivation of the dynamical system in Definition 7.1.5 as the steepest descent direction for the Kempf-Ness function. Recall that $s(g\cdot x) = \tilde{f}_x^G(g) = f_x^P(g^*g)$ by Definition 6.2.7 and Definition 6.2.9. It is intuitively clear by the first order definition of the geodesic gradient in Definition 7.1.1, that for the Kempf-Ness function at point $p \in P$, the geodesic curve $\eta \to \gamma_p(-\eta \nabla f_x^P(p)) = p^{1/2} e^{-\eta \nabla f_x^P(p)} p^{1/2}$ gives the steepest descent (infinitesimally) with respect to the $\mathfrak{p}$-norm in Definition 7.1.2. Recall by the unitary invariance of Fact 6.2.8, $\tilde{f}_x^G(g)$ depends only on the polar part g^*g. Therefore, even though the tangent space for $g \in G$ is the larger set $(i\mathfrak{p} \oplus \mathfrak{p}) \cdot g$ as discussed in Section 2.2.3, the steepest descent direction will only depend on the polar direction in $\mathfrak{p} \cdot g$. As shown in Eq. (2.8), the curve $\eta \to e^{-\eta \nabla f_{g\cdot x}^P(I_V)} \cdot g$ induces the polar curve $\eta \to g^* e^{-\eta \nabla f_{g\cdot x}^P(I_V)} g$, so this also gives the steepest descent direction for $s(g\cdot x) = \tilde{f}_x^G(g)$.

But if we were only interested in the steepest descent for the Kempf-Ness function, we could as well have defined the dynamical system in Definition 7.1.5 just in terms of the polar part. In fact, this seems more natural if we wanted to use geodesic convexity to analyze tensor scaling. The differential equation at time t is always in $\mathfrak{p} \cdot g_t$, so in some sense we are only concerned with the infinitesimal polar direction. But due to the non-commutativity of G, this does not imply that $g_t \in P$ for all time.

The reason we choose to define the gradient flow in terms of $g_t \in G$ comes from Kempf-Ness theory [58], [11], [40]. We can illustrate this using the simple example of matrix scaling. For this setting, we showed in Theorem 4.3.4 that the geodesic gradient flow for the Kempf-Ness function f_A induces the same direction as the Euclidean gradient flow for $\|\nabla_A\|_{\mathfrak{t}}^2$. In order to preserve this natural property for the non-commutative setting, we need to define a dynamical system in terms of $g_t \in G$.

In fact, this gives a principled derivation for the dynamical system for operator scaling that we defined in [62] and [63]. These works were motivated by the Paulsen problem in frame theory described in Chapter 4. Therefore, they defined a dynamical system on frame $U = \{u_1, ..., u_n\} \in \text{Mat}(d, n)$ for the purpose of decreasing a natural notion of error to doubly balanced:

$$\partial_t u_j(t) := \Big(s(U(t)) I_d - dU(t)U(t)^*\Big) u_j(t) + u_j(t)\Big(s(U(t)) - n\|u_j\|_2^2\Big).$$

This is equivalent to the gradient flow given in Definition 7.1.5 for $G = (\text{SL}(d), \text{SL}(n))$ (up to the factor $\frac{1}{2}$) as the terms in parentheses are exactly the geodesic gradients for frame scaling given in Proposition 7.1.3. This is a special case of the Kempf-Ness equivalence [58], [40] from geometric invariant theory which then allows us to use techniques from geodesic convex analysis to further understand this dynamical system.

Another simple feature of the matrix gradient flow in Definition 3.1.14 is that it is defined in terms of $(X_t, Y_t) \in \mathfrak{t}$, which is a vector space. Recall that $P = e^{\mathfrak{p}}$ by the discussion in Section 2.2.3, so we could try to give Definition 7.1.5 in terms of $\mathfrak{p}$ instead. Unfortunately, for non-commutative scalings, this change of variables makes calculus much more difficult as the geodesic structure is only locally defined with respect to $\mathfrak{p}$. Explicitly, for non-commutative scalings $e^Y, e^Z \in P$, $e^{Y/2}e^Z e^{Y/2} \neq e^{Y+Z}$. This makes the curve from $e^Y \to e^Z \in P$ quite difficult to express purely in terms of $\mathfrak{p}$, which is why we define the non-commutative gradient flow on $g_t \in G$. On the other hand, if P is commutative, then everything can be more simply expressed in terms of $\mathfrak{p}$, and this is the approach we take in Section 7.2 to give our improved strong convexity result. Therefore we will further discuss how to generalize Definition 3.1.14 of matrix gradient flow in this simpler commutative setting in Section 7.2.1.

Much of this thesis builds upon the work of [20], which placed the scaling framework into the context of geodesic convex optimization as shown in Chapter 6. That work also used gradient flows similar to Definition 7.1.5 in order to prove convergence results for optimization algorithms for scaling problems. We hope that the gradient flow techniques developed in this thesis will be useful for more questions in the scaling framework.

7.1.3 Strong Convexity

In this subsection, we will use the norm given by Definition 7.1.2 to define geodesic strong convexity for the tensor Kempf-Ness function. We will also present a related spectral condition that will be easier to show for random inputs and will be applied in Chapter 9. With the appropriate geodesic notions of gradient and strong convexity in hand, in Section 7.1.4 we apply standard arguments from convex analysis to give our first quantitative scaling result.

We begin with the natural notion of strong convexity induced by $\| \cdot \|_{\mathfrak{p}}$.

Definition 7.1.7. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces and let $(G, P, \mathfrak{p})$ be the scaling group according to Definition 6.2.3. Then input $x \in V^K$ is α-$\mathfrak{p}$-strongly convex if*

$$\forall Z \in \mathfrak{p} : \quad \partial^2_{\eta=0} f^P_x(\gamma_{I_V}(\eta Z)) = \partial^2_{\eta=0} \langle \rho_x, e^{\eta Z} \rangle \geq \alpha \|Z\|^2_{\mathfrak{p}}.$$

The expression above only depends on the input tuple x. Below, we show that this is equivalent to geodesic strong convexity with respect to the Kempf-Ness function.

Lemma 7.1.8. *Consider tensor product $V = \otimes_{a\in[m]} V_a$ with input $x \in V^K$ and scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3 with $g \in G$. Then $y = g \cdot x$ is α-$\mathfrak{p}$-strongly convex according to Definition 7.1.7 iff f_x^P is α-geodesically strongly convex at g^*g with respect to $\|\cdot\|_{\mathfrak{p}}$ according to Definition 6.2.13.*

Proof. We first rewrite the α-$\mathfrak{p}$-strong convexity of $y = g \cdot x$ as

$$\inf_{Z\in\mathfrak{p}} \frac{\partial^2_{\eta=0}\langle \rho_{g\cdot x}, e^{\eta Z}\rangle}{\|Z\|_{\mathfrak{p}}^2} = \inf_{Z\in\mathfrak{p}} \frac{\langle \rho_x, g^* Z^2 g\rangle}{\|Z\|_{\mathfrak{p}}^2} \geq \alpha,$$

where we used Eq. (6.3) and substituted $\rho_{g\cdot x} = g\rho_x g^*$ by the equivariance property in Lemma 6.2.6(1). Letting $p = g^*g$, we can similarly rewrite the α-geodesic strong convexity condition as

$$\inf_{Z\in\mathfrak{p}} \frac{\partial^2_{\eta=0} f_x^P(\gamma_p(\eta Z))}{\|Z\|_{\mathfrak{p}}^2} = \inf_{Z\in\mathfrak{p}} \frac{\langle \rho_x, p^{1/2} Z^2 p^{1/2}\rangle}{\|Z\|_{\mathfrak{p}}^2} \geq \alpha,$$

where we usd Fact 6.2.11 for the geodesic $\gamma_p(\eta Z) = p^{1/2} e^{\eta Z} p^{1/2}$ and again applied the calculation in Eq. (6.3).

By the polar decomposition in Theorem 2.1.13, we can write $g = up^{1/2}$ where $u \in G \cap \mathrm{SU}(V)$ is an isometry and $p^{1/2} \in P$ is the polar part. This allows us to show the two expressions above are equal, as

$$\inf_{Z\in\mathfrak{p}} \frac{\partial^2_{\eta=0}\langle \rho_{g\cdot x}, e^{\eta Z}\rangle}{\|Z\|_{\mathfrak{p}}^2} = \inf_{Z\in\mathfrak{p}} \frac{\langle \rho_x, g^* Z^2 g\rangle}{\|Z\|_{\mathfrak{p}}^2} = \inf_{Z\in\mathfrak{p}} \frac{\langle \rho_x, p^{1/2}(u^*Zu)^2 p^{1/2}\rangle}{\|Z\|_{\mathfrak{p}}^2} = \inf_{Y\in\mathfrak{p}} \frac{\partial^2_{\eta=0} f_x^P(\gamma_p(\eta Y))}{\|Y\|_{\mathfrak{p}}^2},$$

where the first step was calculated above for the α-$\mathfrak{p}$-strong convexity condition of $y = g\cdot x$, the second step was by the polar decomposition $g = up^{1/2}$, and in the final step we applied the change of variable $Y = u^*Zu$ along with the fact $\|Y\|_{\mathfrak{p}} = \|Z\|_{\mathfrak{p}}$ by unitary invariance of Definition 7.1.2 as well as the calculation above for the α-geodesic strong convexity of f_x^P at p. Therefore, these two expressions are equivalent and $g \cdot x$ is α-$\mathfrak{p}$-strongly convex according to Definition 7.1.7 iff f_x^P is α-geodesically strongly convex at $p = g^*g$ according to Definition 6.2.13. □

This equivalent condition will be helpful in Chapter 8, where we will be able to lift tools from convex optimization to the geodesic setting to prove convergence of algorithms for tensor scaling.

We also present a spectral condition which implies strong convexity and will be easier to prove for random tensors. To motivate this definition, we expand out the second-order derivative calculation in Eq. (6.3):

$$\partial^2_{\eta=0} f_x^P(e^{\eta Z}) = \left\langle \rho_x, \Big(\sum_{a\in[m]} I_{\overline{a}} \otimes Z_a \Big)^2 \right\rangle = \sum_{a\in[m]} \langle \rho_x^{(a)}, Z_a^2\rangle + \sum_{a\neq b\in[m]} \langle \rho_x^{(ab)}, Z_a \otimes Z_b\rangle. \tag{7.4}$$

The cases of interest to us will be when x is a nearly balanced tensor so $d_a\rho_x^{(a)} \approx I_a$ and the diagonal terms $\langle \rho_x^{(a)}, Z_a^2\rangle$ are large. So we define the following spectral condition in order to control the off-diagonal terms.

Definition 7.1.9. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and let $(G, P, \mathfrak{p})$ be the scaling group according to Definition 6.2.3. The input $x \in V^K$ satisfies the λ-$\mathfrak{p}_{ab}$-spectral condition if*

$$\sup_{Z_a\in\mathfrak{p}_a, Z_b\in\mathfrak{p}_b} \frac{|\langle \rho_x^{(ab)}, Z_a\otimes Z_b\rangle|}{\|Z_a\|_F \|Z_b\|_F} \leq \frac{\lambda}{\sqrt{d_a d_b}}.$$

Input x satisfies the λ-$\mathfrak{p}$-spectral condition if the above holds for every pair $a \neq b \in [m]$.

Note that the above condition is symmetric in the sense that the $\mathfrak{p}_{ab}$-spectral condition is equivalent to the $\mathfrak{p}_{ba}$-spectral condition.

A very similar spectral condition was used in [63] to give a fast convergence result for operator scaling, and we discuss the relation of Definition 7.1.9 to [63] at the end of this subsection. We next show how this condition can be simply combined with the balance condition in Definition 6.2.4 to show strong convexity.

Proposition 7.1.10. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and let $(G, P, \mathfrak{p})$ be the scaling group according to Definition 6.2.3. If input $x \in V^K$ is ε-G-balanced according to Definition 6.2.4, and satisfies the λ-$\mathfrak{p}$-spectral condition according to Definition 7.1.9, then x is α-$\mathfrak{p}$-strongly convex for $\alpha \geq s(x)(1-\varepsilon) - (m-1)\lambda$.*

Proof. Our plan is to show that the expression in Eq. (7.4) is lower bounded by using the balanced condition to lower bound the diagonal terms, and the spectral condition to upper bound the off-diagonal terms in absolute value. To show that the diagonal terms in Eq. (7.4) are large, we use the fact that x is ε-balanced so

$$\langle \rho_x^{(a)}, Z_a^2\rangle \geq s(x)\frac{1-\varepsilon}{d_a}\langle I_a, Z_a^2\rangle = s(x)\frac{1-\varepsilon}{d_a}\|Z_a\|_F^2,$$

where the first step was by Definition 6.2.4 and the fact that $Z_a^2 \succeq 0$.

The off-diagonal terms are bounded by Definition 7.1.9 of spectral gap, so we can bound the second order derivative for any $Z \in \mathfrak{p}$ by

$$\begin{aligned}
\partial^2_{\eta=0} f_x(e^{\eta Z}) &= \sum_{a\in[m]} \langle \rho_x^{(a)}, Z_a^2 \rangle + \sum_{a\neq b\in[m]} \langle \rho_x^{(ab)}, Z_a \otimes Z_b \rangle \\
&\geq \sum_{a\in[m]} \frac{s(x)(1-\varepsilon)}{d_a} \|Z_a\|_F^2 - \sum_{a\neq b\in[m]} \frac{\lambda}{\sqrt{d_a d_b}} \|Z_a\|_F \|Z_b\|_F \\
&= \sum_{a\in[m]} (s(x)(1-\varepsilon)+\lambda) \frac{\|Z_a\|_F^2}{d_a} - \lambda \left(\sum_{a\in[m]} \frac{\|Z_a\|_F}{\sqrt{d_a}} \right)^2 \\
&\geq ((s(x)(1-\varepsilon)+\lambda)\|Z\|_{\mathfrak{p}}^2 - m\lambda \|Z\|_{\mathfrak{p}}^2,
\end{aligned}$$

where the first step was given in Eq. (7.4), in the second step we lower bounded the diagonal terms by the calculation above and upper bounded the off-diagonal terms by Definition 7.1.9, and the last step used Cauchy-Schwarz as well as Definition 7.1.2 of $\langle \cdot, \cdot \rangle_{\mathfrak{p}}$. As $Z \in \mathfrak{p}$ was arbitrary, this verifies Definition 7.1.7 of strong convexity. □

Note that the above proof only used the lower bound $\rho_x^{(a)} \succeq \frac{s(x)(1-\varepsilon)}{d_a} I_a$. In the $m = 2$ operator scaling case there is a partial converse using the upper bound $\rho_x^{(a)} \preceq \frac{s(x)(1+\varepsilon)}{d_a} I_a$.

Lemma 7.1.11. *Input $A \in \operatorname{Mat}(d,n)^K$ can be considered as a tuple of elements in the tensor product space $V := \mathbb{F}^d \otimes \mathbb{F}^n$ by the natural isomorphism $A_k \to \operatorname{vec}(A_k)$. Let $(G, P, \mathfrak{p})$ be any choice of scaling group on $V = \mathbb{F}^d \otimes \mathbb{F}^n$ according to Definition 6.2.3. If ε-G-balanced A is satisfies the λ-$\mathfrak{p}$-spectral condition and is α-$\mathfrak{p}$-strongly convex according to Definition 7.1.7, then $\alpha \leq s(A)(1+\varepsilon) - \lambda$.*

Proof. Let $(X, Y) \in \mathfrak{p}$ be the elements that achieve the supremum in Definition 7.1.9. By changing sign and normalizing if necessary, we assume without loss that $\|X\|_F^2 = d$, $\|Y\|_F^2 = n$ and $\langle \rho_A, X \otimes Y \rangle = -\lambda \frac{\|X\|_F \|Y\|_F}{\sqrt{dn}} = -\lambda$. Then, we calculate

$$\begin{aligned}
\partial^2_{\eta=0} f_A(e^{\eta X}, e^{\eta Y}) &= \langle \rho, (X \otimes I_n + I_d \otimes Y)^2 \rangle = \langle \rho^L, X^2 \rangle + \langle \rho^R, Y^2 \rangle + 2\langle \rho, X \otimes Y \rangle \\
&\leq \frac{s(A)(1+\varepsilon)}{d} \|X\|_F^2 + \frac{s(A)(1+\varepsilon)}{n} \|Y\|_F^2 - 2\frac{\lambda}{\sqrt{dn}} \|X\|_F \|Y\|_F \\
&= 2(s(A)(1+\varepsilon) - \lambda) = (s(A)(1+\varepsilon) - \lambda) \|(X,Y)\|_{\mathfrak{p}}^2,
\end{aligned}$$

where the first step was by the calculation in Eq. (6.3), the second was given in Eq. (7.4), in the third step we used the ε-G-balance condition in Definition 6.2.4 to bound the diagonal terms by $\max\{d\|\rho^L\|_{\text{op}}, n\|\rho^R\|_{\text{op}}\} \leq s(A)(1+\varepsilon)$ and used $\langle \rho_A, X \otimes Y\rangle = -\lambda \frac{\|X\|_F \|Y\|_F}{\sqrt{dn}}$ to bound the off-diagonal term, and the final steps were by our assumption that $\frac{\|X\|_F^2}{d} = \frac{\|Y\|_F^2}{n} = 1$. Since Definition 7.1.7 of strong convexity gives a lower bound for every $(X,Y) \in \mathfrak{p}$, this shows the required upper bound for α. □

Therefore, in the case of $m = 2$ scaling (i.e. matrix, frame, or operator), for nearly doubly balanced inputs, strong convexity and the spectral condition are nearly equivalent for analyzing fast convergence.

In [63], we used a similar plan to lower bound the diagonal terms and upper bound the cross term in Eq. (7.4) for $V = \text{Mat}_{\mathbb{R}}(d,n) \simeq \mathbb{R}^d \otimes \mathbb{R}^n$. We showed that the dynamical system in Definition 7.1.5 converged quickly to the solution of operator scaling when the input satisfied a spectral condition. To do so, in [63] we considered the associated operator $\Phi_A : L(\mathbb{R}^n) \to L(\mathbb{R}^d)$ satisfying $\langle X, \Phi_A(Y)\rangle = \langle \rho_A, X \otimes Y\rangle$ according to Proposition 2.4.5. Tuple $A \in \text{Mat}(d,n)^K$ was said to satisfy the σ-spectral gap condition if

$$\sigma_2(\Phi_A) \leq (1-\sigma)\frac{s(A)}{\sqrt{d_a d_b}},$$

where $\sigma_1 \geq \sigma_2 \geq ...$ are the singular values of the linear operator Φ_A written in decreasing order. The main technical work of Section 3.3 of [63] was to control the cross term in Eq. (7.4) by combining the spectral gap and nearly balanced conditions. Specifically, we were able to show that if A is nearly doubly balanced, then the top singular value is $\approx \frac{s(A)}{\sqrt{dn}}$ with singular value pair close to $(\frac{I_a}{\sqrt{d_a}}, \frac{I_b}{\sqrt{d_b}})$. Therefore, if the spectral gap σ is large then the cross term is small. Further, $\sigma \gg \varepsilon$ implies that $\|\nabla_{A_t}\|_{\mathfrak{p}}^2$ decreases exponentially throughout gradient flow.

In this thesis, we give a slightly cleaner analysis using strong convexity. In particular, our definition of the spectral condition is defined on $(\mathfrak{p}_a, \mathfrak{p}_b) \subseteq (V_a, V_b)$ instead of in relation to the top singular vector pair of Φ_A. This allows us to control the cross term in Eq. (7.4) more directly, which then implies strong convexity for nearly balanced inputs. As shown in Lemma 7.1.11, for nearly balanced inputs with λ small enough (or σ large enough), these definitions are nearly equivalent. The value of Definition 7.1.7 is that our results can be applied to inputs that are not nearly balanced if strong convexity is shown by other means.

In the following Section 7.1.4, we will combine these definitions to give strong bounds on tensor scaling.

7.1.4 Strong Convergence Bound

In this part, we will analyze strongly convex tensor inputs that are nearly balanced. The proof can be seen as an extension of standard arguments from strongly convex optimization to the geodesic setting. At the end, we will discuss the exact parameters of the theorem that will be improved in Section 7.2.

Our plan is to show that strong convexity is robust for small scalings $p \approx I_V$. This will allow us to use Lemma 2.3.7 to bound the optimizer of the Kempf-Ness function along each geodesic, which will then allow us to control the scaling solution.

We will need the following version of operator norm, which controls the change in strong convexity with respect to scalings.

Definition 7.1.12. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and let $(G, P, \mathfrak{p})$ be the scaling group according to Definition 6.2.3. Then, for $Z \in \mathfrak{p}$, the operator norm is defined as*

$$\|Z\|_{\text{op}} := \sum_{a\in[m]} \|Z_a\|_{\text{op}},$$

where $\|\cdot\|_{\text{op}}$ refers to the standard Euclidean operator norm on $L(V)$ and $L(V_a)$, respectively.

Similar to the discussion after Definition 7.1.2, note that this definition of the operator norm is motivated by the Euclidean operator norm of Z with respect to the embedding $Z \to \sum_{a\in[m]} I_{\overline{a}} \otimes Z_a \in L(V)$. We note that $\| \sum_{a\in[m]} I_{\overline{a}} \otimes Z_a\|_{\text{op}} \leq \sum_{a\in[m]} \|Z_a\|_{\text{op}}$, and in general, the inequality could be strict (e.g. $X = \text{diag}\{2, -1, -1\} \in \mathfrak{st}(3)$ and $Z := X \otimes I_3 - I_3 \otimes X$). We use the notation $\|Z\|_{\text{op}}$ for the norm in Definition 7.1.12 for simplicity.

This norm allows us to show that strong convexity is maintained along univariate restrictions $h(\eta) := f_x^P(\gamma_p(\eta Z))$ for any $p \in P$. In Lemma 7.2.13, this result is generalized to show that strong convexity is maintained for commutative scalings, and in Section 7.3 it is further generalized to non-commutative scaling groups.

Lemma 7.1.13. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3, and consider $x \in V^K$ with Kempf-Ness function f_x^P according to Definition 6.2.9. If f_x^P is α-geodesically strongly convex at $p \in P$ according to Definition 6.2.13, then for any direction $Z \in \mathfrak{p}$, the univariate restriction $h(\eta) := f_x^P(\gamma_p(\eta Z))$ is $e^{-\|\eta Z\|_{\text{op}}} \cdot \alpha \|Z\|_{\mathfrak{p}}^2$-strongly convex at $\eta \in \mathbb{R}$.*

Proof. By Definition 6.2.13, α-geodesic strong convexity of f_x^P at $p \in P$ implies that

$$\langle p^{1/2} \rho_x p^{1/2}, Z^2 \rangle = \partial_{\eta=0}^2 \langle \rho_x, p^{1/2} e^{\eta Z} p^{1/2} \rangle = \partial_{\eta=0}^2 f_x^P(\gamma_p(\eta Z)) \geq \alpha \|Z\|_{\mathfrak{p}}^2,$$

where the first step was by the derivative calculation $\partial_\eta^2 e^{\eta Z} = e^{\eta Z} Z^2$, in the second step we used Fact 6.2.11 for the geodesic $\gamma_p(\eta Z) = p^{1/2} e^{\eta Z} p^{1/2}$ as well as Definition 6.2.9 of the Kempf-Ness function f_x^P, and the final step was by Definition 6.2.13 of α-geodesic strong convexity in $\|\cdot\|_{\mathfrak{p}}$.

We can use this to bound the second derivative at other points as

$$\partial_\eta^2 f_x^P(\gamma_p(\eta Z)) = \langle p^{1/2}\rho_x p^{1/2}, e^{\eta Z} Z^2\rangle \geq e^{-\|\eta Z\|_{\mathrm{op}}} \langle p^{1/2}\rho_x p^{1/2}, Z^2\rangle \geq e^{-\|\eta Z\|_{\mathrm{op}}} \cdot \alpha \|Z\|_{\mathfrak{p}}^2,$$

where the first step was again by Definition 6.2.9 of the Kempf-Ness function and the derivative calculation, in the second step we used the spectral lower bound $e^{\eta Z} Z^2 \succeq e^{-\|\eta Z\|_{\mathrm{op}}} Z^2$ by Definition 7.1.12 of the operator norm in order to bound the inner product since both terms are positive semi-definite, and the final step was by α-geodesic strong convexity of f_x^P at $p \in P$ as calculated above. This verifies Definition 2.3.2 of strong convexity for $h(\eta) = f_x^P(\gamma_p(\eta Z))$. □

Remark 7.1.14. *The important property used in the proof of Lemma 7.1.13 was that e^Z commuted with the second order term Z^2. In Lemma 7.2.13, we will generalize this to show a similar robustness of strong convexity for commutative scalings. This generalization will be used to give an improved analysis in Section 7.2. In Section 7.3, we will further generalize this result to show that small perturbations maintain $\mathfrak{p}$-strong convexity for non-commutative scaling groups.*

In the following, we give a translation between norms so that we can apply standard gradient based analysis to the strongly convex scaling setting.

Lemma 7.1.15. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. Then for any $Z \in \mathfrak{p}$, the norms $\|\cdot\|_{\mathfrak{p}}$ (Definition 7.1.2) and $\|\cdot\|_{\mathrm{op}}$ (Definition 7.1.12) satisfy*

$$\Big(\sum_{a\in[m]} d_a\Big)^{-1} \|Z\|_{\mathrm{op}}^2 \leq \|Z\|_{\mathfrak{p}}^2 \leq \sum_{a\in[m]} \|Z_a\|_{\mathrm{op}}^2.$$

Proof. By Definition 7.1.2 of $\|\cdot\|_{\mathfrak{p}}$ and Definition 7.1.12 of $\|\cdot\|_{\mathrm{op}}$,

$$\|Z\|_{\mathfrak{p}}^2 = \sum_{a\in[m]} \frac{\|Z_a\|_F^2}{d_a} \leq \sum_{a\in[m]} \|Z_a\|_{\mathrm{op}}^2,$$

where the second step was by the bound $\|Z_a\|_F^2 \leq d_a \|Z_a\|_{\text{op}}^2$ for $Z_a \in L(V_a)$ with $\dim(V_a) = d_a$. To show the reverse bound, we calculate

$$\|Z\|_{\text{op}}^2 = \left(\sum_{a\in[m]} \|Z_a\|_{\text{op}}\right)^2 \leq \left(\sum_{a\in[m]} \|Z_a\|_F\right)^2 \leq \left(\sum_{a\in[m]} d_a\right)\left(\sum_{a\in[m]} \frac{\|Z_a\|_F^2}{d_a}\right) = \left(\sum_{a\in[m]} d_a\right)\|Z\|_{\mathfrak{p}}^2,$$

where the first step was by Definition 7.1.12 of the operator norm, in the second step we used $\|Z_a\|_{\text{op}} \leq \|Z_a\|_F$, the third step was by Cauchy-Schwarz, and the final step was by Definition 7.1.2 of $\|\cdot\|_{\mathfrak{p}}$. □

We can now apply the reduction in Theorem 6.3.1 to bound the tensor scaling solution.

Theorem 7.1.16. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ has size $s(x) = 1$ and is α-$\mathfrak{p}$-strongly convex according to Definition 7.1.7 with $\frac{\alpha}{e} \geq \sqrt{\sum_{a\in[m]} d_a} \cdot \|\nabla f_x^P(I_V)\|_{\mathfrak{p}}$, then there is a G-balanced scaling $x_* := e^{Z_*/2}\cdot x$ with $Z_* \in \mathfrak{p}$ satisfying*

$$\|Z_*\|_{\mathfrak{p}} \leq \frac{e\|\nabla f_x^P(I_V)\|_{\mathfrak{p}}}{\alpha}.$$

As a consequence, $s(x_*) \geq 1 - \frac{e\|\nabla_x\|_{\mathfrak{p}}^2}{2\alpha}$.

Proof. By Proposition 6.2.18(3), in order to find a G-balanced scaling, it is enough to find the global minimum of f_x^P. We decompose the domain P into univariate restrictions using the "sphere" $S_{\mathfrak{p}} := \{Z \in \mathfrak{p} \mid \|Z\|_{\mathfrak{p}} = 1\}$ as

$$P = e^{\mathfrak{p}} = \cup_{Z\in S_{\mathfrak{p}}} e^{\mathbb{R}Z},$$

where the first step is by the discussion in Section 2.2.3 of the Lie algebra $\mathfrak{p} := \log P$. We will apply Theorem 6.3.1 to find the optimizer of f_x^P, so by the partition above it is enough to bound the optimizers of the univariate restrictions $h_{Z\in S_{\mathfrak{p}}}(\eta) := f_x^P(e^{\eta Z})$.

For each univariate restriction, we will show that $|h_Z'(0)|$ is small and h_Z is $\frac{\alpha}{e}$-strongly convex on the interval $|\eta| \leq \frac{|h_Z'(0)|}{\alpha/e}$, which then allows us to apply Lemma 2.3.7 to bound the optimizer. We will use the shorthand $\nabla_x = \nabla f_x^P(I_V)$ as this is the only geodesic gradient we consider in this proof. First we bound the gradient by

$$|h_Z'(0)| = |\partial_{\eta=0} f_x^P(e^{\eta Z})| = |\langle \nabla_x, Z\rangle_{\mathfrak{p}}| \leq \|\nabla_x\|_{\mathfrak{p}}\|Z\|_{\mathfrak{p}} = \|\nabla_x\|_{\mathfrak{p}},$$

where the first step was by our definition of h_Z, the second was by Fact 6.2.11 of geodesics from the identity as well as Definition 7.1.1 of the geodesic gradient $\nabla_x = \nabla f_x^P(I_V)$, the third step was by Cauchy-Schwarz, and in the final step we used that $Z \in S_{\mathfrak{p}}$ so $\|Z\|_{\mathfrak{p}} = 1$.

To show strong convexity of h_Z, we first note that α-strong convexity of input x is equivalent to α-geodesic strong convexity of f_x^P at the identity by Lemma 7.1.8. Therefore, we can apply the robustness property of Lemma 7.1.13 to show $h_Z(\eta) = f_x^P(e^{\eta Z})$ is $e^{-\|\eta Z\|_{\text{op}}} \cdot \alpha$-strongly convex at $\eta \in \mathbb{R}$. In particular, h_Z is $\frac{\alpha}{e}$-strongly convex for all $|\eta| \leq (\sum_{a\in[m]} d_a)^{-1/2}$ as

$$\eta^2 \leq \Big(\sum_{a\in[m]} d_a\Big)^{-1} \implies \|\eta Z\|_{\text{op}}^2 \leq \eta^2 \Big(\sum_{a\in[m]} d_a\Big) \|Z\|_{\mathfrak{p}}^2 \leq 1,$$

where we used Lemma 7.1.15 to transfer between norms and $\|Z\|_{\mathfrak{p}} = 1$ as $Z \in S_{\mathfrak{p}}$. This implies h_Z is $\frac{\alpha}{e}$-strongly convex for $|\eta| \leq \frac{|h_Z'(0)|}{\alpha/e}$ as

$$\frac{|h_Z'(0)|}{\alpha/e} \leq \frac{\|\nabla_x\|_{\mathfrak{p}}}{\alpha/e} \leq \Big(\sum_{a\in[m]} d_a\Big)^{-1/2},$$

where in the first step we used gradient bound $|h_Z'(0)| \leq \|\nabla_x\|_{\mathfrak{p}}$ calculated above and the final step was exactly our assumption $\frac{\alpha}{e} \geq \sqrt{\sum_{a\in[m]} d_a} \cdot \|\nabla_x\|_{\mathfrak{p}}$. Therefore, Lemma 2.3.7 shows that the optimizer η_Z of h_Z satisfies

$$h_Z(\eta_Z) \geq h_Z(0) - \frac{|h_Z'(0)|^2}{2\alpha/e} \geq s(x) - \frac{e\|\nabla_x\|_{\mathfrak{p}}^2}{2\alpha} \qquad \text{and} \qquad |\eta_Z| \leq \frac{|h_Z'(0)|}{\alpha/e} \leq \frac{e\|\nabla_x\|_{\mathfrak{p}}}{\alpha},$$

where we used $h_Z(0) = f_x^P(I_V) = s(x)$ and the bound $|h'(0)| \leq \|\nabla_x\|_{\mathfrak{p}}$ calculated above.

Finally, since each univariate restriction in the partition $P = \cup_{Z\in S_{\mathfrak{p}}} e^{\mathbb{R}Z}$ has bounded optimizer $e^{\eta_Z Z} \in e^{\mathbb{R}Z}$, we can apply Theorem 6.3.1 to show that the global minimum of f_x^P is of the form $e^{Z_*} := e^{\eta_Z Z}$ for some $Z \in S_{\mathfrak{p}}$. By Proposition 6.2.18(3), this shows $e^{Z_*/2} \cdot x$ is G-balanced, and further we can lower bound the function and upper bound $\|Z_*\|_{\mathfrak{p}}$ using the univariate calculations above as

$$f_x^P(e^{Z_*}) \geq \inf_{Z\in S_{\mathfrak{p}}} h_Z(\eta_Z) \geq s(x) - \frac{e\|\nabla_x\|_{\mathfrak{p}}^2}{2\alpha} \qquad \text{and} \qquad \|Z_*\|_{\mathfrak{p}} \leq \sup_{Z\in S_{\mathfrak{p}}} |\eta_Z| \|Z\|_{\mathfrak{p}} \leq \frac{e\|\nabla_x\|_{\mathfrak{p}}}{\alpha}.$$

□

In the following Section 7.2 we will use gradient flow to show stronger bounds for the scaling solution. Specifically, we will be able to use bounds on $\|\nabla\|_{\text{op}}$ to directly analyze the path to the optimizer, instead of just relying on gradient bound $\|\nabla\|_{\mathfrak{p}}$. These results can be compared to the improved strong convexity analysis of Section 3.2 and the pseudorandom analysis of Section 3.3.

7.2 Improvement through Commutative Gradient Flow

In the previous Section 7.1, we gave the appropriate definitions required to lift simple gradient and strong convexity arguments to the geodesic setting. This is analogous to the result in Theorem 3.2.8 on matrix scaling. In this section, we will use structural properties of commutative gradient flow to strengthen these results. Specifically, we will analyze inputs that are ε-G-balanced so that $\|\nabla_x^{(a)}\|_{\text{op}} \leq s(x)\varepsilon$ for every $a \in [m]$. Note that by Fact 7.1.4, this implies $\|\nabla_x\|_{\mathfrak{p}}^2 \leq m \cdot (s(x)\varepsilon)^2$. In this section, we will go beyond standard convexity arguments by directly analyzing the operator norm of the solution through gradient flow. This will allow us to achieve the same conclusions as Theorem 7.1.16 while significantly weakening the assumption on strong convexity. The improved analyses in this section will be applied to give the best-known sample complexity and error bounds for the tensor normal model in Chapter 9.

By Theorem 6.3.1 we will reduce the analysis of general tensor scaling to commutative scaling groups. So in Section 7.2.1 we will review and simplify the definitions of the Kempf-Ness function and gradient flow in these settings. Then, in Section 7.2.2 we present a refined analysis of commutative gradient flow when the input is strongly convex. This is analogous to the improvement from Theorem 3.2.8 to Theorem 3.2.19 for strongly convex matrix scaling, which also came from directly analyzing the ∞-norm of the scaling solution instead of the $\mathfrak{t}$-norm. Finally in Section 7.2.3, we will define a "pseudorandom" condition on tensors and use it to show even faster convergence of gradient flow.

7.2.1 Simplified Setup for Commutative Tensors

In this subsection, we review the definitions of Section 6.2 and Section 7.1.4 for the simpler setting of commutative tensor scaling. We begin by repeating Definition 6.2.3 for commutative scaling groups.

Definition 7.2.1 (Commutative Tensor Scaling Group)**.** *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces $\{V_a\}_{a\in[m]}$ all over field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. A commutative tensor* *scaling group on V is defined as $T = (T_1, ..., T_m)$, where for each $a \in [m]$, $T_a = \mathrm{ST}_{\mathbb{F}}^{\Xi^a}(V_a)$ for some isometry $\Xi^a \in \mathrm{SU}(V_a)$ if $\mathbb{F} = \mathbb{C}$ and $\Xi^a \in \mathrm{SO}(V_a)$ if $\mathbb{F} = \mathbb{R}$. The polar part is denoted by $T_+ = \{\mathrm{ST}_+^{\Xi^a}(V_a)\}_{a\in[m]}$ along with associated vector space $\mathfrak{t} = \oplus_{a\in[m]}\mathfrak{st}_+^{\Xi^a}(V_a)$. We will sometimes refer to $(T, T_+, \mathfrak{t})$ as the commutative scaling group diagonal in the $\Xi = \{\Xi^a\}_{a\in[m]}$ basis.*

A commutative scaling group can be viewed as a set of diagonal matrices in the Ξ^a basis. As a consequence, all the elements of $(T, T_+, \mathfrak{t})$ commute. We will often reduce to the standard basis $T = (\mathrm{ST}(d_1), ..., \mathrm{ST}(d_m))$ by a change of basis on the input. This is only to reduce clutter, and our analysis will hold for any commutative scaling group.

Next, we give a simpler definition of the Kempf-Ness function for commutative scaling groups.

Definition 7.2.2. *Let $V = \otimes_{a\in[m]}V_a$ be a tensor product of inner product spaces with commutative scaling group $(T, T_+, \mathfrak{t})$ according to Definition 7.2.1. Then for tuple $x := \{x_1, ..., x_K\} \in V^K$, the Kempf-Ness function from Definition 6.2.9 can be equivalently defined as $f_x^{\mathfrak{t}} : \mathfrak{t} \to \mathbb{R}_+$ where*

$$f_x^{\mathfrak{t}}(Z) := f_x^{T_+}(e^Z) = \langle \rho_x, e^Z \rangle.$$

We could have given this form of the Kempf-Ness function in Definition 6.2.9 by the change of variable $P = e^{\mathfrak{p}}$. But for non-commutative scaling groups, the geodesic structure is more difficult to understand from this perspective as $e^{Y/2}e^Z e^{Y/2} \neq e^{Y+Z}$. Therefore we chose to give the definitions in Section 6.2 with respect to the polar P. In this section, we will focus on the commutative setting, and so the definition of geodesic gradient and geodesic gradient flow can also be simplified by this change of variable.

Proposition 7.2.3. *Let $V = \otimes_{a\in[m]}V_a$ be a tensor product of inner product spaces and consider commutative scaling group $(T, T_+, \mathfrak{t})$ diagonal in the Ξ basis according to Definition 7.2.1. Then for input $x = \{x_1, ..., x_K\} \in V^K$, the gradient of $f_x^{\mathfrak{t}}$ at $Y \in \mathfrak{t}$ satisfies*

$$\nabla f_x^{\mathfrak{t}}(Y) = \nabla f_x^{T_+}(e^Y) = \nabla f_{e^{Y/2}\cdot x}^{T_+}(I_V) = \nabla f_{e^{Y/2}\cdot x}^{\mathfrak{t}}(0),$$

and is defined component-wise as

$$(\nabla f_x^{\mathfrak{t}}(Y))^{(a)} = (\nabla f_x^{T_+}(e^Y))^{(a)} = \mathrm{diag}^{\Xi^a}\left(d_a \rho_{e^{Y/2}\cdot x}^{(a)} - s(e^{Y/2}\cdot x)I_a\right).$$

We will often use shorthand $\nabla_x = \nabla f_x^{\mathfrak{t}}(0)$ and $\nabla_x = \{\nabla_x^{(a)}\}_{a\in[m]}$ for the marginals.

Proof. The explicit expression for $(\nabla f_x^{\mathfrak{t}}(Y))^{(a)}$ follows directly from the analogous statement in Proposition 7.1.3 by the equality $(\nabla f_x^{\mathfrak{t}}(Y))^{(a)} = (\nabla f_x^{T_+}(e^Y))^{(a)}$. Therefore, we focus on proving the first statement. Note that the equality $\nabla f_x^{T_+}(e^Y) = \nabla f_{e^{Y/2}\cdot x}^{T_+}(I_V)$ was already shown in Proposition 7.1.3, and $\nabla f_{e^{Y/2}\cdot x}^{T_+}(I_V) = \nabla f_{e^{Y/2}\cdot x}^{\mathfrak{t}}(0)$ follows from the first equality by the change of variable $x' = e^{Y/2} \cdot x$ and $Y' = 0$.

We show the first equality $\nabla f_x^{\mathfrak{t}}(Y) = \nabla f_x^{T_+}(e^Y)$ by relating Definition 2.3.12 of the gradient and Definition 7.1.1 of the geodesic gradient. For fixed $Y \in \mathfrak{t}$, we can calculate the directional derivative for arbitrary $Z \in \mathfrak{t}$ as

$$\langle \nabla f_x^{\mathfrak{t}}(Y), Z\rangle_{\mathfrak{t}} = \partial_{\eta=0} f_x^{\mathfrak{t}}(Y + \eta Z) = \partial_{\eta=0} f_x^{T_+}(e^{Y+\eta Z}) = \partial_{\eta=0} f_x^{T_+}(\gamma_{e^Y}(\eta Z)) = \langle \nabla f_x^{T_+}(e^Y), Z\rangle_{\mathfrak{t}},$$

where the first step was by Definition 2.3.12 of the gradient of $f_x^{\mathfrak{t}}$, in the second step we used the equivalence $f_x^{\mathfrak{t}}(\cdot) = f_x^{T_+}(e^{\cdot})$ according to Definition 7.2.2, in the third step we used Fact 6.2.11 for the geodesic $\gamma_{e^Y}(\eta Z) = e^{Y/2} e^{\eta Z} e^{Y/2} = e^{Y+\eta Z}$ by commutativity of $(T, T_+, \mathfrak{t})$, and the final step was by Definition 7.1.1 of the geodesic gradient of $f_x^{T_+}$. The statement follows since $Z \in \mathfrak{t}$ was arbitrary. □

Now that we have an expression for the gradient of the commutative Kempf-Ness function, we can show that the gradient flow from Definition 7.1.5 can also be written as a dynamical system on vector space $\mathfrak{t}$ in the commutative case.

Proposition 7.2.4. *With the same assumptions Proposition 7.2.3, the gradient flow in Definition 7.1.5 can be written in terms of $Z_t \in \mathfrak{t}$ with initial condition $Z_0 = 0$ satisfying differential equation*

$$\partial_t Z_t = -\nabla f_{e^{Z_t/2}\cdot x}^{\mathfrak{t}}(0) = -\nabla f_x^{\mathfrak{t}}(Z_t).$$

This induces the dynamical system $x_t := e^{Z_t/2} \cdot x$.

Proof. Note that we have already shown $\nabla f_{e^{Z_t/2}\cdot x}^{\mathfrak{t}}(0) = \nabla f_x^{\mathfrak{t}}(Z_t)$ in Proposition 7.2.3. Let $x_t := g_t \cdot x$ be the dynamical system from Proposition 7.2.4 and $y_t := e^{Z_t/2} \cdot x$ be the dynamical system in this proposition. We will show that $x_t = y_t$ for all time so that these two equations define the same dynamical system. By the initial conditions $g_0 = I_V = e^{Z_0/2}$, the statement is true at time $t = 0$.

To show the equivalence for all time, we rewrite the differential equation for x_t as

$$\partial_t x_t = \partial_t(g_t \cdot x) = -\frac{1}{2}(\nabla f_{g_t\cdot x}^{T_+}(I_V)) \cdot g_t \cdot x = -\frac{1}{2}(\nabla f_{x_t}^{T_+}(I_V)) \cdot x_t,$$

where in the first and last steps we substituted $x_t = g_t \cdot x$, and in the second step we used Definition 7.1.5 for the gradient flow for g_t. Similarly, we rewrite

$$\partial_t y_t = \partial(e^{Z_t/2} \cdot x) = \frac{1}{2}(\partial_t Z_t) e^{Z_t/2} \cdot x = -\frac{1}{2}(\nabla f^{\mathfrak{t}}_{e^{Z_t/2}\cdot x}(0))(e^{Z_t/2} \cdot x) = -\frac{1}{2}(\nabla f^{\mathfrak{t}}_{y_t}(0)) \cdot y_t,$$

where in the first and last steps we substituted $y_t = e^{Z_t/2} \cdot x$, the second step was by the chain rule, and the third step was by the defining equation $\partial_t Z_t = -\nabla f^{\mathfrak{t}}_{e^{Z_t/2}\cdot x}(0)$.

The calculation above, along with the equivalence $\nabla f^{\mathfrak{t}}_{y_t}(0) = \nabla f^{T_+}_{y_t}(I_V)$ as given in Proposition 7.2.3, shows that if $x_t = y_t$, then the differential equations are also the same at time t. Therefore, since the initial conditions are the same, this implies that the two dynamical systems are equivalent and $x_t = g_t \cdot x = e^{Z_t/2} \cdot x = y_t$ for all time. □

Below, we state the formula for change in size over gradient flow which follows directly from the non-commutative version in Lemma 7.1.6 by the equivalence of gradient flows shown in Proposition 7.2.4.

Lemma 7.2.5. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and consider commutative scaling group $(T, T_+, \mathfrak{t})$ according to Definition 7.2.1. Then for input $x = \{x_1, ..., x_K\} \in V^K$ and $x_t = e^{Z_t/2} \cdot x$ the solution to gradient flow given in Proposition 7.2.4, the change in size is $\partial_t s(x_t) = -\|\nabla_{x_t}\|^2_{\mathfrak{t}}$. As a consequence*

$$s(x_T) - s(x) = -\int_0^T \|\nabla_{x_t}\|^2_{\mathfrak{p}}.$$

We can also simplify Definition 7.1.7 of strong convexity in the commutative setting.

Definition 7.2.6. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and consider commutative scaling group $(T, T_+, \mathfrak{t})$ according to Definition 7.2.1. Then input $x = \{x_1, ..., x_K\} \in V^K$ is α-$\mathfrak{t}$-strongly convex if $f^{\mathfrak{t}}_x$ is α-strongly convex at the origin:*

$$\forall Z \in \mathfrak{t} : \quad \partial^2_{\eta=0} f^{\mathfrak{t}}_x(\eta Z) \geq \alpha \|Z\|^2_{\mathfrak{t}} = \alpha \sum_{a\in[m]} \frac{\|Z_a\|^2_F}{d_a}.$$

By this change of variable (from T_+ to $\mathfrak{t}$), we can use tools from standard convex analysis on the vector function $f^{\mathfrak{t}}_x$. In particular, we can show that strong convexity implies fast convergence of gradient flow. This follows by lifting the argument in the proof of Proposition 3.2.2 to the tensor setting.

Proposition 7.2.7. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces and commutative scaling group $(T, T_+, \mathfrak{t})$ according to Definition 7.2.1. If input $x = \{x_1, ..., x_K\} \in V^K$ is α-$\mathfrak{t}$-strongly convex according to Definition 7.2.6, then the trajectory of gradient flow $x_t = e^{Z_t/2} \cdot x$ given in Proposition 7.2.4 satisfies $-\partial_{t=0}\|\nabla_{x_t}\|_{\mathfrak{t}}^2 \geq 2\alpha\|\nabla_x\|_{\mathfrak{t}}^2$. As a consequence, if x_t is α-$\mathfrak{t}$-strongly convex for $t \in [0, T]$ then*

$$\|\nabla_{x_T}\|_{\mathfrak{t}}^2 \leq e^{-2\alpha T}\|\nabla_x\|_{\mathfrak{t}}^2 \qquad \textit{and} \qquad \|Z_T\|_{\mathfrak{t}} \leq \frac{\|\nabla_x\|_{\mathfrak{t}}}{\alpha}.$$

Proof. This proof is a simple generalization of Proposition 3.2.2 in the matrix setting. So we first show $-\partial_{t=0}\|\nabla_{x_t}\|_{\mathfrak{t}}^2 = 2\partial_{\eta=0}^2 f_x^{\mathfrak{t}}(-\eta\nabla_x)$, which will imply the first statement by strong convexity. Starting from the left hand side, we calculate

$$\begin{aligned}
-\frac{1}{2}\partial_{t=0}\|\nabla_{x_t}\|_{\mathfrak{t}}^2 &= \langle \partial_{t=0}\nabla_{x_t}, -\nabla_x\rangle_{\mathfrak{t}} = \lim_{t\to 0} t^{-1}\langle \nabla_{x_t} - \nabla_x, -\nabla_x\rangle_{\mathfrak{t}} \\
&= \lim_{t\to 0} t^{-1}\Big(\partial_{\eta=0} f_{x_t}^{\mathfrak{t}}(-\eta\nabla_x) - \partial_{\eta=0} f_x^{\mathfrak{t}}(-\eta\nabla_x)\Big) \\
&= \lim_{t\to 0} t^{-1}\Big(\partial_{\eta=0} f_x^{\mathfrak{t}}(Z_t - \eta\nabla_x) - \partial_{\eta=0} f_x^{\mathfrak{t}}(-\eta\nabla_x)\Big) \\
&= \partial_{t=0}\partial_{\eta=0} f_x^{\mathfrak{t}}\Big(Z_t - \eta\nabla_x\Big),
\end{aligned}$$

where the first two steps are by calculus, in the third step we used the Definition 2.3.12 of the gradient to translate $\langle \nabla_y, Z\rangle = \partial_{\eta=0} f_y^{\mathfrak{t}}(\eta Z)$ for $y = x_t$ and $y = x$ in direction $Z = -\nabla_x = -\nabla f_x^{\mathfrak{t}}(0)$, and the fourth equality was by equivariance property of Lemma 6.2.6(1) applied with $x_t := e^{Z_t/2} \cdot x$ so $f_{x_t}^{\mathfrak{t}}(-\eta\nabla_x) = f_x^{\mathfrak{t}}(Z_t - \eta\nabla_x)$. To show this is equal to the right hand side, we calculate

$$\begin{aligned}
\partial_{\eta=0}^2 f_x^{\mathfrak{t}}(-\eta\nabla_x) &= \partial_\eta\Big(\partial_\eta f_x^{\mathfrak{t}}(-\eta\nabla_x)\Big)|_{\eta=0} = \partial_{\eta=0}\langle \nabla f_x^{\mathfrak{t}}(-\eta\nabla_x), -\nabla_x\rangle_{\mathfrak{t}} \\
&= \partial_{\eta=0}\langle \nabla f_x^{\mathfrak{t}}(-\eta\nabla_x), \partial_{t=0} Z_t\rangle_{\mathfrak{t}} = \partial_{\eta=0}\partial_{t=0} f_x^{\mathfrak{t}}\Big(Z_t - \eta\nabla_x\Big),
\end{aligned}$$

where in the second step we used Definition 2.3.12 of the gradient to replace $\partial_\eta f_x^{\mathfrak{t}}(-\eta\nabla_x) = \partial_{\nu=0} f_x^{\mathfrak{t}}(-\eta\nabla_x - \nu\nabla_x) = \langle \nabla f_x^{\mathfrak{t}}(-\eta\nabla_x), -\nabla_x\rangle_{\mathfrak{t}}$, the third step was by Definition 7.1.5 of commutative gradient flow $\partial_{t=0} Z_t = -\nabla f_x^{\mathfrak{t}}(Z_0) = -\nabla_x$ for initial condition $Z_0 = 0$, and the final step was by the Definition 2.3.12 of the gradient of $f_x^{\mathfrak{t}}$ as well as the chain rule for ∂_t with initial condition $Z_0 = 0$.

Therefore, we have the lower bound $-\partial_{t=0}\|\nabla_{x_t}\|_{\mathfrak{t}}^2 = 2\partial_{\eta=0}^2 f_x^{\mathfrak{t}}(-\eta\nabla_x) \geq 2\alpha\|\nabla_x\|_{\mathfrak{t}}^2$ by Definition 7.2.6 of strong convexity.

Equivalently, $-\partial_{t=0} \log \|\nabla_{x_t}\|_{\mathfrak{t}}^2 \geq 2\alpha$ by the chain rule. This implies

$$\log \|\nabla_{x_T}\|_{\mathfrak{t}}^2 - \log \|\nabla_x\|_{\mathfrak{t}}^2 = \int_{t=0}^{T} \partial_t \log \|\nabla_{x_t}\|_{\mathfrak{t}}^2 \leq -2\alpha T,$$

where the first step was by the fundamental theorem of calculus, and the second was by fast convergence inequality just derived. Exponentiating both sides gives the second statement.

The final statement is also a consequence of the fundamental theorem of calculus, as

$$\|Z_T\|_{\mathfrak{t}} = \left\| \int_0^T -\nabla_{x_t} \right\|_{\mathfrak{t}} \leq \int_0^T \|\nabla_{x_t}\|_{\mathfrak{t}} \leq \|\nabla_x\|_{\mathfrak{t}} \int_0^T e^{-\alpha t} \leq \frac{\|\nabla_x\|_{\mathfrak{t}}}{\alpha},$$

where in the first step we used $Z_0 = 0$ and $\partial_t Z_t = -\nabla_{x_t}$ according to Proposition 7.2.4 of gradient flow, in the second step is we used the triangle inequality on $\|\cdot\|_{\mathfrak{t}}$, and in the third step we used $\|\nabla_{x_t}\|_{\mathfrak{t}} \leq e^{-\alpha T} \|\nabla_x\|_{\mathfrak{t}}$ as shown in the second statement. □

This result can also be shown straightforwardly for non-commutative scalings by formally defining the geodesic Hessian. We do not give this definition as commutative gradient flow is sufficient for our analysis.

Next, we simplify Definition 7.1.12 of the operator norm for commutative scaling groups.

Definition 7.2.8. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with commutative scaling group $(T, T_+, \mathfrak{t})$ that is diagonal in basis Ξ according to Definition 7.2.1. The operator norm for $Z \in \mathfrak{t}$ is defined as*

$$\|Z\|_\infty := \sum_{a\in[m]} \|\operatorname{diag}^{\Xi_a}(Z_a)\|_{\mathrm{op}} = \sum_{a\in[m]} \max_{j_a\in[d_a]} |\langle \xi_{j_a}\xi_{j_a}^*, Z_a\rangle|.$$

This commutative version of the operator norm satisfies the same relations with $\|\cdot\|_{\mathfrak{t}}$ as the non-commutative version, so we repeat Lemma 7.1.15 for the commutative setting without proof.

Lemma 7.2.9. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with commutative scaling group $(T, T_+, \mathfrak{t})$ according to Definition 7.2.1. Then, the norms $\|\cdot\|_\infty$ (Definition 7.2.8) and $\|\cdot\|_{\mathfrak{t}}$ (Definition 7.1.2) satisfy the two-sided relations*

$$\Big(\sum_{a\in[m]} d_a\Big)^{-1} \|Z\|_\infty^2 \leq \|Z\|_{\mathfrak{t}}^2 \leq \sum_{a\in[m]} \|Z_a\|_\infty^2.$$

With these definitions in hand, we will make some structural observations about gradient flow for commutative scalings when the input is ε-T-balanced. This will allow us to give an improved analysis in the strongly convex setting in Section 7.2.2, as well as another strong bound on the scaling solution when the input satisfies a pseudorandom condition. As mentioned previously, both of these results will be lifted to the non-commutative scaling setting using Theorem 6.3.1.

7.2.2 Strongly Convex Analysis

In this subsection, we will use commutative gradient flow to analyze the scaling solution for strongly convex inputs. In Theorem 7.2.16 at the end of this subsection, we will use Theorem 6.3.1 to lift this result to the general non-commutative tensor scaling problem. This gives a small polynomial improvement as compared to Theorem 7.1.16 and will be applied to give a better sample complexity result for the tensor normal model in Chapter 9.

In Theorem 7.1.16, our analysis required $\alpha \gtrsim \sqrt{\sum_{a\in[m]} d_a}\|\nabla_x\|_{\mathfrak{p}}$ in order to show that every univariate restriction $\eta \to f_x^P(e^{\eta Z})$ is strongly convex for a large enough interval. The extra factor is due to the inequality $\|\cdot\|_{\text{op}} \leq \sqrt{\sum_{a\in[m]} d_a} \cdot \|\cdot\|_{\mathfrak{p}}$ which is tight in general. In this part, our inputs will be ε-T-balanced according to Definition 6.2.4, which implies the gradient bound $\|\nabla_x\|_{\mathfrak{t}}^2 \leq m\varepsilon^2$ by Fact 7.1.4. To bound the scaling solution, we will follow commutative gradient flow and directly analyze $\|Z_t\|_\infty$ instead of analyzing $\|Z_t\|_{\mathfrak{t}}$ via bounds on the gradient. This will allow us to show that strong convexity is maintained for all time until the optimum is reached, which implies the main theorem of this section by a similar strategy to Theorem 3.2.19.

In this and the following Section 7.2.3, we will focus on the case when $V = \otimes_{a\in[m]}\mathbb{F}^{d_a}$ with standard diagonal scaling groups $T = (\text{ST}(d_1), \dots \text{ST}(d_m))$. The results generalize to an arbitrary choice of commutative scaling group in a straightforward manner by change of basis, and will be lifted to the non-commutative setting at the end of this subsection.

Recall by Definition 7.2.8 and Proposition 7.2.3 for the commutative T-scaling problem,

$$\|\nabla_x\|_\infty = \sum_{a\in[m]} \|(\nabla f_x^{\mathfrak{t}}(0))^{(a)}\|_{\text{op}} = \sum_{a\in[m]} \left\| \operatorname{diag}\left(d_a\rho_x^{(a)} - s(x)I_a\right)\right\|_{\text{op}}, \tag{7.5}$$

where we restricted to the diagonal basis for $T = (\text{ST}(d_1), ..., \text{ST}(d_m))$. Below, we explicitly calculate the change in the quantities $s(x)$ and $\rho_x^{(a)}$ to show that $\|\nabla_{x_t}\|_\infty$ changes somewhat slowly under gradient flow. Once again, we ignore questions of differentiability for the infinity norm (see Remark 3.2.11 for this technical detail).

Lemma 7.2.10. *Let $V = \otimes_{a\in[m]}\mathbb{F}^{d_a}$ with scaling group $T = (\mathrm{ST}(d_1), ..., \mathrm{ST}(d_m))$ and associated polar $(T_+, \mathfrak{t})$ according to Definition 7.2.1. Then for input $x \in V^K$ and x_t the solution to gradient flow according to Proposition 7.2.4, the change in the diagonals of ρ_{x_t} can be explicitly calculated as*

$$-\partial_{t=0}\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle = \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle\langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle + \sum_{b\neq a\in[m]} \langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes \nabla_x^{(b)}\rangle,$$

where $E_{ii}^{(a)}$ are the diagonals matrices in the standard basis of $V_a = \mathbb{F}^{d_a}$.

Proof. By Proposition 7.2.4 of gradient flow, we can calculate

$$\begin{aligned}
-\partial_{t=0}\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle &= -\partial_{t=0}\langle E_{ii}^{(a)} \otimes I_{\overline{a}}, e^{Z_t/2}\rho_x e^{Z_t/2}\rangle = -\langle(\partial_{t=0}Z_t)(E_{ii}^{(a)} \otimes I_{\overline{a}}), \rho_x\rangle \\
&= \sum_{b\in[m]} \langle(\nabla_x^{(b)} \otimes I_{\overline{b}})(E_{ii}^{(a)} \otimes I_{\overline{a}}), \rho_x\rangle \\
&= \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle\langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle + \sum_{b\neq a\in[m]} \langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes \nabla_x^{(b)}\rangle,
\end{aligned}$$

where the first step was by Definition 6.2.2 of the partial trace and marginals as well as the equivariance property in Lemma 6.2.6(1), in the second step we used the initial condition $Z_0 = 0$ along with the fact that Z_a commutes with $E_{ii}^{(a)}$, in the third step we applied Proposition 7.2.4 of gradient flow, and the final step was once again by Definition 6.2.2 of the marginals along with the fact that $\nabla \in \mathfrak{t}$ so $\nabla_x^{(a)}$ commutes with $E_{ii}^{(a)}$. □

We will use the above formula to show that $\|\nabla_{x_t}\|_\infty$ changes somewhat slowly under gradient flow. More precisely, we will bound the rate of increase for small t, and then show that it decreases exponentially for large t. If x_t is α-$\mathfrak{t}$-strongly convex for $t \in [0, T]$, then

$$\|\nabla_{x_T}^{(a)}\|_\infty^2 \leq d_a\|\nabla_{x_T}^{(a)}\|_{\mathfrak{t}}^2 \leq d_a\|\nabla_{x_T}\|_{\mathfrak{t}}^2 \leq d_a\|\nabla_x\|_{\mathfrak{t}}^2 e^{-2\alpha T}, \tag{7.6}$$

where the first step is by the relation between $\|\cdot\|_\infty$ and $\|\cdot\|_{\mathfrak{t}}$ in Definition 7.2.8, and the final step was by Proposition 7.2.7. For input x that is ε-T-balanced, we can use Lemma 7.2.10 to improve this d_a-factor loss near the beginning of gradient flow.

Lemma 7.2.11. *Let $V = \otimes_{a\in[m]}\mathbb{F}^{d_a}$ with scaling group $T = (\mathrm{ST}(d_1), ..., \mathrm{ST}(d_m))$ and associated polar $(T_+, \mathfrak{t})$ according to Definition 7.2.1. Then for input $x \in V^K$ with x_t the solution to gradient flow according to Proposition 7.2.4,*

$$\partial_t \max_{a\in[m]} \log\|\nabla_{x_t}^{(a)}\|_\infty \leq (m-2)s(x_t) + (2m-2)\max_{a\in[m]}\|\nabla_{x_t}^{(a)}\|_\infty.$$

Proof. We will show that for every $y \in V^K$, $\partial_{t=0} \max_{a\in[m]} \log \|\nabla_{yt}^{(a)}\|_\infty \leq (m-2)s(y) + (2m-2)\max_{a\in[m]} \|\nabla_y^{(a)}\|_\infty$. The lemma follows for arbitrary t by considering gradient flow starting at $y = x_t$.

Let $a = \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_\infty$ and further let $i \in \arg\max_{j\in[d_a]} |\langle E_{jj}^{(a)}, d_a \cdot \rho_x^{(a)} - s(x) I_a\rangle|$ be the diagonal with the worst error. We will separate into two cases depending on the sign of this error. We first bound the change in size:

$$-\partial_{t=0} s(x_t) = \|\nabla_x\|_{\mathfrak{t}}^2 \leq \sum_{b\in[m]} \|\nabla_x^{(b)}\|_\infty^2 \leq \|\nabla_x^{(a)}\|_\infty \|\nabla_x\|_\infty, \tag{7.7}$$

where the first step was by Lemma 7.2.5, the second was by Lemma 3.2.7, and the third was by our case assumption $a = \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_\infty$.

Now consider the case $\|\nabla_x^{(a)}\|_\infty = \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle = d_a \langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle - s(x)$, meaning this diagonal is larger than average. We bound its increase by Lemma 7.2.10:

$$\begin{aligned}
\partial_{t=0}\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle &= -\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - \sum_{b\neq a\in[m]} \langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes \nabla_x^{(b)}\rangle \\
&\leq -\|\nabla_x^{(a)}\|_\infty \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle + \sum_{b\neq a\in[m]} \|\nabla_x^{(b)}\|_\infty \langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes I_b\rangle \\
&= \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle \Big(- \|\nabla_x^{(a)}\|_\infty + \sum_{b\neq a\in[m]} \|\nabla_x^{(b)}\|_\infty \Big) \\
&\leq (m-2)\|\nabla_x^{(a)}\|_\infty \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle,
\end{aligned}$$

where the first step was by Lemma 7.2.10, in the second step we used our case assumption $\|\nabla_x^{(a)}\|_\infty = \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle$ to bound the first term and $|\nabla_x^{(b)}| \preceq \|\nabla_x^{(b)}\|_\infty \cdot I_b$ to bound the second term, and in the final step we used the case assumption $\|\nabla_x^{(a)}\|_\infty \geq \|\nabla_x^{(b)}\|_\infty$ for all $b \in [m]$. Ignoring questions of differentiability for $\|\cdot\|_\infty$ (see Remark 3.2.11), this allows us to bound the change in $\nabla_x^{(a)}$ by

$$\partial_{t=0} \log \|\nabla_{x_t}^{(a)}\|_\infty = \frac{\partial_{t=0} d_a \langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle - \partial_{t=0} s(x_t)}{\|\nabla_x^{(a)}\|_\infty} \leq (m-2) d_a \langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle + \|\nabla_x\|_\infty,$$

where in the last step we used the bound derived above for change in $\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle$ and Eq. (7.7) for the change in size.

In the other case $-\|\nabla_x^{(a)}\|_\infty = \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle$, we bound the decrease of the diagonal $\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle$:

$$\begin{aligned} -\partial_{t=0}\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle &= \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle + \sum_{b\neq a\in[m]} \langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes \nabla_x^{(b)}\rangle \\ &\leq -\|\nabla_x^{(a)}\|_\infty \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle + \sum_{b\neq a\in[m]} \|\nabla_x^{(b)}\|_\infty \langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes I_b\rangle \\ &= \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle \Big(\sum_{b\neq a\in[m]} \|\nabla_x^{(b)}\|_\infty - \|\nabla_x^{(a)}\|_\infty \Big) \\ &\leq (m-2)\|\nabla_x^{(a)}\|_\infty \langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle, \end{aligned}$$

where the first step was by Lemma 7.2.10, in the second step we used our case assumption $\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle = -\|\nabla_x^{(a)}\|_\infty$ to bound the first term and $|\nabla_x^{(b)}| \preceq \|\nabla_x^{(b)}\|_\infty \cdot I_b$ to bound the second term, and in the final step we used that $a = \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_\infty$. This allows us to bound the change in $\nabla_x^{(a)}$ as

$$\partial_{t=0} \log \|\nabla_{x_t}^{(a)}\|_\infty = \frac{\partial_{t=0} s(x_t) - \partial_{t=0} d_a \langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle}{\|\nabla_x^{(a)}\|_\infty} \leq (m-2) d_a \langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle - 0,$$

where in the last step we used the bounds derived above for change in $\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle$ and the equality $\partial_{t=0} s(x_t) = -\|\nabla_x\|_{\mathfrak{t}}^2$ given in Eq. (7.7).

In both cases, we can bound the change by

$$\partial_{t=0} \log \|\nabla_{x_t}^{(a)}\|_\infty \leq (m-2)(s(x) + \|\nabla_x^{(a)}\|_\infty) + \|\nabla_x\|_\infty \leq (m-2)s(x) + (2m-2)\|\nabla_x^{(a)}\|_\infty,$$

where the first step was by $d_a\langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle = s(x) \pm \langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle \leq s(x) + \|\nabla_x^{(a)}\|_\infty$, and in the final step we used $a \in \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_\infty$ so $\|\nabla_x\|_\infty \leq m\|\nabla_x^{(a)}\|_\infty$ according to Definition 7.2.8. □

This generalizes the bound on error shown in Lemma 3.2.10 for the matrix case. Note that for $m = 2$, the first term vanishes, and so the bound in Lemma 7.2.11 depends only on the error $\|\nabla_x\|_\infty$. This allows us to bound $\|(X_T, Y_T)\|_\infty$ in Proposition 3.2.13 and Proposition 3.2.18 for strongly convex inputs. For the $m \geq 3$ tensor case, we have a much worse bound on the error over time. Below, we combine Lemma 7.2.11 with Eq. (7.6) to improve the bound on the scaling solution in Theorem 7.1.16 for strongly convex inputs.

Proposition 7.2.12. *Let $V = \otimes_{a\in[m]}V_a$ be a tensor product of $m \geq 3$ inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$, and let $(T, T_+, \mathfrak{t})$ be a commutative scaling group according to Definition 7.2.1. Let $x \in V^K$ be an input of size $s(x) = 1$ and x_t be the solution of gradient flow according to Proposition 7.2.4. Assume x is ε-T-balanced and x_t is α-$\mathfrak{t}$-strongly convex and satisfies $\max_{a\in[m]} \|\nabla^{(a)}_{x_t}\|_\infty \leq \frac{1}{2m}$ for all $t \in [0, T]$. Then*

$$\|Z^{(a)}_T\|_\infty \leq \varepsilon\sqrt{md_a}^{1-\frac{\alpha}{m-1+\alpha}}\left(\frac{1}{m-1} + \frac{1}{\alpha}\right) \leq \frac{3}{2}\cdot\frac{\varepsilon\sqrt{md_a}^{1-\frac{\alpha}{m}}}{\alpha}.$$

Proof. We can assume without loss that $V_a = \mathbb{F}^{d_a}$ and $T_a = \mathrm{ST}(d_a)$ for each $a \in [m]$ by a change of basis if necessary.

We first show that for any $a \in [m]$ and any $t \in [0, T]$, the gradient is bounded by

$$\|\nabla^{(a)}_{x_t}\|_\infty \leq \min\{\varepsilon e^{(m-1)t}, \varepsilon\sqrt{md_a}e^{-\alpha t}\} \leq \varepsilon\sqrt{md_a}^{1-\frac{\alpha}{m-1+\alpha}}. \tag{7.8}$$

The bound on $\|Z\|_\infty$ will follow by the fundamental theorem of calculus. This bound will also be useful in the proof of Theorem 7.2.15.

For the first term in the min, we use Lemma 7.2.11 to show

$$\begin{aligned}\max_{a\in[m]}\log\|\nabla^{(a)}_{x_t}\|_\infty &\leq \max_{a\in[m]}\log\|\nabla^{(a)}_{x}\|_\infty + \int_0^t \partial_\tau \max_{a\in[m]}\log\|\nabla^{(a)}_{x_\tau}\|_\infty \\ &\leq \log\varepsilon + \int_0^t \Big((m-2)s(x_\tau) + (2m-2)\|\nabla^{(a)}_{x_\tau}\|_\infty\Big) \\ &\leq \log\varepsilon + \int_0^t (m-1) = \log(\varepsilon e^{(m-1)t}),\end{aligned}$$

where the first step was by the fundamental theorem of calculus, in the second step we used the bound $\|\nabla^{(a)}_x\|_\infty \leq s(x)\varepsilon \leq \varepsilon$ for the first term as x has size $s(x) = 1$ and is ε-T-balanced according to Definition 6.2.4 and Lemma 7.2.11 to bound the rate of change in the second term, and in the third step we used $s(x_\tau) \leq s(x) = 1$ to bound the first term in the integral (since $\partial_t s(x_t) \leq 0$ by Lemma 7.2.5) and the assumption $\|\nabla^{(a)}_{x_\tau}\|_\infty \leq \frac{1}{2m}$ for all $\tau \in [0, T]$ to bound the second term.

For the second term in the min expression in Eq. (7.8), we use the exponential convergence of Proposition 7.2.7 to show

$$\|\nabla^{(a)}_{x_t}\|_\infty \leq \sqrt{d_a}\|\nabla^{(a)}_{x_t}\|_{\mathfrak{t}} \leq \sqrt{d_a}\|\nabla_{x_t}\|_{\mathfrak{t}} \leq \sqrt{d_a}\|\nabla_x\|_{\mathfrak{t}}e^{-\alpha t} \leq \varepsilon\sqrt{md_a}\cdot e^{-\alpha t},$$

where the first step was by the relation in Lemma 7.2.9, the third step was by Proposition 7.2.7 applied with α-$\mathfrak{t}$-strong convexity till time t, and in the final step we used the bound $\|\nabla_x\|_{\mathfrak{t}}^2 \leq m\varepsilon^2$ by Fact 7.1.4 for ε-T-balanced x of size $s(x) = 1$.

Therefore, we have shown both bounds in the first inequality of Eq. (7.8). This is all that is necessary to bound $\|Z_T^{(a)}\|_\infty$. We show the second inequality for use in the proof of Theorem 7.2.15. We optimize the upper bound in Eq. (7.8) by balancing terms:

$$\varepsilon e^{(m-1)t} = \varepsilon\sqrt{md_a}e^{-\alpha t} \iff (m-1)t = \log\sqrt{md_a} - \alpha t \iff t = \frac{\log\sqrt{md_a}}{m-1+\alpha}.$$

Substituting this in to Eq. (7.8) gives

$$\max_{t\in[0,T]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \max_{t\geq 0} \min\{\varepsilon e^{(m-1)t}, \varepsilon\sqrt{md_a}e^{-\alpha t}\} \leq \varepsilon\sqrt{md_a}^{1-\frac{\alpha}{m-1+\alpha}},$$

where we plugged in $t = \frac{\log\sqrt{md_a}}{m-1+\alpha}$ for the last step.

To show the bound on $\|Z_T^{(a)}\|_\infty$, we apply the gradient bound in Eq. (7.8) up till time T. We assume $T \geq \kappa := \frac{\log\sqrt{md_a}}{m-1+\alpha}$, as the argument below is only stronger otherwise. We can bound

$$\begin{aligned}\|Z_T^{(a)}\|_\infty = \left\|\int_0^T -\nabla_{x_t}^{(a)}\right\|_\infty &\leq \int_0^\kappa \|\nabla_{x_t}^{(a)}\|_\infty + \int_\kappa^T \|\nabla_{x_t}^{(a)}\|_\infty \leq \int_0^\kappa \varepsilon e^{(m-1)t} + \varepsilon\sqrt{md_a}\int_\kappa^T e^{-\alpha t} \\ &\leq \frac{\varepsilon e^{(m-1)\kappa}}{m-1} + \frac{\varepsilon\sqrt{md_a}\cdot e^{-\alpha\kappa}}{\alpha} = \varepsilon\sqrt{md_a}^{1-\frac{\alpha}{m-1+\alpha}}\left(\frac{1}{m-1} + \frac{1}{\alpha}\right),\end{aligned}$$

where the first step was by the fundamental theorem of calculus for gradient flow defined by $Z_0 = 0$ and $\partial_t Z_t = -\nabla_{x_t}$ according to Proposition 7.2.4, the second step was by the triangle inequality for $\|\cdot\|_\infty$, in the third step we used Eq. (7.8) to bound the first stage by $\|\nabla_{x_t}^{(a)}\|_\infty \leq \varepsilon e^{(m-1)t}$ and the second stage by $\|\nabla_{x_t}^{(a)}\|_\infty \leq \exp\sqrt{md_a}\cdot e^{-\alpha t}$, the fourth step was by integration, and in the final step we plugged in $\kappa = \frac{\log\sqrt{md_a}}{m-1+\alpha}$. The final inequality in the proposition follows by the bounds $m-1+\alpha \leq m$ as $\alpha \leq s(x) \leq 1$ by Proposition A.5.2, and $\frac{1}{m-1} \leq \frac{1}{2\alpha}$ as $m \geq 3$. □

The above proposition improves upon the bound on Z from Theorem 7.1.16 as shown below. For ε-T-balanced input x of size $s(x) = 1$, we have $\|\nabla_x\|_{\mathfrak{t}}^2 \leq m\varepsilon^2$ by Fact 7.1.4. Therefore, if x is an $\alpha \geq \Omega(1)$-$\mathfrak{t}$-strongly convex input with $1 \gtrsim \varepsilon\sqrt{m\sum_{a\in[m]} d_a}$, then

$$\|Z_T^{(a)}\|_\infty \leq \|Z_T^{(a)}\|_F \leq \sqrt{d_a}\|Z_T\|_{\mathfrak{t}} \leq \sqrt{d_a}\frac{\|\nabla_x\|_{\mathfrak{t}}}{\alpha/e} \lesssim \varepsilon\sqrt{md_a}, \tag{7.9}$$

where the first step was by the inequality $\|\cdot\|_{\text{op}} \leq \|\cdot\|_F$, the second step was by Definition 7.1.2 of $\|\cdot\|_{\mathfrak{t}}$, the third step was by the conclusion of Theorem 7.1.16, and the final step was by the calculation $\|\nabla_x\|_{\mathfrak{t}}^2 \leq m\varepsilon^2$ as shown in Fact 7.1.4. Comparing this to Proposition 7.2.12, we get an improvement of $d_a^{\Omega(1/m)}$ in the conclusion. This will be useful to give improved error bounds for our statistical application in Chapter 9.

The other valuable part of this analysis is that it allows us to weaken the strong convexity assumption of Theorem 7.1.16. Specifically, we will eventually show that the conditions required throughout gradient flow in Proposition 7.2.12 are implied by sufficient strong convexity of the initial input. To this end, we generalize Lemma 7.1.13 to show robustness of $\mathfrak{t}$-strong convexity.

Lemma 7.2.13. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$, and let $(T, T_+, \mathfrak{t})$ be a commutative scaling group according to Definition 7.2.1. If input $x \in V^K$ is α-$\mathfrak{t}$-strongly convex, then for any $Y \in \mathfrak{t}$ the scaling $e^{Y/2}\cdot x$ is $e^{-\|Y\|_{\text{op}}}\cdot\alpha$-$\mathfrak{t}$-strongly convex.*

Proof. We will lower bound $\partial^2_{\eta=0} f^{\mathfrak{t}}_{e^{Y/2}\cdot x}(\eta Z)$ for arbitrary $Z \in \mathfrak{t}$ to verify Definition 7.2.6 of strong convexity. Note that $(T, T_+, \mathfrak{t})$ are all commutative, so $Y, Z \in \mathfrak{t}$ implies Y, Z commute. Therefore we can lower bound the second-order derivative by

$$\partial^2_{\eta=0} f_{e^{Y/2}\cdot x}(e^{\eta Z}) = \partial^2_{\eta=0}\langle \rho_{e^{Y/2}\cdot x}, e^{\eta Z}\rangle = \langle \rho_x, e^{Y/2} Z^2 e^{Y/2}\rangle \geq e^{-\|Y\|_\infty}\langle \rho_x, Z^2\rangle \geq e^{-\|Y\|_\infty}\alpha\|Z\|_{\mathfrak{t}}^2,$$

where the first step was by Definition 7.2.2 of the Kempf-Ness function, the second was by equivariance from Lemma 6.2.6(1) as well as standard matrix calculus $\partial_x e^x = xe^x$, in the third step we used the spectral lower bound $e^{Y/2}Z^2e^{Y/2} \succeq e^{-\|Y\|_\infty} Z^2$ for commuting $Y, Z \in \mathfrak{t}$ to lower bound the inner product since $\rho_x \succeq 0$ and $Z \in \mathfrak{t} \subseteq \mathrm{H}(V)$ so $Z^2 \succeq 0$, and the final step was by Definition 7.2.6 of α-strong convexity of x. Since $Z \in \mathfrak{t}$ was arbitrary, this verifies $e^{-\|Y\|_\infty}\cdot\alpha$-$\mathfrak{t}$-strong convexity according to Definition 7.2.6. □

Remark 7.2.14. *In Section 7.3, we will further generalize this result to show that small perturbations maintain $\mathfrak{p}$-strong convexity for non-commutative scaling groups. In this case we will only be able to show additive bounds of the form $\alpha - O(\delta)$ for scaling $\|Y\|_{\text{op}} \lesssim \delta$, which becomes vacuous for $\delta \approx \alpha$. For the purpose of our sample complexity results in Chapter 9, this only induces a constant factor loss, as in that setting we analyze random inputs which are shown to satisfy $\alpha \approx 1$ strong convexity. But as discussed in Section 4.2.3, the multiplicative robustness result was crucial for our work on the Paulsen problem, and we believe it is of independent interest, since Appendix A.2 shows that it is a tight result.*

At this point, we can refine our analysis to simultaneously weaken the assumption of Theorem 7.1.16 by a polynomial factor and give slightly stronger bounds on the scaling. The theorem below will be applied in Chapter 9 to get a small polynomial improvement in sample complexity for the tensor normal model.

Theorem 7.2.15. *For $m \geq 3$, let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for $a \in [m]$, and let $(T, T_+, \mathfrak{t})$ be a commutative scaling group according to Definition 7.2.1. If input $x \in V^K$ of size $s(x) = 1$ is ε-T-balanced and α-$\mathfrak{t}$-strongly convex with $\frac{\alpha}{\sqrt{e}} \geq 6m \cdot \varepsilon\sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}$, then:*

1. *For all time $T \geq 0$, the solution Z_T of gradient flow satisfies*

$$\|Z_T\|_{\mathfrak{t}} \leq \frac{\varepsilon\sqrt{m}}{\alpha/\sqrt{e}} \qquad \text{and} \qquad \forall a \in [m] : \|Z_T^{(a)}\|_\infty \leq \frac{3\varepsilon\sqrt{md_a}^{1-\frac{\alpha/\sqrt{e}}{m}}}{2\alpha/\sqrt{e}}.$$

2. *The limit $Z_\infty := \lim_{t\to\infty} Z_t$ exists and $x_\infty := e^{Z_\infty/2} \cdot x$ is a T-balanced scaling solution to the T-tensor scaling problem in Definition 6.2.5;*

3. *The size of the solution can be lower bounded by*

$$s(x_*) = f_x^{\mathfrak{t}}(Z_*) \geq 1 - \frac{m\varepsilon^2}{2\alpha/\sqrt{e}}.$$

Proof. We claim that for all time, x_t satisfies the gradient bound $\max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \frac{1}{2m}$ and is at least $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strongly convex. For contradiction, assume that T is the last time all of these conditions hold. Let us first consider the case that $\|\nabla_{x_T}^{(a)}\|_\infty = \frac{1}{2m}$ and the gradient condition does not hold after time T. Then, we can apply Proposition 7.2.12 for ε-T-balanced input x with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity up till time T to show

$$\|\nabla_{x_T}^{(a)}\|_\infty \leq \varepsilon\sqrt{md_a}^{1-\frac{\alpha/\sqrt{e}}{m}} \leq \frac{\alpha/\sqrt{e}}{6m} \leq \frac{1}{6m},$$

where the first step was by the bound in Eq. (7.8) applied with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity, the second step was by our assumption $\frac{\alpha}{\sqrt{e}} \geq 6m \cdot \varepsilon\sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}$, and the final step was by the bound $\alpha \leq s(x) = 1$ by Proposition A.5.2. By continuity, T cannot be the last time $\max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \frac{1}{2m}$, so the balance condition cannot fail first.

Then, we can consider the case that T is the last time x_T is $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strongly convex. By Lemma 7.2.13 in the contrapositive, this implies $\|Z_T\|_\infty \geq \frac{1}{2}$. But we can apply Proposition 7.2.12 with gradient bound $\|\nabla^{(a)}_{x_t}\|_\infty \leq \frac{1}{2m}$ and $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity up till time T to bound

$$\|Z_T\|_\infty \leq m \cdot \max_{a\in[m]} \|Z_T^{(a)}\|_\infty \leq m \cdot \frac{3m \cdot \varepsilon\sqrt{md_{\max}}^{\,1-\frac{\alpha/\sqrt{e}}{m}}}{2\alpha/\sqrt{e}} \leq \frac{1}{4},$$

where the first step was by Definition 7.2.8 of the infinity norm, the second step was by Proposition 7.2.12 applied with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity, and the final bound was by our assumption $\frac{\alpha}{\sqrt{e}} \geq 6m \cdot \varepsilon\sqrt{md_{\max}}^{\,1-\frac{\alpha/\sqrt{e}}{m}}$. This gives the required contradiction, so the assumptions must hold for all time.

Now that we have the claim, we can apply Proposition 7.2.12 for all time. The bound on $\|Z_T^{(a)}\|_\infty$ in item (1) follows from the conclusion of Proposition 7.2.12 with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity for any $T \geq 0$. The bound in item (1) also follows by strong convexity as

$$\|Z_T\|_{\mathfrak{t}} = \left\| \int_0^T -\nabla_{x_t} \right\|_{\mathfrak{t}} \leq \int_0^T \|\nabla_{x_t}\|_{\mathfrak{t}} \leq \|\nabla_x\|_{\mathfrak{t}} \int_0^T e^{-\alpha t/\sqrt{e}} \leq \frac{\varepsilon\sqrt{m}}{\alpha/\sqrt{e}},$$

where the first step was by the fundamental theorem of calculus as $Z_0 = 0$ and $\partial_t Z_t = -\nabla_{x_t}$ according to Proposition 7.2.4, the second step was by the triangle inequality on $\|\cdot\|_{\mathfrak{t}}$, in the third step we applied Proposition 7.2.7 with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity, and in the final step we used $\|\nabla_x\|_{\mathfrak{t}}^2 \leq m\varepsilon^2$ by Fact 7.1.4 applied to ε-T-balanced input x of size $s(x) = 1$.

To show item (2), we can again use strong convexity for all time to bound

$$\lim_{T\to\infty} \int_{t\geq T} \|\partial_t Z_t\|_{\mathfrak{t}} = \lim_{T\to\infty} \int_{t\geq T} \|\nabla_{x_t}\|_{\mathfrak{t}} \leq \lim_{T\to\infty} \|\nabla_A\|_{\mathfrak{t}} \int_{t\geq T} e^{-\alpha t/\sqrt{e}} = 0,$$

where the first step was by Proposition 7.2.4 of gradient flow, and the second was by Proposition 7.2.7 with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity for all time. This implies that the limit $Z_\infty \in \mathfrak{t}$ exists, and $x_\infty = e^{Z_\infty/2} \cdot x$ satisfies $\nabla_{x_\infty} = \nabla f_x^{\mathfrak{t}}(Z_\infty) = 0$, so x_∞ is T-balanced by Proposition 6.2.18(2).

To show item (3), we can bound the change in size over gradient flow by

$$s(x) - s(x_*) = \int_0^\infty \|\nabla_{x_t}\|_{\mathfrak{t}}^2 \leq \int_0^\infty \|\nabla_x\|_{\mathfrak{t}}^2 e^{-2\alpha t/\sqrt{e}} \leq \frac{m\varepsilon^2}{2\alpha/\sqrt{e}},$$

where the first step was by Lemma 7.2.5, the second step was by Proposition 7.2.7 applied with $\frac{\alpha}{\sqrt{e}}$-$\mathfrak{t}$-strong convexity throughout, and in the final step we used Fact 7.1.4 to bound $\|\nabla_x\|_{\mathfrak{t}}^2 \le m\varepsilon^2$ as $s(x) = 1$ and x is ε-T-balanced by assumption. Item (3) follows by using the fact that $s(x) = 1$ and rearranging. □

Finally, we can lift this result to the non-commutative setting using the reduction given in Theorem 6.3.1.

Theorem 7.2.16. *For $m \ge 3$, let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ of size $s(x) = 1$ is ε-G-balanced and α-$\mathfrak{p}$-strongly convex with $\frac{\alpha}{\sqrt{e}} \ge 6m \cdot \varepsilon\sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}$, then there is a scaling $x_* = p_*^{1/2} \cdot x = e^{Z_*/2} \cdot x$ with $p_* \in P$ and $Z_* \in \mathfrak{p}$ satisfying:*

1. *x_* is a G-balanced tensor;*

2. *$\|Z_*\|_{\mathfrak{p}} \le \frac{\varepsilon\sqrt{m}}{\alpha/\sqrt{e}}$, and $\|Z_*^{(a)}\|_{\text{op}} \le \frac{3\varepsilon\sqrt{md_a}^{1-\frac{\alpha/\sqrt{e}}{m}}}{2\alpha/\sqrt{e}}$ for every $a \in [m]$;*

3. *The size of the scaling solution is lower bounded by $s(x_*) \ge 1 - \frac{m\varepsilon^2}{2\alpha/\sqrt{e}}$.*

Proof. Our plan is to apply the reduction in Theorem 6.3.1. Let $(T^\Xi, T_+^\Xi, \mathfrak{t}^\Xi)$ denote the commutative scaling group that is diagonal in the basis $\Xi = \{\Xi^a\}$ according to Definition 7.2.1, and consider the decomposition $P = \cup_{\Xi\in\mathcal{X}} T_+^\Xi = \cup_{\Xi\in\mathcal{X}} e^{\mathfrak{t}^\Xi}$ where $\Xi \in \mathcal{X}$ runs over all tuples of orthonormal bases such that $T_+^\Xi \subseteq P$, i.e. $\mathfrak{st}_+^{\Xi^a} \subseteq \mathfrak{p}_a$ for each $a \in [m]$. Explicitly, if $\mathfrak{p}_a = \mathfrak{st}_+^{\Xi^a}(V_a)$, then the a-th component just contains the singleton Ξ^a, and if $\mathfrak{p}_a = \mathfrak{spd}(V_a)$, then Ξ^a runs over all orthonormal bases of V_a.

Since $T^\Xi \subseteq G$, the ε-G balance condition of Definition 6.2.4 implies ε-T^Ξ-balance for every $\Xi \in \mathcal{X}$. Similarly, $\mathfrak{t}^\Xi \subseteq \mathfrak{p}$, so the α-$\mathfrak{p}$-strong convexity condition of Definition 7.1.7 implies α-$\mathfrak{t}^\Xi$-strong convexity for every $\Xi \in \mathcal{X}$. Therefore, Theorem 7.2.15 applied to each commutative restriction $(T^\Xi, T_+^\Xi, \mathfrak{t}^\Xi)$ produces scaling solution $Z_\Xi \in \mathfrak{t}^\Xi$ such that

$$f_x^{t^\Xi}(Z_\Xi) = f_x^{T_+^\Xi}(e^{Z_\Xi}) = \inf_{t\in T_+^\Xi} f_x^P(t),$$

where the global minimum property is by Proposition 6.2.18(2) for the $(T^\Xi, T_+^\Xi, \mathfrak{t}^\Xi)$-scaling problem on input x. The bound on $\|Z_\Xi\|_\infty$ given in item (1) of Theorem 7.2.15 clearly implies that $\cup_{\Xi\in\mathcal{X}} e^{Z_\Xi}$ is contained in a compact set, so we can apply Theorem 6.3.1 to get a global minimizer of f_x^P of the form $e^{Z_*} = e^{Z_\Xi}$ for some Ξ.

By Proposition 6.2.18(3), this global minimum is a G-balanced scaling solution. Therefore the guarantees in items (2) and (3) follows exactly from Theorem 7.2.15. □

When $m = 2$, we can apply the $\alpha \gtrsim \varepsilon \log d$ result of Theorem 3.2.19 instead of Theorem 7.2.15, and lift it to the non-commutative setting using the same reduction strategy as the proof above. We show this explicitly for operator scaling in Section 7.3.3, where we combine it with a robustness result to show strong convexity of the optimizer.

7.2.3 Pseudorandom Analysis

In this subsection, we analyze tensor scaling when the input satisfies a pseudorandom property which we show in Section 7.4 to be stronger than the strong convexity condition used in Section 7.1.3. We will use this condition to directly analyze $\|\nabla_{x_t}\|_{\text{op}}$ through gradient flow. This allows us to remove the dimension-dependent factor in the condition on $\frac{\alpha}{\varepsilon}$ as compared to the strongly convex convergence analysis of Theorem 7.2.16.

In order to understand how this analysis compares to the strong convexity analysis in Section 7.1.3, we give a brief overview of the argument. Recall that in Proposition 7.2.12, we bounded the error and scaling solution by breaking the evolution of these quantities into two stages, showing the error $\|\nabla_{x_t}\|_{\text{op}}$ grows slowly in the first stage, and then using exponential convergence of $\|\nabla_{x_t}\|_{\mathfrak{p}}$ by strong convexity for the second stage. In this subsection, we will use pseudorandomness to show that the error $\|\nabla_{x_t}\|_{\text{op}}$ converges exponentially for all time. This avoids the first phase where the error grows, which gives a much stronger dimension independent bound on the scaling solution.

In more detail, the main consequence of pseudorandomness that we use in our analysis is given in Lemma 7.2.19. Here, we show that the error $\|\nabla_{x_t}^{(a)}\|_{\infty}$ in a fixed marginal is not significantly affected by the other marginals through gradient flow. This is used to show that the marginal with the largest error decreases exponentially for all time. Then we show that pseudorandomness is robust to small scalings, which implies that it is maintained throughout gradient flow. The analysis in this subsection can be thought of as the appropriate tensor generalization of Section 3.3.

We first define the pseudorandom condition for general scaling groups. Afterwards, we focus on the commutative case and then lift the result to non-commutative groups by decomposing into commutative subgroups and applying the reduction in Theorem 6.3.1.

Definition 7.2.17 (Tensor Pseudorandom Condition). *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces with* $\dim(V_a) = d_a$ *for each $a \in [m]$ along with scaling* *group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. For any pair $a \neq b \in [m]$, let*

$$\mathcal{S}_a := \begin{cases} \{\xi \in V_a \mid \|\xi\|_2 = 1\} & \text{if } G_a = \mathrm{SL}(V_a) \\ \{\xi_i \in \Xi^a \mid i \in [d_a]\} & \text{if } G_a = ST^{\Xi^a}(V_a) \end{cases}, \quad \text{and}$$

$$\mathcal{P}_b := \begin{cases} \{Q \in L(V_b) \mid Q^* = Q, Q^2 = Q, \mathrm{rk}(Q) = \frac{d_b}{2}\} & \text{if } G_b = \mathrm{SL}(V_b) \\ \{\sum_{j \in T} \psi_j \psi_j^* \mid T \in \binom{[d_b]}{d_b/2}, \psi_j \in \Xi^b\} & \text{if } G_b = ST^{\Xi^b}(V_b). \end{cases}$$

i.e. the first set is the sphere of V_a restricted to domain of G_a, and the second set is a subset of orthogonal projections restricted to the domain of G_b. Then $x \in V^K$ is γ-$\mathfrak{p}_{a \leftarrow b}$-pseudorandom if, for every $\xi \in \mathcal{S}_a$ and $Q \in \mathcal{P}_b$,

$$\langle \rho_x^{(ab)}, \xi\xi^* \otimes Q \rangle \geq e^{-\gamma} \frac{\mathrm{Tr}[\xi\xi^*]}{d_a} \cdot \frac{\mathrm{Tr}[Q]}{d_b} = \frac{e^{-\gamma}}{2d_a}.$$

x is γ-$\mathfrak{p}$-pseudorandom if the above is satisfied for every pair $a \neq b \in [m]$.

Remark 7.2.18. *Technically, if d_b is odd, then we should use $\lfloor \frac{d_b}{2} \rfloor$. We ignore this discrepancy in future for simplicity so as not to further clutter the notation.*

For intuition, the value of $\langle \rho_x^{(ab)}, \xi\xi^* \otimes Q \rangle$ for a uniformly random $\xi \in \mathcal{S}_a, Q \in \mathcal{P}_b$ is exactly $\frac{s(x)}{2d_a}$. Therefore, if x is γ-$\mathfrak{p}_{a \leftarrow b}$-pseudorandom, then $\rho_x^{(ab)}$ has significant weight in every direction $\xi \in \mathcal{S}_a, Q \in \mathcal{P}_b$. We point out that γ-$\mathfrak{p}$-pseudorandomness according to Definition 7.2.17 is a stronger condition for smaller γ, whereas (α, β)-pseudorandomness for matrices according to Definition 3.3.1 is a stronger condition for larger α.

We can directly compare the two conditions for matrix scaling: tuple $A \in \mathrm{Mat}(d, n)^K$ can be viewed as a tensor tuple by the isomorphism $A_k \to \mathrm{vec}(A_k) \in \mathbb{F}^d \otimes \mathbb{F}^n = V_L \otimes V_R$ with scaling group $T = (T_L, T_R) = (\mathrm{ST}(d), \mathrm{ST}(n))$. Then $\mathrm{vec}(A) \in (\mathbb{F}^d \otimes \mathbb{F}^n)^K$ is γ-$\mathfrak{t}_{L \leftarrow R}$-pseudorandom iff A is an $(e^{-\gamma}, \frac{1}{2})$-pseudorandom matrix according to Definition 3.3.1. Similarly, input frame $U \in \mathrm{Mat}(d, n)$ can be viewed as a tensor by the same isomorphism, and frame scaling corresponds to the scaling group $G = (G_L, G_R) = (\mathrm{SL}(d), \mathrm{ST}(n))$ with associated infinitesimal vector space $(\mathfrak{p}_L, \mathfrak{p}_R) = (\mathfrak{spd}(d), \mathfrak{st}_+(n))$. Then $\mathrm{vec}(U) \in \mathbb{F}^d \otimes \mathbb{F}^n$ is γ-$\mathfrak{p}_{L \leftarrow R}$-pseudorandom iff U is an $(e^{-\gamma}, \frac{1}{2})$-pseudorandom frame according to Definition 4.2.11. Note that the pseudorandomness condition in Definition 4.2.11 always corresponds to the $L \leftarrow R$ direction. This asymmetry was useful in Chapter 4 when $n \gg d$ for the Paulsen problem.

In Lemma 7.2.19, we show a consequence of pseudorandomness that will be more directly useful in analyzing gradient flow for tensors. We define pseudorandomness by Definition 7.2.17 because this property enjoys stronger (multiplicative) robustness, which means that less initial pseudorandomness is required for our fast convergence analysis.

In this subsection, we will mostly focus on the case $V = \otimes_{a\in[m]}\mathbb{F}^{d_a}$ and scaling group $T = (\mathrm{ST}(d_1), ..., \mathrm{ST}(d_m))$. This analysis generalizes to an arbitrary choice of commutative scaling group by a simple change of basis, and the results will be lifted to non-commutative scaling groups by Theorem 6.3.1 at the end of the subsection.

Recall that Lemma 7.2.10 gives an expression for the change in any marginal under gradient flow involving terms of the form $\langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes \nabla_x^{(b)}\rangle$. In the following lemma, we will use the pseudorandom condition to show that these terms do not exert too much influence on the a-th marginal, which will be used to show fast convergence.

Lemma 7.2.19. *Let $V = \otimes_{a\in[m]}\mathbb{F}^{d_a}$ with scaling group $T = (\mathrm{ST}(d_1), ..., \mathrm{ST}(d_m))$ and associated polar $(T_+, \mathfrak{t})$ according to Definition 7.2.1. If $x \in V^K$ is γ-$\mathfrak{t}_{a\leftarrow b}$-pseudorandom, then for any $Z_b \in \mathfrak{t}_b = \mathfrak{st}_+(d_b)$,*

$$|d_a\langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes Z_b\rangle| \leq \|Z_b\|_\infty (d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - e^{-\gamma}).$$

Proof. This is a simple application of Fact 2.6.4, which shows that the vertices of $\{Z_b \in \mathfrak{st}_+(d_b), \|Z_b\|_\infty \leq 1\}$ are of the form $I_b - 2P$, where $P = P_T \in \mathcal{P}_b$ is the coordinate projection onto some $T \in \binom{[d_b]}{d_b/2}$ (we leave out the case of odd d_b for simplicity). We can assume without loss that $\|Z_b\|_\infty = 1$ and bound the inner product by

$$d_a\langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes Z_b\rangle \leq d_a\langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes I_b\rangle - 2\min_{P\in\mathcal{P}_b} d_a\langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes P\rangle \leq d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - e^{-\gamma},$$

where in the first step we used Fact 2.6.4 to upper bound by the vertices of $\{Z_b \in \mathfrak{t}_b, \|Z_b\|_\infty \leq 1\}$, and in the final step we applied Definition 6.2.2 of the marginal for the first term and bounded the second term by Definition 7.2.17 of pseudorandomness. □

Lemma 7.2.19 is the main consequence of pseudorandomness that we use in our analysis to show exponential convergence of the error through gradient flow. The pseudorandom property is helpful because it is multiplicatively robust, shown in Lemma 7.2.23, whereas it is more difficult to control the change in the bound in Lemma 7.2.19 for scalings of x.

This allows us to bound the change in $\|\nabla_x\|_\infty$ as follows.

Lemma 7.2.20. *Consider $V = \otimes_{a\in[m]}\mathbb{F}^{d_a}$ with scaling group $T = (\mathrm{ST}(V_1), ..., \mathrm{ST}(V_m))$ and associated infinitesimal vector space $\mathfrak{t} := \oplus_{a\in[m]}\mathfrak{st}_+(d_a)$ according to Definition 6.2.3. If $x \in V^K$ is γ-$\mathfrak{t}$-pseudorandom, then*

$$-\partial_{t=0} \max_{a\in[m]} \log \|\nabla_{x_t}^{(a)}\|_\infty \geq s(x) - (m-1)\Big(s(x) + 2\max_{a\in[m]} \|\nabla_x^{(a)}\|_\infty - e^{-\gamma}\Big).$$

Proof. Let $a \in \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_\infty$ and $i \in \arg\max_{i\in[d_a]} |\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle|$ be the diagonal with the worst error. We will show this error is decreasing by examining the terms in Lemma 7.2.10 and bounding them using the pseudorandom condition in Definition 7.2.17. We first recall the following bound on the change in size:

$$-\partial_{t=0} s(x_t) = \|\nabla_x\|_{\mathfrak{t}}^2 \leq \sum_{b\in[m]} \|\nabla_x^{(b)}\|_\infty^2 \leq \|\nabla_x^{(a)}\|_\infty \|\nabla_x\|_\infty, \tag{7.10}$$

where the first step was by Lemma 7.1.6, the second was by Lemma 7.2.9, and the last step was by our case assumption that $\|\nabla_x^{(a)}\|_\infty \geq \|\nabla_x^{(b)}\|_\infty$ for $b \in [m]$. Now we separate into two cases depending on the sign of the error in the i-th diagonal.

First consider the case $\|\nabla_x^{(a)}\|_\infty = \max_{i\in[d_a]}\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle = d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - s(x)$, meaning this diagonal is larger than average. We show it is decreasing by Lemma 7.2.10, as

$$\begin{aligned}
-\partial_{t=0} d_a\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle &= d_a\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle + \sum_{b\neq a\in[m]} d_a\langle \rho_x^{(ab)}, E_{ii}^{(a)} \otimes \nabla_x^{(b)}\rangle \\
&\geq \|\nabla_x^{(a)}\|_\infty d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - \sum_{b\neq a\in[m]} \|\nabla_x^{(b)}\|_\infty (d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - e^{-\gamma}) \\
&\geq \|\nabla_x^{(a)}\|_\infty (s(x) + \|\nabla_x^{(a)}\|_\infty) - (m-1)\|\nabla_x^{(a)}\|_\infty (s(x) + \|\nabla_x^{(a)}\|_\infty - e^{-\gamma}),
\end{aligned}$$

where the first step was by Lemma 7.2.10 (see Remark 3.2.11 for questions of differentiability of $\|\cdot\|_\infty$), in the second step we used our case assumption $\|\nabla_x^{(a)}\|_\infty = d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle - s(x)$ to bound the first term and Lemma 7.2.19 applied with $Z_b := \nabla_x^{(b)}$ to bound each term in the sum, and in the final step we used $d_a\langle \rho_x^{(a)}, E_{ii}^{(a)}\rangle = s(x) + \|\nabla_x^{(a)}\|_\infty$ by our case assumption on i and $\|\nabla_x^{(a)}\|_\infty \geq \|\nabla_x^{(b)}\|_\infty$ by our case assumption on a. This allows us to bound the change in $\|\nabla_x^{(a)}\|_\infty$ as

$$\begin{aligned}
-\partial_{t=0} \log \|\nabla_{x_t}^{(a)}\|_\infty &= \frac{-\partial_{t=0} d_a\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle + \partial_{t=0} s(x_t)}{\|\nabla_x^{(a)}\|_\infty} \\
&\geq (s(x) + \|\nabla_x^{(a)}\|_\infty) - (m-1)(s(x) + \|\nabla_x^{(a)}\|_\infty - e^{-\gamma}) - \|\nabla_x\|_\infty \\
&\geq s(x) - (m-1)(s(x) + 2\|\nabla_x^{(a)}\|_\infty - e^{-\gamma}),
\end{aligned}$$

where the first step is by the assumption that $\|\nabla_x^{(a)}\|_\infty = d_a\langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle - s(x)$, in the second step we used the bounds derived above for change in $\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle$ as well as Eq. (7.10) for the size, and in the final step we used $\|\nabla_x\|_\infty \leq m\|\nabla_x^{(a)}\|_\infty$ since $a \in \arg\max_{b\in[m]} \|\nabla_x^{(b)}\|_\infty$.

Now consider the case $\|\nabla_x^{(a)}\|_\infty = -\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle = s(x) - d_a\langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle$. We show it is increasing by Lemma 7.2.10, as

$$\begin{aligned}\partial_{t=0} d_a\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle &= d_a\langle E_{ii}^{(a)}, (-\nabla_x^{(a)})\rangle\langle\rho_x^{(a)}, E_{ii}^{(a)}\rangle + \sum_{b\neq a\in[m]} d_a\langle\rho_x^{(ab)}, E_{ii}^{(a)}\otimes(-\nabla_x^{(b)})\rangle \\ &\geq \|\nabla_x^{(a)}\|_\infty d_a\langle\rho_x^{(a)}, E_{ii}^{(a)}\rangle - \sum_{b\neq a\in[m]} \|\nabla_x^{(b)}\|_\infty(d_a\langle\rho_x^{(a)}, E_{ii}^{(a)}\rangle - e^{-\gamma}) \\ &\geq \|\nabla_x^{(a)}\|_\infty(s(x) - \|\nabla_x^{(a)}\|_\infty) - (m-1)\|\nabla_x^{(a)}\|_\infty(s(x) - \|\nabla_x^{(a)}\|_\infty - e^{-\gamma})\end{aligned}$$

where the first step was by Lemma 7.2.10, in the second step we used our case assumption $\|\nabla_x^{(a)}\|_\infty = -\langle E_{ii}^{(a)}, \nabla_x^{(a)}\rangle$ to bound the first term and Lemma 7.2.19 with $Z_b = \nabla_x^{(b)}$ to bound each term in the sum, and in the final step we used $d_a\langle\rho_x^{(a)}, E_{ii}^{(a)}\rangle = s(x) - \|\nabla_x^{(a)}\|_\infty$ by our case assumption on i and $\|\nabla_x^{(a)}\|_\infty \geq \|\nabla_x^{(b)}\|_\infty$ by our case assumption on a. This allows us to bound the change in $\nabla_x^{(a)}$ as

$$\begin{aligned}-\partial_{t=0}\log\|\nabla_{x_t}^{(a)}\|_\infty &= \frac{\partial_{t=0} d_a\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle - \partial_{t=0} s(x_t)}{\|\nabla_x^{(a)}\|_\infty} \\ &\geq (s(x) - \|\nabla_x^{(a)}\|_\infty) - (m-1)(s(x) - \|\nabla_x^{(a)}\|_\infty - e^{-\gamma}) + 0\end{aligned}$$

where the first step is by the assumption that $\|\nabla_x^{(a)}\|_\infty = d_a\langle E_{ii}^{(a)}, \rho_x^{(a)}\rangle - s(x)$, in the second step we used the bounds derived above for change in $\langle E_{ii}^{(a)}, \rho_{x_t}^{(a)}\rangle$ and for the size we simply used $-\partial_{t=0}s(x_t) \geq 0$ by Lemma 7.2.5.

Combining the two cases, the lemma is shown as we have

$$-\partial_{t=0}\log\|\nabla_{x_t}^{(a)}\|_\infty \geq s(x) - (m-1)(s(x) + 2\|\nabla_x^{(a)}\|_\infty - e^{-\gamma}).$$

□

Remark 7.2.21. *As mentioned previously in the analyses of Chapter 3 and Section 7.2.2, we ignore questions of differentiability for the infinity norm and discuss the technical solution in Remark 3.2.11 using e.g. the envelope theorem of Milgrom and Segal [70].*

Recall that in Section 7.2.2, we used Lemma 7.2.11 to show that the error grew slowly for small t, and then we were able to apply the fast convergence of Proposition 7.2.7 to show the error remained bounded. In this analysis, we can use the above Lemma 7.2.20 to show that $\|\nabla_x\|_\infty$ is always exponentially decreasing, which will allow us to bound the scaling $\|Z_T\|_\infty$ from gradient flow directly. This is similar to our analysis in Proposition 3.3.9 for the matrix setting.

Proposition 7.2.22. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$, and let $(T, T_+, \mathfrak{t})$ be a commutative scaling group according to Definition 7.2.1. Let $x \in V^K$ be an input of size $s(x) = 1$ that is ε-T-balanced, and assume that for all $t \in [0,T]$ the solution of gradient flow x_t according to Proposition 7.2.4 satisfies $s(x_t) \geq \frac{3}{4}$, (2) γ-$\mathfrak{t}$-pseudorandomness with $\gamma \leq \frac{1}{16m}$, and (3) $\max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \frac{1}{16m}$. Then*

$$\|\nabla_{x_T}^{(a)}\|_\infty \leq \varepsilon e^{-T/2} \qquad \textit{and} \qquad \|Z_T^{(a)}\|_\infty \leq 2\varepsilon.$$

Proof. We can assume without loss that $V_a = \mathbb{F}^{d_a}$ and $T_a = \mathrm{ST}(d_a)$ by a change of basis if necessary. This is only to reduce clutter, and we emphasize that these results hold for general commutative scaling groups. Under these conditions, we have for $t \in [0,T]$ that

$$\begin{aligned} -\partial_t \max_{a\in[m]} \log \|\nabla_{x_t}^{(a)}\|_\infty &\geq s(x_t) - (m-1)\Big(s(x_t) + 2\max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty - e^{-\gamma}\Big) \\ &\geq \frac{3}{4} - (m-1)\left(1 + \frac{1}{8m} - \left(1 - \frac{1}{8m}\right)\right) \\ &\geq \frac{3}{4} - \frac{m-1}{4m} \geq \frac{1}{2}, \end{aligned}$$

where the first step was by Lemma 7.2.20 applied to x_t, in the second step we used the assumptions $1 = s(x) \geq s(x_t) \geq \frac{3}{4}$ and $\max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \frac{1}{16m}$ as well as the Taylor approximation $e^{-\gamma} \geq 1 - 2\gamma$ for $0 \leq \gamma \leq \frac{1}{16m}$.

Therefore by the fundamental theorem of calculus, we have

$$\max_{a\in[m]} \log \|\nabla_{x_T}^{(a)}\|_\infty = \max_{a\in[m]} \log \|\nabla_{x}^{(a)}\|_\infty + \int_0^T \partial_t \max_{a\in[m]} \log \|\nabla_{x_t}^{(a)}\|_\infty \leq \log \varepsilon - \frac{T}{2},$$

where the first step is by the fundamental theorem of calculus, and the second is by the derivative bound $\partial_t \max_{a\in[m]} \log \|\nabla_{x_t}^{(a)}\|_\infty \geq \frac{1}{2}$. The first statement follows by exponentiating both sides.

For the second statement, we again use the fundamental theorem of calculus to show

$$\|Z_T^{(a)}\|_\infty = \left\| \int_0^T -\nabla_{x_t}^{(a)} \right\| \leq \int_0^T \max_{b\in[m]} \|\nabla_{x_t}^{(b)}\|_\infty \leq \varepsilon \int_0^T e^{-t/2} \leq 2\varepsilon,$$

where the first step was by Proposition 7.2.4 of gradient flow, the second was by triangle inequality on $\|\cdot\|_\infty$, and the third step was by the gradient bound in the first statement. □

Our plan is to remove the assumptions of Proposition 7.2.22 and replace them with sufficient pseudorandomness of the initial tensor. To this end, the next lemma shows that pseudorandomness is maintained by scalings.

Lemma 7.2.23. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$, and let $(T, T_+, \mathfrak{t})$ be a commutative scaling group according to Definition 7.2.1. If input $x \in V^K$ is γ-$\mathfrak{t}_{a\leftarrow b}$-pseudorandom, then for any $Y \in \mathfrak{t}$ the scaling $e^{Y/2}\cdot x$ is $(\|Y\|_\infty + \gamma)$-$\mathfrak{t}_{a\leftarrow b}$-pseudorandom.*

Before proving this statement, note that this is a multiplicative robustness bound on pseudorandomness, as Definition 7.2.17 has the parameter in the exponent. In Theorem 7.3.4, we will show that this kind of multiplicative bound holds for the non-commutative pseudorandom condition as well.

Proof. Choose an arbitrary $\xi \in \mathcal{S}_a$ and orthogonal projection $P \in \mathcal{P}_b$ according to Definition 7.2.17. Then we can bound the quantity

$$\begin{aligned}\langle \rho^{(ab)}_{e^{Y/2}\cdot x}, \xi\xi^* \otimes P\rangle &= \langle e^{Y/2}\rho_x e^{Y/2}, \xi\xi^*\otimes P\otimes I_{\overline{ab}}\rangle = \langle \rho_x, e^Y\cdot(\xi\xi^*\otimes P\otimes I_{\overline{ab}})\rangle \\ &\geq e^{-\|Y\|_\infty}\langle \rho_x, \xi\xi^*\otimes P\otimes I_{\overline{ab}}\rangle \geq e^{-\|Y\|_\infty}\frac{e^{-\gamma}}{2d_a},\end{aligned}$$

where in the first step was by Definition 6.2.2 of marginals as well as the equivariance property of Lemma 6.2.6(1), in the second step we used that fact that $Y \in \mathfrak{t}$ and $\xi \in \mathcal{S}_a$ and $P \in \mathcal{P}_b$ so $e^{Y/2}$ commutes with $\xi\xi^*\otimes P\otimes I_{\overline{ab}}$, in the third step we used the fact that both terms in the inner product are positive semidefinite, so we can bound the inner product by the spectral lower bound $e^Y \succeq e^{-\|Y\|_\infty} I_V$ by Definition 7.2.8, and in the final step the lower bound was by the γ-$t_{a\leftarrow b}$-pseudorandom condition of x according to Definition 7.2.17. Since $\xi \in \mathcal{S}_a$ and $P \in \mathcal{P}_b$ were arbitrary, this verifies Definition 7.2.17 for $e^{Y/2}\cdot x$. □

Remark 7.2.24. *Unlike the strong convexity analysis in Proposition 7.2.12, the pseudorandom condition is only useful when $\gamma \lesssim \frac{1}{m}$, regardless of how small the initial error is. This is in contrast with the $m = 2$ matrix scaling analysis of Section 3.3 where pseudorandomness requirement only depended on the initial error and could be taken as small as $\Omega(\varepsilon)$ for ε-doubly balanced input.*

Now we can prove the main result of this subsection by bounding the scaling solution of nearly balanced and pseudorandom inputs.

Theorem 7.2.25. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$, and consider commutative scaling group $(T, T_+, \mathfrak{t})$ according to Definition 7.2.1. If $x \in V^K$ of size $s(x) = 1$ is ε-T-balanced for $\varepsilon \leq \frac{1}{100m^2}$ and γ-$\mathfrak{t}$-pseudorandom for $\gamma \leq \frac{1}{32m}$, then:*

1. *For all time $t \geq 0$, the solution Z_t of gradient flow satisfies*
$$\max_{a\in[m]} \|Z_t^{(a)}\|_\infty \leq 2\varepsilon;$$

2. *The limit $Z_\infty := \lim_{t\to\infty} Z_t$ exists and $x_\infty := e^{Z_\infty/2} \cdot x$ is a T-balanced scaling solution to the T-tensor scaling problem in Definition 6.2.5;*

3. *The size of the solution can be lower bounded*
$$s(x_*) = f_x^{\mathfrak{t}}(Z_*) \geq 1 - m\varepsilon^2.$$

Proof. We can assume without loss that $V_a = \mathbb{F}^{d_a}$ and $T_a = \mathrm{ST}(d_a)$ by a change of basis if necessary. This is only to reduce clutter, and we emphasize that these results hold for general commutative scaling groups.

We claim that for all time x_t satisfies the assumptions of Proposition 7.2.22, i.e. for all $t \geq 0$: (1) $s(x_t) \geq \frac{3}{4}$, (2) $\max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \frac{1}{16m}$, and (3) $\gamma' = \frac{1}{16m}$-$\mathfrak{t}$-pseudorandomness. For contradiction, assume that T is the last time these conditions hold, and in particular assume that the size condition fails first. Up till this time, we can bound

$$s(x) - s(x_T) = \int_0^T \|\nabla_{x_t}\|_{\mathfrak{t}}^2 \leq \int_0^T m \max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty^2 \leq m\varepsilon^2 \int_0^T e^{-t} < m\varepsilon^2 < \frac{1}{4},$$

where the first step was by Lemma 7.2.5, the second was by Lemma 7.2.9, the third step was by the conclusion of Proposition 7.2.22 showing for all $t \in [0, T] : \max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty \leq \varepsilon e^{-t/2}$, and the final step was by our assumption $\varepsilon \leq \frac{1}{100m^2}$. Therefore the size condition cannot fail first.

We use Proposition 7.2.22 to show that the error condition (2) cannot fail first, as

$$\max_{a\in[m]} \|\nabla_{x_T}^{(a)}\|_\infty \leq \varepsilon e^{-T/2} \leq \varepsilon < \frac{1}{16m},$$

where the first step was by the conclusion of Proposition 7.2.22 applied up till time T, and the final step was by the assumption $\varepsilon \leq \frac{1}{100m^2}$.

Finally, assume that the pseudorandom condition fails first. Then by Lemma 7.2.23 in the contrapositive, this implies $\|Z_T\|_\infty \geq \frac{1}{16m} - \gamma \geq \frac{1}{32m}$ as $\gamma \leq \frac{1}{32m}$. But the conditions for Proposition 7.2.22 are satisfied up to time T, so we can bound

$$\|Z_T\|_\infty \leq m \max_{a\in[m]} \|Z_T^{(a)}\|_\infty \leq 2m\varepsilon \leq \frac{1}{50m},$$

where the first step was by Definition 7.2.8 of the $\|\cdot\|_\infty$, the second was by the conclusion of Proposition 7.2.22 applied up till time T, and the final step was by our assumption $\varepsilon \leq \frac{1}{100m^2}$. This gives the desired contradiction, so we have fast convergence for all time.

This allows us to apply Proposition 7.2.22 for all time, and conclusion (1) follows. To show (2), note that we have already shown that $\max_{a\in[m]} \|\nabla_{x_T}^{(a)}\|_\infty \leq \varepsilon e^{-T/2}$ for all time by Proposition 7.2.22. This allows us to bound

$$\lim_{T\to\infty} \int_T^\infty \|\partial_t Z_t^{(a)}\|_\infty = \lim_{T\to\infty} \int_T^\infty \|\nabla_{x_t}^{(a)}\|_\infty \leq \lim_{T\to\infty} \varepsilon \int_{t\geq T} e^{-t/2} = 0,$$

where the first step was by Proposition 7.2.4 of gradient flow, and the second was by exponential convergence $\|\nabla_{x_t}^{(a)}\|_\infty \leq \varepsilon e^{-t/2}$. Therefore the limit Z_∞ exists, and further $\nabla_{x_\infty} = 0$ implies that x_∞ is T-balanced by Proposition 6.2.18(2).

To show the size lower bound in (3), we can repeat the calculation above:

$$s(x) - s(x_*) = \int_0^\infty \|\nabla_{x_t}\|_{\mathfrak{t}}^2 \leq \int_0^\infty m \max_{a\in[m]} \|\nabla_{x_t}^{(a)}\|_\infty^2 \leq m\varepsilon^2 \int_0^\infty e^{-t} = m\varepsilon^2,$$

where the first step was by the fundamental theorem of calculus and Lemma 7.2.5 of the change in size, in the second step we applied Lemma 7.2.9, and the third step was by the conclusion of Proposition 7.2.22. Item (3) follows as $s(x) = 1$ by assumption. □

Finally, this theorem can be lifted to the non-commutative setting by using the decomposition technique given in Theorem 6.3.1.

Theorem 7.2.26. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$ and scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ of size $s(x) = 1$ is ε-G-balanced for $\varepsilon \leq \frac{1}{100m^2}$ and γ-$\mathfrak{p}$-pseudorandom for $\gamma \leq \frac{1}{32m}$, then there is a scaling $x_* = p_*^{1/2} \cdot x = e^{Z_*/2} \cdot x$ with $p_* \in P, Z_* \in \mathfrak{p}$ that satisfies:*

1. *x_* is a G-balanced tensor scaling solution to Definition 6.2.5;*

2. *$\max_{a\in[m]} \|Z_*^{(a)}\|_{\mathrm{op}} \leq 2\varepsilon$;*

3. The size of the scaling solution is lower bounded by $s(x_*) \geq 1 - m\varepsilon^2$.

Proof. This proof is exactly the same as the proof of Theorem 7.2.16, except that we apply the pseudorandom analysis of Theorem 7.2.25 for the conclusions. □

7.3 Non-Commutative Robustness

A key part of our analysis of tensor scaling in Section 7.2 as well as the analysis of matrix scaling in Chapter 3 relied on showing that the convergence properties (strong convexity or pseudorandomness) were maintained throughout gradient flow. In the proofs of those results, we crucially used commutativity of the scaling group to show that the convergence parameter was multiplicatively robust under scalings. Using the reduction argument of Theorem 6.3.1, we were then able to leverage this multiplicative robustness in the commutative setting to show strong bounds on the solution for general non-commutative tensor scaling problems. However, one drawback is that this leads to non-constructive existential results on the optimizer in the non-commutative setting.

In this section, we will show robustness results for the non-commutative setting. These general robustness results will be valuable in our study of geodesically convex optimization algorithms that will be used to give constructive results for scaling problems in Chapter 8. In Section 7.3.1, we show multiplicative robustness bounds for the frame pseudorandom property in Definition 4.2.11. This will be applied in Section 8.5 in order to make the frame scaling results of Chapter 4 constructive. Similarly, in Section 7.3.2, we show multiplicative robustness bounds for the tensor pseudorandom property in Definition 7.2.17. The final two subsections will study strong convexity for tensors and will be quantitatively weaker. In Section 7.3.3, we show that multiplicative robustness of strong convexity is impossible in the non-commutative setting. Therefore, we settle for a weaker additive robustness result, which is sufficient to reproduce the work of [63] on operator scaling. Then, in Section 7.3.4, we lift these results to higher-order tensors and show additive robustness of tensor strong convexity for non-commutative scalings. This will be used in Chapter 9 to give strong algorithmic convergence guarantees for the Flip-Flop algorithm (Definition 8.4.1) for the tensor normal model.

7.3.1 Robustness of Frame Pseudorandomness

In this subsection, we show that Definition 4.2.11 of frame pseudorandomness is multiplicatively preserved under arbitrary scalings. For $\beta = \frac{1}{2}$, this is a special case of the robustness result for tensors proven in the following Section 7.3.2. We repeat the proof of the simpler frame case here for clarity, as the tensor setting comes with a bit more notation.

Remark 7.3.1. *A very similar robustness property was shown in Prop 4.3.5 of [62], and was crucial to the fast convergence results. This robustness was only proved for the commutative matrix case, so in order to apply these results to the Paulsen problem in that work, we had to use a much more complicated perturbation argument and analysis. Here we show that a simple non-commutative robustness statement follows the same proof and allows us to greatly simplify the original argument of [62], even bypassing the frame to matrix reduction of Theorem 4.2.13 we use in this thesis.*

In the proof of Lemma 3.3.4 we showed a more refined statement that gave multiplicative robustness for every unit vector and projection in Definition 3.3.1 of pseudorandomness. The key feature of this robustness result was that it produced a multiplicative bound $\alpha \to e^{-\delta} \cdot \alpha$ for any scaling $\|(X,Y)\|_{\text{op}} \leq \delta$. Therefore, in the proof of Theorem 3.3.10, we used the fact that pseudorandomness was maintained up to constant factors even for scalings with $\|(X,Y)\|_\infty \approx \Theta(1)$, regardless of the pseudorandomness of the initial input.

In the following, we give a robustness result for frame pseudorandomness according to Definition 4.2.11. This result will also be multiplicative, but will only hold for the infimum and not individually for every unit vector and projection.

Lemma 7.3.2. *Let $U = \{u_1, ..., u_n\} \in \text{Mat}(d,n)$ be a frame that is (α,β)-pseudorandom according to Definition 4.2.11. Then, for any $L \in L(d)$, $R \in \text{diag}(n)$, the scaling $V := LUR$ is (α',β)-pseudorandom as a frame with $\alpha' \geq \sigma_{\min}^2(L) \cdot \sigma_{\min}^2(R) \cdot \alpha$.*

Proof. Recall that Lemma 5.1.1 gives the following equivalent formulation for (α,β)-pseudorandomness: for any $\xi \in S^{d-1}$ and $T \in \binom{[n]}{\beta n}$,

$$\|\xi^* U P_T\|_F^2 = \sum_{j \in T} |\langle \xi, u_j \rangle|^2 \geq \alpha \frac{\beta}{d},$$

where P_T is the orthogonal projection onto the coordinates $T \subseteq [n]$. We verify this holds for V with the bounds given in the lemma:

$$\|\xi^* V P_T\|_F = \|\xi^* LURP_T\|_F \geq \|L^*\xi\|_2 \Big(\inf_{\psi \in S^{d-1}} \|\psi^* U P_T\|_F \Big) \min_{j \in T} |R_{jj}| \geq \sigma_{\min}(L)\sigma_{\min}(R)\sqrt{\alpha \frac{\beta}{d}},$$

where in the first step we substituted $V = LUR$, in the second step we used the change of variable $\psi = \frac{L^*\xi}{\|L^*\xi\|_2} \in S^{d-1}$ for the left scaling and the fact that R commutes with P_T since they are both diagonal, and in the final step we used the definitions of minimum singular value for the lower bounds $\sigma_{\min}(L)$ and $\sigma_{\min}(R)$, and the pseudorandomness of U to lower bound $\|\psi^* U P_T\|_F$. Since $\xi \in S^{d-1}, T \in \binom{[n]}{\beta n}$ were arbitrary, squaring both sides verifies the pseudorandomness property of V according to Lemma 5.1.1 and gives the result. □

We reiterate that this is not a bound for every unit vector and projection individually, as in Lemma 3.3.4. But this is still a multiplicative lower bound, and so gives non-trivial pseudorandomness bounds for $O(1)$-scalings regardless of the initial value of α. This will be useful for our algorithmic results for the Paulsen problem in Section 8.5.

As an illustration, we combine the above robustness with our convergence analysis in Theorem 4.2.14 for pseudorandom frames to show strong convexity of the solution to frame scaling in this case. This is the main consequence of robustness that we use as our algorithmic results in Chapter 8 rely on strong convexity of the geodesic convex formulation in Proposition 6.2.18.

Theorem 7.3.3. *If frame $U \in \mathrm{Mat}(d,n)$ of size $s(U) = 1$ is ε-doubly balanced and (α,β)-pseudorandom for $\frac{1}{5} \geq \alpha \geq 16e \cdot \varepsilon$ and $\beta \leq \frac{1}{2}$, then there is a scaling $U_* = e^{X_*/2} U e^{Y_*/2}$ with $(X_*, Y_*) \in \mathfrak{p}$ satisfying:*

1. *$U_* := e^{X_*/2} U e^{Y_*/2}$ is a doubly balanced frame;*
2. *$\max\{\|X_*\|_{\mathrm{op}}, \|Y_*\|_{\mathrm{op}}\} \leq \frac{9\varepsilon}{\alpha}$;*
3. *The size of the scaling solution is lower bounded by $s(U_*) \geq 1 - \frac{10\varepsilon^2}{\alpha}$.*
4. *U_* is $(\frac{\alpha}{e}, \beta)$-pseudorandom according to Definition 4.2.11;*
5. *If $\beta \leq \frac{1}{16}$, then U_* is an α_*-strongly convex frame for $\alpha_* \geq e^{-12} \cdot \alpha$.*

Proof. The first three items are exactly the content of Theorem 4.2.14. For item (4), we use the fact that V is (α,β)-pseudorandom and apply Lemma 7.3.2 to show $U_* = e^{X_*/2} U e^{Y_*/2}$ is (α', β)-pseudorandom for

$$\alpha' \geq \alpha \cdot \sigma_{\min}(e^{X_*/2})^2 \cdot \sigma_{\min}(e^{Y_*/2})^2 \geq \alpha \cdot e^{-\|X_*\|_{\mathrm{op}} - \|Y_*\|_{\mathrm{op}}} \geq \alpha \cdot \exp\left(-\frac{18\varepsilon}{\alpha}\right) \geq \frac{\alpha}{e},$$

where the first step was by the robustness result in Lemma 7.3.2, the second step used the bound $\sigma_{\min}(e^{X_*/2})^2 \geq e^{\lambda_{\min}(X_*)} \geq e^{-\|X_*\|_{\mathrm{op}}}$ and a similar calculation for Y_*, the third step is by the bound in item (2), and the last step is by our assumption $\alpha \geq 16e \cdot \varepsilon$.

For the final item, we have the stronger assumption that $\beta \leq \frac{1}{16}$, so we can simply combine item (4) with Corollary 4.2.12 to show U_* is an α_*-strongly convex frame for $\alpha_* \geq e^{-11} \cdot \alpha' \geq e^{-12} \cdot \alpha$. □

Note that (X_*, Y_*) is a global minimum of the geodesically convex formulation for frame scaling according to item (2) of Proposition 6.2.18. The above result can be combined with Lemma 7.3.2 to show that there is a large region near the optimizer (X_*, Y_*) where the Kempf-Ness function f_U^P is $\Omega(\alpha)$-geodesically strongly convex. In Section 8.5, we will use this strongly convex region to show fast convergence of frame scaling algorithms for random frames as well as our solution to the Paulsen problem.

7.3.2 Tensor Pseudorandomness

In this subsection, we show that tensor pseudorandomness is multiplicatively preserved under arbitrary scalings. This is a generalization of the robustness of frame pseudorandomness shown in Section 7.3.1, and the proof will be nearly the same, but requires a bit more notation in the tensor case.

We will show that the infimum over unit vectors and projections given in Definition 7.2.17 of pseudorandomness is multiplicatively robust. Note that this is weaker than multiplicative robustness with respect to every individual unit vector and projection.

Theorem 7.3.4. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3, and consider tuple $x \in V^K$ that is γ-$\mathfrak{p}_{a \leftarrow b}$-pseudorandom according to Definition 7.2.17. Then, for any $(g_1, ..., g_m) \in G$, the scaling $y := g \cdot x$ is γ'-$\mathfrak{p}_{a \leftarrow b}$-pseudorandom for*

$$\gamma' \leq \gamma - \sum_{c \in [m]} \log \lambda_{\min}(g_c^* g_c).$$

Note that there are no restrictions on g, i.e. the scalings are not restricted to positive definite elements P, which is the domain of our geodesically convex formulation f_x^P.

Proof. If $g_a^* g_a$ is non-invertible for any $a \in [m]$, then the bound is trivial. So for the remainder of the proof, we assume $\lambda_{\min}(g_a^* g_a) > 0$ for all $a \in [m]$.

To show pseudorandomness according to Definition 7.2.17, we would like to bound $\langle \rho_{g \cdot x}^{(ab)}, \xi\xi^* \otimes P \rangle$ for every $\xi \in \mathcal{S}_a, P \in \mathcal{P}_b$. Our plan is to bound this in terms of $\langle \rho_x^{(ab)}, \psi\psi^* \otimes Q \rangle$ for some $\psi \in \mathcal{S}_a, Q \in \mathcal{P}_b$, and then apply pseudorandomness of x.

First consider arbitrary $\xi \in \mathcal{S}_a, P \in \mathcal{P}_b$ and note

$$\langle \rho_{g\cdot x}^{(ab)}, \xi\xi^* \otimes P \rangle = \langle g\rho_x g^*, \xi\xi^* \otimes P \otimes I_{\overline{ab}} \rangle = \langle \rho_x, (g_a^* \xi\xi^* g_a) \otimes (g_b^* P g_b) \otimes (g_{\overline{ab}}^* g_{\overline{ab}}) \rangle,$$

where we used Definition 6.2.2 of the marginals and the equivariance property $\rho_{g\cdot x} = g\rho_x g^*$ according to Lemma 6.2.6(1).

In order to use the fact that x is γ-$\mathfrak{p}$-pseudorandom, we note that both terms in the inner product above are positive semi-definite, so in order to lower bound the inner product, it is enough to give a spectral lower bound. To this end, note $\|g_a^*\xi\|_2^2 \geq \lambda_{\min}(g_a^* g_a)$ as $\|\xi\|_2 = 1$, so $g_a^*\xi\xi^* g_a \succeq \|(g_a^* g_a)^{-1}\|_{\text{op}} \psi\psi^*$ where $\psi := \frac{g_a^*\xi}{\|g_a^*\xi\|_2} \in \mathcal{S}_a$. Similarly, we can exhibit $Q \in \mathcal{P}_b$ (i.e. Q is an orthogonal projection with $\text{rk}(Q) = \text{rk}(P) = \frac{d_b}{2}$) such that $g_b^* P g_b \succeq \lambda_{\min}(g_b^* g_b) \cdot Q$. If G_b is commutative, then this claim is clear as P and g_b commute. The non-commutative case follows from Claim 7.3.5. With these two spectral lower bounds, we have

$$\begin{aligned}
\langle \rho_{g\cdot x}^{(ab)}, \xi\xi^* \otimes P \rangle &= \langle \rho_x, (g_a^* \xi\xi^* g_a) \otimes (g_b^* P g_b) \otimes (g_{\overline{ab}}^* g_{\overline{ab}}) \rangle \\
&\geq \langle \rho_x, (\lambda_{\min}(g_a^* g_a) \cdot \psi\psi^*) \otimes (\lambda_{\min}(g_b^* g_b) \cdot Q) \otimes (\lambda_{\min}(g_{\overline{ab}}^* g_{\overline{ab}}) \cdot I_{\overline{ab}}) \rangle \\
&\geq \lambda_{\min}(g_a^* g_a) \lambda_{\min}(g_b^* g_b) \lambda_{\min}(g_{\overline{ab}}^* g_{\overline{ab}}) \cdot \frac{e^{-\gamma}}{2d_a} = \prod_{c\in[m]} \lambda_{\min}(g_c^* g_c) \cdot \frac{e^{-\gamma}}{2d_a},
\end{aligned}$$

where the first step was by the calculation above, in the second step we applied our spectral lower bounds, and the third step was by γ-$\mathfrak{p}_{a\leftarrow b}$-pseudorandomness of x applied to $\psi \in \mathcal{S}_a, Q \in \mathcal{P}_b$. This verifies Definition 7.2.17 of pseudorandomness of $g \cdot x$. □

To finish the proof, we show the spectral lower bound in the claim below.

Claim 7.3.5. *For any orthogonal projection $P \in \mathrm{H}(d)$ and any invertible $g \in \mathrm{GL}(d)$, there is an orthogonal projection $Q \in \mathrm{H}(d)$ such that $\text{rk}(P) = \text{rk}(Q)$ and*

$$g^* P g \succeq \lambda_{\min}(g^* g) \cdot Q.$$

Proof. First note that $\text{rk}(g^* P g) = \text{rk}(P)$ as $g \in \mathrm{GL}(d)$ is invertible. To show the spectral lower bound, we will use Sylvester's Law of Intertia (Theorem 4.5.8 in [52]), which says that for any Hermitian A and invertible g, A and $g^* A g$ have the same number of positive, 0, and negative eigenvalues.

Now consider $A := P - (1-\delta) I_d$ for $\delta > 0$ arbitrary. Note that A has $\text{rk}(P)$ positive eigenvalues and the remaining $d - \text{rk}(P)$ are negative (see Definition 2.1.12 of the spectrum of orthogonal projections). Therefore, this must also be the signature of $g^* A g = g^* P g -$

$(1-\delta)g^*g$ by Sylvester's Law of intertia. Therefore, there must be a subspace $S \subseteq \mathbb{F}^d$ with $\dim(S) = \mathrm{rk}(P)$ such that

$$\forall v \in S : \langle vv^*, g^*Pg \rangle > (1-\delta)\langle vv^*, g^*g \rangle \geq (1-\delta) \cdot \lambda_{\min}(g^*g)\|v\|_2^2,$$

where the first step is by the variational principle for Hermitian eigenvalues applied to A which has $\mathrm{rk}(P)$-many positive eigenvalues (see e.g. Corollary III.1.2 in [12]), and the last step was by definition of the minimum eigenvalue of Hermitian g^*g. Therefore, for Q the orthogonal projection onto S according to Definition 2.1.12, we have

$$g^*Pg \succ (1-\delta)Q,$$

as $g^*Pg \succeq 0$ and Q is 0 on the orthogonal complement $\overline{S}$. The statement follows by taking the limit $\delta \to 0$. □

Note that in Lemma 7.2.23, we were able to give a lower bound for each individual $\xi \in \mathcal{S}_a, P \in \mathcal{P}_b$, whereas in Theorem 7.3.4 the scaling was arbitrary, so we could only bound the infimum. But this robustness is still multiplicative, meaning that we get a non-trivial result for arbitrary scalings. But our analysis in Theorem 7.2.26 requires $\gamma \lesssim \frac{1}{m}$ pseudorandomness, so this robustness bound will stop being useful when $\|g - I\|_{\mathrm{op}} \gtrsim \frac{1}{m}$. In Theorem 7.4.11, we will use the robustness of Theorem 7.3.4 to show that the pseudorandom analysis of Theorem 7.2.26 also implies pseudorandomness and strong convexity for the output G-balanced scaling. This will be useful in Chapter 9 to show fast algorithmic convergence for random inputs to the tensor scaling problem.

7.3.3 Strong Convexity for Operators

The subject of this and the following subsection will be robustness of strong convexity for non-commutative tensor scaling. It turns out that unlike the pseudorandomnes results of the previous subsections, for strong convexity, it is not possible to lift the multiplicative robustness bounds from Theorem 7.3.14 to the non-commutative setting as we show using a small frame example. Therefore, the remainder of this subsection will be devoted to setting up the proper definitions for our additive robustness results for strong convexity. At the end of this subsection, we will have enough tools to reproduce a strengthening of the main theorem of [63].

In the proof of Lemma 7.2.13, we crucially used that the direction $Z \in \mathfrak{t}$ and the scaling $Y \in \mathfrak{t}$ commute. It turns out that this is not merely an artifact of the proof, but is necessary for any multiplicative robustness bound. The following example was found during joint work [36] with Cole Franks, Rafael Oliveira, and Michael Walter.

Example 7.3.6. *Let $G = \mathrm{SL}(2) \otimes \mathrm{ST}(2)$ be the frame scaling group with associated polar and infinitesimal vector space $(P, \mathfrak{p})$ according to Definition 6.2.3. Consider input $V \in \mathrm{Mat}(2,2)$ and scaling $V' \in G \cdot V$ defined below:*

$$V := \begin{pmatrix} \sqrt{2} & 1 \\ 1 & \sqrt{2} \end{pmatrix}, \qquad V' := V^{-1}V = I_2.$$

Then V is $\Omega(1)$-$\mathfrak{p}$-strongly convex as a frame, but $V' = I_2$ is not strictly convex, i.e. it is not α-$\mathfrak{p}$-strongly convex for any $\alpha > 0$.

This is proved by a straightforward calculation in Appendix A.4. Note that one consequence of multiplicative robustness is that if x is α-strongly convex for any $\alpha > 0$, then any scaling $e^{Y/2} \cdot x$ is $e^{-\|Y\|_{\mathrm{op}}} \cdot \alpha > 0$ strongly convex. The example shows that it is impossible to derive a general multiplicative robustness bound for arbitrary scalings.

Therefore, in the following, we will focus on proving weaker robustness bounds which show that strong convexity is preserved up to additive error $\alpha \to \alpha - O(\|Y\|_{\mathrm{op}})$.

In the remainder of this section we will fix vector space $V = \otimes_{a \in [m]} V_a$ with $\dim(V_a) = d_a$ for each $a \in [m]$ and $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$, as this is the only scaling group we use for our algorithmic results in Chapter 9. The case of arbitrary scaling groups can be thought of as a (subgroup) restriction of this group G, and the following proofs of robustness can be extended straightforwardly.

By Definition 7.1.7, input $x \in V^K$ is shown to be strongly convex if for every $Z \in \mathfrak{p}$ we can lower bound

$$\partial^2_{\eta=0} f^P_x(e^{\eta Z}) = \sum_{a \in [m]} \langle \rho^{(a)}_x, Z_a^2 \rangle + \sum_{a \neq b \in [m]} \langle \rho^{(ab)}_x, Z_a \otimes Z_b \rangle,$$

where the decomposition is given in Eq. (7.4). Our plan is to use matrix perturbation results to bound each term when $x \to g \cdot x$ for scalings $g \approx I_V$.

Recall that in Proposition 7.1.10, we lower bounded the diagonal terms of the above decomposition for nearly balanced inputs, and upper bounded the off-diagonal term using the spectral condition for tensors. In the $m = 2$ operator scaling case, Lemma 7.1.11 shows that for nearly doubly balanced operators, strong convexity is nearly equivalent to the spectral condition, so in the following proof, we will actually analyze robustness of the spectral condition. We first use Definition 2.4.4 and Proposition 2.4.5 so that we can apply simple operator inequalities to show robustness for the associated quantum maps.

Most of the results in this subsection were already achieved in [63], and we view our contribution mostly as a principled approach via geodesic convexity. At the end, we will be able to combine the robustness result with our improved analysis of matrix scaling in Chapter 3 in order to produce a slight strengthening of the main results of [63] on strongly convex operator scaling. The next subsection lifts these results to the tensor case and requires some new ideas.

Our plan is to relate strong convexity to certain norms on quantum maps. Recall that according to Definition 2.4.4, the bipartite tensor tuple $x \in (U \otimes V)^K$ has associated state ρ_x and quantum map $\Phi_x : \mathrm{H}(V) \to \mathrm{H}(U)$ defined by

$$\langle Z, \Phi_x(Y) \rangle = \langle \rho_x, Z \otimes Y \rangle = \sum_{k=1}^{K} \langle x_k x_k^*, Z \otimes Y \rangle$$

for arbitrary $Z \in \mathrm{H}(U)$ and $Y \in \mathrm{H}(V)$. Also recall that the spectral condition in Definition 7.1.9 is defined with respect to the infinitesimal vector space $\mathfrak{spd}(d) \oplus \mathfrak{spd}(n)$ given in Definition 2.1.10.

We first give preliminary definitions relating the spectral condition to quantum maps.

Definition 7.3.7. *For $\Phi : \mathrm{H}(V) \to \mathrm{H}(U)$, we define two measures*

$$\|\Phi\|_{F\to F} := \sup_{X \in \mathrm{H}(U), Y \in \mathrm{H}(V)} \frac{|\langle X, \Phi(Y) \rangle|}{\|X\|_F \|Y\|_F}, \quad \textit{and} \quad \|\Phi\|_0 := \sup_{X \in \mathfrak{spd}(U), Y \in \mathfrak{spd}(V)} \frac{|\langle X, \Phi(Y) \rangle|}{\|X\|_F \|Y\|_F}.$$

Recall that $\mathfrak{spd}(U) = \{X \in \mathrm{H}(U) \mid \mathrm{Tr}[X] = 0\}$, so the 0 in $\|\cdot\|_0$ represents the trace 0 condition for the supremum.

Remark 7.3.8. *We could have equivalently defined $\Phi : L(V) \to L(U)$ and then $\|\Phi\|_{F\to F} = \sup_{Y \in L(V)} \frac{\|\Phi(Y)\|_F}{\|Y\|_F}$ as is standard. The restriction to $\mathrm{H}(V)$ is without loss of generality since Φ is Hermitian preserving. This can be explicitly shown in a similar fashion to Theorem 4.27 of [98] by using e.g. the Cartesian decomposition $L(V) = H(V) \oplus iH(V)$.*

Also note that the absolute value in the definitions of both $\|\cdot\|_{F\to F}$ and $\|\cdot\|_0$ are not necessary as we can always assume the optimum value is positive by symmetry.

We show some simple properties of these two measures that will be used repeatedly.

Proposition 7.3.9. *For arbitrary inner product spaces U, V,*

1. *$\|\cdot\|_{F\to F}$ is the standard operator norm induced by $\|\cdot\|_F$ according to Eq. (2.4);*
2. *$\|\cdot\|_{F\to F}$ is sub-multiplicative, i.e. for any $\Psi : \mathrm{H}(W) \to \mathrm{H}(V), \Phi : \mathrm{H}(V) \to \mathrm{H}(U)$,*

$$\|\Psi \circ \Phi\|_{F\to F} \leq \|\Psi\|_{F\to F} \|\Phi\|_{F\to F};$$

3. *For any* $\Phi : \mathrm{H}(V) \to \mathrm{H}(U)$*,* $\|\Phi\|_0 \leq \|\Phi\|_{F\to F}$*.*

4. $\|\cdot\|_0$ *is a semi-norm, i.e. it is homogenous and convex.*

Proof. Since $\|\cdot\|_F$ is the standard Euclidean norm on $\mathrm{H}(U)$ and $\mathrm{H}(V)$, the first statement is clear by the definition as

$$\|\Phi\|_{F\to F} = \sup_{X\in\mathrm{H}(U), Y\in\mathrm{H}(V)} \frac{\langle X, \Phi(Y)\rangle}{\|X\|_F\|Y\|_F} = \sup_{Y\in\mathrm{H}(V)} \frac{\|\Phi(Y)\|_F}{\|Y\|_F},$$

where we used Proposition 2.1.17 for $\|\cdot\|_F$ and the fact that Φ is Hermitian-preserving. The second item also follows from this formula as, for arbitrary $Y \in \mathrm{H}(W)$,

$$\|\Psi(\Phi(Y))\|_F \leq \|\Psi\|_{F\to F}\|\Phi(Y)\|_F \leq \|\Psi\|_{F\to F}\|\Phi\|_{F\to F}\|Y\|_F$$

where we repeatedly used the definition of $\|\cdot\|_{F\to F}$ as well as the fact that both Φ and Ψ are Hermitian preserving.

The third item follows simply as the domain of optimization for $\|\cdot\|_0$ is strictly contained in that of $\|\cdot\|_{F\to F}$.

For the fourth item, it is clear that $\|\cdot\|_0$ is homogenous under scalars, so to show it is a semi-norm, we verify the triangle inequality:

$$|\langle X, (\Psi+\Phi)(Y)\rangle| = |\langle X, \Psi(Y)\rangle + \langle X, \Phi(Y)\rangle| \leq |\langle X, \Psi(Y)\rangle| + |\langle X, \Phi(Y)\rangle|.$$

Finally, we note that $\|\cdot\|_0$ is not a norm by the following example:

$$\Phi(Y) := \langle I_n, Y\rangle I_d \implies \|\Phi\|_0 = \sup_{X\in\mathfrak{spd}(d), Y\in\mathfrak{spd}(n)} \frac{\langle X, \Phi(Y)\rangle}{\|X\|_F\|Y\|_F} = 0,$$

where the last step is by Definition 2.1.10 of $\mathfrak{spd}(n) = \{Y \in \mathrm{H}(n) \mid Tr[Y] = 0\}$. Clearly $\Phi \neq 0$ so $\|\cdot\|_0$ is not positive-definite. □

We will use these measures to bound the change in the off-diagonal terms of Eq. (7.4) under scaling. To this end, we show how $\|\cdot\|_0$ and $\|\cdot\|_{F\to F}$ relate to the spectral condition and strong convexity for tensors.

Lemma 7.3.10. *Consider* $V = \otimes_{a\in[m]} V_a$ *with* $\dim(V_a) = d_a$ *for each* $a \in [m]$ *and scaling group* $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$ *along with polar* $(P, \mathfrak{p})$ *according to Definition 6.2.3. For arbitrary pair* $a \neq b \in [m]$*:*

1. *$x \in V^K$ satisfies the λ-$\mathfrak{p}_{ab}$-spectral condition according to Definition 7.1.9 iff*

$$\|\Phi_x^{(ab)}\|_0 \leq \frac{\lambda}{\sqrt{d_a d_b}},$$

where $\Phi_x^{(ab)}$ is the map associated to $\rho_x^{(ab)}$ according to Proposition 2.4.5.

2. *For any $x \in V^K$, $\|\Phi_x^{(ab)}\|_{F\to F} \leq \sqrt{\frac{\|\rho_x^{(a)}\|_{\mathrm{op}}\|\rho_x^{(b)}\|_{\mathrm{op}}}{d_a d_b}}$.*

Proof. The first item follows from Definition 7.1.9 of the spectral condition for $\mathfrak{p}_a = \mathfrak{spd}(V_a)$ and $\mathfrak{p}_b = \mathfrak{spd}(V_b)$ and Definition 7.3.7 of $\|\cdot\|_0$ as $\langle Z_a, \Phi_x^{(ab)}(Z_b)\rangle = \langle \rho_x^{(ab)}, Z_a \otimes Z_b\rangle$ by the correspondence in Proposition 2.4.5. The second item is exactly Lemma 3.6 in [63]. Below, we give another proof inspired by geodesic convexity.

For any $Z \in \mathrm{H}(V_a)$ and $Y \in \mathrm{H}(V_b)$, we have

$$\partial^2_{\eta=0}\langle \rho_x^{(ab)}, e^{\eta Z} \otimes e^{\eta Y}\rangle = \langle \rho_x^{(ab)}, (Z \otimes I_b + I_a \otimes Y)^2\rangle \geq 0,$$

where the first step was by the product rule, and in the last step we used that $\rho_x^{(ab)} \succeq 0$ and $Z \in \mathrm{H}(V_a)$ and $Y \in \mathrm{H}(V_b)$ so $(Z \otimes I_b + I_a \otimes Y)^2 \succeq 0$. We expand this expression as

$$0 \leq \langle \rho_x^{(ab)}, (Z \otimes I_b + I_a \otimes Y)^2\rangle = \langle \rho_x^{(a)}, Z^2\rangle + \langle \rho_x^{(b)}, Y^2\rangle + 2\langle Z, \Phi_x^{(ab)}(Y)\rangle,$$

where in the last step we used $\langle \rho_x, Z \otimes Y\rangle = \langle Z, \Phi_A(Y)\rangle$ by Proposition 2.4.5. Rearranging the terms gives

$$\begin{aligned} 2\langle Z, \Phi_x^{(ab)}(Y)\rangle &= \langle \rho_x^{(a)}, Z^2\rangle + \langle \rho_x^{(b)}, Y^2\rangle - \langle \rho_x^{(ab)}, (Z \otimes I_b - I_a \otimes Y)^2\rangle \\ &\leq \|\rho_x^{(a)}\|_{\mathrm{op}}\|Z^2\|_1 + \|\rho_x^{(b)}\|_{\mathrm{op}}\|Y^2\|_1 = \frac{\|\rho_x^{(a)}\|_{\mathrm{op}}\|Z\|_F^2}{d_a} + \frac{\|\rho_x^{(b)}\|_{\mathrm{op}}\|Y\|_F^2}{d_b}, \end{aligned} \tag{7.11}$$

where the second step was by Schatten norm duality $\langle A, B\rangle \leq \|A\|_{\mathrm{op}}\|B\|_1$ given in Proposition 2.1.17 for the diagonal terms and the lower bound $\langle \rho_x^{(ab)}, (Z \otimes I_b + I_a \otimes Y)^2 \geq 0$ for the second term, and in the third step we used $\|Z^2\|_1 = \|Z\|_F^2$ and $\|Y^2\|_1 = \|Y\|_F^2$ for Hermitian $Z \in \mathrm{H}(V_a)$ and $Y \in \mathrm{H}(V_b)$.

We complete the proof by using AM-GM to bound the $F \to F$ norm:

$$\begin{aligned}
\|\Phi_x^{(ab)}\|_{F\to F} &= \sup_{Z\in \mathrm{H}(V_a), Y \in \mathrm{H}(V_b)} \inf_{\eta>0} \frac{|\langle \eta Z, \Phi_x^{(ab)}(\eta^{-1}Y)\rangle|}{\|Z\|_F\|Y\|_F} \\
&\leq \sup_{Z\in \mathrm{H}(V_a), Y \in \mathrm{H}(V_b)} \inf_{\eta>0} \frac{1}{\|Z\|_F\|Y\|_F}\left(\eta^2 \frac{\|\rho_x^{(a)}\|_{\mathrm{op}}\|Z\|_F^2}{2d_a} + \eta^{-2}\frac{\|\rho_x^{(b)}\|_{\mathrm{op}}\|Y\|_F^2}{2d_b}\right) \\
&= \sqrt{\frac{\|\rho_x^{(a)}\|_{\mathrm{op}}\|\rho_x^{(b)}\|_{\mathrm{op}}}{d_a d_b}}
\end{aligned}$$

where the first step was by Definition 7.3.7 of the $F \to F$ norm, in the second step we used the bound in Eq. (7.11) applied to $(\eta Z, \eta^{-1} Y)$, and the final step was by choosing η in the infimum so that we can replace the arithmetic mean by the geometric mean (i.e. the setting when AM-Gm is tight). □

We defer the more complicated $m \geq 3$ tensor case to Section 7.3.4, and in the rest of this subsection, we focus on showing robustness for the $m = 2$ case of operator scaling. With the previous definitions, we can bound the change in the off-digaonal term $\langle \rho_A, X \otimes Y\rangle$ of Eq. (7.4) using the above semi-norms.

Lemma 7.3.11. *Given tuple $A \in \mathrm{Mat}(d,n)^K$ and $L \in \mathrm{Mat}(d), R \in \mathrm{Mat}(n)$ such that $\max\{\|L - I_d\|_{\mathrm{op}}, \|R - I_n\|_{\mathrm{op}}\} \leq \frac{1}{2}$, scaling $B := LAR$ satisfies*

$$\|\Phi_B - \Phi_A\|_0 \leq \|\Phi_A\|_{F\to F}\Big((1 + 2.5\|L - I_d\|_{\mathrm{op}})(1 + 2.5\|R - I_n\|_{\mathrm{op}}) - 1\Big).$$

As a consequence, if A is ε-G-balanced with $G = (\mathrm{SL}(d), \mathrm{SL}(n))$ according to Definition 6.2.4 and satisfies the λ-$\mathfrak{p}$-spectral condition according to Definition 7.1.9 for $\mathfrak{p} = \mathfrak{spd}(d) \oplus \mathfrak{spd}(n)$, then B satisfies the λ'-$\mathfrak{p}$-spectral condition with

$$\lambda' \leq \lambda + s(A)(1+\varepsilon)\Big((1 + 2.5\|L - I_d\|_{\mathrm{op}})(1 + 2.5\|R - I_n\|_{\mathrm{op}}) - 1\Big).$$

Proof. We first rewrite Φ_B in terms of Φ_A as, for arbitrary $Y \in \mathrm{H}(n)$,

$$\Phi_B(Y) = \sum_{k=1}^K B_k Y B_k^* = \sum_{k=1}^K (LA_kR)Y(LA_kR)^* = L(\Phi_A(RYR^*))L^*,$$

where the first and last steps were by Definition 2.4.4 of Φ_B and Φ_A, and in the second step we substituted $B = LAR$.

Now we can write the map Φ_B as a perturbation of Φ_A. To this end, let $\mathcal{R} : \mathrm{Mat}(n) \to \mathrm{Mat}(n)$ be the linear operator defined by

$$\mathcal{R}(Y) := RYR^* = Y + (R - I_n)Y + Y(R - I_n)^* + (R - I_n)Y(R - I_n).$$

Note that if R is small, we can show $\mathcal{R}$ is close to the identity operator $\mathcal{I}_n : \mathrm{Mat}(n) \to \mathrm{Mat}(n)$, as for any $Y \in \mathrm{H}(n)$,

$$\begin{aligned}\|\mathcal{R}(Y) - \mathcal{I}_n(Y)\|_F &\leq \|(R - I_n)Y\|_F + \|Y(R - I_n)^*\|_F + \|(R - I_n)Y(R - I_n)^*\|_F \\ &\leq \|Y\|_F(2\|R - I_n\|_{\mathrm{op}} + \|R - I_n\|_{\mathrm{op}}^2),\end{aligned}$$

where the first step was by the triangle inequality on $\|\cdot\|_F$ and the last step was by the bound $\|XY\|_F \leq \|X\|_{\mathrm{op}}\|Y\|_F$. Since $Y \in \mathrm{H}(n)$ is arbitrary, this implies $\|\mathcal{R} - \mathcal{I}_n\|_{F\to F} \leq 2\|R - I_n\|_{\mathrm{op}} + \|R - I_n\|_{\mathrm{op}}^2$ by Definition 7.3.7. For $\mathcal{L} : \mathrm{Mat}(d) \to \mathrm{Mat}(d)$ defined as $\mathcal{L}(X) = LXL^*$, and the identity operator $\mathcal{I}_d : \mathrm{Mat}(d) \to \mathrm{Mat}(d)$, we can use a similar calculation to show $\|\mathcal{L} - \mathcal{I}_d\|_{F\to F} \leq 2\|L - I_d\|_{\mathrm{op}} + \|L - I_d\|_{\mathrm{op}}^2$.

Now we observe that Φ_B for scaling $B = LAR$ is close to Φ_A as

$$\begin{aligned}\Phi_B - \Phi_A &= \mathcal{L} \circ \Phi_A \circ \mathcal{R} - \mathcal{I}_d \circ \Phi_A \circ \mathcal{I}_n \\ &= (\mathcal{L} - \mathcal{I}_d)\Phi_A + \Phi_A(\mathcal{R} - \mathcal{I}_n) + (\mathcal{L} - \mathcal{I}_d)\Phi_A(\mathcal{R} - \mathcal{I}_n).\end{aligned}$$

This decomposition allows us to bound the difference by

$$\begin{aligned}\|\Phi_B - \Phi_A\|_0 &\leq \|(\mathcal{L} - \mathcal{I}_d)\Phi_A\|_0 + \|\Phi_A(\mathcal{R} - \mathcal{I}_n)\|_0 + \|(\mathcal{L} - \mathcal{I}_d)\Phi_A(\mathcal{R} - \mathcal{I}_n)\|_0 \\ &\leq \|\Phi_A\|_{F\to F}(\|\mathcal{L} - \mathcal{I}_d\|_{F\to F} + \|\mathcal{R} - \mathcal{I}_n\|_{F\to F} + \|\mathcal{L} - \mathcal{I}_d\|_{F\to F}\|\mathcal{R} - \mathcal{I}_n\|_{F\to F}) \\ &\leq \|\Phi_A\|_{F\to F}\Big((1 + 2.5\|L - I_d\|_{\mathrm{op}})(1 + 2.5\|R - I_n\|_{\mathrm{op}}) - 1\Big)\end{aligned}$$

where in the first step we used the triangle inequality on semi-norm $\|\cdot\|_0$ by Proposition 7.3.9(4), in the second step we used the bounds $\|\cdot\|_0 \leq \|\cdot\|_{F\to F}$ and sub-multiplicativity of $\|\cdot\|_{F\to F}$ from items (2) and (3) of Proposition 7.3.9, and in the final step we used the bounds on $\mathcal{L}, \mathcal{R}$ derived above as well as the assumption that $\max\{\|L - I_d\|_{\mathrm{op}}, \|R - I_n\|_{\mathrm{op}}\} \leq \frac{1}{2}$ to bound $\|L - I_d\|_{\mathrm{op}}^2 \leq \frac{1}{2}\|L - I_d\|_{\mathrm{op}}$ and $\|R - I_n\|_{\mathrm{op}}^2 \leq \frac{1}{2}\|R - I_n\|_{\mathrm{op}}$. This is exactly the perturbation bound in the lemma.

For the last statement, item (1) in Lemma 7.3.10 shows that A satisfies the λ-$\mathfrak{p}$-spectral condition iff $\|\Phi_A\|_0 \leq \frac{\lambda}{\sqrt{dn}}$. Therefore, we can bound

$$\|\Phi_B\|_0 \leq \|\Phi_A\|_0 + \|\Phi_B - \Phi_A\|_0,$$

where we can apply the triangle inequality for semi-norm $\| \cdot \|_0$ by Proposition 7.3.9(4). Further, we can use the balance condition of A to bound

$$\|\Phi_A\|_0 \leq \|\Phi_A\|_{F \to F} \leq \sqrt{\|\rho_A^L\|_{\text{op}} \|\rho_A^R\|_{\text{op}}} \leq \frac{s(A)(1+\varepsilon)}{\sqrt{dn}},$$

where the first step is by item (2) of Lemma 7.3.10 and the final step is by the assumption that A is ε-G-balanced (see Definition 6.2.4). Combining this with the perturbation bound on $\|\Phi_B - \Phi_A\|_0$ shows

$$\|\Phi_B\|_0 \leq \|\Phi_A\|_0 + \|\Phi_B - \Phi_A\|_0 \leq \frac{\lambda}{\sqrt{dn}} + \frac{s(A)(1+\varepsilon)}{\sqrt{dn}} \Big((1 + \|L - I_d\|_{\text{op}})(1 + \|R - I_n\|_{\text{op}}) - 1 \Big),$$

which again by item (1) of Lemma 7.3.10 implies that B satisfies the λ'-$\mathfrak{p}$-spectral condition for λ' as given in the lemma. □

This can be combined with the analysis of matrix scaling in Section 3.2.3 to show strong convexity of the optimizer for operators satisfying the spectral condition. In the next subsection, we will lift this result to the tensor setting.

Theorem 7.3.12. *Consider matrix tuple $A \in \mathrm{Mat}(d,n)^K \simeq (\mathbb{F}^d \otimes \mathbb{F}^n)^K$ along with scaling group $G = (\mathrm{SL}(d), \mathrm{SL}(n))$ and polar $(P, \mathfrak{p})$ according to Definition 6.2.3. If A of size $s(A) = 1$ is ε-G-balanced and satisfies the λ-$\mathfrak{p}$-spectral condition according to Definition 7.1.9 with $\frac{1}{5} \geq 1 - \lambda \geq \varepsilon(4 \log d + 21)$, then there is a scaling $A_* = e^{X_*/2} A e^{Y_*/2}$ with $(X_*, Y_*) \in \mathfrak{p}$ satisfying:*

1. *$A_* := e^{X_*/2} A e^{Y_*/2}$ is a doubly balanced operator;*
2. *The scaling solution $(X_*, Y_*) \in \mathfrak{p}$ satisfies*

$$\|(X_*, Y_*)\|_{\mathfrak{p}} \lesssim \frac{\varepsilon}{1-\lambda} \qquad \text{and} \qquad \max\{\|X_*\|_{\text{op}}, \|Y_*\|_{\text{op}}\} \lesssim \frac{\varepsilon \log d}{1-\lambda};$$

3. *The size of the solution can be lower bounded by*

$$s(A_*) = f_A(e^{X_*}, e^{Y_*}) \geq 1 - O\Big(\frac{\varepsilon^2}{1-\lambda}\Big);$$

4. *A_* is α_*-$\mathfrak{p}$-strongly convex with*

$$\alpha_* \geq 1 - \lambda - O\Big(\frac{\varepsilon \log d}{1-\lambda}\Big).$$

Note item (4) only gives a non-trivial lower bound for $(1-\lambda)^2 \gtrsim \varepsilon \log d$.

Proof. By item (2) of Proposition 6.2.18, if

$$(e^{X_*}, e^{Y_*}) := \arg\inf_{p \in \mathfrak{p}} f_A^P(p)$$

is the global optimizer of the Kempf-Ness function f_A^P given in Definition 6.2.9, then $e^{X_*/2} A e^{Y_*/2}$ is a G-balanced scaling of A according to Definition 6.2.4. We will show that this global minimum is attained and satisfies the bounds above. Our plan is to use the decomposition result of Theorem 6.3.1 in order to reduce to the simpler matrix scaling setting where we can apply the strongly convex analysis of Theorem 3.2.19.

First note that by Proposition 7.1.10, we have that A is α-$\mathfrak{p}$-strongly convex with

$$\alpha \geq s(A)(1-\varepsilon) - \lambda = 1 - \lambda - \varepsilon \geq \varepsilon(4\log d + 20),$$

where the last step is by our assumption $1 - \lambda \geq \varepsilon(4\log d + 21)$.

Now for any choice of orthonormal bases (Ξ, Ψ) for $(\mathbb{F}^d, \mathbb{F}^n)$ respectively, let $(T^{\Xi,\Psi}, T_+^{\Xi,\Psi}, \mathfrak{t}^{\Xi,\Psi})$ be the commutative matrix scaling groups that are diagonal in the (Ξ, Ψ) bases according to Definition 7.2.1. Explicitly, we have $\mathfrak{t}_+^{\Xi,\Psi} = \mathfrak{st}_+^{\Xi}(d) \oplus \mathfrak{st}_+^{\Psi}(n)$ (see Section 2.2.2), and this definition allows us to decompose

$$P = \cup_{\Xi,\Psi} T_+^{\Xi,\Psi}, \qquad \text{and} \qquad \mathfrak{p} = \cup_{\Xi,\Psi} \mathfrak{t}^{\Xi,\Psi}$$

by Eq. (2.6) and Eq. (2.7) for each component.

As $T^{\Xi,\Psi} \subseteq G$, the ε-G-balance condition of A implies the ε-$T^{\Xi,\Psi}$-balance condition according to Definition 6.2.4. Similarly, as $\mathfrak{t}^{\Xi,\Psi} \subseteq \mathfrak{p}$, α-$\mathfrak{p}$-strong convexity of A implies α-$\mathfrak{t}^{\Xi,\Psi}$-strong convexity according to Definition 7.1.7. Therefore, each matrix representation $M^{\Xi,\Psi} := \Xi^* A \Psi$ satisfies the balance and strong convexity conditions of Theorem 3.2.19, which gives matrix scaling solution $(X_{\Xi,\Psi}, Y_{\Xi,\Psi}) \in \mathfrak{t}$. By Proposition 3.1.10(3), this implies that $(X_{\Xi,\Psi}, Y_{\Xi,\Psi})$ is the global minimizer of the matrix Kempf-Ness function $f_{M^{\Xi,\Psi}}$ in Definition 3.1.6. For $(X, Y) \in \mathfrak{t}$, we can rewrite this as

$$f_{M^{\Xi,\Psi}}(X,Y) := \sum_{k=1}^{K} \|e^{X/2} M_k^{\Xi,\Psi} e^{Y/2}\|_F^2 = \sum_{k=1}^{K} \|(\Xi e^{X/2} \Xi^*) A_k (\Psi e^{Y/2} \Psi^*)\|_F^2 = f_A^P(\Xi e^{X/2} \Xi^*, \Psi e^{Y/2} \Psi^*),$$

where the first step was by Definition 3.1.6 of the matrix Kempf-Ness function, in the second step we used invariance of $\|\cdot\|_F$ under isometries Ξ, Ψ, and the final step was by Definition 6.2.9 of the Kempf-Ness function on domain P.

Therefore, $f_{M^{\Xi,\Psi}}$ is equivalent to the restriction of f_A^P to the subset $T_+^{\Xi,\Psi} \subseteq P$, and $(\Xi e^{X/2}\Xi^*)A_k(\Psi e^{Y/2}\Psi^*) \in T_+^{\Xi,\Psi}$ is the global minimizer on this restriction. Since each $(X_{\Xi,\Psi}, Y_{\Xi,\Psi})$ is bounded, we can apply Theorem 6.3.1 to the decomposition $P = \cup_{\Xi,\Psi} T^{\Xi,\Psi}$ to find a global minimizer of f_A^P of the form $(e^{X_*}, e^{Y_*}) \in \cup_{\Xi,\Psi}(\Xi e^{X_{\Xi,\Psi}}\Xi^*, \Psi e^{Y_{\Xi,\Psi}}\Psi^*)$, which then implies that $A_* := e^{X_*/2}Ae^{Y_*/2}$ is a doubly balanced operator by Proposition 6.2.18(3). The bounds in items (1)-(3) are implied by the analogous bounds on the matrix scaling solutions from Theorem 3.2.19 applied with $\alpha \geq 1 - \lambda - \varepsilon \gtrsim 1 - \lambda$ by the assumption $1 - \lambda \gtrsim \varepsilon \log d$.

To prove item (4), we first apply Lemma 7.3.11 to show that A_* satisfies the λ'-$\mathfrak{p}$-spectral condition with

$$\begin{aligned}\lambda' - \lambda &\leq s(A)(1+\varepsilon)\Big((1 + 2.5\|e^{X_*/2} - I_d\|_{\text{op}})(1 + 2.5\|e^{Y_*/2} - I_n\|_{\text{op}}) - 1\Big) \\ &\leq (1+\varepsilon)\Big((1 + 2.5\|X_*\|_{\text{op}})(1 + 2.5\|Y_*\|_{\text{op}}) - 1\Big) \lesssim \|X_*\|_{\text{op}} + \|Y_*\|_{\text{op}} \lesssim \frac{\varepsilon \log d}{1-\lambda},\end{aligned}$$

where the first step was by Lemma 7.3.11 applied to A, which is ε-doubly balanced and satisfies the λ-$\mathfrak{p}$-spectral condition, along with scaling $A_* := e^{X_*/2}Ae^{Y_*/2}$, and the final steps used the assumption $s(A) = 1$ as well as the Taylor approximation $|e^z - 1| \leq 2|z|$ for $|z| \leq \frac{1}{2}$ applied to $\max\{\|X_*\|_{\text{op}}, \|Y_*\|_{\text{op}}\} \lesssim \frac{\varepsilon \log d}{1-\lambda} \leq \frac{1}{2}$ by item (2) and our assumption $1 - \lambda \geq \varepsilon(4\log d + 21)$.

Finally, we apply Proposition 7.1.10 to show that A_* is α_*-$\mathfrak{p}$-strongly convex with

$$\alpha_* \geq s(A_*) - \lambda' \geq 1 - O\Big(\frac{\varepsilon^2}{1-\lambda}\Big) - \lambda - O\Big(\frac{\varepsilon \log d}{1-\lambda}\Big),$$

where we used Proposition 7.1.10 and the fact that A_* is doubly balanced in the first step, and in the second step we used the size lower bound $s(A_*) \geq 1 - O(\frac{\varepsilon^2}{1-\lambda})$ given in item (3) as well as the spectral upper bound on λ' calculated above. □

This should be compared with the proof of Theorem 1.5 in [63] which analyzed the scaling solution using the more complicated non-commutative gradient flow described in Definition 7.1.5. This required robustness of strong convexity in the non-commutative setting, which meant that we were only able to prove existence of the operator scaling solution with the stronger assumption $(1-\lambda)^2 \gtrsim \varepsilon \log d$. We improve this existence part of the result by our reduction to matrix scaling. But as shown in Example 7.3.6, it is not possible to show non-commutative multiplicative robustness of strong convexity, so our result on the strong convexity of the operator scaling solution is identical to the one given in [63] up to constants.

7.3.4 Strong Convexity for Tensors

In this subsection, we will show that strong convexity for non-commutative tensor scaling is additively robust. We will follow the plan laid out in Section 7.3.3 by showing that each term in the block decomposition in Eq. (7.4) does not change much for small scalings $g \approx I_V$. For this subsection, we will fix vector space $V = \otimes_{a \in [m]} V_a$ with $\dim(V_a) = d_a$ for each $a \in [m]$ and $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$, as this is the only scaling group we use for our algorithmic results in Chapter 9. The case of arbitrary scaling groups can be thought of as a (subgroup) restriction of this group G, and the following proofs of robustness can be extended straightforwardly.

We first rephrase Lemma 7.3.11 in the more general tensor scaling setting.

Lemma 7.3.13. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces and consider scaling group $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$ along with polar $(P, \mathfrak{p})$ according to Definition 6.2.3. For input $x \in V^K$ and pair $a \neq b \in [m]$, if $x' := (g_c \otimes I_{\overline{c}}) \cdot x$ for $c = a$ or $c = b$ and $\|g_c - I_c\|_{\mathrm{op}} \leq \frac{1}{2}$, then the scaling x' satisfies*

$$\|\Phi_{x'}^{(ab)} - \Phi_x^{(ab)}\|_0 \leq 2.5 \|g_c - I_c\|_{\mathrm{op}} \|\Phi_x^{(ab)}\|_{F \to F}.$$

Consequently, for any $Z \in \mathfrak{spd}(V_a)$ and $Y \in \mathfrak{spd}(V_b)$

$$|\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y \rangle| \leq \left(2.5 \|g_c - I_c\|_{\mathrm{op}} \sqrt{\|\rho_x^{(a)}\|_{\mathrm{op}} \|\rho_x^{(b)}\|_{\mathrm{op}}} \right) \|Z\|_F \|Y\|_F.$$

Proof. Since the scaling $g_c \in G_c$ for $c \in \{a, b\}$, the first statement follows by Lemma 7.3.11 applied to the map $\Phi_x^{(ab)}$. To show the second statement, we rewrite this in terms of the marginals: for any $Z \in \mathfrak{spd}(V_a)$ and $Y \in \mathfrak{spd}(V_b)$ with $\|Z\|_F = \|Y\|_F = 1$ we have

$$\begin{aligned} |\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y \rangle| &= |\langle Z, \Phi_{x'}^{(ab)}(Y) - \Phi_x^{(ab)}(Y) \rangle| \leq \|\Phi_{x'}^{(ab)} - \Phi_x^{(ab)}\|_0 \\ &\leq 2.5 \|g_c - I_c\|_{\mathrm{op}} \|\Phi_x^{(ab)}\|_{F \to F} \leq 2.5 \|g_c - I_c\|_{\mathrm{op}} \sqrt{\|\rho_x^{(a)}\|_{\mathrm{op}} \|\rho_x^{(b)}\|_{\mathrm{op}}}, \end{aligned}$$

where the first step was by Proposition 2.4.5, the second was by Definition 7.3.7 of $\|\cdot\|_0$ with the assumption $\|Z\|_F = \|Y\|_F = 1$, the third step was by the first part of the lemma, and the final step was by item (2) of Lemma 7.3.10. □

Recall that for input $x \in V^K$ and $Z \in \mathfrak{p}$, Eq. (7.4) gives a decomposition

$$\partial_{\eta=0}^2 f_x^P(e^{\eta Z}) = \sum_{a \in [m]} \langle \rho_x^{(a)}, Z_a^2 \rangle + \sum_{a \neq b \in [m]} \langle \rho_x^{(ab)}, Z_a \otimes Z_b \rangle.$$

Our plan is to consider the perturbation $g \cdot x$ as a composition of perturbations, $\{g_1 \otimes I_{\overline{1}}, ..., g_m \otimes I_{\overline{m}}$, one for each part. We will show that each term in Eq. (7.4) changes only a small amount for each part-wise perturbation. The above Lemma 7.3.13 allows us to control the change in the ab-term by scaling $g_a \otimes I_{\overline{a}}$ or $g_b \otimes I_b$. But this bound contains terms of the form $\|\Phi_{x'}^{(ab)}\|_{F \to F}^2 \leq \|\rho_{x'}^{(a)}\|_{\text{op}} \|\rho_{x'}^{(b)}\|_{\text{op}}$, where $x' \in G \cdot x$ come from the scalings of x. In order to control these terms, we first show in Proposition 7.3.17 that $\rho_{x'}^{(a)} \approx \rho_x^{(a)}$ for small scalings $x' = g \cdot x$. We will then bound the change in two-body marginals $\rho^{(ab)}$ which will allow us to control the off-diagonal terms in Proposition 7.3.19. The main result in this subsection will be the following theorem.

Theorem 7.3.14. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, and consider scaling group $G = (\text{SL}(V_1), ..., \text{SL}(V_m))$ along with polar $(P, \mathfrak{p})$ according to Definition 6.2.3. Consider tuple $x \in V^K$ of size $s(x) = 1$ that is ε-G-balanced and α-$\mathfrak{p}$-strongly convex along with any $g \in G$ such that $\delta := \sum_{c \in [m]} \|g_c - I_c\|_{\text{op}} \leq \frac{1}{20}$. Then the scaling $x' := g \cdot x$ is α'-$\mathfrak{p}$-strongly convex for $\alpha' \geq \alpha - (4 + 7.5(m-1))\delta \cdot s(x)(1+\varepsilon)$.*

We begin by showing each one-body marginal $\rho^{(a)}$ does not change much under a single scaling $g_c \otimes I_{\overline{c}}$. We separate into two cases depending on whether $c = a$ or $c \neq a$.

Lemma 7.3.15. *For tensor product $V = \otimes_{a \in [m]} V_a$, consider tuple $x \in V^K$ and perturbation $g_a \in G_a$ for $a \in [m]$ such that $\|g_a - I_a\|_{\text{op}} \leq \frac{1}{2}$. Then scaling $x' := (g_a \otimes I_{\overline{a}}) \cdot x$ satisfies*

$$\|\rho_{x'}^{(a)} - \rho_x^{(a)}\|_{\text{op}} \leq 2.5 \|g_a - I_a\|_{\text{op}} \|\rho_x^a\|_{\text{op}}.$$

Proof. We first rewrite the marginal $\rho_{x'}^{(a)}$ as a perturbation of $\rho_x^{(a)}$:

$$\rho_{x'}^{(a)} = g_a \rho_x^{(a)} g_a^* = \rho_x^{(a)} + (g_a - I_a)\rho_x^{(a)} + \rho_x^{(a)}(g_a - I_a)^* + (g_a - I_a)\rho_x^{(a)}(g_a - I_a)^*,$$

where we used the equivariance property in Lemma 6.2.6(2) for the $S = \{a\}$ marginal. This allows us to bound

$$\begin{aligned}\|\rho_{x'}^{(a)} - \rho_x^{(a)}\|_{\text{op}} &\leq \|(g_a - I_a)\rho_x^{(a)}\|_{\text{op}} + \|\rho_x^{(a)}(g_a - I_a)^*\|_{\text{op}} + \|(g_a - I_a)\rho_x^{(a)}(g_a - I_a)^*\|_{\text{op}} \\ &\leq \|\rho_x^{(a)}\|_{\text{op}}(2\|g_a - I_a\|_{\text{op}} + \|g_a - I_a\|_{\text{op}}^2) \leq 2.5\|g_a - I_a\|_{\text{op}} \|\rho_x^{(a)}\|_{\text{op}},\end{aligned}$$

where the first step was by the triangle inequality, the second step was by sub-multiplicativity of $\|\cdot\|_{\text{op}}$, and in the final step we used the assumption $\|g_a - I_a\|_{\text{op}} \leq \frac{1}{2}$. □

The proof for perturbation g_b with $b \neq a$ is slightly more complicated, as we have to expand out the marginals.

Lemma 7.3.16. *For tensor product $V = \otimes_{a\in[m]} V_a$, consider tuple $x \in V^K$ and perturbation $g_b \in G_b$ for $b \neq a$ such that $\|g_b - I_b\|_{\text{op}} \leq \frac{1}{10}$. Then scaling $x' := (g_b \otimes I_{\bar{b}}) \cdot x$ satisfies*

$$\|\rho_{x'}^{(a)} - \rho_x^{(a)}\|_{\text{op}} \leq 2.5\|g_b - I_b\|_{\text{op}}\|\rho_x^{(a)}\|_{\text{op}}.$$

Proof. By Eq. (2.5), we can rewrite the operator norm of the difference as $\|\rho_{x'}^{(a)} - \rho_x^{(a)}\|_{\text{op}} = \sup_{\xi\in S^{d_a-1}} |\langle \xi\xi^*, \rho_{x'}^{(a)} - \rho_x^{(a)}\rangle|$ since both $\rho_x^{(a)}$ and $\rho_{x'}^{(a)}$ are Hermitian. To bound each such inner product, we first rewrite $\rho_{x'}^{(ab)}$ as a perturbation of $\rho_x^{(ab)}$:

$$\begin{aligned}\rho_{x'}^{(ab)} - \rho_x^{(ab)} &= (g_b \otimes I_{\bar{b}})\rho_x^{(ab)}(g_b \otimes I_{\bar{b}})^* - \rho_x^{(ab)} \\ &= (I_a \otimes (g_b - I_b))\rho_x^{(ab)} + \rho_x^{(ab)}(I_a \otimes (g_b - I_b))^* + (I_a \otimes (g_b - I_b))\rho_x^{(ab)}(I_a \otimes (g_b - I_b))^*,\end{aligned}$$

where we applied the equivariance property from Lemma 6.2.6(2) to $\rho_x^{(ab)}$.

Now for arbitrary $\xi \in S^{d_a-1}$, we can bound the inner product as

$$\begin{aligned}|\langle \xi\xi^*, \rho_{x'}^{(a)} - \rho_x^{(a)}\rangle| &= \langle \xi\xi^* \otimes I_b, (I_a \otimes g_b)\rho_x^{(ab)}(I_a \otimes g_b)^* - \rho_x^{(ab)}\rangle \\ &= \left|\left\langle \xi\xi^* \otimes \Big((g_b - I_b) + (g_b - I_b)^* + (g_b - I_b)^*(g_b - I_b)\Big), \rho_x^{(ab)}\right\rangle\right| \\ &\leq |\langle \xi\xi^* \otimes I_b, \rho_x^{(ab)}\rangle|(2\|g_b - I_b\|_{\text{op}} + \|g_b - I_b\|_{\text{op}}^2),\end{aligned}$$

where the first step was by the equivariance property from Lemma 6.2.6(2), the second step was shown above, and in the final step we used the fact that $|g_b - I_b| \preceq \|g_b - I_b\|_{\text{op}} \cdot I_b$ and $\rho_x^{(ab)} \succeq 0$ so we can bound the inner product using this spectral upper bound. The lemma follows as

$$|\langle \xi\xi^* \otimes I_b, \rho_x^{(ab)}\rangle| = |\langle \xi\xi^*, \rho_x^{(a)}\rangle| \leq \|\rho_x^{(a)}\|_{\text{op}},$$

where the first step was by Definition 6.2.2 of marginals, and the final step was by the fact that $\xi \in S^{d_a-1}$ was arbitrary. □

We can now combine the above lemmas to bound the diagonal term in Eq. (7.4).

Proposition 7.3.17. *For tensor product $V = \otimes_{a\in[m]} V_a$, consider tuple $x \in V^K$ and perturbation $g \in G = (\text{SL}(V_1), ..., \text{SL}(V_m))$ such that $\sum_{b\in[m]} \|g_b - I_b\|_{\text{op}} \leq \frac{1}{10}$. Then scaling $x' := \otimes_{b\in[m]} g_b \cdot x$ satisfies*

$$\forall Z \in H(V_a): \quad |\langle \rho_{x'}^{(a)} - \rho_x^{(a)}, Z^2\rangle| \leq 4\|\rho_x^{(a)}\|_{\text{op}}\|Z\|_F^2 \sum_{b\in[m]} \|g_b - I_b\|_{\text{op}}.$$

Proof. We treat the scaling as the composition of m perturbations

$$x_0 := x, \quad \to \quad x_1 := (g_1 \otimes I_{\overline{1}}) \cdot x_0, \quad \to \quad \ldots \to \quad x_m := (g_m \otimes I_{\overline{m}}) \cdot x_{m-1} = x'.$$

For Hermitian $Z \in H(V_a)$ so that $Z^2 \succeq 0$, so we can bound

$$|\langle \rho_{x'}^{(a)} - \rho_x^{(a)}, Z^2 \rangle| \leq \sum_{b=1}^{m} \|\rho_{x_b}^{(a)} - \rho_{x_{b-1}}^{(a)}\|_{\text{op}} \|Z^2\|_1 \leq \sum_{b=1}^{m} 2.5 \|g_b - I_b\|_{\text{op}} \|\rho_{x_{b-1}}^{(a)}\|_{\text{op}} \|Z\|_F^2, \quad (7.12)$$

where the first step is by a telescoping sum and the triangle inequality for $\langle X, Y \rangle \leq \|X\|_{S_1} \|Y\|_{S_\infty}$ as shown in Proposition 2.1.17, and in the final step we used Lemma 7.3.15 to bound the $b = a$ term and Lemma 7.3.16 for the rest. It can further be shown that

$$\|\rho_{x_b}^{(a)}\|_{\text{op}} \leq \|\rho_{x_{b-1}}^{(a)}\|_{\text{op}} + \|\rho_{x_b}^{(a)} - \rho_{x_{b-1}}^{(a)}\|_{\text{op}} \leq \|\rho_{x_{b-1}}^{(a)}\|_{\text{op}} (1 + 2.5 \|g_b - I_b\|_{\text{op}}),$$

where the first step was by the triangle inequality, and the final step was by the perturbation bounds in Lemma 7.3.15 and Lemma 7.3.16. By induction, this implies $\|\rho_{x_b}^{(a)}\|_{\text{op}} \leq \|\rho_x^{(a)}\|_{\text{op}} \prod_{j=1}^{b} (1 + 2.5 \|g_j - I_j\|_{\text{op}})$, so we can collect terms to show

$$\begin{aligned} |\langle \rho_{x'}^{(a)} - \rho_x^{(a)}, Z^2 \rangle| &\leq \|\rho_x^{(a)}\|_{\text{op}} \|Z\|_F^2 \sum_{b=1}^{m} 2.5 \|g_b - I_b\|_{\text{op}} \prod_{j=1}^{b-1} (1 + 2.5 \|g_j - I_j\|_{\text{op}}) \\ &= \|\rho_x^{(a)}\|_{\text{op}} \|Z\|_F^2 \left(\prod_{b=1}^{m} (1 + 2.5 \|g_b - I_b\|_{\text{op}}) - 1 \right) \\ &\leq \|\rho_x^{(a)}\|_{\text{op}} \|Z\|_F^2 \left(\exp\left(2.5 \sum_{b=1}^{m} \|g_b - I_b\|_{\text{op}} \right) - 1 \right) \\ &\leq 4 \|\rho_x^{(a)}\|_{\text{op}} \|Z\|_F^2 \sum_{b=1}^{m} \|g_b - I_b\|_{\text{op}}, \end{aligned}$$

where the first step was by Eq. (7.12) where we substituted $\|\rho_{x_{b-1}}^{(a)}\|_{\text{op}} \leq \|\rho_x^{(a)}\|_{\text{op}} \prod_{j=1}^{b-1} (1 + 2.5 \|g_j - I_j\|_{\text{op}})$, the second step was by the formula $1 + \sum_{i=1}^{m} a_i \prod_{j<i} (1 + a_j) = \prod_{i=1}^{m} (1 + a_i)$ which can be shown by a simple induction, in the third step we used the bound $1 + x \leq e^x$, and the final step was by Taylor approximation $e^x \leq 1 + \frac{3}{2} x$ for $0 \leq x \leq \frac{1}{4}$ which is satisfied by our assumption $\sum_{b \in [m]} \|g_b - I_b\|_{\text{op}} \leq \frac{1}{10}$. □

Recall that in Lemma 7.3.13, we bounded the difference $\rho_{x'}^{(ab)} - \rho_x^{(ab)}$ when the scaling was on one of the parts a or b. In the next lemma, we will show a similar perturbation bound when the scaling acts on another part $c \notin \{a, b\}$. This will allow us to bound the change in the off-diagonal term in Proposition 7.3.19.

Lemma 7.3.18. *For tensor product $V = \otimes_{a\in[m]} V_a$, consider tuple $x \in V^K$, fixed $a, b, c \in [m]$ all distinct, and perturbation $g_c \in \mathrm{Mat}(d_c)$ such that $\|g_c - I_c\|_{\mathrm{op}} \leq \frac{1}{2}$. Then scaling $x' := (g_c \otimes I_{\overline{c}}) \cdot x$ satisfies, for any $Z \in \mathrm{H}(V_a), Y \in \mathrm{H}(V_b)$:*

$$|\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y\rangle| \leq \left(5\|g_c - I_c\|_{\mathrm{op}}\sqrt{\|\rho_x^{(a)}\|_{\mathrm{op}}\|\rho_x^{(b)}\|_{\mathrm{op}}}\right)\|Z\|_F\|Y\|_F.$$

Proof. We first rewrite $\rho_{x'}^{(abc)} = (I_{ab} \otimes g_c)\rho_x^{(abc)}(I_{ab} \otimes g_c)^*$ as a perturbation of $\rho_x^{(abc)}$:

$$\rho_{x'}^{(abc)} = \rho_x^{(abc)} + (I_{ab}\otimes(g_c - I_c))\rho_x^{(abc)} + \rho_x^{(abc)}(I_{ab}\otimes(g_c-I_c))^* + (I_{ab}\otimes(g_c-I_c))\rho_x^{(abc)}(I_{ab}\otimes(g_c-I_c))^*.$$

To bound the inner product in the lemma, we first consider $Z \in H(V_a), Y \in H(V_b)$ such that $Z \succeq 0$ and $Y \succeq 0$:

$$\begin{aligned}
\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y\rangle &= \langle (I_{ab} \otimes g_c)\rho_x^{(abc)}(I_{ab} \otimes g_c)^* - \rho_x^{(abc)}, Z \otimes Y \otimes I_c\rangle \\
&= \left\langle \rho_x^{(abc)}, Z \otimes Y \otimes \Big((g_c - I_c) + (g_c - I_c)^* + (g_c - I_c)^*(g_c - I_c)\Big)\right\rangle \\
&\leq \langle \rho_x^{(abc)}, Z \otimes Y \otimes I_c\rangle(2\|g_c - I_c\|_{\mathrm{op}} + \|g_c - I_c\|_{\mathrm{op}}^2) \\
&\leq \|\Phi_x^{(ab)}\|_{F\to F}\|Z\|_F\|Y\|_F \cdot 2.5\|g_c - I_c\|_{\mathrm{op}},
\end{aligned}$$

where the first step was by the equivariance property in Lemma 6.2.6(2) for marginal $S = \{a, b, c\}$, the second step was by the decomposition above, in the third step we used the fact that $\rho_x^{(abc)}$ as well as Z, Y are positive semi-definite so we can apply spectral upper bounds $|g_c - I_c| \preceq \|g_c - I_c\|_{\mathrm{op}} I_c$ to bound the inner product, and the final step was by Definition 6.2.2 of the marginal $\rho_x^{(ab)}$, the relation $\langle \rho_x^{(ab)}, Z \otimes Y\rangle = \langle Z, \Phi_x^{(ab)}(Y)\rangle$ byProposition 2.4.5, and Definition 7.3.7 for the $F \to F$ norm.

Now consider arbitrary $Z \in H(d_a), Y \in H(d_b)$ with decomposition $Z = Z_+ - Z_-$ and $Y = Y_+ - Y_-$ where $Y_\pm, Z_\pm \succeq 0$. Then

$$\begin{aligned}
|\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y\rangle| &= |\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, (Z_+ - Z_-) \otimes (Y_+ - Y_-)\rangle| \\
&\leq 2.5\|g_c - I_c\|_{\mathrm{op}}\|\Phi_x^{(ab)}\|_{F\to F}(\|Z_+\|_F + \|Z_-\|_F)(\|Y_+\|_F + \|Y_-\|_F) \\
&\leq 5\|g_c - I_c\|_{\mathrm{op}}\|\Phi_x^{(ab)}\|_{F\to F}\|Z\|_F\|Y\|_F,
\end{aligned}$$

where the second step was by the bounds derived above for $Z_\pm, Y_\pm \succeq 0$, and the final step was because

$$(\|Z_+\|_F + \|Z_-\|_F)^2 \leq 2(\|Z_+\|_F^2 + \|Z_-\|_F^2) = 2\|Z\|_F^2,$$

by Cauchy-Schwarz and the fact that the decomposition $Z = Z_+ - Z_-$ into positive and negative parts is orthogonal for Hermitian Z. The same calculation holds for $Y = Y_+ - Y_-$. The lemma then follows as $\|\Phi_x^{(ab)}\|_{F\to F} \leq \sqrt{\|\rho_x^{(a)}\|_{\mathrm{op}}\|\rho_x^{(b)}\|_{\mathrm{op}}}$ by Lemma 7.3.10(2). □

At this point, we can collect the above bounds to show that the off-diagonal term does not change much under scaling. We will follow a similar inductive strategy as in the proof of Proposition 7.3.17.

Proposition 7.3.19. *For tensor product $V = \otimes_{a\in[m]} V_a$, consider tuple $x \in V^K$, fixed $a, b \in [m]$, and perturbation $g \in G$ such that $\delta := \sum_{c\in[m]} \|g_c - I_c\|_{\mathrm{op}} \leq \frac{1}{20}$. For scaling $x' := g \cdot x$ and any $Z \in \mathfrak{spd}(V_a), Y \in \mathfrak{spd}(V_b)$:*

$$|\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y\rangle| \leq \left(7.5\delta\sqrt{\|\rho_x^{(a)}\|_{\mathrm{op}}\|\rho_x^{(b)}\|_{\mathrm{op}}}\right)\|Z\|_F\|Y\|_F.$$

As a consequence, if x is ε-G-balanced and satisfies the λ-$\mathfrak{p}_{ab}$-spectral condition, then scaling x' satisfies the λ'-$\mathfrak{p}_{ab}$-spectral condition with

$$\lambda' \leq \lambda + 7.5\delta \cdot s(x)(1+\varepsilon).$$

Proof. We treat the perturbation as the composition of m perturbations

$$x_0 := x, \quad \to \quad x_1 := (g_1 \otimes I_{\overline{1}}) \cdot x_0, \quad \to \quad ... \to \quad x_m := (g_m \otimes I_{\overline{m}}) \cdot x_{m-1} = x'.$$

To show the first statement, for arbitrary $Z \in \mathfrak{spd}(d_a), Y \in \mathfrak{spd}(d_b)$, we bound

$$|\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Z \otimes Y\rangle| \leq \sum_{c=1}^{m} |\langle \rho_{x_c}^{(ab)} - \rho_{x_{c-1}}^{(ab)}, Z \otimes Y\rangle| \tag{7.13}$$

$$\leq \sum_{c=1}^{m} 5\|g_c - I_c\|_{\mathrm{op}}\sqrt{\|\rho_{x_{c-1}}^{(a)}\|_{\mathrm{op}}\|\rho_{x_{c-1}}^{(b)}\|_{\mathrm{op}}}\|Z\|_F\|Y\|_F \tag{7.14}$$

where the first step is by the triangle inequality applied to the telescoping sum, and in the second step we apply Lemma 7.3.13 for $c \in \{a, b\}$ and Lemma 7.3.18 for the rest. It can further be shown that

$$\|\rho_{x_c}^{(a)}\|_{\mathrm{op}} \leq \|\rho_{x_{c-1}}^{(a)}\|_{\mathrm{op}} + \|\rho_{x_c}^{(a)} - \rho_{x_{c-1}}^{(a)}\|_{\mathrm{op}} \leq \|\rho_{x_{c-1}}^{(a)}\|_{\mathrm{op}}(1 + 2.5\|g_c - I_c\|_{\mathrm{op}}),$$

where the first step was by the triangle inequality, and the final step was by the perturbation bounds in Lemma 7.3.15 and Lemma 7.3.16. This implies $\|\rho_{x_c}^{(a)}\|_{\mathrm{op}} \leq \|\rho_x^{(a)}\|_{\mathrm{op}} \prod_{j=1}^{c}(1 + 2.5\|g_j - I_j\|_{\mathrm{op}})$, and same for $\|\rho_{x_c}^{(b)}\|_{\mathrm{op}}$. Therefore, we can substitute this bound into

Eq. (7.13) to show

$$
\begin{aligned}
|\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_{x}, Z \otimes Y\rangle| &\le \|Z\|_F \|Y\|_F \sum_{c=1}^{m} 5\|g_c - I_c\|_{\mathrm{op}} \sqrt{\|\rho^{(a)}_{x_{c-1}}\|_{\mathrm{op}} \|\rho^{(b)}_{x_{c-1}}\|_{\mathrm{op}}} \\
&\le \sqrt{\|\rho^{(a)}_{x}\|_{\mathrm{op}} \|\rho^{(b)}_{x}\|_{\mathrm{op}}}\, \|Z\|_F \|Y\|_F \sum_{c=1}^{m} 5\|g_c - I_c\|_{\mathrm{op}} \prod_{j=1}^{c-1} (1 + 5\|g_j - I_j\|_{\mathrm{op}}) \\
&= \sqrt{\|\rho^{(a)}_{x}\|_{\mathrm{op}} \|\rho^{(b)}_{x}\|_{\mathrm{op}}}\, \|Z\|_F \|Y\|_F \left(\prod_{c=1}^{m} (1 + 5\|g_c - I_c\|_{\mathrm{op}}) - 1 \right) \\
&\le 7.5 \sqrt{\|\rho^{(a)}_{x}\|_{\mathrm{op}} \|\rho^{(b)}_{x}\|_{\mathrm{op}}}\, \|Z\|_F \|Y\|_F \sum_{c=1}^{m} \|g_c - I_c\|_{\mathrm{op}},
\end{aligned}
$$

where the first two steps were shown above, the third step was by the formula $1 + \sum_{i=1}^{m} a_i \prod_{j<i} (1 + a_j) = \prod_{i=1}^{m} (1 + a_i)$ which can be shown by a simple induction, in the fourth step we used the bound $1 + x \le e^x$, and the final step was by Taylor approximation $e^x \le 1 + \frac{3}{2}x$ for $0 \le x \le \frac{1}{4}$ which is satisfied by our assumption $\delta = \sum_{c \in [m]} \|g_c - I_c\|_{\mathrm{op}} \le \frac{1}{15}$.

To show the second statement in the proposition, we recall Definition 7.1.9 showing that the λ'-$\mathfrak{p}_{ab}$-spectral condition is equivalent to

$$
\sup_{Z \in \mathfrak{spd}(V_a), Y \in \mathfrak{spd}(V_b)} \frac{\langle \rho^{(ab)}_{x'}, Z \otimes Y \rangle}{\|Z\|_F \|Y\|_F} \le \frac{\lambda'}{\sqrt{d_a d_b}}.
$$

Therefore the statement follows as

$$
\begin{aligned}
\frac{\langle \rho^{(ab)}_{x'}, Z \otimes Y \rangle}{\|Z\|_F \|Y\|_F} &\le \frac{|\langle \rho^{(ab)}_{x}, Z \otimes Y \rangle|}{\|Z\|_F \|Y\|_F} + \frac{|\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_{x}, Z \otimes Y \rangle|}{\|Z\|_F \|Y\|_F} \\
&\le \frac{\lambda}{\sqrt{d_a d_b}} + 7.5\delta \cdot \sqrt{\|\rho^{(a)}_{x}\|_{\mathrm{op}} \|\rho^{(b)}_{x}\|_{\mathrm{op}}} \le \frac{\lambda + 7.5\delta \cdot s(x)(1+\varepsilon)}{\sqrt{d_a d_b}},
\end{aligned}
$$

where in the second step we bounded the first term by the fact that x satisfies the λ-spectral condition and the second term by the bound on the difference $\rho^{(ab)}_{x'} - \rho^{(ab)}_{x}$ in terms of $\delta := \sum_{c \in [m]} \|g_c - I_c\|_{\mathrm{op}}$ derived above, and in the last step we used $d_a \|\rho^{(a)}_{x}\|_{\mathrm{op}} \le s(x)(1+\varepsilon)$ for every $a \in [m]$ by the ε-G-balance condition. □

Finally, we can combine the bounds on each individual term to show robustness of strong convexity.

Proof of Theorem 7.3.14. We follow the plan outlined above by bounding the difference of each term in the decomposition in Eq. (7.4):

$$\begin{aligned}\left|\left\langle \rho_{x'}-\rho_x, \Big(\sum_{a\in[m]} Z_a\otimes I_{\overline{a}}\Big)^2\right\rangle\right| &\le \sum_{a\in[m]} |\langle \rho_{x'}^{(a)}-\rho_x^{(a)}, Z_a^2\rangle| + \sum_{a\neq b\in[m]} |\langle \rho_{x'}^{(ab)}-\rho_x^{(ab)}, Z_a\otimes Z_b\rangle| \\ &\le 4\delta \sum_{a\in[m]} \|\rho_x^{(a)}\|_{\mathrm{op}}\|Z_a\|_F^2 + 7.5\delta \sum_{a\neq b\in[m]} \sqrt{\|\rho_x^{(a)}\|_{\mathrm{op}}\|\rho_x^{(b)}\|_{\mathrm{op}}}\|Z_a\|_F\|Z_b\|_F \\ &\le s(x)(1+\varepsilon)\cdot\delta\left(4\sum_{a\in[m]}\frac{\|Z_a\|_F^2}{d_a} + 7.5\sum_{a\neq b}\frac{\|Z_a\|_F\|Z_b\|_F}{\sqrt{d_a d_b}}\right) \\ &\le s(x)(1+\varepsilon)\delta\cdot(4+7.5(m-1))\|Z\|_{\mathfrak{p}}^2,\end{aligned}$$

where the first step was by the decomposition in Eq. (7.4), in the second step we bounded the diagonal terms by Proposition 7.3.17 and the off-diagonal terms by Proposition 7.3.19, in the third step we used the bound $d_a\|\rho_x^{(a)}\|_{\mathrm{op}} \le s(x)(1+\varepsilon)$ for every $a\in[m]$ by the ε-G-balance condition, and the final step was by the bound

$$\sum_{a\neq b}\frac{\|Z_a\|_F\|Z_b\|_F}{\sqrt{d_a d_b}} = \left(\sum_{a\in[m]}\frac{\|Z_a\|_F}{\sqrt{d_a}}\right)^2 - \left(\sum_{a\in[m]}\frac{\|Z_a\|_F^2}{d_a}\right) \le m\left(\sum_{a\in[m]}\frac{\|Z_a\|_F^2}{d_a}\right) - \left(\sum_{a\in[m]}\frac{\|Z_a\|_F^2}{d_a}\right)$$

by Cauchy-Schwarz, which exactly matches $(m-1)\|Z\|_{\mathfrak{p}}^2$ by Definition 7.1.2.

Combining this with the initial strong convexity of x, for arbitrary $Z\in\mathfrak{p}$, we get

$$\begin{aligned}\partial_{\eta=0}^2 f_{x'}^P(e^{\eta Z}) = \langle \rho_{x'}, Z^2\rangle &\ge \langle \rho_x, Z^2\rangle - |\langle \rho_{x'}-\rho_x, Z^2\rangle| \\ &\ge \Big(\alpha - (4+7.5(m-1))\delta\cdot s(x)(1+\varepsilon)\Big)\|Z\|_{\mathfrak{p}}^2,\end{aligned}$$

where the first step was by Eq. (6.3), and in the final step we lower bounded the first term by α-strong convexity of x and upper bounded the difference by the calculation above. Since $Z\in\mathfrak{p}$ was arbitrary, this verifies Definition 7.1.7 of strong convexity for x'. □

Remark 7.3.20. *Surprisingly, all of the proofs in Section 7.3.3 and Section 7.3.4 go through just as well for Schatten norms $S_p \to S_p$ operator norm bound instead of $F\to F$. This suggests that such a robustness theorem may be useful for future analyses of tensor scaling using different forms of the spectral condition.*

This robustness result will be key to our algorithmic guarantees in Chapter 8, as we will be able to show a strongly convex region around the optimizer $p_* \in P$ of the Kempf-Ness function f_x^P for sufficiently strongly convex input x.

As an illustration, we combine the robustness result shown above with the convergence analysis of strongly convex inputs Theorem 7.2.16 to show that the tensor scaling solution is also strongly convex. This will be helpful in Chapter 9 to show algorithmic guarantees for the tensor normal model, as our convergence analysis of the Flip-Flop algorithm in Theorem 8.4.8 relies on strong convexity of the optimizer.

Theorem 7.3.21. *For $m \geq 3$, let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with $\dim(V_a) = d_a$ for each $a \in [m]$ with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ of size $s(x) = 1$ is ε-G-balanced and α-$\mathfrak{p}$-strongly convex according to Definition 7.1.7 with $\frac{\alpha^2}{\sqrt{e}} \geq 30m^2 \cdot \varepsilon\sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}$, then there is a scaling $x_* = p_*^{1/2} \cdot x = e^{Z_*/2} \cdot x$ that satisfies:*

1. *x_* is a G-balanced tensor;*
2. *$\|Z_*\|_{\mathfrak{p}} \leq \frac{\varepsilon\sqrt{m}}{\alpha/\sqrt{e}}$ and $\|Z_*^{(a)}\|_{\mathrm{op}} \leq \frac{3\varepsilon\sqrt{md_a}^{1-\frac{\alpha/\sqrt{e}}{m}}}{2\alpha/\sqrt{e}}$ for each $a \in [m]$;*
3. *The size is lower bounded by $s(x_*) \geq 1 - \frac{m\varepsilon^2}{2\alpha/\sqrt{e}}$;*
4. *x_* is $\alpha_* \geq \frac{\alpha}{\sqrt{e}}$-$\mathfrak{p}$-strongly convex.*

Proof. Note that the condition is stronger than that of Theorem 7.2.16 as $\alpha \leq s(x) = 1$ by Proposition A.5.2, so $\frac{\alpha}{\sqrt{e}} \geq \frac{\alpha^2}{\sqrt{e}} \geq 6m \cdot \varepsilon\sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}$. Therefore, the first three items follow exactly from the conclusions of Theorem 7.2.16.

To show item (4), we show that $x_* = e^{Z_*/2} \cdot x$ is a small perturbation of x so that we can apply the robustness result of Theorem 7.3.14. We first bound the operator norm of the scaling $e^{Z_*/2}$ as

$$\delta := \sum_{a\in[m]} \|e^{Z_*^{(a)}/2} - I_a\|_{\mathrm{op}} \leq \sum_{a\in[m]} \|Z_*^{(a)}\|_{\mathrm{op}} \leq m \cdot \frac{3\varepsilon\sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}}{2\alpha/\sqrt{e}} \leq \frac{\alpha}{20m},$$

where the second step was by Taylor approximation $|e^z - 1| \leq 2|z|$ for $|z| \leq \frac{1}{2}$, in the third step we used the bound on $\|Z_*^{(a)}\|_{\mathrm{op}}$ given in item (2), and the final step was by the assumption $\frac{\alpha^2}{\sqrt{e}} \geq 30m^2 \cdot \varepsilon \sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}}$. By Theorem 7.3.14, this implies that $x_* = e^{Z_*/2} \cdot x$ is α_*-$\mathfrak{p}$-strongly convex with

$$\alpha_* \geq \alpha - (4 + 7.5(m-1))\delta \cdot s(x)(1+\varepsilon) \geq \alpha - 8m \cdot \delta \geq \frac{\alpha}{\sqrt{e}},$$

where the first step was by the robustness result of Theorem 7.3.14 applied to α-$\mathfrak{p}$-strongly convex input x, in the second step we used the fact that $s(x) = 1$ and $\varepsilon \leq \frac{1}{m^2}$, and in the final step we substituted in the bound $\delta \leq \frac{\alpha}{20m}$ calculated above. □

7.4 Relation between Strong Convexity and Pseudorandomness

In this section, we will prove that the pseudorandom condition in Definition 7.2.17 for $\mathfrak{p}_{a\leftarrow b}$ and $\mathfrak{p}_{b\leftarrow a}$ implies the spectral condition for $\mathfrak{p}_{ab}$. At the end of this section, we use this to show that an input satisfying the pseudorandom assumptions of Theorem 7.2.26 produces a tensor scaling solution which is strongly convex, which will be useful for our algorithmic results for the tensor normal model in Chapter 9.

Once again, we will fix vector space $V = \otimes_{a\in[m]} V_a$ with $\dim(V_a) = d_a$ for each $a \in [m]$ and $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$, as this is the only scaling group we use in Chapter 9. The case of arbitrary scaling groups can be thought of as a (subgroup) restriction of this group G, and the following proofs can be extended straightforwardly.

Our main tool will be the following interpolation inequality for Schatten norms which we repeat from the preliminaries.

Theorem 7.4.1 (Corollary 3.1 of [61])**.** *Let $\Phi : \mathrm{H}(m) \to \mathrm{H}(n)$ or $\Phi : \mathrm{S}(m) \to \mathrm{S}(n)$ be a linear operator between two spaces of self-adjoint operators such that, for given $p, q \in [1, \infty]$, the operator norms induced by the Schatten norms (Definition 2.1.16) $\|\Phi\|_{p\to p}, \|\Phi\|_{q\to q}$ are bounded. For any $\theta \in [0,1]$ and p_θ defined by $\frac{1}{p_\theta} := \frac{1-\theta}{p} + \frac{\theta}{q}$, Φ satisfies*

$$\|\Phi\|_{p_\theta \to p_\theta} \leq \|\Phi\|_{p\to p}^{1-\theta} \|\Phi\|_{q\to q}^{\theta}.$$

We suspect that there is a more direct way to relate pseudorandomness and strong convexity. We take this slightly more convoluted route because the interpolation technique gives a family of bounds on a given pseudorandom input (parametrized by the Schatten-p-norm) which we believe could be useful for future analyses of tensor scaling.

Our plan is to view both the spectral and pseudorandom conditions as particular induced norms of the map Φ_x associated to a tensor x according to Proposition 2.4.5, and then use Theorem 7.4.1 to interpolate between them. The operator we consider is defined as follows.

Definition 7.4.2. *For tensor product $V = \otimes_{a\in[m]} V_a$ of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, let $Q_{I_a^\perp} : \mathrm{H}(V_a) \to \mathrm{H}(V_a)$ be the orthogonal projection onto the subspace orthogonal to I_a. For input $x \in V^K$ and fixed $a \neq b \in [m]$, the off-diagonal operator $M_x^{a\leftarrow b} : H(V_b) \to H(V_a)$ is defined as $M_x^{a\leftarrow b} := Q_{I_a^\perp} \circ \Phi_x^{(ab)} \circ Q_{I_b^\perp}$. Note that for $Z \in H(V_a)$ and $Y \in H(V_b)$,*

$$\langle Z, M_x^{a\leftarrow b}(Y)\rangle = \langle Q_{I_a^\perp}(Z), \Phi_x^{(ab)}(Q_{I_b^\perp}(Y))\rangle = \langle \Phi_x^{(ba)}(Q_{I_a^\perp}(Z)), Q_{I_b^\perp}(Y)\rangle = \langle M_x^{b\leftarrow a}(Z), Y\rangle,$$

i.e. $M_x^{a\leftarrow b}$ and $M_x^{b\leftarrow a}$ are adjoint linear maps (see Section 2.1.2).

The off-diagonal operator defined above can be thought of as the map $\Phi^{(ab)}$ restricted to the traceless subspaces $\mathfrak{spd}(V_a) \subseteq \mathrm{H}(V_a)$ and $\mathfrak{spd}(V_b) \subseteq \mathrm{H}(V_b)$ given in Definition 2.1.10. Therefore we expect its norm to be related to the spectral condition. This is formalized below.

Lemma 7.4.3. *Let $V = \otimes_{a\in[m]} V_a$ with scaling group $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$ and associated polar and infinitesimal vector space $(P, \mathfrak{p})$ according to Definition 6.2.3. Then input $x \in V^K$ satisfies the λ-$\mathfrak{p}_{ab}$-spectral condition according to Definition 7.1.9 iff*

$$\|M_x^{a\leftarrow b}\|_{F\to F} = \|M_x^{b\leftarrow a}\|_{F\to F} \leq \frac{\lambda}{\sqrt{d_a d_b}}.$$

Proof. The equality holds because adjoints $(M_x^{a\leftarrow b})^* = M_x^{b\leftarrow a}$ are adjoints and therefore have the same induced operator norm, which is the $F \to F$ norm by Proposition 7.3.9(1).

We will show $\|M_x^{a\leftarrow b}\|_{F\to F} = \|\Phi_x^{(ab)}\|_0$ from which the inequality follows by Lemma 7.3.10(1). By Definition 7.3.7 of the $F \to F$ norm, we can rewrite

$$\|M_x^{a\leftarrow b}\|_{F\to F} = \sup_{Z\in H(V_a), Y\in H(V_b)} \frac{\langle Z, M_x^{a\leftarrow b}(Y)\rangle}{\|Z\|_F \|Y\|_F} = \sup_{Z\in H(V_a), Y\in H(V_b)} \frac{\langle Q_{I_a^\perp}(Z), \Phi_x^{(ab)}(Q_{I_b^\perp}(Y))\rangle}{\|Z\|_F \|Y\|_F},$$

where in the last step we substituted Definition 7.4.2 of the off diagonal operator. We have $Q_{I_a^\perp}(Z) = Z$ for $Z \in \mathfrak{spd}(V_a)$, and similarly $Q_{I_b^\perp}(Y) = Y$ for $Y \in \mathfrak{spd}(V_b)$. Since $\mathfrak{spd}(V_a) \subseteq H(V_a)$, we have $\|M_x^{a\leftarrow b}\|_{F\to F} \geq \|\Phi_x^{(ab)}\|_0$.

To show the reverse inequality, we write

$$\begin{aligned}
\|M_x^{a\leftarrow b}\|_{F\to F} &= \sup_{Z\in H(V_a), Y\in H(V_b)} \frac{\langle Q_{I_a^\perp}(Z), \Phi_x^{(ab)}(Q_{I_b^\perp}(Y))\rangle}{\|Z\|_F\|Y\|_F} \\
&= \sup_{Z'\in\mathfrak{spd}(V_a), Q_{I_a^\perp}(Z)=Z';} \sup_{Y'\in\mathfrak{spd}(V_b), Q_{I_b^\perp}(Y)=Y';} \frac{\langle Z', \Phi_x^{(ab)}(Y')\rangle}{\|Z\|_F\|Y\|_F} \\
&\leq \sup_{Z'\in\mathfrak{spd}(V_a), Y'\in\mathfrak{spd}(V_b)} \frac{\langle Z', \Phi_x^{(ab)}(Y')\rangle}{\|Z'\|_F\|Y'\|_F} = \|\Phi_x^{(ab)}\|_0,
\end{aligned}$$

where the second step was by the change of variable $Q_{I_a^\perp}(Z) = Z'$ and $Q_{I_b^\perp}(Y) = Y'$, in the third step we bounded $\|Z'\|_F = \|Q_{I_a^\perp}(Z)\|_F \leq \|Z\|_F$ as $Q_{I_a}^\perp$ is an orthogonal projection and similarly $\|Y'\|_F \leq \|Y\|_F$, and the final step was by Definition 7.3.7 of $\|\cdot\|_0$. Therefore, we have $\|M_x^{a\leftarrow b}\|_{F\to F} = \|\Phi_x^{(ab)}\|_0$, and the inequality follows by Lemma 7.3.10(1) for x satisfying the λ-$\mathfrak{p}_{ab}$-spectral condition. □

Next we will define a condition related to pseudorandomness that looks more like an induced norm. In the subsequent lemma, we make this connection formal.

Definition 7.4.4. *$x \in V^K$ is a λ-$(\mathfrak{p}_{a\leftarrow b}, \infty)$-expander if*

$$\sup_{Z\in\mathfrak{spd}(V_b)} \frac{\|\Phi_x^{(ab)}(Z)\|_{\text{op}}}{\|Z\|_{\text{op}}} = \sup_{\xi\in\mathcal{S}_a} \sup_{Z\in\mathfrak{spd}(V_b)} \frac{|\langle \xi\xi^*, \Phi_x^{(ab)}(Z)\rangle|}{\|Z\|_{\text{op}}} = \sup_{\xi\in\mathcal{S}_a} \sup_{Z\in\mathfrak{spd}(V_b)} \frac{|\langle \rho_x^{(ab)}, \xi\xi^*\otimes Z\rangle|}{\|Z\|_{\text{op}}} \leq \frac{\lambda}{d_a}.$$

It is a λ-$(\mathfrak{p}, \infty)$-expander if the above holds for every $a \neq b \in [m]$.

We could drop the absolute value in the definition as the optimizer can be assumed to be positive by switching signs. Note that just like Definition 7.2.17, this condition is not symmetric, i.e. $(\mathfrak{p}_{a\leftarrow b}, \infty)$-expansion differs from $(\mathfrak{p}_{b\leftarrow a}, \infty)$-expansion.

Looking back at Lemma 7.2.19, we can see that this lemma is really showing a bound on this $(\mathfrak{p}_{a\leftarrow b}, \infty)$-expansion condition when the input is pseudorandom. Further, as we discussed in Section 7.2.3, this was the main consequence of pseudorandomness that we use in our analysis to show fast convergence. In the following lemma, we show that this property is equivalent to pseudorandomness for nearly balanced tensors (up to ε factors).

Lemma 7.4.5. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with dimension* $\dim(V_a) = d_a$ *for each $a \in [m]$, and consider scaling group* $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$

along with polar $(P, \mathfrak{p})$ according to Definition 6.2.3. If $x \in V^K$ is ε-G-balanced (Definition 6.2.4) and satisfies the γ-$\mathfrak{p}_{a \leftarrow b}$-pseudorandom condition according to Definition 7.2.17, then x is a λ-$(\mathfrak{p}_{a \leftarrow b}, \infty)$-expander according to Definition 7.4.4 with $\lambda \leq s(x)(1+\varepsilon) - e^{-\gamma}$.

Conversely, if x is ε-G-balanced and a λ-$(\mathfrak{p}_{a \leftarrow b}, \infty)$-expander, then it satisfies the γ-$\mathfrak{p}_{a \leftarrow b}$-pseudorandom condition for $e^{-\gamma} \geq s(x)(1-\varepsilon) - \lambda$.

Proof. To show the first direction, consider fixed $\xi \in \mathcal{S}_a$ and note

$$\sup_{Z_b \in \mathfrak{spd}(V_b)} \frac{\langle \rho_x^{(ab)}, \xi\xi^* \otimes Z_b \rangle}{\|Z_b\|_{\mathrm{op}}} = \sup_{P \in \mathcal{P}_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes (I_b - 2P) \rangle \leq \frac{s(x)(1+\varepsilon) - e^{-\gamma}}{d_a},$$

where the first step was by Fact 2.6.4 which characterizes the vertices of $\{Z \in \mathfrak{spd}(V_b) \mid \|Z\|_{\mathrm{op}} \leq 1\}$ along with the bound $\|I_b - 2P\|_{\mathrm{op}} = 1$ for $P \in \mathcal{P}_b$, and in the final step we used $\langle \rho_x^{(ab)}, \xi\xi^* \otimes I_b \rangle = \langle \rho_x^{(a)}, \xi\xi^* \rangle$ by Definition 6.2.2 of marginals as well as the ε-G-balance condition to upper bound the first term, and Definition 7.2.17 of pseudorandomness to lower bound the second term. Since $\xi \in \mathcal{S}_a$ was arbitrary, this verifies Definition 7.4.4 of $(\mathfrak{p}_{a \leftarrow b}, \infty)$-expansion.

To show the reverse direction, consider arbitrary $\xi \in \mathcal{S}_a$ and note that the λ-$(\mathfrak{p}_{a \leftarrow b}, \infty)$-expansion condition gives a bound

$$\begin{aligned} \frac{\lambda}{d_a} &\geq \sup_{Z \in \mathfrak{spd}(V_b)} \frac{\langle \rho_x^{(ab)}, \xi\xi^* \otimes Z \rangle}{\|Z\|_{\mathrm{op}}} = \sup_{P \in \mathcal{P}_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes (I_b - 2P) \rangle \\ &= \langle \rho_x^{(a)}, \xi\xi^* \rangle - 2 \inf_{P \in \mathcal{P}_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes P \rangle, \end{aligned}$$

where the first step was by Definition 7.4.4 of the ∞-expansion condition, in the second step we used Fact 2.6.4 which characterizes the vertices of $\{Z \in \mathfrak{spd}(V_b) \mid \|Z\|_{\mathrm{op}} \leq 1\}$ in terms of $P \in \mathcal{P}_b$, and in the final step we used Definition 6.2.2 of marginals. Finally, we can use the fact that $s(x) = 1$ and x is ε-G-balanced in order to bound $d_a \langle \rho_x^{(a)}, \xi\xi^* \rangle \geq s(x)(1-\varepsilon)$ for arbitrary $\xi \in \mathcal{S}_a$. So in total we can rearrange to show

$$\inf_{\xi \in \mathcal{S}_a} \inf_{P \in \mathcal{P}_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes P \rangle \geq \frac{s(x)(1-\varepsilon) - \lambda}{2d_a},$$

which verifies Definition 7.2.17 of pseudorandomness. □

Finally, we can connect this expansion condition to the off-diagonal operator in Definition 7.4.2. This will allow us to use interpolation to bound the spectral condition in Definition 7.1.9 in terms of the ∞-expansion condition.

Lemma 7.4.6. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces of dimension* $\dim(V_a) = d_a$ *for each $a \in [m]$, and consider scaling group $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$ with polar $(P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ satisfies the λ-$(\mathfrak{p}_{a\leftarrow b}, \infty)$-expansion condition, then*

$$\|M_x^{a\leftarrow b}\|_{\mathrm{op}\to\mathrm{op}} = \|M_x^{b\leftarrow a}\|_{1\to 1} \le 4\frac{\lambda}{d_a},$$

where $\|\cdot\|_1$ denotes the Schatten-1-norm according to Definition 2.1.16 and $M_x^{a\leftarrow b}$ is the off-diagonal operator given in Definition 7.4.2.

Proof. The equality follows since $M_x^{a\leftarrow b}, M_x^{b\leftarrow a}$ are adjoints so

$$\begin{aligned}\|M_x^{a\leftarrow b}\|_{\mathrm{op}\to\mathrm{op}} &= \sup_{Y\in L(V_b)} \frac{\|M_x^{a\leftarrow b}(Y)\|_{\mathrm{op}}}{\|Y\|_{\mathrm{op}}} = \sup_{Y\in L(V_b)} \sup_{Z\in L(V_a)} \frac{\langle Z, M_x^{a\leftarrow b}(Y)\rangle}{\|Z\|_1\|Y\|_{\mathrm{op}}} \\ &= \sup_{Z\in L(V_a)} \sup_{Y\in L(V_b)} \frac{\langle M_x^{b\leftarrow a}(Z), Y\rangle}{\|Z\|_1\|Y\|_{\mathrm{op}}} = \sup_{Z\in L(V_a)} \frac{\|M_x^{b\leftarrow a}(Z)\|_1}{\|Z\|_1} = \|M_x^{b\leftarrow a}\|_{1\to 1},\end{aligned}$$

where in the first and last steps we considered the operator norms induced by the Schatten-∞-norm and Schatten-1-norm respectively, the second and fourth step was by the dual characterization of Proposition 2.1.17, and the third step was by the observation in Definition 7.4.2 showing $M_x^{a\leftarrow b}, M_x^{b\leftarrow a}$ are adjoints. To show the inequality, we will use the following claim.

Claim 7.4.7. *For any $a \in [m]$, $\|Q_{I_a^\perp}\|_{\mathrm{op}\to\mathrm{op}} \le 2$.*

Proof. Note that for any $Z \in L(V_a)$ we have $Q_{I_a^\perp}(Z) = Z - \langle Z, I_a\rangle \frac{I_a}{d_a}$. Therefore

$$\|Q_{I_a^\perp} Z\|_{\mathrm{op}} \le \|Z\|_{\mathrm{op}} + \frac{|\langle Z, I_a\rangle|}{d_a}\|I_a\|_{\mathrm{op}} \le \|Z\|_{\mathrm{op}} + \frac{\|Z\|_{\mathrm{op}}\|I_a\|_1}{d_a} = 2\|Z\|_{\mathrm{op}},$$

where the first step was by triangle inequality, and in the second step we used Proposition 2.1.17 for $|\langle Z, I_a\rangle| \le \|Z\|_{\mathrm{op}}\|I_a\|_1$. □

Now we compute the induced norm

$$\begin{aligned}\|M_x^{a\leftarrow b}\|_{\mathrm{op}\to\mathrm{op}} &= \sup_{Y\in H(V_b)} \frac{\|M_x^{a\leftarrow b}(Y)\|_{\mathrm{op}}}{\|Y\|_{\mathrm{op}}} = \sup_{Y\in H(V_b)} \frac{\|Q_{I_a^\perp} \circ \Phi_x^{(ab)} \circ Q_{I_b^\perp}(Y)\|_{\mathrm{op}}}{\|Y\|_{\mathrm{op}}} \\ &\le \|Q_{I_a^\perp}\|_{\mathrm{op}\to\mathrm{op}}\|Q_{I_b^\perp}\|_{\mathrm{op}\to\mathrm{op}} \sup_{Y'\in\mathfrak{spd}(V_b)} \frac{\|\Phi_x^{(ab)}(Y')\|_{\mathrm{op}}}{\|Y'\|_{\mathrm{op}}} \le \frac{4\lambda}{d_a},\end{aligned}$$

where the first step is by definition of $\|\cdot\|_{\mathrm{op}\to\mathrm{op}}$, in the second step we unravel Definition 7.4.2 of $M_x^{a\leftarrow b}$, in the third step we perform a change of variable $Y' := Q_{I_b^\perp}(Y)$, and in the final step we use the the claim to bound $\|Q_{I_a^\perp}\|_{\mathrm{op}\to\mathrm{op}} \leq 2$ and $\|Q_{I_b^\perp}\|_{\mathrm{op}\to\mathrm{op}} \leq 2$ as well as Definition 7.4.4 of the ∞-expansion condition to bound the supremum. □

Putting these together, we can interpolate between induced norms to show that the ∞-expansion condition in Definition 7.4.4 implies the spectral condition in Definition 7.1.9.

Theorem 7.4.8. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, and consider scaling group $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$ with polar $(P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ is a λ-$(\mathfrak{p}_{a\to b}, \infty)$-expander and a λ-$(\mathfrak{p}_{b\to a}, \infty)$-expander, then x satisfies the 4λ-$\mathfrak{p}_{ab}$-spectral condition.*

Proof. We will apply the interpolation result of Theorem 7.4.1 to the operator $M_x^{a\leftarrow b} := Q_{I_a^\perp} \circ \Phi_x^{(ab)} \circ Q_{I_b^\perp}$. By Lemma 7.4.6 we can bound

$$\|M_x^{a\leftarrow b}\|_{\mathrm{op}\to\mathrm{op}} \leq \frac{4\lambda}{d_a} \qquad \text{and} \qquad \|M_x^{a\leftarrow b}\|_{1\to 1} \leq \frac{4\lambda}{d_b},$$

where the first and second inequalities follow from $(\mathfrak{p}_{a\leftarrow b}, \infty)$-expansion and $(\mathfrak{p}_{b\leftarrow a}, \infty)$-expansion respectively. Recalling that $\|\cdot\|_{\mathrm{op}}$ is the Schatten-∞-norm, $\|\cdot\|_F$ is the Schatten-2-norm, and we have used $\|\cdot\|_1$ to denote the Schatten-1-norm, we can apply Theorem 7.4.1 with $p = \infty, q = 1$ and $\theta = \frac{1}{2}$ so that $p_\theta = (\frac{1/2}{p} + \frac{1/2}{2q})^{-1} = (0 + \frac{1}{2})^{-1} = 2$ to bound

$$\|\Phi_x^{(ab)}\|_0 = \|M_x^{a\leftarrow b}\|_{F\to F} \leq \sqrt{\|M_x^{a\leftarrow b}\|_{\mathrm{op}\to\mathrm{op}} \|M_x^{a\leftarrow b}\|_{1\to 1}} \leq \frac{4\lambda}{\sqrt{d_a d_b}},$$

where the first step was shown in the proof of Lemma 7.4.3. The theorem follows by item (1) of Lemma 7.3.10 as this verifies Definition 7.1.9 of the spectral condition. □

Remark 7.4.9. *This interpolation technique in fact gives the following family of inequalities by choosing $\theta = \frac{1}{p}$ for $p \in [1, \infty]$ so that $p_\theta = (\frac{(p-1)/p}{\infty} + \frac{1/p}{1})^{-1} = p$:*

$$\|M_x^{a\leftarrow b}\|_{p\to p} \leq \|M_x^{a\leftarrow b}\|_{\mathrm{op}\to\mathrm{op}}^{1-1/p} \|M_x^{a\leftarrow b}\|_{1\to 1}^{1/p} \leq \frac{4\lambda}{d_a}\left(\frac{d_a}{d_b}\right)^{1/p},$$

where $M_x^{a\leftarrow b}$ is the off-diagonal operator given in Definition 7.4.2, $p \to p$ denotes the norm on $M_x^{a\leftarrow b}$ induced by the Schatten p-norm according to Definition 2.1.16, and $q = \frac{p-1}{p}$ is the conjugate exponent.

We believe this result is of independent interests and suggests future strategies to analyze tensor scaling using these different expansion conditions on $\|M_x^{a\leftarrow b}\|_{p\to p}$.

We can now collect all these results to show that pseudorandomness implies strong convexity of tensors.

Proposition 7.4.10. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ is ε-G-balanced and γ-$\mathfrak{p}$-pseudorandom, then x is α-$\mathfrak{p}$-strongly convex for*

$$\alpha \geq s(x)(1-\varepsilon) - 4(m-1)(s(x)(1-\varepsilon) - e^{-\gamma}).$$

Proof. We focus on the case $G = (\mathrm{SL}(d_1), ..., \mathrm{SL}(d_m))$, as the statement for other scaling groups follow by restricting to the appropriate diagonal entries. Our plan is to use following chain of implications: pseudorandomness $\implies$ ∞-expansion $\implies$ spectral condition $\implies$ strong convexity.

Now we carry out this plan quantitatively. First, since x is ε-G-balanced according to Definition 6.2.4 and γ-$\mathfrak{p}$-pseudorandom according to Definition 7.2.17, Lemma 7.4.5 shows that x satisfies the λ-$(\mathfrak{p}, \infty)$-expansion condition for $\lambda \leq s(x)(1-\varepsilon) - e^{-\gamma}$. Next, we can use Theorem 7.4.8 to show that this implies the 4λ-$\mathfrak{p}$-spectral condition according to Definition 7.1.9. Finally, we use Proposition 7.1.10 to show x is α-$\mathfrak{p}$-strongly convex with

$$\alpha \geq s(x)(1-\varepsilon) - (m-1)(4\lambda) \geq s(x)(1-\varepsilon) - 4(m-1)(s(x)(1-\varepsilon) - e^{-\gamma}),$$

where the first step was by Proposition 7.1.10 applied to ε-G-balanced input x satisfying the 4λ-$\mathfrak{p}$-spectral condition, and the final step was by the bound $\lambda \leq s(x)(1-\varepsilon) - e^{-\gamma}$ derived above. □

Before we move on to showing how we can apply Proposition 7.4.10 to analyze tensor scaling, we compare it to the matrix result in Section 3.4.

For illustration, consider matrix scaling with doubly balanced input $A \in \mathrm{Mat}(d, n)$ with $s(A) = 1$. This is equivalent to $\mathrm{vec}(A) \in \mathbb{F}^d \otimes \mathbb{F}^n$ that is T-balanced for scaling group $T = (\mathrm{ST}(d), \mathrm{ST}(n))$ and polar $\mathfrak{t} = \mathfrak{t}_L \oplus \mathfrak{t}_R = \mathfrak{st}_+(d) \oplus \mathfrak{st}_+(n)$. Note that both Theorem 3.4.7 and Proposition 7.4.10 use pseudorandomness to show strong convexity, though the proof of Theorem 3.4.7 is much more direct. Let compare the conditions required to show $\Omega(1)$-strong convexity of A: Theorem 3.4.7 requires A to be an $(\Omega(1), \frac{1}{16})$-pseudorandom matrix according to Definition 3.3.1, whereas Proposition 7.4.10 requires A to be $O(1)$-$\mathfrak{t}$-pseudorandom according to Definition 7.2.17. As discussed in Section 7.2.3, the $(\Omega(1), \frac{1}{16})$-pseudorandom matrix is strictly stronger than the $(\Omega(1), \frac{1}{2})$-pseudorandom matrix condition, which is equivalent to $O(1)$-$\mathfrak{t}_{L\leftarrow R}$-pseudorandomness. On the other hand, $\mathfrak{t}$-pseudorandomness is really equivalent to simultaneous $\mathfrak{t}_{L\leftarrow R}$ and $\mathfrak{t}_{R\leftarrow L}$-pseudorandomness. Therefore, the two conditions are incomparable. Theorem 3.4.7 was especially useful for our application to the Paulsen problem, as $n \gg d$ in that setting so it was easier to show pseudorandomness of one side. We can also compare the conclusions of the two theorems: Theorem 3.4.7 shows that $(e^{-\gamma}, \frac{1}{16})$-pseudorandomness implies $(e^{-11} \cdot e^{-\gamma})$-strong convexity when $\gamma \gtrsim 1$ is a large enough constant, whereas Proposition 7.4.10 shows that γ-$\mathfrak{t}$-pseudorandomness implies α-strong convexity for

$$\alpha \geq s(A) - 4(2-1)(s(A) - e^{-\gamma}) = 1 - 4(1 - e^{-\gamma}),$$

where the first step was by Proposition 7.4.10 applied to T-balanced and γ-$\mathfrak{t}$-pseudorandom x, and in the final step we substituted in $s(A) = 1$. Note that this is only non-trivial when $\gamma \lesssim 1$ is small enough, whereas Theorem 3.4.7 gives some strong convexity for arbitrarily large γ. On the other hand, the best possible conclusion of Theorem 3.4.7 gives at most e^{-12}-strong convexity, whereas Proposition 7.4.10 gives α-strong convexity for α arbitrarily close to 1 when γ is small enough. In our application in Chapter 9, we will have pseudorandomness for all pairs, and will be concerned with showing small constants to give improved sample complexity bounds, so it will be advantageous to apply Proposition 7.4.10.

In the following Chapter 8, we will study the convergence of algorithms for tensor scaling. We will mostly use tools from standard convex optimization lifted to this geodesic setting. In particular, we will show that algorithms converge quickly in the presence of strong convexity. Therefore, the results in this section will be useful in showing pseudorandom inputs also enjoy this fast convergence.

Below is an illustration of how we will connect the fast convergence analysis of Section 7.2.3 to strong convexity.

Theorem 7.4.11. *Let $V = \otimes_{a\in[m]} \mathbb{F}^{d_a}$ be a tensor product of inner product spaces with scaling group $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$ and polar $(P, \mathfrak{p})$ according to Definition 6.2.3. If input $x \in V^K$ of size $s(x) = 1$ is ε-G-balanced for $\varepsilon \leq \frac{1}{100m^2}$ and γ-$\mathfrak{p}$-pseudorandom for $\gamma \leq \frac{1}{32m}$, then there is a scaling $x_* = p_*^{1/2} \cdot x = e^{Z_*/2} \cdot x$ with $p_* \in P, Z_* \in \mathfrak{p}$ that satisfies:*

1. *x_* is a G-balanced tensor scaling solution to Definition 6.2.5;*
2. *$\max_{a\in[m]} \|Z_*^{(a)}\|_{\mathrm{op}} \leq 2\varepsilon$;*
3. *The size of the scaling solution is lower bounded by $s(x_*) \geq 1 - m\varepsilon^2$;*
4. *x_* is $\alpha_* \geq \frac{1}{2}$-$\mathfrak{p}$-strongly convex.*

Proof. The first three items are exactly the content of Theorem 7.2.26 with scaling group $G = (\mathrm{SL}(V_1), ..., \mathrm{SL}(V_m))$.

For the fourth item, we first note that $x_* = e^{Z_*/2} \cdot x$ is a small scaling of x, and in particular, $\lambda_{\min}(e^{Z_*}) = e^{\lambda_{\min}(Z_*)} \geq e^{-\|Z_*\|_{\mathrm{op}}}$ as $Z_* \in \mathfrak{p}$ is Hermitian. Since x is γ-$\mathfrak{p}$-pseudorandom, we can use Theorem 7.3.4 to show that x_* is γ'-$\mathfrak{p}$-pseudorandom with $\gamma' \leq \gamma + \|Z_*\|_{\mathrm{op}}$. Now, we can apply Proposition 7.4.10 to show that

$$\alpha_* \geq s(x_*) - 4(m-1)(s(x_*) - e^{-\gamma'}) \geq 1 - m\varepsilon^2 - 4(m-1)(1 - e^{-\gamma - \|Z_*\|_{\mathrm{op}}}),$$

where the first step was by applying Proposition 7.4.10 to G-balanced and γ'-$\mathfrak{p}$-pseudorandom x_*, and in the second step we substituted $1 = s(x) \geq s(x_*) \geq 1 - m\varepsilon^2$ by item (3) of this theorem and $\gamma' \leq \gamma + \|Z_*\|_{\mathrm{op}}$ as derived above. Now, we can bound the scaling by

$$\|Z_*\|_{\mathrm{op}} = \sum_{a \in [m]} \|Z_*^{(a)}\|_{\mathrm{op}} \leq 2m \cdot \varepsilon \leq \frac{1}{50m},$$

where the first step was by Definition 7.1.12 of $\| \cdot \|_{\mathrm{op}}$ for $\mathfrak{p}$, and the last step was by our assumption $\varepsilon \leq \frac{1}{100m^2}$. Therefore, plugging in our assumptions to give

$$\alpha_* \geq 1 - m\varepsilon^2 - 4(m-1)(1 - e^{-\gamma - \|Z_*\|_{\mathrm{op}}}) \geq 1 - m \cdot \frac{1}{100m^2} - 4m \cdot 2\Big(\frac{1}{32m} + \frac{1}{50m}\Big) \geq \frac{1}{2},$$

where the first step was shown above, in the second step we applied the Taylor approximation $1 - e^{-x} \leq 2x$ for $0 \leq x \leq \frac{1}{2}$ along with the assumptions $\varepsilon \leq \frac{1}{100m^2}$, $\gamma \leq \frac{1}{32m}$, and $\|Z_*\|_{\mathrm{op}} \leq \frac{1}{50m}$ as calculated above. □

Chapter 8

Algorithms for Geodesic Convex Optimization and Scaling

In Chapter 3, we showed a convex formulation for the matrix scaling problem which allowed us to apply tools from standard convex analysis to give strong bounds on the matrix scaling solution. In the subsequent chapters, we lifted these results to frame scaling (Chapter 4) and general tensor scaling (Chapter 7). By leveraging the geodesic convexity of Proposition 6.2.18 and using the reduction in Theorem 6.3.1, we were able to reduce the analysis of these non-commutative scaling problems to their simpler commutative counterparts, which we could then approach using standard convex analysis.

However, the reduction in Theorem 6.3.1 is non-constructive, so one drawback is that our analysis only provides existential results for the scaling solution. In this chapter, we will be able to make these results algorithmic by showing exponential convergence for many natural tensor scaling algorithms when the inputs satisfy a strong convexity assumption. Our techniques will be based on lifting standard convergence results for strongly convex function optimization to the geodesic setting. These results will be especially valuable for our statistical application to the tensor normal model in Chapter 9.

In Section 8.1, we present a review of the literature on algorithms for scaling. There are many important scaling problems that have each been rediscovered in a variety of communities, so we only present a small selection of the results here. In Section 8.2 we use our convex optimization perspective to re-prove the results of Linial, Samorodnitsky, and Wigderson [66] analyzing convergence of the Sinkhorn algorithm for matrix scaling. In Section 8.3, we discuss how strong convexity can help us improve the analysis of Sinkhorn scaling. These results will then be formalized and generalized in Section 8.4 to show fast convergence of the Flip-Flop algorithm for tensor scaling for strongly convex inputs. Consequently, in Section 8.5, we will make our results on the Paulsen problem algorithmic by showing fast convergence of frame scaling for the random inputs studied in Section 5.1, as well as the perturbations in Section 4.5.

8.1 Previous Work

In this section, we discuss the previous algorithms in the scaling framework. We begin with classical algorithms for matrix scaling, then discuss more and more general settings culminating in the large class of geodesic convex optimization algorithms presented and analyzed in [20].

As mentioned in Section 3.1, the original matrix scaling problem involves finding positive diagonal matrices $L, R \in \mathrm{diag}_+(d)$ to scale a non-negative input $A \in \mathbb{R}_+^{d\times d}$ to doubly stochastic ($A1_d = A^T 1_d = 1_d$). Matrix balancing is a similar problem where the requirement is to conjugate non-negative $A \in \mathbb{R}_+^{d\times d}$ by positive diagonal $X \in \mathrm{diag}_+(d)$ such that for every $i \in [d]$, the i-th row and i-th column sum of XAX^{-1} is equal. These are basic problems in numerical linear algebra that are used as subroutines for a variety of applications in mathematics and statistics, e.g. optimal transport [27], matrix preconditioning [76], and approximation of the permanent [66]. The most well-known algorithm for matrix scaling is the Sinkhorn algorithm [83], which iteratively fixes the row and column condition. A similar method for matrix balancing is known as Osborne's iteration [76], which fixes a single row/column pair in each iteration. Both of these produce solutions whose iteration complexity scales as $\mathrm{poly}(\frac{1}{\delta})$, where δ is the desired error bound. In Section 8.2, we show that Sinkhorn scaling can be viewed as a natural descent method for the convex formulation of matrix scaling presented in Proposition 3.1.10. This allows us to prove convergence of the algorithm via standard convex optimization techniques.

The alternating scaling algorithm was generalized to solve the operator scaling problem in the work of Gurvits [45]. In [38], this algorithm was shown to converge in polynomial time to decide whether an operator is scalable. Note that this analysis parallels the matrix Sinkhorn analysis, and therefore the convergence is once again $\mathrm{poly}(\frac{1}{\delta})$. This was used in [38] to give the first polynomial time algorithm for a variety of problems in algebraic complexity, including a non-commutative version of polynomial identity testing.

For matrix scaling, there are also many algorithms which require only $\mathrm{poly}\log(\frac{1}{\delta})$ many iterations to produce an δ-approximate solution. While the iterative Sinkhorn algorithm is incredibly simple to implement and only requires first order information, these results tend to use more complicated optimization procedures, such as the ellipsoid method [57] or interior point methods [21], [88]. Recently, two independent groups [26], [2] used trust-region methods as well as techniques from fast Laplacian solvers in order to give nearly linear time algorithms for matrix scaling with $\operatorname{poly}\log(\frac{1}{\delta})$ error convergence. As a side note, the algorithm of Cohen et al. [26] depends linearly on the condition number of the scaling solution $\|(X_*, Y_*)\|_\infty$. Therefore, our fast convergence analysis of Chapter 3 gives sufficient conditions for this algorithm to converge in nearly linear time.

Trust region methods were used in [3] to give the first algorithm with $\operatorname{poly}\log(\frac{1}{\delta})$ convergence for operator scaling. This gave the first known polynomial time algorithms for some more complicated versions of the polynomial identity testing problem, which were the main motivation of [45] and [38].

In a grand generalization, Bürgisser et al. [19] proposed a non-trivial extension of the alternating algorithm which was able to solve a wide class of tensor scaling problems, sometimes called scaling with prescribed marginals. The analysis was quite technical as it relied on some machinery from the representation theory of Lie groups. Further, this resulted in a runtime which was polynomial in $\frac{1}{\delta}$, but crucially also depended on the binary description of the desired marginals. This left open whether there were $\log\frac{1}{\delta}$ convergent algorithms in this significantly more general setting.

As discussed in Section 6.1.2, these scaling problems can be seen from the lens of geometric invariant theory. Therefore, there are some known algebraic algorithms [72] which rely on this invariant theory connection, but these are usually quite expensive in terms of runtime. Interestingly, the main result of [18] shows that the most general version of this scaling problem, moment polytope membership testing (described in Section 6.1.1), is in NP $\cap$ coNP. This gives some evidence for tractability even in this general setting.

As a first step towards polynomial time scaling algorithms, the work of [20] gave a foundation to unify the various (sometimes ad-hoc) techniques which were used to prove fast convergence in each individual scaling setting. Specifically, they presented a geodesic convex formulation for general scaling problems and gave quantitative analyses for a variety of optimization algorithms (see Theorem 6.1.7). In particular, they defined natural geometric quantities (the weight norm and weight margin) which depended on each scaling problem and controlled the convergence of natural geodesic convex optimization algorithms. This gave an explanation for previously known algorithms for individual scaling problems. On the other hand, in many cases of interest, these parameters only have exponential bounds which only leads to exponential time algorithms. In fact, in the simplest open case of 3-tensor scaling, the work of Kravtsov [60] shows that the exponential dependence of these geometric parameters is necessary. These obstructions have recently been generalized to higher order tensor scaling problems by Franks and Reichenbach [37].

Therefore, in order to find polynomial time algorithms for these difficult problems, new ideas are required which bypass the standard convex optimization techniques of [20]. In this thesis, we show that strong convexity is one such natural assumption which leads to beyond worst case bounds for tensor scaling. Specifically, for $m \geq 3$ tensor scaling inputs that satisfy special assumptions (strong convexity and pseudorandomness), we are able to show that even the simplest iterative algorithms have linear convergence ($\log \frac{1}{\delta}$) to high quality solutions. This suggests that a natural first step towards understanding the complexity of more general scaling problems would be to analyze inputs satisfying similar sufficient conditions for fast convergence.

8.2 Sinkhorn's Algorithm for Matrix Scaling

In this section, we present Sinkhorn's algorithm for matrix scaling [83]. This algorithm has been extensively studied for both its theoretical guarantees as well as its practical performance for applications of matrix scaling (see survey [54]). We will be studying the work of [66], where the goal was to design a deterministic approximation algorithm for the permanent of non-negative matrices. Using the framework of Chapter 3 (heavily inspired by [20]), we can re-interpret the work of [66] as a convergence analysis of Sinkhorn scaling viewed as a natural convex optimization algorithm for the Kempf-Ness function for matrix scaling. In the following Section 8.3, we will discuss how to improve these convergence results when the input is strongly convex, as studied in Section 3.2.

Recall that by Proposition 3.1.10, we have shown that for input tuple $A \in \mathrm{Mat}(d,n)^K$, the matrix scaling problem on A in Definition 3.1.3 can be equivalently solved by optimizing the convex function

$$\inf_{(X,Y)\in \mathfrak{t}} f_A(X,Y) := s(e^{X/2} A e^{Y/2}) = \sum_{k=1}^{K} \sum_{i=1}^{d} e^{X_i} |(A_k)_{ij}|^2 e^{Y_j},$$

where $\mathfrak{t}$ is given in Definition 3.1.5 and f_A is given in Definition 3.1.6.

The following algorithm for matrix scaling is quite natural, easy to implement, and performs well in practice. For this reason, it has been rediscovered and studied in a variety of fields (see survey [54]).

Definition 8.2.1 (Sinkhorn Scaling)**.** *For matrix tuple $A \in \mathrm{Mat}(d,n)^K$, the Sinkhorn scaling algorithm for matrix scaling alternates between the following operations*

$$A \leftarrow \left(\frac{d \cdot R}{\det(d \cdot R)^{1/d}}\right)^{-1/2} A, \qquad \text{and} \qquad A \leftarrow A \left(\frac{n \cdot C}{\det(n \cdot C)^{1/n}}\right)^{-1/2},$$

where $R = \mathrm{diag}\{r_i(A)\}_{i=1}^{d}$ and $C = \mathrm{diag}\{c_j(A)\}_{j=1}^{n}$ are the diagonal transformations containing the row and column sums given in Definition 3.1.1.

The iterations can be equivalently defined as an optimization method on $\mathfrak{t}$ with alternating steps

$$e^{X_{t+1}} := \left(\frac{d \cdot R}{\det(d \cdot R)^{1/d}}\right)^{-1} e^{X_t} \qquad \text{and} \qquad e^{Y_{t+1}} := e^{Y_t} \left(\frac{n \cdot C}{\det(n \cdot C)^{1/n}}\right)^{-1},$$

where the normalization by det *implies $(X_t, Y_t) \in \mathfrak{t}$ for all steps.*

Observe that the transformations produce a left-balanced and right-balanced matrix tuple in alternating iterations. In the following, we show that if the current iterate is far from doubly balanced, then the Sinkhorn scaling step makes significant progress in terms of the Kempf-Ness function.

Lemma 8.2.2. *For matrix tuple $A \in \mathrm{Mat}(d,n)^K$, let $A \to A'$ represent one iteration of Sinkhorn scaling. Then size decreases as*

$$\log s(A') \leq \log s(A) - \frac{1}{6} \begin{cases} \min\left\{\frac{\|(\nabla_A^L, 0)\|_{\mathfrak{t}}^2}{s(A)^2}, \frac{1}{d}\right\} & \text{for left normalization} \\ \min\left\{\frac{\|(0, \nabla_A^R)\|_{\mathfrak{t}}^2}{s(A)^2}, \frac{1}{n}\right\} & \text{for right normalization.} \end{cases}$$

In terms of the Kempf-Ness function, this can be written as

$$\log f_A(X_{t+1}, Y_{t+1}) - \log f_A(X_t, Y_t) \leq -\frac{1}{6} \begin{cases} \min\{\|(\nabla \log f_A(X_t, Y_t))^L\|_{\mathfrak{t}}^2, \frac{1}{d}\} & t = 0 \mod 2 \\ \min\{\|(\nabla \log f_A(X_t, Y_t))^R\|_{\mathfrak{t}}^2, \frac{1}{n}\} & t = 1 \mod 2 \end{cases}.$$

Proof. We follow the proofs of Lemma 3.1 in [66] along with the approximation given in Lemma 5.2 of [38].

The second statement on the Kempf-Ness function follows by applying the first statement on size to input A_t. This is because $f_A(X,Y) = s(e^{X/2} A e^{Y/2})$ by Definition 3.1.6 of the Kempf-Ness function, as well as the simple fact (by chain rule) that $\nabla \log f = \frac{\nabla f}{f}$, so the progress terms are the same.

So we first analyze the change in size for row-normalization step $A \to A'$:

$$\begin{aligned} s(A') = \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{K} |(A'_k)_{ij}|^2 &= \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{K} \left(\frac{d \cdot r_i(A)}{\det(d \cdot R)^{1/d}} \right)^{-1} |(A_k)_{ij}|^2 \\ &= \det(d \cdot R)^{1/d} \sum_{i=1}^{d} \frac{r_i(A)}{d \cdot r_i(A)} = \left(\prod_{i=1}^{d} d \cdot r_i(A) \right)^{1/d}, \end{aligned} \tag{8.1}$$

where the first step was by Definition 3.1.1 of size, in the second step we plugged in Definition 8.2.1 of a Sinkhorn step, and in the third step we simply used Definition 3.1.1 of the row marginal $r_i(A)$.

To bound this value in terms of ∇_A^L, we note that $\sum_{i=1}^{d} r_i(A) = s(A)$, so we can apply Claim 8.2.3 below with $x_i := \frac{d \cdot r_i(A)}{s(A)}$ to show

$$-\log \prod_{i=1}^{d} \frac{d \cdot r_i(A)}{s(A)} \geq \frac{1}{6} \min \left\{ \sum_{i=1}^{d} \left(\frac{d \cdot r_i(A)}{s(A)} - 1 \right)^2, 1 \right\} = \frac{1}{6} \min \left\{ \frac{d \|\nabla_A^L\|_{\mathfrak{t}}^2}{s(A)^2}, 1 \right\}, \tag{8.2}$$

where the first inequality is by Claim 8.2.3, and the last step is by Proposition 3.1.12 of ∇^L and Definition 3.1.11 of $\|\cdot\|_{\mathfrak{t}}$ on the left part. Combining this with the bound on size above gives

$$\log s(A') - \log s(A) = \frac{1}{d} \log \prod_{i=1}^{d} \frac{d \cdot r_i(A)}{s(A)} = \frac{1}{d} \log \prod_{i=1}^{d} \frac{d \cdot r_i(A)}{s(A)} \leq \frac{-1}{6} \min \left\{ \frac{\|\nabla_A^L\|_{\mathfrak{t}}^2}{s(A)^2}, \frac{1}{d} \right\},$$

where the first step was by Eq. (8.1), and the final step was by Eq. (8.2). The calculation for column-normalization is the same with ∇_A^L replaced by ∇_A^R and d replaced by n. □

For the proof of Lemma 8.2.2, we need the following robust version of AM-GM. We omit the proof, which is given in [38].

Claim 8.2.3 (Lemma 5.1 of [38]). *For $x \in \mathbb{R}_{++}^d$ satisfying $\sum_{i=1}^{d} x_i = d$,*

$$-\log \prod_{i=1}^{d} x_i \geq \frac{1}{6} \min \left\{ 1, \sum_{i=1}^{d} (x_i - 1)^2 \right\}.$$

The main goal of Linial et al. [66] was to give an approximation algorithm for the permanent. Therefore, they first applied a simple preprocessing step to recognize the case when the permanent was 0. They complemented this with an exponential lower bound for the permanent of nearly doubly stochastic matrices. Their strategy was to apply Sinkhorn scaling to transform any non-negative matrix to a nearly doubly stochastic one, which could be approximated effectively. The key to this algorithm was the following polynomial iteration bound for Sinkhorn scaling.

Theorem 8.2.4. *Consider matrix tuple $A \in \mathrm{Mat}(d,n)^K$ with $f^* := \inf_{(X,Y)\in\mathfrak{t}} f_A(X,Y) > -\infty$. Then for any $\delta > 0$, Sinkhorn scaling with starting point $A_0 = e^{X_0/2} A e^{Y_0/2}$ produces an iterate $(X_T, Y_T) \in \mathfrak{t}$ satisfying $\|\nabla \log f_A(X_T,Y_T)\|_{\mathfrak{t}} \leq \delta$ for some iteration*

$$T \lesssim \frac{\log f_A(X_0,Y_0) - \log f^*}{\min\{\delta^2, \frac{1}{n}\}}.$$

Proof. Assume, by applying a single Sinkhorn step if necessary, that $\nabla_A^R = 0$. This way we are alternating between left and right balanced matrices. Let T be the first time $\|\nabla \log f_A(X_T,Y_T)\|_{\mathfrak{t}} \leq \delta$. Then until this time we make significant progress:

$$\log f_A(X_T,Y_T) - \log f_A(X_0,Y_0) = \sum_{t<T} \Big(\log f_A(X_{t+1},Y_{t+1}) - \log f_A(X_t,Y_t) \Big) < \frac{-T}{6} \min\{\delta^2, \frac{1}{n}\},$$

where the final inequality was by Lemma 8.2.2 as $\|\nabla \log f_A(X_t,Y_t)\|_{\mathfrak{t}} > \delta$ for every step $t < T$. The theorem follows by applying the simple bound $f^* \leq f_A(X_T,Y_T)$ and rearranging. □

Remark 8.2.5. *In the case when $n \gg d$, the column normalization step may make much less progress due to the bottle-neck $\frac{1}{n}$ term. To remedy this, we can apply two iterations successively so our iterates are always right balanced and in each two steps we make significant progress. In particular, for $\frac{1}{n} \leq \delta^2 \leq \frac{1}{d}$, this two step algorithm allows us to replace $\frac{1}{n} \to \frac{1}{d}$ in the denominator of Theorem 8.2.4. The same observation carries over to the analogous alternating algorithms for frame and operator scaling.*

In Section 8.4, we will generalize Theorem 8.2.4 to show progress of the Flip-Flop algorithm for tensor scaling. Many of these ideas have been greatly generalized in [20], where they use use the framework of geodesic convex optimization to show convergence of simple iterative methods for more general scaling problems.

Recall that one of the main results of Theorem 3.2.19 was the stronger lower bound

$$\inf_{(X,Y)\in\mathfrak{t}} f_A(X,Y) = s(A_*) \geq 1 - O\left(\frac{\|\nabla_A\|_{\mathfrak{t}}^2}{\alpha}\right)$$

for input tuples $A \in \mathrm{Mat}(d,n)^K$ of size $s(A) = 1$ that were sufficiently close to doubly balanced and sufficiently strongly convex. Plugging this into Theorem 8.2.4 would give a strong bound on the leading constant term but the convergence of the gradient would still be at a rate of $\frac{1}{\delta^2}$. In the following section, we give an improved convergence rate of $\log(\frac{1}{\delta})$ by better utilizing strong convexity.

8.3 Algorithms for Strongly Convex Matrix Scaling

In this section, we will improve the result of Theorem 8.2.4 when the input is strongly convex by showing fast convergence of the Sinkhorn algorithm in Definition 8.2.1. We follow the analysis of [35] by using standard tools from strong convexity to effectively analyze the progress made by Sinkhorn scaling. Many of the proofs in this section are omitted or just sketched as we will cover them formally in the more general tensor scaling setting in Section 8.4.

We first present the following standard result for strongly convex optimization. This will give some intuition for how we will apply strong convexity in our matrix scaling setting.

Proposition 8.3.1. *Let V be an inner product space with function $F : V \to \mathbb{R}$ is α-strongly convex in norm $\|\cdot\|$. Consider sequence $\{x_t\}_{t\in\mathbb{N}}$ satisfying*

$$F(x_{t+1}) \leq F(x_t) - \|\nabla F(x_t)\|^2$$

for all $t \in \mathbb{N}$. This is known as descent sequence, and an algorithm that produces such iterates is known as a descent method. Then, for any $\delta \leq \|\nabla F(x_0)\|$, an element of the sequence satisfies the bound $\|\nabla F(x_T)\| \leq \delta$ for some $T \lesssim \frac{1}{\alpha}\log\frac{\|\nabla F(x_0)\|}{\delta}$.

Proof. We will show by induction that the $\|\nabla F(x_t)\|^2$ halves every $O(\frac{1}{\alpha})$ steps. Let T be the first time that $\|\nabla F(x_T)\|^2 \leq 2^{-1}\|\nabla F(x_0)\|^2$. We calculate the progress as

$$F(x_T) - F(x_0) = \sum_{t<T}\Big(F(x_{t+1}) - F(x_t)\Big) \leq -\sum_{t<T}\|\nabla F(x_t)\|^2 \leq -\frac{T}{2}\|\nabla F(x_0)\|^2,$$

where the first step was by a telescoping sum, the second step was by the assumption on descent sequence $\{x_t\}$, and the final step was by our choice of T so that $\|\nabla F(x_t)\|^2 > 2^{-1}\|\nabla F(x_0)\|^2$ for all previous steps.

By strong convexity, at any point $x \in V$ we have the following strong lower bound in terms of the gradient:

$$F^* := \inf_{y\in V} F(y) \geq F(x) - \frac{\|\nabla F(x)\|^2}{2\alpha}.$$

This can be shown by applying Lemma 2.3.7 to the univariate restriction between x and the optimizer, which is α-strongly convex by assumption.

Therefore, applying this to $x = x_0$ gives the lower bound $F(x_T) \geq F^* \geq F(x_0) - \frac{\|\nabla F(x_0)\|^2}{2\alpha}$. Combining this with the progress shown above, we get

$$\frac{-\|\nabla F(x_0)\|^2}{2\alpha} \leq F(x_T) - F(x_0) \leq \frac{-T\|\nabla F(x_0)\|^2}{2},$$

which gives $T \lesssim \frac{1}{\alpha}$ by rearranging.

Continuing this way, we define T_k to be the first time $\|\nabla F(x_{T_k})\|^2 \leq 2^{-k}\|\nabla F(x_0)\|^2$. Then for each k, we can repeat the argument above to show

$$\frac{-\|\nabla F(x_0)\|^2}{2^k \cdot 2\alpha} \leq \frac{-\|\nabla F(x_{T_k})\|^2}{2\alpha} \leq F(x_{T_{k+1}}) - F(x_{T_k}) \leq -(T_{k+1} - T_k)\frac{\|\nabla F(x_0)\|^2}{2^{k+1} \cdot 2},$$

which shows $T_{k+1} - T_k \lesssim \frac{1}{\alpha}$ by rearranging. Applying this for $k = \log(\frac{1}{\delta})$ gives the convergence bound in the proposition. □

In Section 3.2, we studied matrix scaling for strongly convex tuples $A \in \mathrm{Mat}(d,n)^K$. According to Definition 3.2.1, this only implies strong convexity of the Kempf-Ness function f_A at the origin, whereas the above Proposition 8.3.1 assumed that the function F is strongly convex everywhere. Examining the proof, we note that strong convexity was only used on the univariate restrictions between x_t and the optimizer. Therefore, in the following theorem we show how to leverage strong convexity of A to prove fast convergence of Sinkhorn scaling.

Theorem 8.3.2. *Consider matrix tuple $A \in \mathrm{Mat}(d,n)^K$ such that the Kempf-Ness function f_A from Definition 3.1.6 is α strongly convex at the optimizer (X_*, Y_*). If starting point $(X_0, Y_0) \in \mathfrak{t}$ satisfies $\|\nabla f_A(X_0, Y_0)\|_{\mathfrak{t}} \lesssim \frac{\alpha}{\sqrt{d+n}}$, then Sinkhorn scaling produces iterate $A_T := e^{X_T/2} A_* e^{Y_T/2}$ with $\|\nabla \log f_A(X_T, Y_T)\|_{\mathfrak{t}} \leq \delta$ for some $T \lesssim \frac{f_A(X_0,Y_0)}{\alpha} \log \frac{\|\nabla \log f_A(X_0,Y_0)\|_{\mathfrak{t}}}{\delta}$.*

This is a special case of the corresponding theorem for tensor scaling given in Theorem 8.4.8, and so we only sketch the proof here.

Proof Overview. In order to apply the analysis of Proposition 8.3.1, it suffices to show that for every iterate (X_t, Y_t) produced by Sinkhorn scaling, f_A is $\Omega(\alpha)$-strongly convex on the restriction to the line $(X_t, Y_t) \to (X_*, Y_*)$. By assumption, f_A is α-strongly convex at the optimizer (X_*, Y_*), and by the robustness property in Lemma 3.2.4, f_A is $\Omega(\alpha)$-strongly convex at every point in some neighborhood around (X_*, Y_*). So our goal will be to show that (X_0, Y_0) and every subsequent iterate is within this neighborhood. This is proven by using strong convexity to show that every point $(X, Y) \in \mathfrak{t}$ with small gradient must be near the optimizer (X_*, Y_*). The theorem then follows by applying the convergence analysis of Proposition 8.3.1. $\square$

In the next section, we will formalize and generalize Theorem 8.3.2 to setting of tensor scaling by lifting the above analysis of strongly convex optimization to the geodesic setting using structural properties of the Kempf-Ness function.

8.4 Algorithms for Geodesic Strongly Convex Optimization and Tensor Scaling

In this section, we lift tools from standard convex analysis to the geodesic setting to analyze natural algorithms for tensor scaling. The main contribution of this section is to show linear convergence when the input is a strongly convex tensor. These results will be applied in Section 8.5 for the Paulsen problem and Chapter 9 in order to give fast algorithmic guarantees for the tensor normal model.

We first give the appropriate generalization of Sinkhorn scaling to the tensor setting. This is a very natural iterative algorithm which performs quite well in practice. In Section 9.2.2, we discuss some background and motivate this algorithm in the context of statistical estimation.

Definition 8.4.1 (Flip-Flop Algorithm). *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, and let $(G, P, \mathfrak{p})$ be a choice of scaling group according to Definition 6.2.3. Then for input $x \in V^K$, one iteration of the Flip-Flop algorithm chooses $a := \arg\max_{b \in [m]} \|\nabla_x^{(b)}\|_{\mathfrak{p}}$, and normalizes this marginal*

$$x \leftarrow I_{\overline{a}} \otimes \begin{cases} \left(\frac{d_a \rho_x^{(a)}}{\det(d_a \rho_x^{(a)})^{1/d_a}} \right)^{-1/2} \cdot x & \text{if } G_a = \mathrm{SL}(V_a) \\ \left(\frac{d_a \operatorname{diag}^{\Xi^a}(\rho_x^{(a)})}{\det(d_a \operatorname{diag}^{\Xi^a}(\rho_x^{(a)}))^{1/d_a}} \right)^{-1/2} \cdot x & \text{if } G_a = \mathrm{ST}^{\Xi^a}(V_a) \end{cases}.$$

Applying this iteratively gives the sequence of scalings $\{g_t \in G\}_{t \geq 0}$ such that $x_t := g_t \cdot x$.

This can be equivalently defined with respect to the Kempf-Ness function f_x^P (Definition 6.2.9) by the sequence

$$p_{t+1} = \begin{cases} p_t^{1/2}\left(I_{\overline{a}} \otimes \frac{d_a \rho_x^{(a)}}{\det(d_a \rho_x^{(a)})^{1/d_a}}\right)^{-1} p_t^{1/2} & \text{if } G_a = \mathrm{SL}(V_a) \\ p_t \cdot I_{\overline{a}} \otimes \left(\frac{d_a \operatorname{diag}^{\Xi^a}(\rho_x^{(a)})}{\det(d_a \operatorname{diag}^{\Xi^a}(\rho_x^{(a)}))^{1/d_a}}\right)^{-1} & \text{if } G_a = \mathrm{ST}^{\Xi^a}(V_a) \end{cases}.$$

Note $p_t \in P$ for all steps and $(\nabla f_x^P(p_t))^{(a)} = 0$ after normalizing the a-th marginal.

Remark 8.4.2. *By Lemma 6.2.6(3), the balance condition of the a-th marginal is unaffected by unitaries, so we could as well have chosen any $h \in G_a$ with the same polar part $(h_a^* h_a)^{-1} = \frac{d_a \rho_x^{(a)}}{\det(d_a \rho_x^{(a)})^{1/d_a}}$ to normalize the a-th marginal. This choice is unimportant for the purpose of geodesic convex optimization with respect to f_x^P, as the value only depends on the polar part of the scaling $g_t^* g_t$, and we choose the positive definite square root in Definition 8.4.1 for convenience.*

In [19], the authors solve a more general scaling problem, and for their analysis it was necessary to choose scaling $h \in G_a$ such that g_t is upper triangular for each iteration. Other choices of h may be useful for the purposes of numerical stability or bit-complexity.

This natural normalization algorithm also satisfies a progress bound which generalizes the result of Lemma 8.2.2 for matrix Sinkhorn.

Proposition 8.4.3. *Let $V = \otimes_{a \in [m]} V_a$ be a tensor product of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, and let $(G, P, \mathfrak{p})$ be a choice of scaling group according to Definition 6.2.3. Then for input $x \in V^K$, if x' denotes the normalization of the a-th marginal by Definition 8.4.1 of the Flip-Flop algorithm, then*

$$\log s(x') - \log s(x) \leq -\frac{1}{6} \min\left\{ \frac{\|\nabla_x^{(a)}\|_{\mathfrak{p}}^2}{s(x)^2}, \frac{1}{d_a} \right\}.$$

This can be rewritten in terms of the Kempf-Ness function as

$$\log f_x^P(p_{t+1}) \leq \log f_x^P(p_t) - \frac{1}{6} \min\left\{ \left\| \left(\nabla \log f_x^P(p_t)\right)^{(a)} \right\|_{\mathfrak{p}}^2, \frac{1}{d_a} \right\},$$

where we have normalized the a-th marginal in step t.

Proof. Note that the second statement on the Kempf-Ness function follows by applying the first statement on size to $g_t \cdot x$. This is because $f_x^P(p) = s(p^{1/2} \cdot x)$ by Definition 6.2.9, as well as the simple fact (by chain rule) that $\nabla \log f = \frac{\nabla f}{f}$, so the progress terms are the same.

To show the progress in size, let a be marginal we are normalizing, and consider the case $G_a = \mathrm{SL}(V_a)$. The other case $G_a = \mathrm{ST}^{\Xi}(V_a)$ follows by the same calculation applied to the diag$^{\Xi}$ entries. We can write

$$s(x') = \langle I_V, \rho_{x'} \rangle = \langle I_a, \rho_{x'}^{(a)} \rangle = \left\langle \left(\frac{d_a \rho_x^{(a)}}{\det(d_a \rho_x^{(a)})^{1/d_a}} \right)^{-1}, \rho_x^{(a)} \right\rangle = \det(d_a \rho_x^{(a)})^{1/d_a}, \tag{8.3}$$

where the first step was by Definition 6.2.1 of size, in the second step we used Definition 6.2.2 of marginals, the third step was by the equivariance property for marginals shown in Lemma 6.2.6(2), and in the final step we substituted $\langle (\rho_x^{(a)})^{-1}, \rho_x^{(a)} \rangle = \mathrm{Tr}[I_a] = d_a$.

From this point, we can use the same argument as in the proof of Lemma 8.2.2 to show

$$\log s(x') - \log s(x) = \frac{1}{d_a} \log\det\left(\frac{d_a \rho_x^{(a)}}{s(x)} \right) \leq -\frac{1}{6} \min\left\{ \frac{\|\nabla_x^{(a)}\|_{\mathfrak{p}}^2}{s(x)^2}, \frac{1}{d_a} \right\},$$

where we have applied Claim 8.2.3 to the eigenvalues of $\frac{d_a \rho_x^{(a)}}{s(x)}$ in the same way as the proof of Lemma 8.2.2. □

At this point, we can show that the norm of the geodesic gradient for tensor scaling decreases under the Flip-Flop algorithm. This is a special case of the result in [19], which applies to much more general scaling problems.

Theorem 8.4.4. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, and let $(G, P, \mathfrak{p})$ be a choice of scaling group according to Definition 6.2.3. Consider input tuple $x \in V^K$ with $f^* := \inf_{p\in P} f_x^P(p) > -\infty$. Then for any $\delta > 0$, the Flip-Flop algorithm with starting point $x_0 := g_0 \cdot x$ produces $x_T := g_T \cdot x$ satisfying $\|\nabla \log f_x^P(g_T^* g_T)\|_{\mathfrak{p}} \leq \delta$ for some iteration*

$$T \lesssim \frac{\log f_x^P(g_0^* g_0) - \log f^*}{\min\{\frac{\delta^2}{m}, \frac{1}{d_{\max}}\}}.$$

Proof. The argument is exactly the same as the proof of Theorem 8.2.4 except that we apply Proposition 8.4.3 to bound the progress of Flip-Flop.

We will bound the number of iterations using the Kempf-Ness function f_x^P as a progress measure, so let $\{g_t \in G\}_{t\geq 0}$ be the iterates of the Flip-Flop algorithm according to Definition 8.4.1 with corresponding polar $\{p_t := g_t^* g_t \in P\}_{t\geq 0}$, and let T be the first time that $\|\nabla \log f_x^P(p_T)\|_{\mathfrak{p}} \leq \delta$. Then we make significant progress until this time:

$$\log f_x^P(p_T) - \log f_x^P(p_0) = \sum_{t<T} \Big(\log f_x^P(p_{t+1}) - \log f_x^P(p_t) \Big) \leq -\frac{T}{6} \min \left\{ \frac{\delta^2}{m}, \frac{1}{d_{\max}} \right\},$$

where the final inequality was by Proposition 8.4.3 as $\|\nabla \log f_x^P(p_T)\|_{\mathfrak{p}} > \delta$ for $t < T$, and $\|(\nabla f_x^P(p_t))^{(a)}\|_{\mathfrak{p}}^2 \geq \frac{1}{m}\|\nabla f_x^P(p_t)\|_{\mathfrak{p}}^2$ as we normalize the marginal with the largest error in each step. The statement follows by applying the bound $f^* \leq f_x^P(p_T)$ and rearranging. □

The main result of Section 8.3 was to show linear convergence of Sinkhorn scaling for strongly convex inputs. We next define the property of geodesic strong convexity that is required to generalize the argument in Theorem 8.3.2 to tensor scaling.

Definition 8.4.5. *Let $(G, P, \mathfrak{p})$ be a scaling group according to Definition 6.2.3 and let $F : P \to \mathbb{R}$ be a geodesically convex function with optimizer $p_* := \arg\inf_{p\in P} F(p)$. Then $p \in P$ is α-strongly convergent with respect to F if the restriction of F to the geodesic between p and p_* is α-strongly convex. Explicitly, if $Z = \log(p_*^{-1/2} p p_*^{-1/2})$ so that $\gamma_{p_*,p}(\eta) = \gamma_{p_*}(\eta Z) = p_*^{1/2} e^{\eta Z} p_*^{1/2}$ according to Fact 6.2.11, then p is α-strongly convergent if $h(\eta) := F(\gamma_{p_*,p}(\eta))$ is $\alpha\|Z\|_{\mathfrak{p}}^2$-strongly convex for $\eta \in [0, 1]$.*

Note that this property only requires strong convexity on the geodesic from p to p_*, which is much weaker than the condition that F is α-geodesically strongly convex at p according to Definition 6.2.13.

This was the key property used in Proposition 8.3.1 to show fast convergence of a descent sequence. In the following lemma, we derive the properties of strongly convergent points that will be useful for our fast convergence results for tensor scaling.

Lemma 8.4.6. *Let $(G, P, \mathfrak{p})$ be a scaling group according to Definition 6.2.3 and let $F : P \to \mathbb{R}$ be a geodesically convex function with optimizer $F(p_*) = \inf_{p\in P} F(p)$. Then for α-strongly convergent $p \in P$:*

1. *(Function): $F(p_*) \geq F(p) - \frac{\|\nabla F(p)\|_{\mathfrak{p}}^2}{2\alpha}$;*
2. *(Distance): for $Z := \log(p^{-1/2} p_* p^{-1/2})$, $\|Z\|_{\mathfrak{p}} \leq \frac{\|\nabla F(p)\|_{\mathfrak{p}}}{\alpha}$;*

where $\nabla F(p)$ is the geodesic gradient according to Definition 7.1.1.

Proof. Consider the univariate restriction $h(\eta) := F(\gamma_{p,p_*}(\eta)) = F(\gamma_p(\eta Z))$. For the function bound in item (1), our plan is to apply Lemma 2.3.7 to bound function gap between optimizer $h(1) = F(p_*)$ and $h(0) = F(p)$. For this purpose, we bound

$$|h'(0)| = |\partial_{\eta=0} F(\gamma_p(\eta Z))| = |\langle \nabla F(p), Z\rangle_{\mathfrak{p}}| \leq \|\nabla F(p)\|_{\mathfrak{p}} \|Z\|_{\mathfrak{p}},$$

where the first step was by definition of h, the second step was by Definition 7.1.1 of the geodesic gradient on $(P, \mathfrak{p})$, and the final step was by Cauchy-Schwarz for $\langle \cdot, \cdot \rangle_{\mathfrak{p}}$.

Recall that by Fact 6.2.12, the geodesics between p and p_* are related by $\gamma_{p,p_*}(\eta) = \gamma_{p_*,p}(1-\eta)$ for $\eta \in [0,1]$, and $\gamma_{p_*,p}(\eta) = \gamma_{p_*}(\eta Y)$ for some $Y \in \mathfrak{p}$ satisfying $\|Y\|_{\mathfrak{p}} = \|Z\|_{\mathfrak{p}}$. Therefore, the α-strong convergent property of p in Definition 8.4.5 implies that $h(\eta) = F(\gamma_{p,p_*}(\eta)) = F(\gamma_{p_*,p}(1-\eta))$ is $\alpha\|Z\|_{\mathfrak{p}}^2$-strongly convex for $\eta \in [0,1]$ with optimizer $\eta_* = 1$, so we can lower bound

$$F(p_*) = h(1) \geq h(0) - \frac{|h'(0)|^2}{2\alpha\|Z\|_{\mathfrak{p}}^2} \geq F(p) - \frac{\|\nabla F(p)\|_{\mathfrak{p}}^2}{2\alpha},$$

where in the first step we used $h(1) = F(\gamma_p(Z)) = F(p_*)$ by definition of h and Z, in the second step we applied Lemma 2.3.7 to h with $\alpha\|Z\|_{\mathfrak{p}}^2$-strong convexity from point $\eta = 0$ to the optimizer $\eta_* = 1$, and in the final step we used $h(0) = F(p)$ and the bound $|h'(0)| \leq \|\nabla F(p)\|_{\mathfrak{p}} \|Z\|_{\mathfrak{p}}$ derived above.

To show the distance bound on $\|Z\|_{\mathfrak{p}}$ in item (2), we use the strong convexity of h along with the bound on $|h'(0)|$ derived above to show

$$\|\nabla F(p)\|_{\mathfrak{p}} \|Z\|_{\mathfrak{p}} \geq |h'(0)| = \left| h'(1) + \int_{\eta=0}^{1} h''(1-\eta) \right| \geq \alpha\|Z\|_{\mathfrak{p}}^2,$$

where the first inequality is by the bound $|h'(0)| \leq \|\nabla F(p)\|_{\mathfrak{p}} \|Z\|_{\mathfrak{p}}$ derived above, the second step is by the fundamental theorem of calculus, and in the final step we used $h'(1) = 0$ for the first term by optimality of $h(1) = F(p_*)$ and lower bounded the second term by $h''(1-\eta) = \partial_\eta^2 F(\gamma_{p_*}(\eta Y)) \geq \alpha\|Y\|_{\mathfrak{p}}^2 = \alpha\|Z\|_{\mathfrak{p}}^2$ according to the α-strong convergent property of p in Definition 8.4.5 and the fact that $\|Y\|_{\mathfrak{p}} = \|Z\|_{\mathfrak{p}}$ by Fact 6.2.12. The bound follows by rearranging. □

Using this property of strong convergence in the geodesic setting, we can lift the analysis of Proposition 8.3.1 and Theorem 8.3.2 to formally prove linear convergence of the Flip-Flop algorithm for strongly convex tensors. Our plan is to use geodesic strong convexity at the optimizer and the multiplicative robustness of Lemma 7.1.13 to derive the strong convergent property for points sufficiently close to the optimizer. To this end, we first show a bound on the gradient which is sufficient for the strong convergent property. This argument is heavily inspired by the analysis in [35].

Lemma 8.4.7. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces of dimension* $\dim(V_a) = d_a$ *for each $a \in [m]$ along with scaling group $(G, P, \mathfrak{p})$ according to Definition 6.2.3 and input tuple $x \in V^K$. If the Kempf-Ness function f_x^P in Definition 6.2.9 is α-geodesically strongly convex at optimizer $p_* := \arg\inf_{p\in P} f_x^P(p)$, then $p \in P$ with*

$$f_x^P(p)\|\nabla \log f_x^P(p)\|_{\mathfrak{p}} = \|\nabla f_x^P(p)\|_{\mathfrak{p}} \le \frac{\alpha}{e\sqrt{\sum_{a\in[m]} d_a}}$$

satisfies the bound $\|\log(p^{-1/2}p_*p^{-1/2})\|_{\mathrm{op}} \le 1$. As a consequence, any such p is $\frac{\alpha}{e}$-strongly convergent with respect to f_x^P according to Definition 8.4.5.

Proof. Let $Z := \gamma_p^{-1}(p_*) = \log(p^{-1/2}p_*p^{-1/2})$ so that $p_* = \gamma_p(Z) = p^{1/2}e^Z p^{1/2}$ according to Fact 6.2.11, and consider the univariate restriction $h(\eta) := f_x^P(\gamma_{p,p_*}(\eta)) = f_x^P(\eta Z)$. We will show that the bound on the gradient $\|\nabla f_x^P(p)\|_{\mathfrak{p}}$ implies $\|Z\|_{\mathrm{op}} \le 1$. This is combined with the robustness of strong convexity shown in Lemma 7.1.13 to prove the strong convergent property.

In order to bound $\|Z\|_{\mathrm{op}}$, we give upper and lower bounds for $|h'(0)|$ using Definition 7.1.1 of the geodesic gradient and strong convexity respectively. For the upper bound, we use a similar argument to the one in Lemma 8.4.6, showing

$$|h'(0)| = |\partial_{\eta=0} f_x^P(\gamma_p(\eta Z))| = |\langle \nabla f_x^P(p), Z\rangle_{\mathfrak{p}}| \le \|\nabla f_x^P(p)\|_{\mathfrak{p}}\|Z\|_{\mathfrak{p}}, \tag{8.4}$$

where the first step was by definition of $h(\eta) = f_x^P(\gamma_p(\eta Z))$, in the second step we used Definition 7.1.1 of the geodesic gradient, and the final step was by Cauchy-Schwarz.

For the lower bound, we will use strong convexity to show that $|h'(\eta)|$ grows rapidly. Let $Y := \log(p_*^{-1/2}pp_*^{-1/2})$, and note that $\gamma_{p,p_*}(\eta) = \gamma_{p_*,p}(1-\eta) = \gamma_{p_*}((1-\eta)Y)$ and $\|Y\|_{\mathfrak{p}} = \|Z\|_{\mathfrak{p}}$ by Fact 6.2.12 applied to unitarily invariant norm $\|\cdot\|_{\mathfrak{p}}$. Therefore, we can apply the robustness property of Lemma 7.1.13 to show

$$h''(1-\eta) = \partial_\eta^2 f_x^P(\gamma_p((1-\eta)Z)) = \partial_\eta^2 f_x^P(\gamma_{p_*}(\eta Y)) \ge e^{-\|\eta Y\|_{\mathrm{op}}}\cdot\alpha\|Y\|_{\mathfrak{p}}^2 = e^{-\|\eta Z\|_{\mathrm{op}}}\cdot\alpha\|Z\|_{\mathfrak{p}}^2,$$

where the first step was by definition $h(\eta) = f_x^P(\gamma_p(\eta Z))$, in the second step we used $\gamma_{p,p_*}(1-\eta) = \gamma_{p_*,p}(\eta) = \gamma_{p_*}(\eta Y)$ by Fact 6.2.12, the third step was by Lemma 7.1.13 since

f_x^P is α-geodesically strongly convex at p_* according to Definition 6.2.13, and in the final step we used that $\|Y\|_{\text{op}} = \|Z\|_{\text{op}}$ and $\|Y\|_{\mathfrak{p}} = \|Z\|_{\mathfrak{p}}$ by Fact 6.2.12.

This allows us to lower bound $|h'(0)|$ by the following calculation:

$$|h'(0)| = \left|h'(1) + \int_0^1 h''(1-\eta)\right| \geq \alpha\|Z\|_{\mathfrak{p}}^2 \int_0^1 e^{-\|\eta Z\|_{\text{op}}} = \alpha\frac{\|Z\|_{\mathfrak{p}}^2}{\|Z\|_{\text{op}}}(1 - e^{-\|Z\|_{\text{op}}}), \qquad (8.5)$$

where the first step was by the fundamental theorem of calculus, and in the second step we used $h'(1) = 0$ for the first term by the optimality of $h(1) = f_x^P(p_*)$ and the lower bound $h''(1-\eta) \geq e^{-\|\eta Z\|_{\text{op}}} \cdot \alpha\|Z\|_{\mathfrak{p}}^2$ derived above.

Combining the upper and lower bounds, we can rearrange to get

$$\|\nabla f_x^P(p)\|_{\mathfrak{p}} \geq \frac{|h'(0)|}{\|Z\|_{\mathfrak{p}}} \geq \alpha\frac{\|Z\|_{\mathfrak{p}}}{\|Z\|_{\text{op}}}(1 - e^{-\|Z\|_{\text{op}}}) \geq \frac{\alpha(1 - e^{-\|Z\|_{\text{op}}})}{\sqrt{\sum_{a\in[m]} d_a}},$$

where the first step was by the bound $|h'(0)| \leq \|\nabla f_x^P(p)\|_{\mathfrak{p}}\|Z\|_{\mathfrak{p}}$ from Eq. (8.4), the second step was by the lower bound calculated in Eq. (8.5), and the last step was by the relation in Lemma 7.1.15. Therefore, $\|Z\|_{\text{op}} \geq 1$ implies $\|f_x^P(p)\|_{\mathfrak{p}} \geq \frac{\alpha(1-e^{-1})}{\sqrt{\sum_{a\in[m]} d_a}}$, which gives the claim by contrapositive.

To prove the strong convergent property for p, we can apply Lemma 7.1.13 to show $h(\eta) = f_x^P(\gamma_{p_*,p}(1-\eta)) = f_x^P(\gamma_{p_*}((1-\eta)Y))$ is $\frac{\alpha}{e}\|Y\|_{\mathfrak{p}}^2$-strongly convex for $\eta \in [0,1]$ since $\|Y\|_{\text{op}} = \|Z\|_{\text{op}} \leq 1$ by Fact 6.2.12 and the first statement. □

Now that we have a simple gradient condition which implies the strong convergent property, we can lift the strongly convex analysis of Proposition 8.3.1 to the geodesic setting to analyze the convergence of the Flip-Flop algorithm in Definition 8.4.1 for the tensor scaling problem.

Theorem 8.4.8. *Let $V = \otimes_{a\in[m]} V_a$ be a tensor product of inner product spaces of dimension $\dim(V_a) = d_a$ for each $a \in [m]$, and let $(G, P, \mathfrak{p})$ be a choice of scaling group according to Definition 6.2.3. Consider input tuple $x \in V^K$ such that the Kempf-Ness function f_x^P is α-geodesically strongly convex at the optimizer $p_* := \arg\inf_{p\in P} f_x^P(p)$. Then, for starting point $x_0 := g_0 \cdot x$ and any $\delta \leq \min\{\|\nabla \log f_x^P(g_0^* g_0)\|_{\mathfrak{p}}, \delta_0\}$ with $\delta_0 := \frac{\alpha/f_x^P(g_0^* g_0)}{e\sqrt{\sum_{a\in[m]} d_a}}$, the Flip-Flop algorithm produces output $x_T := g_T \cdot x$ satisfying*

1. *(Gradient): $\|\nabla \log f_x^P(g_T^* g_T)\|_{\mathfrak{p}} \leq \delta$;*

2. *(Function):* $f_x^P(p_*) \geq f_x^P(g_T^* g_T) - \frac{e\|\nabla f_x^P(g_T^* g_T)\|_{\mathfrak{p}}^2}{2\alpha} \geq f_x^P(g_T^* g_T)\Big(1 - f_x^P(g_T^* g_T)\frac{e\delta^2}{2\alpha}\Big)$;

3. *(Distance):* $\|\log(p_*^{-1/2} g_T^* g_T p_*^{-1/2})\|_{\mathfrak{p}} \leq f_x^P(g_T^* g_T)\frac{e\cdot\delta}{\alpha}$;

for some number of iterations bounded by

$$T \lesssim \frac{m}{\delta_0^2} \cdot \log \frac{f_x^P(g_0^* g_0)}{f^*} + f_x^P(g_0^* g_0) \cdot \frac{m}{\alpha} \log \frac{\delta_0}{\delta}.$$

Proof. Let $\{g_t \in G\}_{t\geq 0}$ be the iterates of the Flip-Flop algorithm according to Definition 8.4.1 with $\{p_t := g_t^* g_t \in P\}_{t\geq 0}$ the corresponding polar parts. Note that for any $p \in P$ with $f_x^P(p) \leq f_x^P(p_0)$,

$$\|\nabla \log f_x^P(p)\|_{\mathfrak{p}} \leq \delta_0 = \frac{\alpha / f_x^P(g_0^* g_0)}{e\sqrt{\sum_{a\in[m]} d_a}} \leq \frac{\alpha / f_x^P(p)}{e\sqrt{\sum_{a\in[m]} d_a}}$$

is a sufficient condition for p to be $\frac{\alpha}{e}$-strongly convergent by Lemma 8.4.7. Therefore, our plan is to break the iterations of the Flip-Flop algorithm into two stages based on the first time $\|\nabla \log f_x^P(p)\|_{\mathfrak{p}} \leq \delta_0$: we analyze the first stage using Theorem 8.4.4, and in the second stage we use the strong convergent property to show exponential convergence by an argument similar to the one in Proposition 3.2.2.

For the first stage, let T_0 be the first time that $\|\nabla \log f_x^P(p_{T_0})\|_{\mathfrak{p}} \leq \delta_0$. If $\|\nabla \log f_x^P(p_0)\|_{\mathfrak{p}} \leq \delta_0$, then $T_0 = 0$ and this part can be skipped. Otherwise, by Theorem 8.4.4, we have

$$T_0 \lesssim \frac{\log f_x^P(p_0) - \log f_x^P(p_*)}{\min\{\delta_0^2/m, d_{\max}^{-1}\}} = \frac{m(\log f_x^P(p_0) - \log f_x^P(p_*))}{\delta_0^2},$$

where the first step was by Theorem 8.4.4 applied with δ_0, and in the second step we used

$$\delta_0^2 = \frac{\alpha^2 / f_x^P(p_0)^2}{e^2 \sum_{a\in[m]} d_a} \leq \frac{1}{d_{\max}},$$

where the first step was by definition of δ_0, and in the second step we used Proposition A.5.2 to bound $\alpha \leq f_x^P(p_*) \leq f_x^P(p_0)$.

Now that $\|\nabla \log f_x^P(p_{T_0})\|_{\mathfrak{p}} \leq \delta_0$, Lemma 8.4.7 implies that p_{T_0} is $\frac{\alpha}{e}$-strongly convergent as f_x^P is α-geodesically strongly convex at p_* by assumption. Therefore, we analyze the subsequent iterations of the Flip-Flop algorithm by following the strategy in Proposition 8.3.1 and showing $\|\nabla \log f_x^P(p)\|_{\mathfrak{p}}^2$ halves every $O(\frac{m}{\alpha})$ iterations.

Let T_1 be the first time that $\|\nabla \log f_x^P(p_T)\|_{\mathfrak{p}}^2 \leq \frac{1}{2}\delta_0^2$. We first use the $\frac{\alpha}{e}$-strong convergent property of p_{T_0} to lower bound the optimizer by

$$f_x^P(p_*) \geq f_x^P(p_{T_0}) - \frac{\|\nabla f_x^P(p_{T_0})\|_{\mathfrak{p}}^2}{2\alpha/e} = f_x^P(p_{T_0})\Big(1 - \frac{e\|\nabla \log f_x^P(p_{T_0})\|_{\mathfrak{p}}^2}{2\alpha/f_x^P(p_{T_0})}\Big),$$

where the first step is by Lemma 8.4.6(1) with $\frac{\alpha}{e}$-strong convergence, and in the final step we used $\nabla \log f = \frac{\nabla f}{f}$. Since $\|\nabla \log f_x^P(p_{T_0})\|_{\mathfrak{p}} \leq \delta_0$ by assumption of T_0, we rewrite this as

$$\log f_x^P(p_{T_0}) - \log f_x^P(p_*) \leq \log\left(1 - \frac{e \cdot \delta_0^2}{2\alpha/f_x^P(p_{T_0})}\right)^{-1} \leq \frac{e \cdot \delta_0^2}{\alpha/f_x^P(p_{T_0})}, \tag{8.6}$$

where the last step was by the Taylor approximation $-\log(1-x) \leq 2x$ for $0 \leq x \leq \frac{2}{3}$ applied to

$$\frac{e\delta_0^2}{2\alpha/f_x^P(p_{T_0})} = \left(\sum_{a\in[m]} d_a\right)^{-1} \cdot \frac{e \cdot \alpha^2/f_x^P(p_0)^2}{2\alpha/f_x^P(p_{T_0})} \leq \left(\sum_{a\in[m]} d_a\right)^{-1} \frac{e \cdot \alpha/f_x^P(p_0)}{2} \leq \frac{2}{3},$$

where in the first step we substituted in the definition of δ_0, in the second step we used $f_x^P(p_{T_0}) \leq f_x^P(p_0)$ by the descent property of the Flip-Flop algorithm shown in Proposition 8.4.3, in the third step we used $\alpha \leq f_x^P(p_*) \leq f_x^P(p_0)$ by Proposition A.5.2, and the final step was by the assumption that $m \geq 2$.

We can now apply Theorem 8.4.4 with this stronger lower bound to show

$$T_1 - T_0 \lesssim \frac{\log f_x^P(p_{T_0}) - \log f_x^P(p_*)}{\min\{\delta_0^2/2m, d_{\max}^{-1}\}} \lesssim \frac{m}{\delta_0^2} \cdot \frac{\delta_0^2}{\alpha/f_x^P(p_{T_0})} \lesssim f_x^P(p_0)\frac{m}{\alpha},$$

where the first step was by Theorem 8.4.4 applied with $\delta_0^2/2$, in the second step we used $\delta_0^2 \leq d_{\max}^{-1}$ as shown above and substituted in the lower bound from Eq. (8.6), and in the final step we used $f_x^P(p_{T_0}) \leq f_x^P(p_0)$ as Flip-Flop is a descent method according to Proposition 8.4.3.

Continuing this way, we can define T_k to be the first time $\|\nabla \log f_x^P(p_{T_k})\|_{\mathfrak{p}}^2 \leq 2^{-k}\delta_0^2$ and bound

$$T_{k+1} - T_k \lesssim f_x^P(p_{T_k})\frac{m}{\alpha} \leq f_x^P(p_0)\frac{m}{\alpha}.$$

The gradient bound in item (1) now follows by applying this argument inductively until $2^{-k}\delta_0^2 \leq \delta^2$.

Since $\|\nabla \log f_x^P(p_T)\|_{\mathfrak{p}} \leq \delta \leq \delta_0$, we have that p_T is $\frac{\alpha}{e}$-strongly convergent by Lemma 8.4.7. Therefore, items (2) and (3) now follow from the gradient bound in item (1) as simple consequences of Lemma 8.4.6(1) and (2), respectively. □

Remark 8.4.9. *Much of this algorithmic analysis goes through for any so-called descent method which outputs a sequence $\{p_t\}$ satisfying*

$$f(p_{t+1}) \leq f(p_t) - \frac{\|\nabla f(p_t)\|_{\mathfrak{p}}^2}{2L}$$

for some constant L which will then show up in the numerator of the iteration bound for fast convergence. This more general approach was taken by Franks and Moitra [35] to give a unified analysis for many natural scaling algorithms. In this thesis, we only require the analysis of Sinkhorn and Flip-Flop algorithms.

In the following Section 8.5, we will show that the results on frame scaling given in Chapter 4 can be made constructive. This will allow us to give fast algorithmic guarantees for the statistical application of random frame scaling studied in [35].

8.5 Algorithms for the Paulsen Problem and Frame Scaling

In this section, we will apply the results of Section 8.4 to give fast algorithmic guarantees for particular cases of frame scaling. Specifically, we first give a slight improvement of the core technical contribution of [35] which studied the convergence of Sinkhorn scaling for random frames. This immediately implies a tight sample complexity bound for the statistical estimation problem studied in [35]. Our second result is a randomized algorithm which, given an input to the Paulsen problem, i.e. an ε-doubly balanced frame, converges quickly to some exactly doubly balanced frame. This is really a constructive version of Section 4.5 and so only the perturbation part will be randomized, whereas the remaining scaling algorithm will be deterministic.

Our first result gives a strong bound on the solution of frame scaling for random unit vectors. We apply Theorem 8.4.8 to give fast algorithmic guarantees for the frame version of Sinkhorn/Flip-Flop. Note that for the statistical application of [35], we are mainly concerned with bounding the error of the current iterate from the true scaling solution.

Theorem 8.5.1. *Let $U \in \mathrm{Mat}(d,n)$ be a random matrix where the columns are independent and uniformly distributed as $u_j \sim n^{-1/2} S^{d-1}$. Then there exists a universal constant C such that if $n \geq Cd$, the following hold simultaneously with probability at least $1 - \exp(-\Omega(n))$:*

1. *U has size $s(U) = 1$ and is an ε-doubly balanced frame with $\varepsilon^2 \lesssim \frac{d}{n}$.*

2. *There is a doubly balanced scaling $U_* := e^{X_*/2} U e^{Y_*/2}$ with $X_* \in \mathfrak{spd}(d), Y_* \in \mathfrak{st}_+(n)$ such that*
$$\|(X_*, Y_*)\|_{\mathfrak{p}} \lesssim \varepsilon \qquad \text{and} \qquad \|(X_*, Y_*)\|_{\text{op}} \lesssim \varepsilon.$$

Further, in this event, for every $\delta \lesssim \frac{1}{\sqrt{d+n}}$, the Flip-Flop algorithm for frame scaling given in Definition 8.4.1 produces iterate $U_T := L_T U R_T$ and polar $p_T := (L_T^* L_T, R_T^* R_T)$ such that, for $p_* := (e^{X_*}, e^{Y_*})$,

$$\|\nabla_{U_T}\|_{\mathfrak{p}} \lesssim \delta \qquad \text{and} \qquad \|\log p_*^{-1/2} p_T p_*^{-1/2}\|_{\mathfrak{p}} \lesssim \delta$$

in at most $T \lesssim d + \log \frac{1}{\delta\sqrt{d+n}}$ iterations.

Proof. We will first verify the first two items concerning initial error and bounds on the doubly balanced solution using our pseudorandom analysis in Theorem 7.3.3. Then we will use strong convexity along with Theorem 8.4.8 to prove fast algorithmic convergence.

To bound the frame scaling solution, our plan is to apply the pseudorandom analysis in Theorem 7.3.3. So below, we show that U is nearly doubly balanced according to Definition 4.1.2 and satisfies the pseudorandom condition in Definition 4.2.11. The random vectors are distributed as $u_j \sim n^{-1/2} S^{d-1}$, so by construction U is equal-norm and has size

$$s(U) = \sum_{j=1}^{n} \|u_j\|_2^2 = \frac{n}{n} = 1.$$

To show that U is ε-doubly balanced, we bound the error of the left marginal: Theorem 4.4.1 shows that with probability at least $1 - \exp(-\Omega(n))$:

$$\left\| d \sum_{j=1}^{n} u_j u_j^* - I_d \right\|_{op} \lesssim \sqrt{\frac{d}{n}} \leq \varepsilon$$

where the final step was by our assumption that $n \gtrsim \frac{d}{\varepsilon^2}$ with $\varepsilon \leq O(\sqrt{\frac{d}{n}})$. This verifies Definition 4.1.2 showing U is ε-doubly balanced.

We next show that U is pseudorandom. Specifically, we can apply Theorem 5.1.6 with $\beta = \frac{1}{16}$ (since $n \geq Cd$ for C large enough by assumption) to show that with probability at least $1 - \exp(-\Omega(n))$, U is $(\Omega(1), \frac{1}{16})$-pseudorandom according to Definition 4.2.11.

By the union bound, both these events occur simultaneously with probability at least $1 - \exp(-\Omega(n))$. In particular, U is ε-doubly balanced and $(\alpha, \frac{1}{16})$-pseudorandom with

$\alpha \gtrsim \Omega(1) \gtrsim \varepsilon$. This allows us to apply Theorem 7.3.3(1) to show that $U_* := e^{X_*/2} U e^{Y_*/2}$ is a doubly balanced frame, and bound the scaling solution by

$$\max\{\|X_*\|_{\text{op}}, \|Y_*\|_{\text{op}}\} \lesssim \frac{\varepsilon}{\alpha} \lesssim \varepsilon,$$

where we used the bound in Theorem 7.3.3(2) for $O(\varepsilon)$-doubly balanced and $(\Omega(1), \frac{1}{16})$-pseudorandom frame.

Note the above is the only randomized part of the procedure, and in the event that we can apply Theorem 7.3.3, the following convergence guarantees are deterministic.

To show the algorithmic convergence of the frame Flip-Flop algorithm, our plan is to use Theorem 8.4.8. For this purpose, we rewrite the above results in terms of the Kempf-Ness formulation described in Proposition 6.2.18. We are given input $U \in \operatorname{Mat}(d,n) \simeq \mathbb{R}^d \otimes \mathbb{R}^n$ with frame scaling group $G = (\operatorname{SL}(d), \operatorname{ST}(n))$ and associated polar $(P, \mathfrak{p})$ according to Definition 6.2.3. $U_* = e^{X_*/2} U e^{Y_*/2}$ is a doubly balanced frame by Theorem 7.3.3(1), which is equivalent to $p_* := (e^{X_*}, e^{Y_*})$ being an optimizer of the Kempf-Ness function f_U^P given in Definition 6.2.9. Finally, Theorem 7.3.3(4) says that U_* is an α_*-strongly convex frame with $\alpha_* \geq e^{-12} \cdot \alpha \geq \Omega(1)$, which translates to α_*-geodesic strong convexity of f_U^P at p_* by to Lemma 7.1.8.

We can also use the size lower bound in Theorem 7.3.3(3) to bound

$$\log \frac{f_U^P(I_d, I_n)}{f_U^P(p_*)} = \log \frac{s(U)}{s(U_*)} \leq -\log\left(1 - \frac{10\varepsilon^2}{\alpha}\right) \lesssim \varepsilon^2, \tag{8.7}$$

where the first step was by the definition $U_* = e^{X_*/2} U e^{Y_*/2}$ so $s(U_*) = f_U^P(e^{X_*}, e^{Y_*})$, the second step was by the size lower bound in Theorem 7.3.3(3) with $(\alpha \geq \Omega(1), \frac{1}{16})$-pseudorandomness of U, and the final step was by Taylor approximation $-\log(1-x) \leq 2x$ for $|x| \leq \frac{1}{2}$ applied to $\varepsilon^2 \lesssim \frac{d}{n} \lesssim 1$ by assumption.

From this perspective, the Flip-Flop algorithm for frame scaling produces iterates $U_t = L_t U R_t$ with $(L_t, R_t) \in (\operatorname{SL}(d), \operatorname{ST}(n))$ and associated polar $p_t := (L_t^* L_t, R_t^* R_t)$. The conditions for algorithmic convergence translate to

$$\|\nabla \log f_U^P(p_T)\|_{\mathfrak{p}} \lesssim \|\nabla f_U^P(p_T)\|_{\mathfrak{p}} \lesssim \delta \qquad \text{and} \qquad \|\log(p_*^{-1/2} p_T p_*^{-1/2})\|_{\mathfrak{p}} \lesssim \delta.$$

These are exactly the conclusions (1) and (3) of Theorem 8.4.8 applied with $\alpha_* \geq \Omega(1)$-geodesic strong convexity at optimizer p_* and $\delta \leq \frac{\alpha_*/f_U^P(I_d, I_n)}{\sqrt{d+n}} = \delta_0$, so we can show show that this occurs by iteration

$$\begin{aligned} T &\lesssim \frac{d+n}{\alpha_*^2/f_U^P(I_d,I_n)^2} \cdot \log\left(\frac{f_U^P(I_d,I_n)}{f_U^P(p_*)}\right) + \frac{f_U^P(I_d,I_n)}{\alpha_*} \log\left(\frac{\alpha_*}{\delta\sqrt{d+n}}\right) \\ &\lesssim (d+n)\varepsilon^2 + \log\frac{1}{\delta\sqrt{d+n}} \lesssim d + \log\frac{1}{\delta\sqrt{d+n}}, \end{aligned}$$

where the first step was by the iteration bound in Theorem 8.4.8 with $\alpha_* \geq \Omega(1)$-geodesic strong convexity and $\delta_0 = \frac{\alpha_*/f_U^P(I_d,I_n)}{\sqrt{d+n}}$, in the second step we used $\log\frac{f_U^P(I_d,I_n)}{f_U^P(p_*)} \lesssim \varepsilon^2$ by Eq. (8.7) as well as the lower bound $\alpha_* \geq \Omega(1)$ derived above, and the final step was by the bound $\varepsilon^2 \lesssim \frac{d}{n}$ as shown in item (1) of this theorem. □

Remark 8.5.2. *This is a technical improvement of the core scaling lemma in [35], which requires $n \gtrsim d\log^2 d$ in order to get the same conclusions. With this improvement, we show an optimal sample complexity result for Tyler's M-estimator as shown in [35]. The analysis of [35] focused on a slightly different convex formulation for frame scaling which only produces equal-norm frames. This function does not necessarily have the multiplicative univariate robustness properties shown in Lemma 7.1.13, which we used to show faster convergence of the Flip-Flop algorithm in Theorem 8.4.8 (by the strong convergent property in Definition 8.4.5).*

The next result in this section is to give algorithmic guarantees for the smoothed analysis strategy of Section 4.5 for the Paulsen problem. For this application, we are mainly concerned with convergence in $\|\cdot\|_F$ to a doubly balanced frame, not on scalings.

Theorem 8.5.3. *There exists a universal constant C such that, for any ε-doubly balanced frame $U \in \mathrm{Mat}(d,n)$ of size $s(U) = 1$, if either of the following two conditions hold:*

$$d \geq C, Cd \leq n \leq e^{d/C}, \varepsilon \leq \frac{1}{C} \qquad \textit{or} \qquad d \geq C, n \geq e^{d/C}, \varepsilon \leq \frac{1}{Cd},$$

then with $\Omega(1)$ probability, there is a doubly balanced frame V_* of size $s(V_*) = 1$ satisfying

$$\|U - V_*\|_F^2 \lesssim \varepsilon.$$

Further in this event, for every $\delta \lesssim \frac{\varepsilon}{\sqrt{d+n}}$, the Flip-Flop algorithm for frame scaling takes at most $T \lesssim \frac{1}{\varepsilon}((n+d) + \log\frac{\varepsilon}{\delta\sqrt{d+n}})$ iterations to produce $V_T \in \mathrm{Mat}(d,n)$ such that $\|\nabla_{V_T}\|_{\mathfrak{p}} \lesssim \delta$ and $\|V_T - W\|_F^2 \lesssim \frac{\delta^2}{\varepsilon}$ for some doubly balanced frame $W \in \mathrm{Mat}(d,n)$.

Proof. We will first verify distance bound on the doubly balanced solution using our pseudorandom analysis in Theorem 7.3.3. Then we will use strong convexity of the solution along with Theorem 8.4.8 to prove fast algorithmic convergence.

By the conditions on (n, d, ε), we are exactly in the two cases covered by Theorem 4.5.1 and Theorem 4.5.2, respectively. Applying these two theorems with $\beta = \frac{1}{16}$, we get perturbation V that with $\Omega(1)$ probability satisfies

1. V has size $s(V) = 1$ and $\|V - U\|_F^2 \lesssim \varepsilon$;
2. V is $O(\varepsilon)$-doubly balanced;
3. V is $(\alpha, \frac{1}{16})$-pseudorandom with $\alpha \geq \Omega(\varepsilon)$.

This is the only randomized part of the algorithm. In the remainder of the proof, we assume these events hold simultaneously.

V now satisfies the conditions of Theorem 4.2.14, so conclusions (1) and (4) imply the frame scaling solution V_* is doubly balanced and that

$$\|V_* - V\|_F^2 \lesssim \frac{\varepsilon^2}{\alpha} \lesssim \varepsilon,$$

where we used $\alpha \gtrsim \varepsilon$. Combining this with the perturbation distance, we get

$$\|V_* - U\|_F^2 \lesssim \|V_* - V\|_F^2 + \|V - U\|_F^2 \lesssim \varepsilon + \varepsilon \lesssim \varepsilon,$$

which verifies this distance bound in this theorem.

Our plan is now to use Theorem 8.4.8 to analyze algorithmic convergence of the frame Flip-Flop algorithm. For this purpose, we rewrite the above results in terms of the Kempf-Ness formulation described in Proposition 6.2.18. We begin with perturbed input $V \in \mathrm{Mat}(d, n) \simeq \mathbb{R}^d \otimes \mathbb{R}^n$ with frame scaling group $G = (\mathrm{SL}(d), \mathrm{ST}(n))$ and associated polar $(P, \mathfrak{p})$ according to Definition 6.2.3. $V_* = e^{X_*/2} V e^{Y_*/2}$ is a doubly balanced frame by Theorem 7.3.3(1), which by Proposition 6.2.18(3) is equivalent to $p_* := (e^{X_*}, e^{Y_*})$ being an optimizer of the Kempf-Ness function f_V^P given in Definition 6.2.9. Finally, Theorem 7.3.3(4) says that V_* is an α_*-strongly convex frame with $\alpha_* \geq e^{-12} \cdot \alpha \geq \Omega(\varepsilon)$, which translates to α_*-geodesic strong convexity of f_V^P at p_* according to Definition 6.2.13.

To bound the iterations, we require the size lower bound in Theorem 7.3.3(3):

$$\log \frac{f_V^P(I_d, I_n)}{f_V^P(p_*)} = \log \frac{s(V)}{s(V_*)} \leq -\log\left(1 - \frac{10\varepsilon^2}{\alpha}\right) \lesssim \varepsilon, \tag{8.8}$$

where the first step was by the definition $V_* = e^{X_*/2} V e^{Y_*/2}$ so $s(V_*) = f_V^P(e^{X_*}, e^{Y_*})$, the second step was by the size lower bound in Theorem 7.3.3(3) with $(\alpha \geq \Omega(\varepsilon), \frac{1}{16})$-pseudorandomness of V, and the final step was by the Taylor approximation $-\log(1-x) \leq 2x$ applied to argument $\frac{\varepsilon^2}{\alpha} \lesssim \varepsilon \lesssim 1$ by assumption.

From this perspective, the Flip-Flop algorithm for frame scaling produces iterates $V_t = L_t V R_t$ with $(L_t, R_t) \in (\mathrm{SL}(d), \mathrm{ST}(n))$ and associated polar $p_t := (L_t^* L_t, R_t^* R_t)$. The first requirement for algorithmic convergence then translates to

$$\|\nabla \log f_V^P(p_T)\|_{\mathfrak{p}} \lesssim \|\nabla f_V^P(p_T)\|_{\mathfrak{p}} \lesssim \delta.$$

This is exactly conclusions (1) of Theorem 8.4.8 applied with $\alpha_* \geq \Omega(\varepsilon)$-geodesic strong convexity and $\delta \leq \frac{\alpha_*/f_U^P(I_d,I_n)}{\sqrt{d+n}} = \delta_0$, and we can show show that this occurs by iteration

$$\begin{aligned} T &\lesssim \frac{d+n}{\alpha_*^2/f_V^P(I_d,I_n)^2} \cdot \log\left(\frac{f_V^P(I_d,I_n)}{f_U^P(p_*)}\right) + \frac{f_V^P(I_d,I_n)}{\alpha_*} \log\left(\frac{\alpha_*}{\delta\sqrt{d+n}}\right) \\ &\lesssim \frac{(d+n)}{\varepsilon} \cdot \varepsilon + \frac{1}{\varepsilon} \log \frac{\varepsilon}{\delta\sqrt{d+n}}, \end{aligned}$$

where the first step was by the iteration bound in Theorem 8.4.8 with $\alpha_* \geq \Omega(\varepsilon)$-geodesic strong convexity and $\delta_0 = \frac{\alpha_*/f_V^P(I_d,I_n)}{\sqrt{d+n}}$, and in the second step we used $\log \frac{f_V^P(I_d,I_n)}{f_V^P(p_*)} \lesssim \varepsilon$ by Eq. (8.8) as well as the lower bound $\alpha_* \geq \Omega(\varepsilon)$ derived above.

In the remainder of the proof, we focus on showing that there is a doubly balanced frame $W \in \mathrm{Mat}(d,n)$ satisfying the distance bound $\|V_T - W\|_F \lesssim \frac{\delta^2}{\varepsilon}$. We want to use the fact that V_* is a doubly balanced frame that is close to V, and both V_* and V_T are scalings of V. Our plan is to use the analysis of Proposition 4.3.6, which bounds the distance to a doubly balanced matrix via the path length of matrix gradient flow. For this purpose, we will need to perform a change of basis to find the appropriate matrix scaling.

Consider $V_T = L_T V R_T$ and $V_* = e^{X_*/2} V e^{Y_*/2}$. Since we want to use Proposition 4.3.6, we exhibit a simple transformation of V_* is a matrix scaling of V_T:

$$V_* = e^{X_*/2} V e^{Y_*/2} = e^{X_*/2}(L_T^{-1} V_T R_T^{-1}) e^{Y_*/2} = (e^{X_*/2} L_T^{-1}) V_T (R_T^{-1} e^{Y_*/2}),$$

where we substituted $V_* = e^{X_*/2} V e^{Y_*/2}$ in the first step and $V = L_T^{-1} V_T R_T^{-1}$ in the second step. Now let $e^{X_*/2} L_T^{-1} = \Xi A$ be the polar decomposition according to Theorem 2.1.13 where $A = |e^{X_*/2} L_T^{-1}| \in \mathrm{SPD}(d)$ is the polar component and and $\Xi \in \mathrm{SO}(d)$ is the isometry component. Further let $e^{Y/2} := R_T^{-1} e^{Y_*/2}$ for $Y \in \mathfrak{st}_+(n)$ since both R_T and $e^{Y_*/2}$ are in $\mathrm{ST}_+(n)$. We observe that

$$\Xi^* V_* = \Xi^*(e^{X_*/2} L_T^{-1}) V_T (R_T^{-1} e^{Y_*/2}) = A V_T e^{Y/2}$$

is a doubly balanced frame (by Fact 4.2.7), and further that it is a positive definite scaling of V_T. In fact, we can show that it is a matrix scaling of V_T when viewed in the appropriate basis. Therefore, let $\Psi \in \mathrm{SO}(d)$ be the eigenbasis of $A \in \mathrm{SPD}(d)$ according to Theorem 2.1.8 so that $A = \Psi e^{X/2} \Psi^*$ for some diagonal $X \in \mathfrak{st}_+(d)$. Then we consider the matrix representations $M := \Psi^* V_T$ and $M_* := \Psi^* \Xi^* V_*$ so that

$$M_* = \Psi^* \Xi^* V_* = \Psi^* A V_T e^{Y/2} = e^{X/2} \Psi^* V_T e^{Y/2} = e^{X/2} M e^{Y/2},$$

where the first step was by definition of M_*, in the second step we substituted $\Xi^* V_* = A V_T e^{Y/2}$ as shown above, in the third step we substituted $A = \Psi e^{X/2} \Psi^*$ for eigenbasis $\Psi \in \mathrm{SO}(d)$, and the final step was by the definition $M := \Psi^* V_T$.

Since $M_* = \Psi^* \Xi^* V_*$ is a particular matrix representation of doubly balanced frame V_*, Definition 4.2.11 shows that $M_* = e^{X/2} M e^{Y/2}$ is a doubly balanced matrix scaling of M.In fact, M_* is an α_*-strongly convex matrix as V_* is an α_*-strongly convex frame. By Proposition 3.1.10(3), this implies that $(X, Y) \in \mathfrak{t} = \mathfrak{st}_+(d) \oplus \mathfrak{st}_+(n)$ is an optimizer of the matrix Kempf-Ness formulation f_M given in Definition 3.1.6, and further that f_M is α_*-strongly convex at (X, Y). Therefore, Lemma 2.3.11 in fact shows that (X, Y) is the unique optimizer of f_M, which implies $M_* = e^{X_*/2} M e^{Y_*/2}$ is the unique doubly balanced matrix scaling of M.

We have exhibited doubly balanced frame $\Xi^* V_*$ which is a scaling of V_T, and further

$$\|\Xi^* V_* - V_T\|_F = \|\Psi^* \Xi^* V_* - \Psi^* V_T\|_F = \|M_* - M\|_F,$$

where we used invariance of $\| \cdot \|_F$ under isometry $\Psi \in \mathrm{SO}(d)$. Therefore, if we can bound the distance of matrix gradient flow travelling from M to M_*, this also bounds the distance between V_T and doubly balanced frame $W := \Xi^* V_*$.

The analysis of Proposition 4.3.6 requires strong convexity throughout gradient flow and the bound is given in terms of the gradient of M. Note that $\nabla_M := \nabla f_M^{\mathfrak{t}}(0, 0)$, so we can bound this in terms of our iterate V_T by

$$\begin{aligned}
\|\nabla_M\|_{\mathfrak{t}}^2 &= \frac{1}{d} \sum_{i=1}^{d} (\langle E_{ii}, d \cdot M M^* - s(M) I_d \rangle)^2 + \frac{1}{n} \sum_{j=1}^{n} (\langle E_{jj}, n \cdot M^* M - s(M) I_n \rangle)^2 \\
&= \frac{1}{d} \sum_{i=1}^{d} (\langle E_{ii}, d \cdot \Psi^* V_T V_T^* \Psi - s(V_T) I_d \rangle)^2 + \frac{1}{n} \sum_{j=1}^{n} (\langle E_{jj}, n \cdot V_T^* \Psi \Psi^* V_T - s(V_T) I_n \rangle)^2 \\
&\leq \frac{1}{d} \| d \cdot V_T V_T^* - s(V_T) I_d \|_F^2 + \frac{1}{n} \| \operatorname{diag}(n \cdot V_T^* V_T - s(V_T) I_n) \|_F^2 = \|\nabla f_{V_T}^P (I_d, I_n)\|_{\mathfrak{p}}^2,
\end{aligned}$$

where in the first step we used Proposition 3.1.12 for the matrix gradient, in the second step we substituted $M = \Psi^* V_T$, in the third step we used $\|\operatorname{diag}(X)\|_F^2 \leq \|X\|_F^2$ for the first term, and the fourth step was by Proposition 7.1.3 of the geodesic gradient of V_T.

We have already shown that $\|\nabla_{V_T}\|_{\mathfrak{p}} \lesssim \delta$ by Theorem 8.4.8, so the above calculation implies $\|\nabla_M\|_{\mathfrak{t}} \lesssim \delta$. Further, if $M_t = e^{X_t/2} M e^{Y_t/2}$ is the solution to matrix gradient flow given in Definition 3.1.14, then Proposition 3.2.2 (with $\alpha \geq 0$) shows that $\|\nabla_{M_t}\|_{\mathfrak{t}} \leq \|\nabla_M\|_{\mathfrak{t}} \leq \|\nabla_{V_T}\|_{\mathfrak{p}} \lesssim \delta$ for all time.

We can use these gradient bounds to show strong convexity throughout matrix gradient flow. We have already shown $M_* = \Psi^* \Xi^* V_*$ is an α_*-strongly convex matrix. Further, since $\delta \lesssim \frac{\varepsilon}{\sqrt{d+n}} \leq \frac{\alpha_*}{e\sqrt{d+n}} =: \delta_0$ by assumption, we have that M_t satisfies the gradient condition $\|\nabla_{M_t}\|_{\mathfrak{t}} \lesssim \delta \leq \delta_0$ in Lemma 8.4.7. Specifically, rewriting $M_* = e^{X/2} M e^{Y/2}$ as shown above, this allows us to apply Lemma 8.4.7 to show that $\|(X_t, Y_t) - (X, Y)\|_\infty \leq 1$ for all time. Therefore, we can use the robustness property in Lemma 3.2.4 to show M_t is also $\frac{\alpha_*}{e} \geq \Omega(\varepsilon)$-strongly convex as a matrix for all time.

This verifies the conditions of Proposition 4.3.6, which shows

$$\|\Xi^* V_* - V_T\|_F^2 = \|M_* - M\|_F^2 \leq \frac{\|\nabla f_M^{\mathfrak{t}}(0,0)\|_{\mathfrak{t}}^2}{4\alpha_*/e} \lesssim \frac{\delta^2}{\varepsilon},$$

where the first step was by the invariance of $\|\cdot\|_F$ under isometry $\Psi \in \mathrm{SO}(d)$ as $M_* = \Psi^* \Xi^* V_*$ and $M = \Psi^* V_T$, the second step was by the distance bound in Proposition 4.3.6 with $\Delta(M) = \|\nabla_M\|_{\mathfrak{t}}^2$ according to Definition 4.3.2, and the final step was by the fact that $\|\nabla_M\|_{\mathfrak{t}} \leq \|\nabla_{V_T}\|_{\mathfrak{p}} \lesssim \delta$ and $\frac{\alpha_*}{e} \geq \Omega(\varepsilon)$-strong convexity of M_t shown above. This gives the required distance bound to frame $W = \Xi^* V_*$ which is doubly balanced by Fact 4.2.7 as V_* is doubly balanced by the first statement in the theorem. □

At this point, all the results of Chapter 3 and Chapter 4 have been made algorithmic. In the next Chapter 9, we will combine the bounds of Chapter 7 on tensor scaling with Theorem 8.4.8 to give sample complexity bounds and algorithmic guarantees for a statistical estimation problem on tensors.

Chapter 9

Maximum Likelihood Estimator for the Tensor Normal Model

The results of this chapter are based on [36], which is joint work with Cole Franks, Rafael Oliveira, and Michael Walter.

In this chapter, we consider covariance estimation for matrix-variate and tensor-variate Gaussian data. In order to bypass information theoretical sample lower bounds, we consider the well-studied matrix and tensor normal models, where the covariance is assumed to factor into a product of tensor factors. These distributions arise naturally in numerous applications like gene microarrays, spatio-temporal data, and brain imaging. This is the second main application in this thesis, after the Paulsen problem discussed in Chapter 4. It turns out that the maximum likelihood estimator (MLE) for this model is, up to some small reductions, exactly the solution to a tensor scaling problem. In particular, we will study the random tensor scaling problems that arise in this statistical setting and show strong bounds on the MLE as a consequence of the convergence analyses presented in Chapter 7. We will also use the algorithmic framework of Chapter 8 to give the first rigorous convergence analysis of the natural Flip-Flop algorithm for finding the MLE, which explains the fast convergence of this algorithm in practice.

The reader is not required to have any background in statistical estimation, and the only concepts assumed will be linear algebra as covered in Section 2.1. Therefore, in our first Section 9.1, we present the relevant concepts from statistics using the running example of covariance estimation for the Gaussian distribution. In Section 9.2, we present a natural generalization of this problem to matrix or tensor valued data, and give new sample complexity results for covariance estimation in this setting. Our main tool will be the analyses of tensor scaling from Chapter 7. We will also use the framework of Chapter 8 to show the promised estimator can be computed to high accuracy via the Flip-Flop algorithm in Definition 8.4.1. In Section 9.3, we show that random tensors satisfy the strong convexity and pseudorandomness properties required to apply the analyses of Chapter 7. These are stand-alone results on spectral properties of random tensors, and we believe they are of independent interest. We also mention that our proof that random tensors satisfy the spectral condition of Definition 7.1.9 is a small adaptation of a powerful result of Pisier [80], which is stated from the perspective of quantum information theory.

9.1 Statistical Background

In this section, we present the statistical background necessary to state our new results on the matrix and tensor normal model. Specifically, we will define statistical inference in Section 9.1.1, maximum likelihood estimation in Section 9.1.2, and the measure of error we use for our estimators in Section 9.1.3. We will use the running example of Gaussian covariance estimation to illustrate these concepts. Finally, in Section 9.1.4 we present tight sample complexity results for Gaussian covariance estimation.

9.1.1 Statistical Inference

The core problem in statistics is to gain some quantitative knowledge about an unknown distribution based on samples from that distribution. A statistical model is a set of assumptions which constrains the possible family of distributions $\mathcal{F}$ that we are dealing with. For a known model, the task of statistical inference is, given independent samples $X_1, ..., X_n$ chosen uniformly from a fixed unknown distribution $\mathcal{D} \in \mathcal{F}$, to estimate some concrete property Θ of the distribution $\mathcal{D}$. The quality of this estimate can be measured according to various metrics depending on the application requirements. The theoretical goal is to give an estimator $\widehat{\Theta}$ which, with high probability, uses few samples and is as close to the truth as possible. Note that the estimator can depend on the model and the samples, but obviously cannot depend on knowledge of the unknown distribution $\mathcal{D}$.

The following are a few simple examples of statistical estimation problems.

Example 9.1.1 (Bernoulli Estimation). *Given samples $X_1, ..., X_n \sim Ber(p)$ from a Bernoulli distribution, estimate the unknown bias p.*
Output: *The sample mean $\widehat{p} := \frac{1}{n} \sum_{i=1} X_i$ is a natural high-quality estimator.*

Example 9.1.2 (Gaussian Covariance Estimation)**.** *Given samples $X_1, ..., X_n \sim N(0, \Theta^{-1})$, estimate the unknown precision matrix $\Theta \in \mathrm{Mat}(d)$.*
***Output**: The inverse sample covariance matrix $\widehat{\Theta} := \left(\frac{1}{n}\sum_{i=1} X_i X_i^*\right)^{-1}$ is a natural high-quality estimator for the precision matrix.*

This Gaussian covariance estimation problem will be a running example throughout this section. We study the precision matrix Θ instead of the covariance matrix as a small notational convenience due to our choice of error measure and estimator.

In both the previous examples, our choice of estimator seemed quite intuitive given the available information. In the following Section 9.2 on the matrix and tensor normal models, there will not be such a clear choice. Therefore, in the next subsection, we present a formal paradigm which describes a reasonable choice of estimator in general.

9.1.2 Maximum Likelihood Estimation

In this subsection, we will present the maximum likelihood method for statistical estimation. This method does not come with general guarantees, and instead gives a recipe for an estimator, the quality of which depends heavily on the application. In particular, we will compute the maximum likelihood estimator (MLE) for Gaussian covariance estimation, which will present some justification for our choice in Example 9.1.2. We re-iterate that the MLE is not always a good estimator (and in fact is not even required to exist in general), so the analysis in Section 9.1.4 gives the final justification for our choice of estimator in Example 9.1.2.

The method is motivated by the following reasoning. Suppose we are given sample $X \in \mathbb{R}^d$ from some unknown centered Gaussian distribution, and we guess that the true distribution is $N(0, \Theta^{-1})$. According to Definition 2.5.7, the probability density function (pdf) is given by

$$f_\Theta(x \in \mathbb{R}^d) = \sqrt{\frac{\det(\Theta)}{(2\pi)^d}} \exp\left(-\frac{1}{2} x^* \Theta x\right).$$

If $f_\Theta(X)$ is small for our given sample X, then this sample was very unlikely, and in some sense Θ is a bad guess. This inuition is formalized below.

Definition 9.1.3. *Given samples $X_1, ..., X_n \in \mathbb{R}^d$ from some unknown distribution in $\mathcal{F} := \{\mathcal{D}_\omega\}_{\omega \in \Omega}$, the likelihood function of guess $\theta \in \Omega$ is*

$$L_X(\theta) := f_\theta(X_1, ..., X_n) = \prod_{i=1}^n f_\theta(X_i),$$

where f_θ is the pdf of $\mathcal{D}_\theta$. We also consider the log-likelihood function $\ell_X(\theta) := \log L_X(\theta)$.

The maximum likelihood estimator (MLE) is the maximizer

$$\widehat{\theta} := \arg\max_{\omega\in\Omega} L(\omega).$$

It turns out that the estimator in Example 9.1.2 can be derived using this perspective.

Proposition 9.1.4. *Given samples $X_1, ..., X_n \in \mathbb{R}^d$ from an unknown centered Gaussian distribution, the MLE is the inverse sample covariance $\widehat{\Theta} := \left(\frac{1}{n}\sum_{i=1} X_i X_i^*\right)^{-1}$. This can be computed as the solution to the following optimization problem:*

$$\arg\min_{\Theta\in\mathrm{PD}(d)} F_X(\Theta) := \left\langle \frac{1}{n}\sum_{i=1}^{n} X_i X_i^*, \Theta \right\rangle - \log\det(\Theta).$$

Proof. By independence, the log-likelihood of Θ for samples $X_1, ..., X_n$ is just the sum of log-likelihoods for each individual sample. So we compute

$$\log f_\Theta(x) = \frac{1}{2}\log\det(\Theta) - \frac{d}{2}\log(2\pi) - \frac{1}{2}x^*\Theta x$$

$$\implies \ell_X(\Theta) = \frac{n}{2}\log\det(\Theta) - \frac{nd}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\langle X_i X_i^*, \Theta\rangle.$$

The MLE is the maximizer of the function $\ell_X(\Theta)$. We first perform some simplifying transformations to more clearly show the similarity to scaling (Proposition 6.2.18). We can drop the $\frac{nd}{2}\log(2\pi)$ term, since it does not depend on Θ, and renormalize to find the MLE as

$$\arg\min_{\Theta\in\mathrm{PD}(d)} F_X(\Theta) := \frac{-2}{n}\Big(\ell(\Theta) + \frac{nd}{2}\log(2\pi)\Big) = \left\langle \frac{1}{n}\sum_{i=1}^{n} X_i X_i^*, \Theta \right\rangle - \log\det(\Theta),$$

where we have used the natural Frobenius (entrywise) inner product on $\mathrm{Mat}(d)$. We can find the optimizer by solving for critical points:

$$0 = \nabla_\Theta F_X(\Theta) = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^* - \Theta^{-1} \implies \widehat{\Theta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^*\right)^{-1}.$$

Note that this critical point is unique whenever the samples $\{X_1, ..., X_n\}$ are of full rank, which occurs with probability 1 for $n \geq d$. To prove that this is in fact the MLE, we can show that it is the global minimizer of F_X by computing the second derivative, and we leave this folklore result to the reader. □

This optimization formulation enjoys a certain linear invariance properties which will be helpful for our analysis.

Proposition 9.1.5. *For samples $X_1, ..., X_n \in \mathbb{R}^d$ and $A \in \mathrm{GL}(d)$, let $Y_i := AX_i$. Then*

$$F_Y(\Theta) = F_X(A\Theta A^*) + \log\det(AA^*).$$

As a consequence, $\widehat{\Theta}_Y$ is the MLE for Y iff $\widehat{\Theta}_X = A\widehat{\Theta}_Y A^*$ is the MLE for X.

Proof. This is a simple change of variable calculation:

$$\begin{aligned} F_Y(\Theta) &= \frac{1}{n}\sum_{i=1}^{n}\langle Y_i Y_i^*, \Theta\rangle - \log\det(\Theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\langle X_i X_i^*, A\Theta A^*\rangle - \log\det(A\Theta A^*) + \log\det(AA^*) \\ &= F_X(A\Theta A^*) + \log\det(AA^*), \end{aligned}$$

where the first and third steps were by the definition of MLE given in Proposition 9.1.4, and in the second step we used $Y = AX$ and multiplicativity of det for invertible $A \in \mathrm{GL}(d)$. For the the second statement, $\log\det(AA^*)$ does not depend on Θ, so we can drop this term without changing the optimizer. □

We will use this invariance property in Section 9.1.4 to reduce to the case $\Theta = I_d$, where we will be able to use tighter concentration bounds in our analysis.

9.1.3 Quality of Gaussian Covariance Estimator

There are many ways to measure how good an estimator is, and the particular choice depends greatly on the application. One natural measure of error for Gaussian covariance estimation is the following.

Definition 9.1.6. *The relative Frobenius and operator error between $A, B \in \mathrm{PD}(d)$ is defined*

$$d_F(A,B) = \|I_d - B^{-1/2}AB^{-1/2}\|_F, \qquad d_{\mathrm{op}}(A,B) = \|I_d - B^{-1/2}AB^{-1/2}\|_{\mathrm{op}}.$$

Note that these measures are not symmetric, and so not strictly a metric or distance measure. One reason for this choice is that these errors are scale-invariant. In fact, in the next proposition we show that they satisfy a stronger linear invariance property.

Proposition 9.1.7. *For $A, B \in \mathrm{PD}(d)$, $d(A,B) = d(B^{-1/2}AB^{-1/2}, I_d)$ where d denotes both d_F and d_{op}. As a consequence, for $X := \log B^{-1/2}AB^{-1/2}$,*

$$\|X\| \leq 1 \implies d(A,B) = \|I_d - e^X\| \leq 2\|X\|,$$

where $\|\cdot\|$ denotes $\|\cdot\|_F$ for $d = d_F$ and $\|\cdot\|_{\mathrm{op}}$ for $d = d_{\mathrm{op}}$.

Proof. The first statement is clear by definition, and the second statement follows by a simple Taylor approximation $|e^z - 1| \leq 2|z|$ for $|z| \leq 1$ applied to the eigenvalues of X. □

Intuitively, d_F, d_{op} give a multiplicative form of error between A, B. For example

$$d_{\mathrm{op}}(A,B) = \sup_{v \in \mathbb{R}^d} \frac{|\langle vv^*, I_d - B^{-1/2}AB^{-1/2}\rangle|}{\|v\|_2^2} = \sup_{u \in \mathbb{R}^d} \frac{|\langle u, Bu\rangle - \langle u, Au\rangle|}{\langle u, Bu\rangle},$$

where the last line was a change of variable $v = B^{1/2}u$. Therefore $d_{\mathrm{op}}(A,B) \leq \varepsilon$ implies a multiplicative approximation of the quadratic form

$$\forall u \in \mathbb{R}^d : \langle u, Au\rangle \in (1 \pm \varepsilon)\langle u, Bu\rangle.$$

This kind of approximation is common in the literature on Laplacian solvers and graph sparsification (e.g. [87], [84]).

Another reason to measure error this way is that it approximates many other natural statistical error measures such as total variation distance, KL-divergence, and Fisher-Rao distance. Specifically, due to the linear invariance property shown above, all of these measures are the same up to constant factors whenever any one of them is bounded by a small constant.

Our results in Section 9.2 will rely on geodesic convex optimization which will give strong bounds on the geodesic distance $\|X = \log B^{-1/2}AB^{-1/2}\|$ to the optimizer. We will use the second property in Proposition 9.1.7 to show that this also implies strong bounds on d_F and d_{op}.

9.1.4 Analysis of the MLE

In this subsection, we will give explicit sample complexity bounds for high quality Gaussian covariance estimation by bounding the relative error of the MLE given in Proposition 9.1.4. These results are standard in the literature, and are tight up to constant factors due to folklore lower bounds discussed informally at the end of this subsection.

Theorem 9.1.8. *Let $X_1, ..., X_n \in \mathbb{R}^d$ be samples from Gaussian distribution $N(0, \Theta^{-1})$, and let $\widehat{\Theta}$ be the MLE for the precision matrix according to Proposition 9.1.4. For any $\varepsilon \leq \frac{1}{10}$ such that $n \geq \frac{d}{\varepsilon^2}$, the following error bounds are satisfied with probability at least $1 - 2\exp(-\Omega(\varepsilon^2 n))$:*

$$d_{\text{op}}(\widehat{\Theta}, \Theta) \lesssim \varepsilon \qquad \textit{and} \qquad d_F(\widehat{\Theta}, \Theta)^2 \lesssim d\varepsilon^2.$$

Proof. Our plan is to use the linear invariance of the MLE and distance measure to reduce to the case when $\Theta = I_d$. The result will then follow from standard matrix concentration results of Gaussian distributions as given in Theorem 2.5.12.

By Definition 2.5.7, $X \sim N(0, \Theta^{-1})$ is distributed as $\Theta^{-1/2} Y$ for standard Gaussian $Y \sim N(0, I_d)$. By Proposition 9.1.5, the MLE of X and Y are related as follows:

$$\widehat{\Theta}_Y = \Theta^{-1/2} \widehat{\Theta}_X \Theta^{-1/2}.$$

The error measures also satisfy a similar invariance according to Proposition 9.1.7:

$$d(\widehat{\Theta}_X, \Theta) = d(\Theta^{-1/2} \widehat{\Theta}_X \Theta^{-1/2}, I_d) = d(\widehat{\Theta}_Y, I_d).$$

Therefore, in order to prove the theorem, it is enough to show the error bound in the case when $\Theta = I_d$. In this case, the sample covariance has spectrum concentrated close to one. For $t = \varepsilon\sqrt{n}$, Theorem 2.5.12 gives the following bound with probability at least $1 - 2\exp(-\varepsilon^2 n/2)$:

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n Y_i Y_i^*\right) = \left(\frac{\sigma_{\max}(Y)}{\sqrt{n}}\right)^2 \leq \left(1 + \frac{\sqrt{d} + \varepsilon\sqrt{n}}{\sqrt{n}}\right)^2 \leq 1 + 5\varepsilon,$$

where we applied Theorem 2.5.12 to the random matrix of Gaussian samples $Y = [Y_1, ..., Y_n] \in \text{Mat}(d, n)$, and in the last step we used the assumption that $n \geq d/\varepsilon^2$ and $\varepsilon \leq \frac{1}{10}$. Theorem 2.5.12 also gives the following lower bound with the same probability:

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n Y_i Y_i^*\right) \geq 1 - 5\varepsilon.$$

Therefore, when this event occurs, we can bound the error

$$d_{\text{op}}(\widehat{\Theta}_Y, I_d) = \left\|\left(\frac{1}{n}\sum_{i=1}^n Y_i Y_i^*\right)^{-1} - I_d\right\|_{\text{op}} = \max\left\{|\lambda_{\max}^{-1} - 1|, |\lambda_{\min}^{-1} - 1|\right\} \leq 10\varepsilon,$$

where the last step was by Taylor approximation $|\frac{1}{1+x} - 1| \leq 2|x|$ for $|x| \leq \frac{1}{2}$ along with the assumption $\varepsilon \leq \frac{1}{10}$. Similarly, we can calculate

$$d_F(\widehat{\Theta}_Y, I_d)^2 = \left\| \left(\frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^* \right)^{-1} - I_d \right\|_F^2 \leq \sum_{j=1}^{d} (\lambda_j^{-1} - 1)^2 \leq d(10\varepsilon)^2,$$

where the last step was again by the same Taylor approximation. □

This is in fact best possible error bound up to constant factors. In fact, the sample covariance is non-invertible for $n < d$ samples so in this case we cannot have any constant error estimator. Intuitively, if we rewrite the sample complexity requirement as $nd \gtrsim d^2$, then the right hand side represents the degrees of freedom of the unknown precision matrix, and the left hand side represents the information content of n samples of d-dimensional vectors. This reasoning will also give heuristic lower bounds for sample complexity of the matrix and tensor normal model in the following section.

9.2 Matrix and Tensor Normal Model

This section contains the main new sample complexity results for covariance estimation in the matrix and tensor normal models. In Section 9.2.1, we will introduce the matrix and tensor normal model as well as the maximum likelihood estimator and error measure used for our new results. This model can be viewed as a generalization of the Gaussian model of Example 9.1.2 to matrix and tensor-variate data and the MLE therefore reduces to an optimization problem similar to Proposition 9.1.4. In Section 9.2.2, we discuss previous results for this estimator as well as the the natural Flip-Flop algorithm used to compute it in practice. Then, in Section 9.2.3, we state the new results in [36], proving the best-known sample complexity results for the tensor normal model as well as the first rigorous convergence analysis of the Flip-Flop algorithm. In Section 9.2.7, we state and prove two slightly stronger results improving the sample complexity and error bounds, respectively. This is accomplished by a reduction to the tensor scaling problem for random inputs as we show in Section 9.2.4. Therefore, we can prove our new results using the analyses from Chapter 7: specifically, we will show that when the number of samples is large enough, these random inputs have small gradient in Section 9.2.5, and are strongly convex and pseudorandom in Section 9.2.6.

9.2.1 Setup

In the previous section, we saw tight results for Gaussian covariance estimation. In this section we will consider the case when our random data is in the form of a matrix or a tensor. Explicitly, the data X is an element of the vector space $\mathbb{R}^D := \mathbb{R}^{d_a} \otimes ... \otimes \mathbb{R}^{d_m}$ for some $m \geq 2$. The discussion after Theorem 9.1.8 shows that for such distributions $N(0, \Theta^{-1})$ with $\Theta \in \mathrm{PD}(D)$ and no further assumptions on the covariance matrix, it is information-theoretically impossible to get any reasonable estimator unless the number of samples n satisfies $n \gtrsim D = \prod_{a\in[m]} d_a$. To bypass this lower bound, we will consider the following model which imposes a natural structural assumption on the covariance matrix, and show that this reduces the required number of samples for a good estimator.

Definition 9.2.1 (Matrix and Tensor Normal Model)**.** *The tensor normal model with $m \geq 2$ and dimensions $d_1, ..., d_m$ is the family of Gaussian distributions $N(0, \Theta^{-1})$ where*

$$\Theta = \Theta_1 \otimes ... \otimes \Theta_m$$

with $\{\Theta_a \in \mathrm{PD}(d_a)\}_{a\in[m]}$. When $m = 2$, this is known as the matrix normal model. The tensor product structure $\mathbb{R}^D = \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ is specified as part of the input to the model.

Note the decomposition is only unique up to scalars, so we use the convention

$$\Theta = \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$$

where $\Theta_a \in \mathrm{SPD}(d_a)$ for all $a \in [m]$ (i.e. $\det(\Theta_a) = 1$ for all $a \in [m]$), and $\theta = \det(\Theta)^{1/D}$ is the scalar normalization factor.

This allows us to formally define the statistical estimation problem below.

Definition 9.2.2 (Covariance Estimation for Matrix and Tensor Normal Model)**.** *Given samples $X_1, ..., X_n \sim N(0, \Theta^{-1})$ where $\Theta = \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$ with $\Theta_a \in \mathrm{SPD}(d_a)$, find estimator $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes .... \otimes \widehat{\Theta}_m$ such that*

$$\max\left\{|\hat{\theta} - \theta|, \quad \max_{a\in[m]} d(\widehat{\Theta}_a, \Theta_a)\right\} \leq \delta$$

for chosen precision δ according to error measure d_{op} or d_F given in Definition 9.1.6. A weaker requirement is $d(\widehat{\Theta}, \Theta) \leq \delta$.

The above estimation question can be split into a two parts: the theoretical goal is to find an estimator with provably low error using as few samples as possible; the algorithmic goal is to compute a good estimator given a fixed set of samples. Both of these are with high probability over the random samples.

In the Gaussian model in Example 9.1.2, the inverse sample covariance matrix was a natural estimator which had optimal error. In the tensor setting, this is not even a feasible solution as the sample covariance matrix will almost surely not factorize into a tensor product of the required dimensions. If $n < D = d_1 \cdot ... \cdot d_m$, the sample covariance will not be invertible. But each tensor factor Θ_a can be described by $O(d_a^2)$ entries, so the total number of unknown parameters is $\sum_{a\in[m]} d_a^2$.

One may also think of each random sample X_i as taking values in the set of $d_1 \times \cdots \times d_m$ arrays of real numbers. There are m natural ways to "flatten" X_i to a matrix: for example, we may think of it as an element of $\mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2 d_3 ... d_m}$, i.e. a matrix with columns in $\mathbb{R}^{d_1}$ indexed by $(j_2, ..., j_m)$. In the tensor normal model, the $d_2 d_3 ... d_m = \frac{D}{d_1}$ many columns are each distributed as a Gaussian random vector with covariance proportional to Θ_1. In an analogous way we may flatten it to a $d_a \times \frac{D}{d_a}$ for any $a \in [m]$. As such, the columns of the a-th flattening can be used to estimate Θ_a up to a scalar.

As such, another natural estimator is the set of marginals

$$\forall a \in [m]: \quad \widehat{\Theta}_a := \left(\mathrm{Tr}_{\overline{a}} \left[\frac{1}{n} \sum_{i=1}^{n} X_i X_i^* \right] \right)^{-1}.$$

By properties of Gaussian concentration, this estimator has very good error properties when the true covariance is I_D. But in general, this could result in an estimator with very high variance. This is because the columns of the flattenings are not independent and may be arbitrarily correlated. The MLE decorrelates the columns to obtain rates like those one would obtain if the columns were independent.

Before presenting this estimator formally, we give an intuitive derivation of the Flip-Flop algorithm that is used to compute it in practice. Say we are in the setting of the matrix normal model with $X_1, ..., X_n \sim N(0, \Theta_L^{-1} \otimes \Theta_R^{-1})$ so that $X_i = \Theta_L^{-1/2} Y_i \Theta_R^{-1/2}$ for random matrix Y_i independent standard Gaussian entries, where we used $X_i \in \mathrm{Mat}(d_L, d_R) \simeq \mathbb{R}^{d_L} \otimes \mathbb{R}^{d_R}$ by abuse of notation. Now assume that we know Θ_R. Scaling our samples by this factor, we observe that $X_i \Theta_R^{1/2}$ is distributed as $\Theta_L^{-1/2} Y_i$, which has independent columns $y_1, ..., y_{d_R} \sim N(0, \Theta_L^{-1})$. Therefore, we can simply use the inverse sample convariance of the left marginal to estimate the remaining tensor factor Θ_L, as shown in Section 9.1.4.

In general, we do not know Θ_R exactly, so we do not have access to a distribution with independent columns for any marginal. The Flip-Flop algorithm uses our current iterate as the best guess and performs the same procedure, updating one factor at a time. Explicitly, if our current guess is $\overline{\Theta}_R$, then we can update our current guess for $\overline{\Theta}_L$ to the sample covariance of the left marginal of the scaled samples $\{X_i \overline{\Theta}_R\}_{i \in [n]}$. For the general tensor normal model, in each step the flip flop algorithm chooses one of the dimensions $a \in [m]$ and uses the a-th flattening of the samples to update $\overline{\Theta}_a$.

It turns out that this procedure converges to the MLE [31], which is defined by the optimization formulation below.

Proposition 9.2.3. *For samples $X_1, ..., X_n \in \mathbb{R}^D$ where $\mathbb{R}^D = \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ and $m \geq 2$ according to Definition 9.2.1 of the tensor normal model, the MLE $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_m$ is given by the minimizer of the function*

$$F_X(\theta, \Theta_1, ..., \Theta_m) := \frac{\theta}{nD} \left\langle \sum_{i=1}^n X_i X_i^*, \otimes_{a \in [m]} \Theta_a \right\rangle - \log \theta. \tag{9.1}$$

over all $\theta > 0$ and $\{\Theta_a \in \mathrm{SPD}(d_a)\}_{a \in [m]}$.

Proof. The matrix and tensor normal models are a subset of the family of Gaussian distributions, so the pdf and the likelihood function are still of the same form:

$$\ell_X(\Theta) = \frac{n}{2} \log\det(\Theta) - \frac{1}{2} \sum_{i=1}^n \langle X_i X_i^*, \Theta \rangle$$

as calculated in Proposition 9.1.4. Substituting in $\Theta = \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$, and applying some simple transformations gives

$$\begin{aligned} F_X(\Theta) := \frac{-2}{nD} \ell_X(\Theta) &= \left\langle \frac{1}{nD} \sum_{i=1}^n X_i X_i^*, \Theta \right\rangle - \frac{1}{D} \log\det\left(\theta \cdot \Theta_1 \otimes ... \otimes \Theta_m\right) \\ &= \left\langle \frac{1}{nD} \sum_{i=1}^n X_i X_i^*, \Theta \right\rangle - \frac{1}{D} \left(\log \theta^D + \sum_{a \in [m]} \log\det(\Theta_a)^{D/d_a} \right) \\ &= \left\langle \frac{1}{nD} \sum_{i=1}^n X_i X_i^*, \Theta \right\rangle - \log \theta, \end{aligned}$$

where the first step was because $\det(\theta \cdot \Theta_1 \otimes ... \otimes \Theta_m) = \theta^D \det(\Theta_1)^{D/d_1} ... \det(\Theta_m)^{D/d_m}$ which can be shown inductively using Fact 2.4.1, and in the last step we used that $\log\det(\Theta_a) = 0$ by our convention $\Theta_a \in \mathrm{SPD}(d_a)$. □

The above should look very familiar. In fact this is almost exactly the Kempf-Ness function for tensor scaling given in Definition 6.2.9 on input X as $f_X(\Theta) = \langle \rho_X, \Theta \rangle$, and we will formalize this connection between the MLE and the tensor scaling solution in Section 9.2.4. Therefore, if we can show that input X satisfies the strong convergence conditions of the analyses in Chapter 7, we can derive strong bounds on the MLE in terms of the geodesic distance bounds for the tensor scaling solution.

In Section 9.2.4, we will explicitly show the connection between the MLE and tensor scaling and show how to reduce to the case $Y \sim N(0, I_D)$. This will allow us to use properties of Gaussian concentration to show that the input to the tensor normal model satisfies the fast convergence conditions in Chapter 7 with high probability.

9.2.2 Previous Work

In this subsection, we discuss previous results for the matrix and tensor normal models. These are quite natural assumptions for tensor data, and therefore there are many heuristics and algorithms used in practice. Though there has been a large volume of work on estimating the covariance in the matrix and tensor normal models under further assumptions like sparsity and well-conditionedness, some fundamental questions concerning estimation without further assumptions were still open prior to our work. As a natural heuristic to find a good estimator, the Flip-Flop algorithm (see Definition 8.4.1) was proposed and studied for the matrix normal model by [31] and [99]. The authors also showed the MLE converges to the true distribution when the number of samples n goes to ∞. The algorithm was naturally extended to the tensor setting in [68] and [69], but without a convergence analysis. Here we will be interested in non-asymptotic rates. In [92], it was shown that three steps of the Flip-Flop algorithm for the matrix normal model output an estimator with bounded error $d_F \lesssim (d_1^2 + d_2^2)/n$ in expectation, though they did not give bounds for the individual tensor factors. The same authors showed tighter error bounds when the covariance matrix satisfied additional sparsity assumptions. But for the general tensor normal model, there were no known results on high probability error bounds prior to our work in [36].

Even characterizing the existence of the MLE for the matrix and tensor normal model has remained elusive until recently. Améndola, Kohn, Reichenbach, and Seigal in [4] proposed a framework of statistical models known as Gaussian group models. This allowed them to relate natural existence questions about the MLE for these models to algebraic problems about group orbits (discussed in Section 6.1.2). In the special cases of matrix and tensor normal models, these are exactly related to the operator and tensor scaling problems studied in Chapter 7 (as well as the line of works [38], [19], [20]). Independently from [4], Franks and Moitra [35] used our analysis of frame and operator scaling in [63] to give nearly optimal sample complexity bounds for Tyler's M-estimator for elliptical distributions, which is the MLE for the matrix normal model under the additional assumption that the second factor is diagonal.

Recently, using the connection to the left-right action, Derksen and Makam [29] were able compute exact sample size thresholds for the existence of the MLE of the matrix normal model. Subsequently Derksen, Makam, and Walter [30] used similar algebraic techniques to compute the exact sample threshold for the tensor normal model.

In the context of operator scaling, Gurvits [45] showed much earlier that the flip-flop algorithm converges to the matrix normal MLE whenever it exists. As a special case of the analyses of [19] and [20] for tensor scaling, it can be shown that the number of flip-flop steps to obtain a gradient of magnitude δ in the log-likelihood function for the tensor and matrix normal model is polynomial in the input size and $1/\delta$.

It was observed by Wiesel [100] that the negative log-likelihood exhibits a certain variant of convexity known as geodesic convexity. This will be key to both our sample complexity and algorithmic results.

9.2.3 Main Results

In this work, we are able to achieve high probability error bounds for the individual tensor factors when the number of samples is slightly above the existence thresholds recently shown in [29] and [30]. Further, we are also able to analyze the natural Flip-Flop algorithm using techniques from strong geodesic convex optimization, in order to show exponential convergence to the MLE.

We present a version of our main result in order to give an overview of our proof strategy. This is improved in two ways in Section 9.2.7.

Theorem 9.2.4. *Let $X_1, ..., X_n \in \mathbb{R}^D$ be samples from the tensor normal model $\mathbb{R}^D := \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ with $m \geq 2$ and distribution $N(0, \Theta^{-1})$ with $\Theta := \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$ for $\Theta_a \in \mathrm{SPD}(d_a)$ for each $a \in [m]$. If $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$ for some $\varepsilon^2 \lesssim \left(\mathrm{poly}(m) \sum_{a \in [m]} d_a\right)^{-1}$, then the MLE $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_m$ according to Proposition 9.2.3 satisfies*

$$d_F(\widehat{\Theta}, \Theta)^2 \lesssim Dm\varepsilon^2$$

with probability at least $1 - k^2 \exp(-\Omega(d_{\min}))$.

Further, in this event, for any $\delta^2 \lesssim \left(\operatorname{poly}(m)\sum_{a\in[m]} d_a\right)^{-1}$*, the Flip-Flop algorithm in Definition 8.4.1 applied to tensor input* X *outputs estimator* Θ_T *such that* $d_F(\Theta_T, \widehat{\Theta}) \lesssim \sqrt{D}\cdot\delta$ *for some iteration*

$$T \lesssim m\varepsilon^2 \sum_{a\in[m]} d_a + m\log\frac{1}{\delta\sqrt{\sum_{a\in[m]} d_a}}.$$

Proof Overview. We first use invariance properties of the MLE and relative error shown in Section 9.2.4 to reduce the optimization problem to tensor scaling with random input $Y \sim N(0, I_D)$. In order to give strong bounds on the tensor scaling solution, our plan is to apply the convergence analysis of Theorem 7.1.16.

In Proposition 9.2.7 we use Gaussian concentration to bound the gradient, and in Proposition 9.2.8 we apply Pisier's theorem to show that x is $\Omega(1)$-$\mathfrak{p}$-strongly convex according to Definition 7.1.7. Both of these occur with high probability when $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$ is large enough. This allows us to apply Theorem 7.1.16 to bound on the optimal scaling $\|Z_*\|_{\mathfrak{p}}$, which can then be translated to a bound on the relative error $d_F(\widehat{\Theta}, I_D)$.

For the algorithmic guarantees, we first use the robustness property of strong convexity shown in Theorem 7.3.14 to show that $x_* = e^{Z_*/2}\cdot x$ is also $\Omega(1)$-$\mathfrak{p}$-strongly convex. This implies that f_x^P is $\Omega(1)$-geodesically strongly convex at p_* by Lemma 7.1.8. Then, we can apply Theorem 8.4.8 to bound the number of iterations required for $\|\nabla \log f_x^P(p_T)\|_{\mathfrak{p}} \lesssim \delta$, which can again be translated to a bound on the relative error $d_F(\Theta_T, \widehat{\Theta})$. □

These are the first such non-asymptotic guarantees for the tensor normal model without any additional structural assumptions, as well as the first rigorous convergence analysis of the Flip-Flop algorithm, explaining its performance in practice. The sample complexity results should be compared to the heuristic lower bound $nD \gtrsim \sum_{a\in[m]} d_a^2$, where the right hand side refers to the degrees of freedom in the tensor normal model, and the left hand side is the "information content" of n samples $X_i \in \mathbb{R}^D$. Therefore, the above theorem applied with $\frac{1}{\varepsilon^2} \approx \operatorname{poly}(m)\sum_{a\in[m]} d_a$ gives a tight error bound for the tensor normal model that requires only $\operatorname{poly}(m)d_{\max}$ factor more samples than the lower bound.

In Section 9.2.7, we give two improvements of this result: in Theorem 9.2.10, we are able to weaken the requirement on ε in order to improve the best-known sample complexity by a factor of $d_{\max}^{\Omega(1/m)}$; and in Theorem 9.2.13 we are able to refine the error bounds for the same sample complexity requirement by analyzing the tighter d_{op} measure for each individual part $a \in [m]$. Both of these theorems come with the same algorithmic guarantees. We also give a near-optimal improvement for sample complexity and d_{op} error for the matrix normal model in Theorem 9.2.11.

9.2.4 Reduction to Tensor Scaling

In this subsection, we will make clear the relation between scaling and the MLE for tensor normal model. We also discuss natural invariance properties of the MLE and our error, which will later allow us to reduce the error analysis to the simpler $\Theta = I_D$ case. In this simpler setting, we can use tighter results from Gaussian concentration, similar to our analysis in Section 9.1.4 for the Gaussian model.

From this point, fix scaling group $G = (\mathrm{SL}(d_1), ..., \mathrm{SL}(d_m))$ with associated polar $P := (\mathrm{SPD}(d_1), ..., \mathrm{SPD}(d_m))$ and infinitesimal vector space $\mathfrak{p} := \oplus_{a\in[m]}\mathfrak{spd}(d_a)$. We will consider the relation between the MLE in Proposition 9.2.3 and the Kempf-Ness function f^P given in Definition 6.2.9 for G-tensor scaling.

Lemma 9.2.5. *Consider tensor tuple $X = \{X_1, ..., X_n\} \in (\mathbb{R}^D)^n$ with $\mathbb{R}^D = \mathbb{R}^{d_1}\otimes...\otimes\mathbb{R}^{d_m}$. For $x := \frac{1}{\sqrt{nD}}X$, let $P = (\mathrm{SPD}(d_1), ..., \mathrm{SPD}(d_m))$ be the polar scaling group according to Definition 6.2.3. Then, the function F_X given in Proposition 9.2.3 and the Kempf-Ness function f_x^P given in Definition 6.2.9 are related as follows: for any fixed $(\Theta_1, ..., \Theta_m) \in (\mathrm{SPD}(d_1), ..., \mathrm{SPD}(d_m))$,*

$$\inf_{\theta>0} F_X(\theta, \Theta_1, ..., \Theta_m) = 1 + \log f_x^P(\Theta_1, ..., \Theta_m),$$

with optimizer $\theta^{-1} = f_x^P(\Theta_1, ..., \Theta_m)$.

As a consequence, the MLE for X is related to optimizer $p_* := \arg\min_{p\in P} f_x^P(p)$ by

$$\hat{\theta}^{-1} = f_x^P(p_*) \qquad \textit{and} \qquad \forall a \in [m] : \quad \widehat{\Theta}_a = p_*^{(a)}.$$

Proof. We first rewrite the MLE of X in terms of the Kempf-Ness function for $x := \frac{1}{\sqrt{nD}}X$:

$$F_X(\Theta) = \frac{\theta}{nD}\langle\sum_{i=1}^n X_iX_i^*, \otimes_{a\in[m]}\Theta_a\rangle - \log\theta = \theta\cdot f_x^P(\Theta_1, ..., \Theta_m) - \log\theta,$$

where in the first step we substituted Proposition 9.2.3 for F_X, and the last step was by Definition 6.2.9 of the Kempf-Ness function with $\rho_x = \frac{1}{nD}\sum_{i=1}^n X_iX_i^*$.

Letting $\nu := f_x^P(\Theta_1, ..., \Theta_m)$ for brevity, we solve for the optimizer of this univariate function simply as

$$0 = \partial_\theta(\theta \cdot \nu - \log\theta) = \nu - \frac{1}{\theta} \implies \theta_* = \frac{1}{\nu}.$$

Since $\partial_\theta^2(\theta \cdot \nu - \log\theta)|_{\theta_*} = \theta_*^{-2} > 0$, this is the global minimum. Substituting this into F_X gives the value

$$\inf_{\theta > 0} F_X(\theta, \Theta_1, ..., \Theta_m) = \theta_* \cdot \nu - \log\theta_* = 1 + \log f_x^P(\Theta_1, ..., \Theta_m),$$

where the last step was by $\theta_* = \frac{1}{\nu}$ and the definition of ν.

This implies the second statement, as we can optimize F_X over $\Theta = \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$ by first finding the optimizer $p_* := \arg\min_{p \in P} f_x^P(p)$ and then choosing the appropriate minimizing scalar $\hat{\theta}$. □

We can now use that f_x is geodesically convex on its domain P as shown in Proposition 6.2.18(1). It can be shown more directly that F_X is also geodesically convex, and this is the approach taken in [36]. We choose to study f_x because we can analyze it using the results in Chapter 7, and further because it enjoys slightly stronger multiplicative robustness properties than F_X. and because

Below, we collect a few invariance properties of the MLE and error measures which will allow us to reduce our analysis to the case $\Theta = I_D$.

Proposition 9.2.6. *Consider samples $X_1, ..., X_n \sim N(0, \Theta^{-1})$ from the tensor normal model with $\Theta = \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$ for $\Theta_a \in \mathrm{SPD}(d_a)$ for each $a \in [m]$. Let $Y_i := \Theta^{1/2} X_i$ such that $Y \sim N(0, I_D)$.*

1. *The MLE functions for X and Y are related by*

$$F_Y(\Theta') = F_X(\Theta^{1/2}\Theta'\Theta^{1/2}) + \log\det(\Theta).$$

As a consequence, $\widehat{\Theta}_Y$ is the MLE for Y iff $\widehat{\Theta}_X = \Theta^{1/2}\widehat{\Theta}_Y\Theta^{1/2}$ is the MLE for X.

2. *For any $a \in [m]$ and relative error $d = d_{\mathrm{op}}$ or $d = d_F$:*

$$d(\widehat{\Theta}_X^a, \Theta_a) = d(\widehat{\Theta}_Y^a, I_a).$$

3. *Given $\Gamma = \gamma \cdot \Gamma_1 \otimes \ldots \otimes \Gamma_m$ with $\gamma > 0$ and $\Gamma_a \in \mathrm{SPD}(d_a)$ for each $a \in [m]$, and similarly $\Psi = \psi \cdot \Psi_1 \otimes \ldots \otimes \Psi_m$, if $|\log(\gamma^{-1}\psi)| \leq 1$ and $\|\log \Gamma_a^{-1/2}\Psi_a\Gamma_a^{-1/2}\|_{\mathrm{op}} \leq 1$ for each $a \in [m]$, then the relative error is bounded by*

$$d_F(\Psi, \Gamma)^2 \leq 4D\bigg(|\log \gamma^{-1}\psi|^2 + \sum_{a\in[m]} \frac{\|\log \Gamma_a^{-1/2}\Psi_a\Gamma_a^{-1/2}\|_F^2}{d_a}\bigg).$$

Proof. Item (1) follows from the same calculations as Proposition 9.1.5, since the tensor normal model is also comprised of Gaussian distributions, and item (2) follows from Proposition 9.1.7 applied to each part.

For item (3), let $z := \log(\gamma^{-1}\psi)$ and $Z_a := \log \Gamma_a^{-1/2}\Psi_a\Gamma_a^{-1/2}$ for each $a \in [m]$ so that $\Gamma^{-1/2}\Psi\Gamma^{-1/2} = e^z \cdot e^{Z_1} \otimes \ldots \otimes e^{Z_m}$. Then we can bound

$$\begin{aligned} d_F(\Psi, \Gamma)^2 = \Big\| I_D - e^z \cdot e^{Z_1} \otimes \ldots \otimes e^{Z_m} \Big\|_F^2 &\leq 4\Big\| z \cdot I_D + \sum_{a\in[m]} I_{\overline{a}} \otimes Z_a \Big\|_F^2 \\ &= 4z^2\|I_D\|_F^2 + 4\sum_{a\in[m]} \|I_{\overline{a}}\|_F^2 \|Z_a\|_F^2 = 4D\Big(z^2 + \sum_{a\in[m]} \frac{\|Z_a\|_F^2}{d_a}\Big), \end{aligned}$$

where the first step was by Definition 9.1.6 of d_F and $\Gamma^{-1/2}\Psi\Gamma^{-1/2} = e^z \cdot e^{Z_1} \otimes \ldots \otimes e^{Z_m}$, the second step was by Taylor approximation $|e^x - 1| \leq 2|x|$ for $|x| \leq \frac{1}{2}$ applied to the eigenvalues of e^Z, the third step was by the orthogonality of the terms as $Z_a \in \mathfrak{spd}(d_a) \implies \langle I_a, Z^{(a)} \rangle = 0$ for each $a \in [m]$ according to Definition 2.1.10, and in the final step we used $\|I_D\|_F^2 = D$ for the identity on $\mathbb{R}^D$ and $\|I_{\overline{a}}\|_F^2 = \frac{D}{d_a}$ for the identity on $\otimes_{b\neq a}\mathbb{R}^{d_b}$. □

In order to apply the analyses of Chapter 7, we would like to show that random inputs are nearly balanced and satisfy either the strong convexity or pseudorandom property. Just as in the proof of Theorem 9.1.8, we can use items (1) and (2) of Proposition 9.2.6 to reduce to the case where $\Theta = I_D$. Therefore, in the next two sections, we will use Gaussian concentration to show that random inputs $Y \sim N(0, I_D)$ satisfy the fast convergence properties required to apply the tensor scaling analyses in Chapter 7.

9.2.5 Bounding the Gradient

In this subsection we will bound the initial gradient of the Kempf-Ness function for random Gaussian inputs $X \sim N(0, I_D)$. As shown in the proof outline of Theorem 9.2.4, this will be used along with strong convexity or pseudorandomness in order to give geodesic distance bounds on the optimizer of f_x, which will then imply strong error bounds on the MLE by Lemma 9.2.5.

Note that the geodesic gradient in Proposition 7.1.3 for tensor scaling input $x \in V^K$ is exactly the marginals of ρ_x up to a scalar shift. Therefore, in order to bound the gradient of our random Gaussian input, we can follow the analysis of Theorem 9.1.8 and use Gaussian concentration to give spectral bounds on the marginals.

Proposition 9.2.7. *Consider samples $X_1, ..., X_n \sim N(0, I_D)$ with $\mathbb{R}^D = \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$. For any $a \in [m]$ and any $0 < \varepsilon \leq \frac{1}{10}$, if $nD \geq \frac{d_a^2}{\varepsilon^2}$, the following bound holds with probability at least $1 - 2\exp(-\varepsilon^2 \frac{nD}{2d_a})$:*

$$\frac{1-\varepsilon}{d_a} I_a \preceq \frac{1}{nD} \operatorname{Tr}_a \left[\sum_{i=1}^n X_i X_i^* \right] \preceq \frac{1+\varepsilon}{d_a} I_a.$$

As a consequence, with probability at least $1 - 2(m+1)\exp(-\frac{\varepsilon^2 nD}{2d_{\max}})$, the tensor $x := \frac{1}{\sqrt{nD}} X$ has size $|s(x) - 1| \leq 3\varepsilon$, is 8ε-G-balanced according to Definition 6.2.4, and satisfies the gradient bound $\|\nabla_x\|_{\mathfrak{p}}^2 \leq 64m \cdot \varepsilon^2$ for $P = (\mathrm{SPD}(d_1), ..., \mathrm{SPD}(d_m))$ where $\nabla_x = \nabla f_x^P(I_V)$ is given in Proposition 7.1.3 with respect to $P = (\mathrm{SPD}(d_1), ..., \mathrm{SPD}(d_m))$.

Proof. For each $X_i \in \mathbb{R}^D = \mathbb{R}^{d_a} \otimes (\otimes_{b \neq a} \mathbb{R}^{d_b})$, we consider the flattening $\{Y_{i1}, ..., Y_{iN}\}$ of $N = \frac{D}{d_a}$ columns in $\mathbb{R}^{d_a}$ as described in Section 2.4.1. We concatenate these flattenings into the random matrix $Y := \{Y_{i1}, ..., Y_{iN}\}_{i=1}^n \in \mathrm{Mat}(d_a, nN)$, and since $X_1, ..., X_n \sim N(0, I_D)$, Y has independent standard Gaussian entries. This allows us to apply the matrix concentration bound in Theorem 2.5.12 with $t = \varepsilon\sqrt{\frac{nD}{/} d_a}$ to show that with probability at least $1 - \exp(-t^2/2) = 1 - 2\exp(-\varepsilon^2 \frac{nD}{2d_a})$,

$$\lambda_{\max}\left(\frac{1}{nD} \operatorname{Tr}_a \left[\sum_{i=1}^n X_i X_i^* \right] \right) = \lambda_{\max}\left(\frac{1}{nD} \sum_{i=1}^n \sum_{j=1}^N Y_{ij} Y_{ij}^* \right) = \sigma_{\max}\left(\frac{1}{\sqrt{nD}} \{Y_{ij}\}_{i \in [n], j \in [N]} \right)^2$$
$$\leq \left(\frac{\sqrt{nD/d_a} + \sqrt{d_a} + \varepsilon\sqrt{nD/d_a}}{\sqrt{nD}} \right)^2 \leq \frac{1+5\varepsilon}{d_a},$$

where the first step was by considering each $X_i = \{Y_{i1}, ..., Y_{iN}\}$, in the second step we rewrote the maximum eigenvalue in terms of the maximum singular value of the random Gaussian matrix $Y \in \mathrm{Mat}(d_a, \frac{nD}{d_a})$, in the third step we applied the singular value upper bound of Theorem 2.5.12 to Y with $t = \varepsilon\sqrt{\frac{nD}{d_a}}$, and in the final step we used the assumption $nD \gtrsim \frac{d_a^2}{\varepsilon^2}$ and $\varepsilon \leq \frac{1}{10}$. The analogous bound $\lambda_{\min} \geq 1 - 5\varepsilon$ follows by a similar calculation.

To show the second statement, we first rewrite the size of x as a chi-square variable:

$$s(x) = \frac{1}{nD}\sum_{i=1}^{n} \|X_i\|_2^2 = \frac{1}{nD}\chi(nD),$$

where the last step was by $X \sim N(0, I_D)$ and Definition 2.5.9. Note that $\frac{1}{nD}\mathbb{E}\chi(nD) = 1$, so we can use concentration to show

$$Pr\Big[|s(x) - 1| \geq 3\varepsilon\Big] \leq Pr\Big[|\chi(nD) - nD| \leq 3\varepsilon nD\Big] \leq 2\exp(\varepsilon^2 nD),$$

where we applied Theorem 2.5.11 with $\theta = \varepsilon nD$ and the assumption $\varepsilon \leq \frac{1}{10}$.

Now recall that x is ε-G-balanced according to Definition 6.2.4 iff

$$\forall a \in [m]: \quad \|\nabla_x^{(a)}\|_{\text{op}} = \|d_a\rho_x^{(a)} - s(x)I_a\|_{\text{op}} \leq s(x)\varepsilon,$$

where we used Proposition 7.1.3 to substitute in the formula for the gradient. By the two-sided spectral bounds on X calculated in the first statement, we can simply apply the union bound over all marginals to show

$$\begin{aligned}\|\nabla_x^{(a)}\|_{\text{op}} = \|d_a\rho_x^{(a)} - s(x)I_a\|_{\text{op}} &\leq \|d_a\rho_x^{(a)} - I_a\|_{\text{op}} + |s(x) - 1| \\ &\leq \left\|\frac{d_a}{nD}\operatorname{Tr}_a\left[\sum_{i=1}^{n} X_iX_i^*\right] - I_a\right\|_{\text{op}} + |s(x) - 1| \leq 8\varepsilon,\end{aligned}$$

where the second step was by the triangle inequality, in the third step we used the definition $x = \frac{1}{\sqrt{nD}}X$ as well as Definition 6.2.2 of the marginals $\rho_x^{(a)}$, and the final step was by combining the bound on $|s(x) - 1| \leq 3\varepsilon$ with the spectral bound on X derived above. □

9.2.6 Strong Convergence Properties

In this subsection, we show that the tensor scaling problem on standard Gaussian inputs satisfies the strong convexity and pseudorandom conditions with high probability when the number of samples is large enough. Specifically, we will show in Proposition 9.2.8 that it satisfies the spectral condition given in Definition 7.1.9, and in Theorem 9.2.9 we will show it satisfies the ∞-expansion condition given in Definition 7.4.4. Each of these will allows us to use the strong convergence analyses of Chapter 7. We defer the proofs of these statements to Section 9.3.

Recall that we are in the setting of vector space $V = \otimes_{a\in[m]}\mathbb{R}^{d_a}$ and scaling group $G = (\mathrm{SPD}(d_1), ..., \mathrm{SPD}(d_m))$ with associated polar $(P, \mathfrak{p})$ according to Definition 6.2.3. We first show that the random input $\frac{1}{\sqrt{nD}}X$ for $X \sim N(0, I_D)$ satisfies the spectral condition given in Definition 7.1.9.

Proposition 9.2.8. *For random Gaussian tensors $X_1, ..., X_n \sim N(0, I_D)$ with $\mathbb{R}^D = \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ and $m \geq 2$, $x := \frac{1}{\sqrt{nD}}X$ satisfies the λ-$\mathfrak{p}$-spectral condition according to Definition 7.1.9 for $\lambda \lesssim \frac{d_{\max}}{\sqrt{nD}}$ with probability at least $1 - m^2\exp(-\Omega(d_{\min}))$.*

The proof relies on a powerful theorem of Pisier [80] and is given in Section 9.3.1. The trace method technique used in [80] lifts straightforwardly to our tensor setting, but as an artifact, we do not manage to get failure probability which has inverse exponential dependence in the number of samples. We can combine this with the gradient bound of Proposition 9.2.7 to show that x is strongly convex by Proposition 7.1.10. This will allow us to apply Theorem 7.2.16 and Theorem 7.3.12 to give our best known sample complexity results for the MLE in Section 9.2.7.

We can also show x satisfies the ∞-expansion condition with high probability. Note that the sample requirement for this result is larger by a $d_{\max}$ factor, but we do manage to get inverse exponential dependence of the failure probability in n.

Theorem 9.2.9. *For random Gaussian tensors $X_1, ..., X_n \sim N(0, I_D)$ with $\mathbb{R}^D = \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ and $m \geq 2$, if $nD \gtrsim m^2 d_{\max}^3$, then $x := \frac{1}{\sqrt{nD}}X$ satisfies the λ-$(\mathfrak{p}, \infty)$-expansion condition according to Definition 7.4.4 for $\lambda \lesssim \frac{1}{m}$ with probability at least $1 - m^2\exp(\Omega(\frac{nD}{m^2 d_{\max}}))$.*

This result is proved Section 9.3.2 using Gaussian concentration and a net argument. Combined with the gradient bound in Proposition 9.2.7, we will use Lemma 7.4.5 to translate this to the pseudorandom condition, which will allow us to apply the pseudorandom convergence analysis of Theorem 7.2.26 and show strong bounds on d_{op} for each part of the MLE.

In the following Section 9.2.7, we carry out the proof outline of Theorem 9.2.4 using the above strong convergence properties to give our sample complexity results and algorithmic guarantees using the analyses of tensor scaling given in Chapter 7.

9.2.7 Improved Results and Proofs

In this subsection, we apply the analyses from Chapter 7 to prove stronger error bounds on the MLE for the matrix and tensor normal model. We also apply the framework of Chapter 8 to show that the Flip-Flop algorithm converges quickly to the MLE. We will use the gradient bounds of Section 9.2.5 and the strong convergence properties of Section 9.2.6.

We first use strong convexity and the analysis of Theorem 7.2.16 to improve the constraint on ε given in Theorem 9.2.4 as well as give refined error bounds in terms of d_{op}.

Theorem 9.2.10. *Consider random tensors $X_1, ..., X_n \in \mathbb{R}^D = \mathbb{R}^{d_1} \otimes ... \otimes \mathbb{R}^{d_m}$ with $m \geq 3$ that are sampled according to the unknown distribution $N(0, \Theta^{-1})$ from the tensor normal model with $\Theta := \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$ where $\theta > 0$ and $\Theta_a \in \text{SPD}(d_a)$ for each $a \in [m]$. For any $\varepsilon \lesssim (m^2 \cdot \sqrt{md_{\max}}^{1-1/2m})^{-1}$, if $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$, then the MLE $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_m$ satisfies $\hat{\theta} \in (1 \pm O(\varepsilon))\theta$,*

$$d_F(\widehat{\Theta}, \Theta)^2 \lesssim Dm\varepsilon^2, \qquad \text{and} \qquad \forall a \in [m] : \quad d_{\text{op}}(\widehat{\Theta}_a, \Theta_a) \lesssim \varepsilon \sqrt{md_a}^{1-\frac{1}{2m}}$$

with probability at least $1 - O(m^2) \exp(-\Omega(d_{\min}))$.

Also in this event, for any $\delta^2 \lesssim \left(m^2 \sum_{a \in [m]} d_a\right)^{-1}$, the Flip-Flop algorithm outputs estimator $\Theta_T := \theta_T \cdot (\Theta_T)_1 \otimes ... \otimes (\Theta_T)_m$ such that $d_F(\Theta_T, \widehat{\Theta})^2 \lesssim D \cdot \delta^2$ for some iteration

$$T \lesssim m^2 \varepsilon^2 \sum_{a \in [m]} d_a + m \log \frac{1}{\delta \sum_{a \in [m]} d_a}.$$

Proof. We follow the plan laid out in the proof overview of Theorem 9.2.4.

First, rewrite $Y_i = \Theta^{1/2} X_i$ so that $Y_1, ..., Y_n \sim N(0, I_D)$. This allows us to relate the MLE $\widehat{\Theta}_X = \Theta^{1/2} \widehat{\Theta}_Y \Theta^{1/2}$ by Proposition 9.2.6(1) and the relative distance for $d = d_{\text{op}}$ or $d = d_F$ by

$$d(\widehat{\Theta}_X, \Theta) = d(\Theta^{-1/2} \widehat{\Theta}_X \Theta^{-1/2}, I_D) = d(\widehat{\Theta}_Y, I_D),$$

where we used the equivariance property of relative error given in Proposition 9.1.7. Therefore, from this point on we assume that $\Theta = I_D$ so our samples are distributed according to $Y_1, ..., Y_n \sim N(0, I_D)$. We emphasize that this step is only for analysis, and knowledge of the true parameter Θ is not required in order to compute the estimator for our input X.

Now consider $y := \frac{1}{\sqrt{nD}} Y$, and let $p_* := \arg\min_{p \in P} f_y^P(p)$ be the optimizer of the Kempf-Ness function given in Definition 6.2.9. By Lemma 9.2.5, the MLE $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_m$

can be written as

$$\hat{\theta} = f_y^P(p_*)^{-1}, \qquad \text{and} \qquad \forall a \in [m]: \quad \widehat{\Theta}_a = p_*^{(a)}. \tag{9.2}$$

In order to bound the relative error $d(\widehat{\Theta}, I_D)$, our plan is use the convergence results of Theorem 7.3.21, which requires the input y to be nearly G-balanced according to Definition 6.2.4 and strongly convex according to Definition 7.1.7.

First, we use Proposition 9.2.7 to show $|s(y) - 1| \leq O(\varepsilon)$. Technically, the conditions of Theorem 7.3.21 require the input to have size $s(y) = 1$, so we should normalize y before we apply this analysis. We ignore this normalization in the remainder, as this only has negligible $O(\varepsilon)$ effect on all relevant quantities. Therefore, Proposition 9.2.7 shows that y is $O(\varepsilon)$-G-balanced since $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$. Next, we apply Proposition 9.2.8 to show y satisfies the λ-$\mathfrak{p}$-spectral condition according to Definition 7.1.9 with $\lambda \leq O(\varepsilon)$. By the union bound, both of these events occur simultaneously with failure probability at most

$$m \exp\Big(- \Omega(\varepsilon^2 nD/d_{\max})\Big) + m^2 \exp(-\Omega(d_{\min})) \leq O(m^2)\exp(-\Omega(d_{\min})),$$

where the last step was by the assumption $\varepsilon^2 nD \gtrsim d_{\max}^2 \geq d_{\min}^2$. We can now apply Proposition 7.1.10 to show that y is α-$\mathfrak{p}$-strongly convex for

$$\alpha \geq s(y)(1 - O(\varepsilon)) - (m-1)\lambda \geq 1 - O(m \cdot \varepsilon),$$

where in the first step we applied Proposition 7.1.10 to $O(\varepsilon)$-G-balanced input y with size $s(y) \geq 1 - O(\varepsilon)$, and the final step was by the bound $\lambda \leq O(\varepsilon)$ calculated above.

We can explicitly lower bound $\frac{\alpha}{\sqrt{e}} \geq \frac{1-O(m\cdot\varepsilon)}{\sqrt{e}} \geq \frac{1}{2}$ with $\varepsilon \ll \frac{1}{m^2}$, which allows us to verify the condition of Theorem 7.3.21:

$$\frac{\alpha^2}{\sqrt{e}} \geq \Omega(1) \gtrsim m^2 \cdot \varepsilon \sqrt{md_{\max}}^{1-\frac{1}{2m}} \gtrsim m^2 \cdot \varepsilon \sqrt{md_{\max}}^{1-\frac{\alpha/\sqrt{e}}{m}},$$

where the first step was by the lower bound $\frac{\alpha}{\sqrt{e}} \geq \frac{1}{2}$, the second step was by our assumption $\varepsilon \lesssim (m^2\sqrt{md_{\max}}^{1-\frac{1}{2m}})^{-1}$, and in the last step we used the lower bound $\frac{\alpha}{\sqrt{e}} \geq \frac{1}{2}$ for the exponent of $d_{\max}$. Therefore, we can apply Theorem 7.3.21 to find G-balanced scaling $y_* := p_*^{1/2} \cdot y = e^{Z_*/2} \cdot y$. In the sequel, we will use the conclusions on size and the scaling solution to bound the relative error for the MLE.

First note that Theorem 7.3.21(2) bounds the scaling solution by

$$\forall a \in [m]: \quad \|Z_*^{(a)}\|_{\text{op}} \lesssim \frac{\varepsilon\sqrt{md_a}^{1-\frac{1}{2m}}}{\alpha} \lesssim \varepsilon\sqrt{md_a}^{1-\frac{1}{2m}} \leq \frac{1}{2},$$

where we used the lower bound $\alpha \geq \Omega(1)$ in the second step and the assumption $\varepsilon\sqrt{md_a}^{1-\frac{1}{2m}} \lesssim 1$ in the last step. According to Eq. (9.2), this allows us to bound the relative error of the individual tensor factors of the MLE by

$$d_{\mathrm{op}}(\widehat{\Theta}_a, I_a) = \|e^{Z_*^{(a)}} - I_a\|_{\mathrm{op}} \leq 2\|Z_*^{(a)}\|_{\mathrm{op}} \lesssim \varepsilon\sqrt{md_a}^{1-\frac{1}{2m}},$$

where the first two steps were by Proposition 9.1.7 with d_{op}, and the final step was by the bound $\|Z_*^{(a)}\|_{\mathrm{op}} \leq \varepsilon\sqrt{md_a}^{1-\frac{1}{2m}}$ in Theorem 7.4.11(2) as derived above.

In order to bound the relative error d_F, we first need to bound the scalar term $\hat{\theta} = f_y^P(p_*)^{-1}$. This is accomplished by Theorem 7.3.21(3), which bounds

$$\log\frac{f_y^P(I_V)}{f_y^P(p_*)} = \log\frac{s(y)}{s(y_*)} \leq -\log\left(1 - \frac{O(m\cdot\varepsilon^2)}{\alpha}\right) \leq O(m\cdot\varepsilon^2), \tag{9.3}$$

where the first step was by Definition 6.2.9 of the Kempf-Ness function with $y_* = p_*^{1/2}\cdot y$, the second step was by the size lower bound in Theorem 7.3.21(3), and the final step used the lower bound $\alpha \geq \Omega(1)$ and Taylor approximation $-\log(1-x) \leq 2x$ the argument $m\varepsilon^2 \ll 1$ as $\varepsilon \ll \frac{1}{m^2}$. We derived the lower bound $s(y) \geq 1 - O(\varepsilon)$ above, so this shows

$$|\hat{\theta} - 1| = |f_y^P(p_*)^{-1} - 1| \lesssim |\log s(y_*)| \lesssim \varepsilon + m\cdot\varepsilon^2 \lesssim \varepsilon,$$

where in the first step we substituted $\hat{\theta} = f_y^P(p_*)^{-1}$ by Eq. (9.2), in the second step we used the Taylor approximation $|e^x - 1| \leq 2|x|$ for $|x| \leq \frac{1}{2}$, in the third step we applied the lower bound $s(y) \geq 1 - O(\varepsilon)$ and $s(y_*) \geq s(y)(1 - O(m\cdot\varepsilon^2))$, and the final step was by the assumption $\varepsilon \ll \frac{1}{m^2}$.

Now, we use the fact that Theorem 7.4.11(2) bounds the scaling solution by

$$\|Z_*\|_{\mathfrak{p}}^2 \lesssim \frac{m\cdot\varepsilon^2}{\alpha^2} \lesssim m\cdot\varepsilon^2,$$

where we used the lower bound $\alpha \geq \Omega(1)$. According to Eq. (9.2), this allows us to bound the relative error of the MLE by

$$d_F(\widehat{\Theta}, I_D)^2 \lesssim D\left(|\log\hat{\theta}|^2 + \sum_{a\in[m]}\frac{\|Z_a\|_F^2}{d_a}\right) = D\left(|\log s(y_*)|^2 + \|Z_*\|_{\mathfrak{p}}^2\right) \lesssim m\cdot\varepsilon^2,$$

where the first step was by Proposition 9.2.6(3), the second step was by Eq. (9.2) and Definition 7.1.2 of $\|\cdot\|_{\mathfrak{p}}$, and the final step was by the bounds $|\log s(y_*)| \lesssim \varepsilon$ and $\|Z_*\|_{\mathfrak{p}}^2 \lesssim m\cdot\varepsilon^2$.

Now we will show algorithmic convergence of the Flip-Flop algorithm from Definition 8.4.1 to the MLE. By Proposition 8.4.3, the Flip-Flop algorithm is actually a descent method for the Kempf-Ness function f_y^P with starting point $p_0 = I_V$ producing iterates $\{p_t \in P\}_{t\geq 0}$. By Lemma 9.2.5, we can translate this to a sequence of estimators

$$\theta_t = f_y^P(p_t)^{-1} \qquad \text{and} \qquad \forall a \in [m]: \quad (\Theta_t)_a = p_t^{(a)}. \tag{9.4}$$

We can view these iterates as a descent sequence converging to the optimizer of the likelihood function F_Y in Proposition 9.2.3. Therefore, these iterates converge to MLE $\widehat{\Theta}$ and not necessarily to the true value $\Theta = I_D$. Below, we show that the relative error between $\widehat{\Theta}$ and Θ_t converges exponentially.

We first observe that $y_* = p_*^{1/2} \cdot y$ is α_*-$\mathfrak{p}$-strongly convex with $\alpha_* \geq \frac{\alpha}{\sqrt{e}} \geq \frac{1}{2}$ by item (4) of Theorem 7.3.21, where the last step was by the lower bound $\alpha \geq 1 - O(m \cdot \varepsilon)$ and $\varepsilon \ll \frac{1}{m^2}$ calculated above. By Lemma 7.1.8, this is equivalent to f_y^P being $\alpha_* \geq \frac{1}{2}$-geodesically strongly convex at p_* according to Definition 6.2.13. Therefore, letting $\delta_0 := \frac{\alpha_*/f_y^P(p_0)}{\sqrt{\sum_{a\in[m]} d_a}}$, we can apply Theorem 8.4.8 with parameter δ to show that its conclusions hold by iteration

$$\begin{aligned} T &\lesssim \frac{m}{\delta_0^2} \cdot \log \frac{f_x^P(p_0)}{f_x^P(p_*)} + f_x^P(p_0) \cdot \frac{m}{\alpha_*} \log \frac{\delta_0}{\delta} \\ &\lesssim \frac{f_y^P(p_0)^2}{\alpha_*^2} \cdot m \sum_{a\in[m]} d_a \cdot m\varepsilon^2 + f_x^P(p_0) \cdot \frac{m}{\alpha_*} \log \frac{\alpha_*/f_y^P(p_0)}{\delta \sqrt{\sum_{a\in[m]} d_a}} \\ &\lesssim m^2 \varepsilon^2 \sum_{a\in[m]} d_a + m \log \frac{1}{\delta \sqrt{\sum_{a\in[m]} d_a}}, \end{aligned}$$

where we substituted $\delta_0 := \frac{\alpha_*/f_y^P(p_0)}{\sqrt{\sum_{a\in[m]} d_a}}$ and used the following bounds: $\log \frac{f_y^P(p_0)}{f_y^P(p_*)} \lesssim m \cdot \varepsilon^2$ by Eq. (9.3), $\alpha_* \geq \frac{1}{2}$, and $f_y^P(p_0) = s(y) \leq 1 + O(\varepsilon)$.

To translate this to a relative error bound on Θ_T, we note that Theorem 8.4.8(2) gives

$$\log \frac{\theta_T^{-1}}{\hat{\theta}^{-1}} = \log \frac{f_y^P(p_T)}{f_y^P(p_*)} \leq \log \left(1 - f_y^P(p_T) \frac{e \cdot \delta^2}{2\alpha_*}\right)^{-1} \lesssim \delta^2,$$

where in the first step we substituted in $\hat{\theta} = f_y^P(p_*)^{-1}$ from Eq. (9.2) and $\theta_T = f_y^P(p_T)^{-1}$ from Eq. (9.4), the second step was by rearranging the function lower bound in Theorem 8.4.8(2), and in the final step we used the bounds $f_y^P(p_T) \leq s(y) \leq 1 + O(\varepsilon)$,

$\alpha_* \geq \frac{1}{2}$, and the Taylor approximation $-\log(1-x) \leq 2x$ for $|x| \leq \frac{1}{2}$. Further, if we define $Z_T := \log(p_*^{-1/2} p_T p_*^{-1/2})$, then Theorem 8.4.8(3) gives the geodesic distance bound

$$\|Z_T\|_{\mathfrak{p}} = \|\log(p_*^{-1/2} p_T p_*^{-1/2})\|_{\mathfrak{p}} \leq f_y^P(p_T) \frac{e \cdot \delta}{\alpha_*} \lesssim \delta,$$

where again in the last step we used $f_y^P(p_T) \leq s(y) \leq 1 + O(\varepsilon)$ and $\alpha_* \geq \frac{1}{2}$.

Using Eq. (9.2) and Eq. (9.4), we can translate these geodesic bounds to a bound on the relative error, as

$$d_F(\Theta_T, \widehat{\Theta})^2 \lesssim D\left(|\log(\hat{\theta}^{-1}\theta_T)|^2 + \|Z_T\|_{\mathfrak{p}}^2\right) \lesssim D\delta^2,$$

where the first step was by Proposition 9.2.6(3), the second step was by Eq. (9.4) for Θ_T and $\widehat{\Theta}$ as well as the definitions $Z_T := \log(p_*^{-1/2} p_T p_*^{-1/2})$, and the final step was by the bounds $|\log(\hat{\theta}^{-1}\theta_T)| \lesssim \delta^2$ and $\|Z_T\|_{\mathfrak{p}} \lesssim \delta$ derived above. □

At this point, we can set the parameter $\varepsilon \approx (\text{poly}(m) \cdot \sqrt{d_{\max}}^{1-1/2m})^{-1}$ in Theorem 9.2.10 which allows us to take $n \approx \text{poly}(m) d_{\max}^{3-1/2m}/D$ samples and prove tight relative error bounds for d_F. This improves on Theorem 9.2.4 by a factor of $d_{\max}^{1/2m}$ in sample complexity as well as the fact that we can bound the d_{op} error measure. Further, by the discussion after Theorem 9.1.8, even estimating a single marginal requires $n \geq \Omega(d_{\max}^2/D)$ many samples, so this result is $d_{\max}^{1-1/2m}$ factor away from optimal.

In the $m = 2$ case of the matrix normal model, this can be further improved by applying our analysis of strongly convex operator scaling in Theorem 7.3.12 instead of the strongly convex tensor scaling analysis of Theorem 7.3.21. Since this is the only change, we omit the proof. By the discussion after Theorem 9.1.8, no estimator can have constant error bounds even for a specific marginal for $nD < d_{\max}^2$. Therefore, the following result is optimal up to poly log d factors.

Theorem 9.2.11. *Let $X_1, ..., X_n \in \mathbb{R}^D = \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}$ be samples from the matrix normal model with distribution $N(0, \Theta^{-1})$ for $\Theta := \theta \cdot \Theta_1 \otimes \Theta_2$ with $\theta > 0, \Theta_1 \in \text{SPD}(d_1), \Theta_2 \in \text{SPD}(d_2)$. For any $\varepsilon^2 \lesssim \frac{1}{\log^2 d_{\min}}$, if $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$, then the MLE $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes \widehat{\Theta}_2$ satisfies $\hat{\theta} \in (1 \pm O(\varepsilon))\theta$,*

$$d_F(\widehat{\Theta}, \Theta)^2 \lesssim D\varepsilon^2, \qquad \text{and} \qquad \max\left\{d_{\text{op}}(\widehat{\Theta}_1, \Theta_1), d_{\text{op}}(\widehat{\Theta}_2, \Theta_2)\right\} \lesssim \varepsilon \log d_{\min}$$

with probability at least $1 - 4\exp(-\Omega(d_{\min}))$.

Further, in this event, for any $\delta^2 \lesssim \frac{1}{d_{\max}}$, the Flip-Flop algorithm outputs estimator Θ_T such that $d_F(\Theta_T, \widehat{\Theta})^2 \lesssim D \cdot \delta^2$ for some iteration

$$T \lesssim \varepsilon^2(d_1 + d_2) + \log \frac{1}{\delta\sqrt{d_1 + d_2}}.$$

Remark 9.2.12. *Due to the use of the trace method in Theorem 9.3.1, both of the above theorems can only achieve $\exp(-\Omega(d_{\min}))$ failure probability. In section 4 of [36], we are able to improve this to $\exp(-\Omega(\varepsilon^2 n/d_{\max}))$ failure probability by a different approach. Specifically, the proof uses a version of Cheeger's inequality for the matrix normal model to strong convexity. The technique is similar to the approach of [35], and it is an interesting open question to find a similar approach to prove strong convexity for higher order tensors.*

Our final result uses the pseudorandom analysis of Theorem 7.2.26 to give optimal bounds on the relative operator norm error d_{op} for each individual part when the number of samples is large enough.

Theorem 9.2.13. *Let $X_1, ..., X_n \in \mathbb{R}^D = \mathbb{R}^{d_1} \otimes ...\mathbb{R}^{d_m}$ be samples from the tensor normal model with $m \geq 2$ and distribution $N(0, \Theta^{-1})$ with $\Theta := \theta \cdot \Theta_1 \otimes ... \otimes \Theta_m$ with $\Theta_a \in \text{SPD}(d_a)$. For any $\varepsilon \lesssim \frac{1}{m^2}$, if $nD \gtrsim \max\{\frac{d_{\max}^2}{\varepsilon^2}, m^2 d_{\max}^3\}$, then the MLE $\widehat{\Theta} := \hat{\theta} \cdot \widehat{\Theta}_1 \otimes ... \otimes \widehat{\Theta}_m$ satisfies $\hat{\theta} \in (1 \pm O(\varepsilon))\theta$,*

$$d_F(\widehat{\Theta}, \Theta)^2 \lesssim Dm\varepsilon^2, \qquad \textit{and} \qquad \max_{a \in [m]} d_{\text{op}}(\widehat{\Theta}_a, \Theta_a) \lesssim \varepsilon$$

with probability at least $1 - O(m^2)\exp(-\Omega(\frac{\varepsilon^2 nD}{2d_{\max}}))$.

Also in this event, for any $\delta^2 \lesssim \left(m^2 \sum_{a \in [m]} d_a\right)^{-1}$, the Flip-Flop algorithm outputs estimator Θ_T such that $d_F(\Theta_T, \widehat{\Theta})^2 \lesssim D \cdot \delta^2$ for some iteration

$$T \lesssim m^2\varepsilon^2 \sum_{a \in [m]} d_a + m \log \frac{1}{\delta\sqrt{\sum_{a \in [m]} d_a}}.$$

Proof. The proof of the relative error bound for the MLE is quite similar to Theorem 9.2.4 and Theorem 9.2.11 so we focus on the parts that are different.

We rewrite $Y_i = \Theta^{1/2} X_i$ and use property (1) and (2) of Proposition 9.2.6 to reduce our analysis to the case $\Theta = I_D$ without loss of generality. Then we define $y := \frac{1}{\sqrt{nD}} Y$

and consider $p_* := \arg\min_{p \in P} f_y^P(p)$ the optimizer of the Kempf-Ness function given in Definition 6.2.9, which by Lemma 9.2.5, translates to the MLE

$$\hat{\theta} = f_y^P(p_*)^{-1}, \qquad \text{and} \qquad \forall a \in [m] : \quad \widehat{\Theta}_a = p_*^{(a)}. \tag{9.5}$$

In order to bound the relative error $d(\widehat{\Theta}_Y, I_D)$, our plan is use the convergence results of Theorem 7.4.11, which requires the input y to be nearly G-balanced according to Definition 6.2.4 and $\mathfrak{p}$-pseudorandom according to Definition 7.2.17. By the assumption $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$, we can apply Proposition 9.2.7 to show $|s(y) - 1| \leq O(\varepsilon)$ and (ignoring the $O(\varepsilon)$ factors caused by normalization) that y is $O(\varepsilon)$-G-balanced with probability at least $1-2(m+1)\exp(-\frac{\varepsilon^2 nD}{2d_{\max}})$. Next, by the assumption that $nD \gtrsim m^2 d_{\max}^3$, we can apply Theorem 9.2.9 to show y satisfies the λ-$(\mathfrak{p}, \infty)$-expansion condition according to Definition 7.4.4 for $\lambda \lesssim \frac{1}{m}$. By the union bound, both of these events occur simultaneously with failure probability at most

$$m \exp\Big(-\Omega(\varepsilon^2 nD/d_{\max})\Big) + m^2 \exp\Big(-\Omega(nD/m^2 d_{\max})\Big) \leq O(m^2) \exp\Big(-\Omega(\varepsilon^2 nD/d_{\max})\Big),$$

where the last step was by the assumption $\varepsilon \lesssim \frac{1}{m^2}$. We can now apply Lemma 7.4.5 to show that y is γ-$\mathfrak{p}$-pseudorandom for

$$e^{-\gamma} \geq s(y)(1 - O(\varepsilon)) - \lambda \geq 1 - O(\varepsilon) - O(1/m) \geq e^{-O(1/m)},$$

where the first step was by Lemma 7.4.5, the second step was by our bounds on $|s(y)-1| \leq O(\varepsilon) \leq O(\frac{1}{m^2})$ and $\lambda \leq O(\frac{1}{m})$, and the final step was by the bound $1 - z \geq e^{-z}$ for $z \geq 0$.

Since y is $O(\varepsilon)$-G-balanced for $\varepsilon \ll \frac{1}{m^2}$ and γ-$\mathfrak{p}$-pseudorandom with $\gamma \lesssim \frac{1}{m}$, the first three conclusions of Theorem 7.4.11 give scaling $y_* := p_*^{1/2} \cdot y = e^{Z_*/2} \cdot y$ which is G-balanced and satisfies

$$\max_{a \in [m]} \|Z_*^{(a)}\|_{\text{op}} \lesssim \varepsilon, \qquad \text{and} \qquad s(y_*) \geq s(y)(1 - O(m\varepsilon^2)).$$

By Proposition 6.2.18(2), this implies that $p_* = \arg\min_{p \in P} f_y^P(p)$, and therefore we can use Lemma 9.2.5 and Eq. (9.5) to give relative error bounds for the individual tensor factors:

$$d_{\text{op}}(\widehat{\Theta}_a, I_a) = \|e^{Z_*^{(a)}} - I_a\|_{\text{op}} \leq 2\|Z_*^{(a)}\|_{\text{op}} \lesssim \varepsilon,$$

where the first step was by Definition 9.1.6 of d_{op}, the second step was by Proposition 9.1.7, and the final step was by the bound in Theorem 7.4.11(2).

Item (4) of Theorem 7.4.11 shows that y_* is $\Omega(1)$-strongly convex, so the remainder of the proof of fast algorithmic convergence is the same as in the proof of Theorem 9.2.10. □

We conjecture that the error bound $d_{\text{op}}(\widehat{\Theta}_a, \Theta_a) \lesssim \varepsilon$ is achieved for any $\varepsilon \leq \frac{1}{\text{poly}(m)}$ when $nD \gtrsim \frac{d_{\max}^2}{\varepsilon^2}$. Note that if true, this would be the optimal sample complexity (up to $\text{poly}(m)$ factors), as there is a matching lower bound $nD \gtrsim \frac{d_a^2}{\varepsilon^2}$ even for the simpler problem of estimating a single marginal Θ_a.

9.3 Expansion of Random Tensors

In this section, we prove the strong convergence properties of random tensors given in Section 9.2.6. In Section 9.3.1, we use a powerful theorem of Pisier [80] to show that random Gaussian inputs satisfy the spectral condition given in Definition 7.1.9. In Section 9.3.2, we use Gaussian concentration and a net argument to show that our random inputs satisfy the ∞-expansion condition given in Definition 7.4.4.

9.3.1 Spectral Condition via Pisier's Theorem

In this subsection, we present the proof of Proposition 9.2.8. This follows from a result of Pisier [80], whose original theorem dealt with square matrices and gave slightly weaker probabilistic guarantees than Theorem 9.3.1 stated below. We adapt this result to give exponentially small error probability for random rectangular matrices.

Theorem 9.3.1 (Pisier [80]). *Let $A_1, \ldots, A_N, A$ be independent $n \times m$ random matrices with independent standard Gaussian entries. For any $t \geq 2$, with probability at least $1 - t^{-\Omega(m+n)}$,*

$$\left\| \left(\sum_{i=1}^{N} A_i \otimes A_i \right) \circ \Pi \right\|_{\text{op}} \leq O\left(t^2 \sqrt{N}(m+n) \right),$$

where Π denotes the orthogonal projection onto the traceless subspace of $\mathbb{R}^m \otimes \mathbb{R}^m$, that is, onto the orthogonal complement of $\text{vec}(I_m)$.

We emphasize that these minor modifications follow readily from the arguments in [79], [80]. We state the proof below for completeness and claim no originality.

We first explain how Theorem 9.3.1 implies the spectral condition in Proposition 9.2.8 for our random inputs.

Proof of Proposition 9.2.8. By Lemma 7.3.10, for fixed $a \neq b \in [m]$ we can write the spectral condition in terms of

$$\|\Phi_x^{(ab)}\|_0 = \sup_{Y \in \mathfrak{spd}(d_a)} \sup_{Z \in \mathfrak{spd}(d_b)} \frac{\langle Y, \Phi_x^{(ab)}(Z)\rangle}{\|Y\|_F \|Z\|_F} \leq \sup_{Z \in \mathfrak{spd}(d_b)} \frac{\|\Phi_x^{(ab)}(Z)\|_F}{\|Z\|_F} = \|\Phi_x^{(ab)} \circ Q_{I_b^\perp}\|_{F \to F}$$

where the first step was by Definition 7.3.7 of $\|\cdot\|_0$, the second step was by Cauchy-Schwarz, and in the final step we use Definition 7.3.7 of $\|\cdot\|_{F\to F}$ as well as the fact that $Q_{I_b^\perp}$ is a contraction with image $\mathfrak{spd}(d_b)$.

We want to rewrite this in terms of the natural representation of $\Phi_x^{(ab)}$ so that it resembles Theorem 9.3.1. Note that any tensor $x_i \in \mathbb{R}^D = (\mathbb{R}^{d_a} \otimes \mathbb{R}^{d_b}) \otimes (\otimes_{c \notin \{a,b\}} \mathbb{R}^{d_c})$ can be naturally identified with the tuple of $N = \frac{D}{d_a d_b}$ columns $\{x_{ij}^{(ab)} \in \mathbb{R}^{d_a} \otimes \mathbb{R}^{d_b}\}$ for $j = 1, ..., N$. Therefore we can rewrite the marginal

$$\rho_x^{(ab)} = \mathrm{Tr}_{ab}\left[\sum_{i=1}^n x_i x_i^*\right] = \sum_{i=1}^n \sum_{j=1}^N (x_{ij}^{(ab)})(x_{ij}^{(ab)})^*,$$

following the partial trace calculation as in the example in Eq. (2.9). Then, letting $x_{ij}^{(ab)} \in \mathbb{R}^{d_a} \otimes \mathbb{R}^{d_b} \to A_{ij} := \mathrm{Mat}(x_{ij}^{(ab)}) \in \mathrm{Mat}(d_a, d_b)$ gives the Kraus operators of $\Phi_x^{(ab)}$ by Definition 2.4.4. Therefore, we can rewrite Definition 7.1.9 as

$$\|\Phi_x^{(ab)}\|_0 \leq \|\Phi_x^{(ab)} \circ Q_{I_b^\perp}\|_{F \to F} = \left\| \left(\sum_{i=1}^n \sum_{j=1}^N A_{ij} \otimes A_{ij} \right) \circ Q_{I_b^\perp} \right\|_{\mathrm{op}},$$

where the first step was by the calculation above, and in the final step we used Definition 2.4.6 to translate to the natural representation. Since $x = \frac{1}{\sqrt{nD}} X$, we have that each entry of A_{ij} is i.i.d. from $N(0, \frac{1}{nD})$. So we apply Theorem 9.3.1 to show

$$\|\Phi_x^{(ab)}\|_0 \leq \left\| \left(\sum_{i=1}^n \sum_{j=1}^N A_{ij} \otimes A_{ij} \right) \circ Q_{I_b^\perp} \right\|_{\mathrm{op}} \lesssim t^2 \frac{(d_a + d_b)\sqrt{nN}}{nD} \lesssim \frac{d_a + d_b}{\sqrt{nD}} \frac{1}{\sqrt{d_a d_b}},$$

where in the final step we substituted $n = d_a$, $m = d_b$, $t = 2$, and $N = \frac{D}{d_a d_b}$ into Theorem 9.3.1. This verifies that x satisfies the characterization in Lemma 7.3.10(1) of the λ-$\mathfrak{p}_{ab}$-spectral condition with $\lambda \lesssim \frac{d_a + d_b}{\sqrt{nD}}$ with probability $\exp(-\Omega(d_a + d_b))$, so the theorem follows by a union bound over all $a \neq b \in [m]$. □

Remark 9.3.2. *Note that the failure probability can be improved at the cost of worse spectral condition by choosing t larger. But the parameter t is only the base of the exponent so this does not give* $\exp(-\Omega(nD))$ *failure probability. This is in contrast to Theorem 9.2.9 which gives better failure probability but only works when $nD \gtrsim d_{\max}^3$ instead of the $nD \gtrsim d_{\max}^2$ condition of Proposition 9.2.8.*

In the remainder we present the proof of Theorem 9.3.1. The argument consists of a symmetrization trick, followed by the trace method. We first state some relevant bounds on Gaussian random variables.

We will often use the following estimate of the operator norm of a standard Gaussian $n \times m$ random matrix A (see Theorem 5.32 in [94]),

$$\mathbb{E}\|A\|_{\text{op}} \leq \sqrt{n} + \sqrt{m}. \tag{9.6}$$

Theorem 9.3.3. *Let A be a centered Gaussian random variable that takes values in a separable Banach space with norm $\|\cdot\|$. Then $\|A\|$ satisfies the following concentration and moment inequalities with parameter $\sigma^2 := \sup\{\mathbb{E}\langle X, A\rangle^2 \mid \|X\|_* \leq 1\}$, where $\|\cdot\|_*$ denotes the dual norm:*

$$\forall t > 0\colon \quad \mathbb{P}\Big(\big|\|A\| - \mathbb{E}\|A\|\big| \geq t\Big) \leq 2\exp\Big(-\frac{\Omega(t^2)}{\sigma^2}\Big), \qquad \textit{and}$$

$$\forall p \geq 1\colon \quad \mathbb{E}\|A\|^p \leq (2\mathbb{E}\|A\|)^p + O(\sigma\sqrt{p})^p. \tag{9.7}$$

Proof. The first statement on concentration is exactly Theorem 1.5 in [77]. For the second, we consider the random variable $X := \frac{1}{\sigma}(\|A\| - \mathbb{E}\|A\|)$. Then the equivalence in Lemma 5.5 of [94] gives the moment bound

$$\Big(\mathbb{E}|X|^p\Big)^{1/p} = \frac{1}{\sigma}\Big(\mathbb{E}\Big|\|A\| - \mathbb{E}\|A\|\Big|^p\Big)^{1/p} \leq O(\sqrt{p}).$$

The moment bound in the theorem now follows by rearranging as

$$\mathbb{E}\|A\|^p = \mathbb{E}\Big(\mathbb{E}\|A\| + \sigma X\Big)^p \leq 2^p\Big((\mathbb{E}\|A\|)^p + O(\sigma\sqrt{p})^p\Big),$$

where the last step was by the simple inequality $(a+b)^p \leq 2^p(|a|^p + |b|^p)$. □

Below, we calculate the σ^2 parameter in Theorem 9.3.3 with regards to our random matrix setting.

Corollary 9.3.4. *Let A be an $n \times m$ matrix with independent standard Gaussian entries $A_{ij} \sim N(0,1)$. Then $\|A\|_{\mathrm{op}}$ satisfies the conclusions of Theorem 9.3.3 with $\sigma^2 = 1$.*

Proof. Note that the dual norm of $\|\cdot\|_{\mathrm{op}}$ is the trace norm $\|\cdot\|_1$ by Proposition 2.1.17, hence the concentration parameter can be estimated as

$$\sigma^2 = \sup\left\{\mathbb{E}\langle X, A\rangle^2 \;\mid\; \|X\|_1 \le 1\right\} = \sup\left\{\|X\|_F^2 \;\mid\; \|X\|_1 \le 1\right\} = 1,$$

where in the first step we used that random variable $\langle X, A\rangle$ has the same distribution as $\|X\|_F A_{11}$ by orthogonal invariance of Gaussian variables, and in the second step we used that $\|\cdot\|_F \le \|\cdot\|_1$ with equality attained for example by $X = E_{11}$. □

We will also use the multi-argument Hölder inequality given in Theorem 2.1.18:

$$\left|\mathrm{Tr}\prod_{i=1}^{p} A_i\right| \le \prod_{i=1}^{p}\|A_i\|_p, \tag{9.8}$$

where $\|\cdot\|_p = \|\cdot\|_{S_p}$ denotes the Schatten-p-norm according to Definition 2.1.16 and $p \in \mathbb{N}$ with $p \ge 1$ by assumption.

Proof of Theorem 9.3.1. The operator we want to control has entries which are dependent in complicated ways. We first begin with a standard symmetrization trick to linearize the random operator (compare the proof of Lemma 4.1 in [80]). A single entry of $A_i \otimes A_i$ is either a product gg' of two independent standard Gaussians, or the square g^2 of a single standard Gaussian. In expectation, we have $\mathbb{E}gg' = 0$ and $\mathbb{E}g^2 = 1$, so

$$\mathbb{E}\left(\sum_{i=1}^{N} A_i \otimes A_i\right) = N\,\mathrm{vec}(I_n)\,\mathrm{vec}(I_m)^T \quad\Longrightarrow\quad \mathbb{E}\left(\sum_{i=1}^{N} A_i \otimes A_i\right) \circ \Pi = 0,$$

as Π projects orthogonal to $\mathrm{vec}(I_m)$. Therefore we may add an independent copy: let $B_1, \dots, B_N$ be independent $n \times m$ random matrices with standard Gaussian entries that are also independent from $A_1, \dots, A_N$. Then,

$$\left(\sum_{i=1}^{N} A_i \otimes A_i\right) \circ \Pi = \mathbb{E}_B\left(\sum_{i=1}^{N} A_i \otimes A_i - \sum_{i=1}^{N} B_i \otimes B_i\right) \circ \Pi$$

and hence, for any $p \ge 1$,

$$\mathbb{E}_A\left\|\left(\sum_{i=1}^{N} A_i \otimes A_i\right) \circ \Pi\right\|_{\mathrm{op}}^p \le \mathbb{E}_{A,B}\left\|\left(\sum_{i=1}^{N} A_i \otimes A_i - \sum_{i=1}^{N} B_i \otimes B_i\right) \circ \Pi\right\|_{\mathrm{op}}^p$$

by Jensen's inequality applied over $\mathbb{E}_B$ for function $\|\cdot\|_{\text{op}}^p$, which is convex as it is the composition of norm $\|\cdot\|_{\text{op}}$ with the convex and nondecreasing function $x \to x^p$ for $p \geq 1$. Now note (A_i, B_i) has the same joint distribution as $(\frac{A_i+B_i}{\sqrt{2}}, \frac{A_i-B_i}{\sqrt{2}})$ by orthogonal invariance, so the right-hand side is

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{2}\left(\sum_{i=1}^{N}(A_i + B_i) \otimes (A_i + B_i) - \sum_{i=1}^{N}(A_i - B_i) \otimes (A_i - B_i)\right) \circ \Pi\right\|_{\text{op}}^p \\
&= \mathbb{E}\left\|\left(\sum_{i=1}^{N} A_i \otimes B_i + \sum_{i=1}^{N} B_i \otimes A_i\right) \circ \Pi\right\|_{\text{op}}^p \leq 2^p\, \mathbb{E}\left\|\sum_{i=1}^{N} A_i \otimes B_i\right\|_{\text{op}}^p,
\end{aligned}$$

where in the last step we use $\|\Pi\|_{\text{op}} \leq 1$ since it is an orthogonal projection, and use the triangle inequality along with the fact that $A \otimes B$ and $B \otimes A$ are identically distributed. Thus, we have proved that

$$\mathbb{E}\left\|\left(\sum_{i=1}^{N} A_i \otimes A_i\right) \circ \Pi\right\|_{\text{op}}^p \leq 2^p\, \mathbb{E}\left\|\sum_{i=1}^{N} A_i \otimes B_i\right\|_{\text{op}}^p. \tag{9.9}$$

Note that we no longer have the projection, but all products are now independent. Next we use the trace method to bound the right-hand side of Eq. (9.9). That is, we approximate the operator norm by the Schatten p-norm for a large enough p and control these Schatten norms using concentration of moments of Gaussians (compare the proof of Theorem 16.6 in [79]). For any $q \geq 1$,

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{N} A_i \otimes B_i\right\|_{2q}^{2q} &= \mathbb{E}\operatorname{Tr}\left(\sum_{i,j\in[N]} A_i^T A_j \otimes B_i^T B_j\right)^q \\
&= \sum_{i,j\in[N]^q} \mathbb{E}\operatorname{Tr}\left(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q} \otimes B_{i_1}^T B_{j_1} \cdots B_{i_q}^T B_{j_q}\right) \\
&= \sum_{i,j\in[N]^q} \mathbb{E}\operatorname{Tr}\left(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q}\right) \mathbb{E}\operatorname{Tr}\left(B_{i_1}^T B_{j_1} \cdots B_{i_q}^T B_{j_q}\right),
\end{aligned}$$

where the first step was by Definition 2.1.16 of the Schatten norms, and the last step was by independence of $\{A_i\}$ and $\{B_i\}$. Observe that the expectation of a product of independent standard Gaussian random variables is always nonnegative. Thus the same is true for $\mathbb{E}\operatorname{Tr}(A_{i_1}^T A_{j_1}\cdots A_{i_q}^T A_{j_q})$, so we can upper bound the sum term by term as

$$\begin{aligned}
&\sum_{i,j\in[N]^q} \mathbb{E}\operatorname{Tr}\left(A_{i_1}^T A_{j_1}\cdots A_{i_q}^T A_{j_q}\right)\mathbb{E}\operatorname{Tr}\left(B_{i_1}^T B_{j_1}\cdots B_{i_q}^T B_{j_q}\right)\\
&\le \sum_{i,j\in[N]^q} \mathbb{E}\operatorname{Tr}\left(A_{i_1}^T A_{j_1}\cdots A_{i_q}^T A_{j_q}\right)\mathbb{E}\left(\|B_{i_1}\|_{2q}\|B_{j_1}\|_{2q}\cdots\|B_{i_q}\|_{2q}\|B_{j_q}\|_{2q}\right)\\
&\le \mathbb{E}\sum_{i,j\in[N]^q} \operatorname{Tr}\left(A_{i_1}^T A_{j_1}\cdots A_{i_q}^T A_{j_q}\right)\mathbb{E}\left(\|B_1\|_{2q}^{2q}\right)\\
&= \left(\mathbb{E}\|\sum_{i=1}^N A_i\|_{2q}^{2q}\right)\left(\mathbb{E}\|A\|_{2q}^{2q}\right) = N^q\left(\mathbb{E}\|A\|_{2q}^{2q}\right)^2 .
\end{aligned}$$

In the first step we used Hölder's inequality (9.8) for the Schatten norm. The second step holds since $\{B_i\}$ are all mutually independent for $i\neq j$ so we can bound each term in the product by $\mathbb{E}\|B_i\|_{2q}^k \le (\mathbb{E}\|B_i\|_{2q}^{2q})^{\frac{k}{2q}}$ by Jensen's inequality and collect all terms for the product which has total degree $2q$. For the third step, the equality of the first term is by expanding out the sum and considering Definition 2.1.16 of the Schatten-norm, and the equality in the second is because the B_i have the same distribution as A. In the last step, we used that $\sum_{i=1}^N A_i$ has the same distribution as $\sqrt{N}A$. Accordingly, we have proved

$$\mathbb{E}\left\|\sum_{i=1}^N A_i\otimes B_i\right\|_{\mathrm{op}}^{2q} \le \mathbb{E}\left\|\sum_{i=1}^N A_i\otimes B_i\right\|_{2q}^{2q} \le N^q\left(\mathbb{E}\|A\|_{2q}^{2q}\right)^2 \le N^q m^2\Big(\mathbb{E}\|A\|_{\mathrm{op}}^{2q}\Big)^2, \qquad (9.10)$$

where in the third inequality we used that $A\in\operatorname{Mat}(n,m)$ has rank $\le m$, and therefore $\|A\|_{2q}^{2q}\le m\|A\|_{\mathrm{op}}^{2q}$. To bound the right-hand side, we use Theorem 9.3.3 applied to the random variable A in the Banach space $\operatorname{Mat}(n,m)$ with the operator norm $\|\cdot\|_{\mathrm{op}}$. We have $\sigma^2=1$ as computed in Corollary 9.3.4, so we can bound the expectation by

$$\mathbb{E}\left\|\left(\sum_{i=1}^N A_i\otimes A_i\right)\circ\Pi\right\|_{\mathrm{op}}^{2q} \le (4N)^q m^2\Big((2\mathbb{E}\|A\|_{\mathrm{op}})^{2q}+(C\sqrt{q})^{2q}\Big)^2.$$

where $C>0$ is a universal constant implied by the big-O notation in Eq. (9.7). We can use Eq. (9.6) to bound the first term $\mathbb{E}\|A\|_{\mathrm{op}}\le\sqrt{m}+\sqrt{n}$, so choosing $q=2(m+n)$, we can upper bound the mean by

$$\mathbb{E}\left\|\left(\sum_{i=1}^N A_i\otimes A_i\right)\circ\Pi\right\|_{\mathrm{op}}^{2q} \le 4m^2\Big((\max\{2,C\})^2\cdot q\cdot\sqrt{4N}\Big)^{2q}.$$

Finally, we can use Markov's inequality to see that, for $C' = \sqrt{2}\max\{2, C\}$, the event

$$\left\| \left(\sum_{i=1}^{N} A_i \otimes A_i \right) \circ \Pi \right\|_{\mathrm{op}} \leq (C't)^2 \cdot (m+n) \cdot \sqrt{4N} \tag{9.11}$$

holds with failure probability at most

$$4m^2 \left(\frac{(\max\{2, C\})^2 \cdot q \cdot \sqrt{4N}}{(C't)^2 \cdot (m+n) \cdot \sqrt{4N}} \right)^{2q} \leq 4m^2 t^{-2q} \leq t^{-\Omega(m+n)},$$

where the first step was by our choice of $q = 2(m+n)$ and of $C' = \sqrt{2}\max\{2, C\}$, and the final inequality was by the fact that $t \geq 2$, so the $4m^2$ term can be absorbed at the cost of slightly changing the constant in the exponent. □

9.3.2 Net proof of ∞-Expansion

In this subsection, we prove Theorem 9.2.9 showing random Gaussian tensors satisfy the $(\mathfrak{p}, \infty)$-expansion condition of Definition 7.4.4 with high probability. In Theorem 9.2.13, this is combined with the gradient bound in Proposition 9.2.7 to show that the input is pseudorandom, which allows us to apply the fast convergence analysis of Theorem 7.2.26.

The proof uses a slightly non-standard net argument for which we need the following claim.

Claim 9.3.5. *Let $B_{\mathrm{op}} := \{Z \in \mathfrak{spd}(d) \mid \|Z\|_{\mathrm{op}} \leq 1\}$ be the unit ball of $\|\cdot\|_{\mathrm{op}}$ in the subspace $\mathfrak{spd}(d) = \{X \in \mathrm{H}(d) \mid \mathrm{Tr}[X] = 0\}$. Consider the subset of vertices $\mathcal{S}$ described in Fact 2.6.4, i.e. if d is even then $\mathcal{S}$ consists of elements $P - (I_d - P)$ where P is an orthogonal projection of $\mathrm{rk}(P) = \frac{d}{2}$, and if d is odd then $\mathcal{S}$ consists of elements $P - Q$ where both P and Q are orthogonal projections with $\mathrm{rk}(P) = \mathrm{rk}(Q) = \lfloor \frac{d}{2} \rfloor$ such that $PQ = 0$, i.e. their ranges are disjoint.*

For every $\eta > 0$, there is an η-net $N \subseteq \mathcal{S}$ with respect to the operator norm such that $|N| \leq (1 + 2\eta^{-1})^{d^2/2}$. Explicitly, for every $Z \in \mathcal{S}$, there exists $Z' \in N$ such that $\|Z - Z'\|_{\mathrm{op}} \leq \eta$.

Proof. We follow the proof of Lemma 4.10 in [78] given as Fact 2.6.3 in this thesis. Consider $N \subseteq \mathcal{S}$ to be a maximal η-packing of $\mathcal{S}$ with respect to $\|\cdot\|_{\mathrm{op}}$ according to Definition 2.6.2, i.e. for every pair $Z, Z' \in N$, $\|Z - Z'\|_{\mathrm{op}} \geq \eta$. Note that by maximality, this is automatically an η-net of $\mathcal{S}$, as any point not covered by N could be added to the packing, contradicting maximality. Let N_p be a maximum cardinality η-packing of $\mathrm{H}(d)$ with respect to $\|\cdot\|_{\mathrm{op}}$. Then we can bound

$$|N| \leq |N_p| \leq (1+2\eta^{-1})^{d^2/2},$$

where the first step was because $\mathcal{S} \subseteq \mathrm{H}(d)$ so N is an η-packing for $\mathrm{H}(d)$, and the final step was by Fact 2.6.3 applied with $\dim(\mathrm{H}(d)) = \frac{d^2}{2}$. □

For the remainder of the proof, we will focus on the case of even d_b so that $\mathcal{S}_b$ consists of $P-(I_b-P)$ where P is an orthogonal projection in $H(d_b)$ with rank $\mathrm{rk}(P) = \frac{d_b}{2}$. This is to reduce clutter, and the calculation is similar for the odd case.

With this net, we prove the bound on ∞-expansion of random Gaussian tensors.

Proof of Theorem 9.2.9. We will prove the stronger statement that for any $a \neq b \in [m]$ and any $\lambda > 0$ such that $nD \gtrsim \frac{d_a^2 + d_a d_b^2}{\lambda^2}$, x satisfies the λ-$(\mathfrak{p}_{a\leftarrow b}, \infty)$-expansion condition with failure probability at most $\exp(-\Omega(\lambda^2 nD/d_a))$. The theorem follows by setting $\lambda = O(\frac{1}{m})$ and taking a union bound over all pairs $a \neq b \in [m]$.

Fix $a \neq b \in [m]$ and recall that according to Definition 7.4.4, the $(\mathfrak{p}_{a\leftarrow b}, \infty)$-expansion condition is given in terms of a supremum over inner products $\langle \rho_x^{(ab)}, \xi\xi^* \otimes Z\rangle$ for unit vectors $\xi \in S^{d_a-1}$ and elements of $B_{\mathrm{op}} := \{Z \in \mathfrak{spd}(d) \mid \|Z\|_{\mathrm{op}} \leq 1\}$. Further, by Fact 2.6.4, the maximum is achieved at some element of $\mathcal{S}_b$ as described in Claim 9.3.5. Our plan is to use concentration of chi-square variables along with a net argument over S^{d_a-1} and $\mathcal{S}_b$ in order to bound the supremum.

In order to show concentration, first consider a fixed $\xi \in S^{d_a-1}$ and $Z = P-(I_b-P) \in \mathcal{S}_b$ where P and I_b-P both orthogonal projections of $H(d_b)$ with rank $\mathrm{rk}(P) = \mathrm{rk}(I_b-P) = \frac{d_b}{2}$. We will show that the inner product terms with P and $I_b - P$ both concentrate around a common mean, which allows us to bound the difference $Z = P-(I_b-P)$. So for a fixed projection P of rank $\mathrm{rk}(P) = \frac{d_b}{2}$, we can rewrite the term we want to bound as

$$\langle \rho_x^{(ab)}, \xi\xi^* \otimes P\rangle = \frac{1}{nD}\sum_{i=1}^{n} \langle \xi\xi^* \otimes P \otimes I_{\overline{ab}}, X_i X_i^*\rangle,$$

where we used Definition 6.2.2 of the marginal $\rho_x^{(ab)}$ for $x = \frac{1}{\sqrt{nD}}X$. Since $X_1, ..., X_n \sim N(0, I_D)$ are all independent, we can rewrite this in terms of a chi-square random variable $\frac{1}{nD}\chi(\frac{nD}{2d_a})$ according to Definition 2.5.9 as $\xi\xi^* \otimes P \otimes I_{\overline{ab}}$ is an orthogonal projection of rank $\operatorname{rk}(\xi\xi^*)\operatorname{rk}(P)\operatorname{rk}(I_{\overline{ab}}) = 1 \cdot \frac{d_b}{2} \cdot \frac{D}{d_a d_b} = \frac{D}{2d_a}$, so its spectrum is in $\{0,1\}$. Note that $\mathbb{E}\chi(\frac{nD}{2d_a}) = \frac{nD}{2d_a}$ by Definition 2.5.9, so we show concentration for any $0 < \lambda \leq 1$:

$$Pr\Big[|2d_a\langle\rho_x^{(ab)}, \xi\xi^* \otimes P\rangle - 1| \geq 4\lambda\Big] \leq Pr\Big[\Big|\chi\Big(\frac{nD}{2d_a}\Big) - \frac{nD}{2d_a}\Big| \geq 2\lambda\frac{nD}{d_a}\Big] \leq \exp\Big(-\lambda^2\frac{nD}{2d_a}\Big),$$

where we plugged $\theta = \lambda^2\sqrt{\frac{2nD}{d_a}}$ into Theorem 2.5.11 and used $\lambda^2 \leq \lambda \leq 1$.

Now consider an η_a-net $N_a \subseteq S^{d_a-1}$ with respect to $\|\cdot\|_2$ and consider an η_b-net of $\mathcal{S}_b$ with respect to $\|\cdot\|_{\text{op}}$. Choosing $\eta_a = \frac{1}{9}, \eta_b = \frac{1}{3}$, we can bound the size by Fact 2.6.3 and Claim 9.3.5 respectively:

$$|N_a| \leq (1 + 2\eta_a^{-1})^{d_a} \leq e^{3d_a} \qquad \text{and} \qquad |N_b| \leq (1 + 2\eta_b^{-1})^{d_b^2/2} \leq e^{d_b^2}.$$

Now we can apply the union bound to show

$$\begin{aligned}
&Pr\Big[\sup_{\xi\in N_a}\sup_{Z\in N_b}|d_a\langle\rho_x^{(ab)}, \xi\xi^* \otimes Z\rangle| \geq 4\lambda\Big] \\
&\qquad \leq \sum_{\xi\in N_a}\sum_{Z=P-(I_b-P)\in N_b} Pr\Big[\max\Big\{|d_a\langle\rho_x^{(ab)}, \xi\xi^* \otimes P\rangle|, |d_a\langle\rho_x^{(ab)}, \xi\xi^* \otimes (I_b - P)\rangle|\Big\} \geq 2\lambda\Big] \\
&\qquad \leq (|N_a|)(2|N_b|)\exp\Big(-\lambda^2\frac{nD}{2d_a}\Big) \leq 2\exp\Big(3d_a + d_b^2 - \lambda^2\frac{nD}{2d_a}\Big),
\end{aligned}$$

where the first step was by the union bound over all $\xi \in N_a$ and both P and $I_b - P$ for $Z = P - (I_b - P) \in N_b$, in the second step we applied the concentration bound for each individual unit vector and projection calculated above, and in the final step we used the bounds $|N_a| \leq e^{3d_a}$ and $|N_b| \leq e^{d_b^2}$. Note that this is only non-trivial when $nD \gtrsim \frac{d_a(d_a+d_b^2)}{\lambda^2}$.

Now assume we are in the event where the above bound holds for every $\xi \in N_a$ and $P, I_b - P$ for $Z \in N_b$. In order to bound the supremum over all S^{d_a-1} and $B_{\text{op}} := \{Z \in \mathfrak{spd}(d_b) \mid \|Z\|_{\text{op}} \leq 1\}$, we use an approximation argument. We proceed one argument at a time, so first consider fixed Z and note that we can rewrite $\langle\rho_x^{(ab)}, \xi\xi^* \otimes Z\rangle = \langle\xi\xi^*, \Phi_x^{(ab)}(Z)\rangle$ by Proposition 2.4.5. This allows us to apply Lemma 2.6.5 with Hermitian $\Phi_x^{(ab)}(Z)$ and η_a-net N_a to show

$$\sup_{\xi\in S^{d_a-1}}\langle\rho_x^{(ab)}, \xi\xi^* \otimes Z\rangle = \|\Phi_x^{(ab)}(Z)\|_{\text{op}} \leq (1 - 2\eta_a - \eta_a^2)^{-1}\sup_{\xi\in N_a}\langle\xi\xi^*, \Phi_x^{(ab)}(Z)\rangle \leq \frac{3}{2}\langle\rho_x^{(ab)}, \xi\xi^* \otimes Z\rangle,$$

where the first step was by definition of $\|\cdot\|_{\text{op}}$, the second step was by the approximation argument in Lemma 2.6.5, and in the final step we substituted in $\eta_a = \frac{1}{9}$ and again used Proposition 2.4.5 to translate back to $\rho_x^{(ab)}$.

Similarly, fix $\xi \in S^{d_a-1}$ and consider $Z := \arg\max_{Y \in B_{\text{op}}} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Y \rangle$. By Fact 2.6.4 applied to the vertices of B_{op}, we can assume $Z \in \mathcal{S}_b$. Further, by the property of η_b-net N_b, we can decompose $Z = Z' + Z''$ with $Z' \in N_b$ and $Z'' \in \eta_b B_{\text{op}}$, i.e. $\text{Tr}[Z''] = \text{Tr}[Z - Z'] = 0$ and $\|Z''\|_{\text{op}} \leq \eta_b$. Here, we crucially used that the optimizer must be in $\mathcal{S}_b$, so that we can approximate it using an element of $N_b \subseteq \mathcal{S}_b$. This gives a quantitative improvement for the net argument as the inner product term involving elements of $\mathcal{S}_b$ have much better concentration properties than those with an arbitrary element of B_{op}. Then, we bound

$$\langle \rho_x^{(ab)}, \xi\xi^* \otimes Z \rangle = \langle \rho_x^{(ab)}, \xi\xi^* \otimes (Z' + Z'') \rangle \leq \sup_{Y' \in N_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Y' \rangle + \eta_b \sup_{Y \in B_{\text{op}}} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Y \rangle.$$

Since Z is the maximizer over B_{op}, we can rearrange this to give

$$\sup_{Y \in B_{\text{op}}} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Y \rangle \leq (1 - \eta_b)^{-1} \sup_{Y \in N_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Y \rangle. \tag{9.12}$$

Combining both approximation arguments, we can bound the supremum by

$$\sup_{\xi \in S^{d-1}} \sup_{Z \in B_{\text{op}}} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Z \rangle \leq (1 - 2\eta_a - \eta_a^2)^{-1} (1 - \eta_b)^{-1} \sup_{\xi \in N_a} \sup_{Z \in N_b} \langle \rho_x^{(ab)}, \xi\xi^* \otimes Z \rangle \leq \Big(\frac{3}{2}\Big)\Big(\frac{3}{2}\Big) 4\lambda,$$

where the first step was by our two approximations above, in the second step we substituted $\eta_a = \frac{1}{9}, \eta_b = \frac{1}{3}$, and in the final step we used the bound derived above for the supremum over N_a, N_b. Since B_{op} is symmetric around the origin, this verifies Definition 7.4.4 of 9λ-$(\mathfrak{p}_{a \leftarrow b}, \infty)$-expansion, and we are done by the union bound over all $a \neq b \in [m]$. □

Chapter 10

Conclusions and Future Work

In this thesis, we studied problems from the scaling framework and leveraged the perspective of geodesic convex optimization [20] in order to give stronger analyses of instances satisfying certain strong convexity and pseudorandom conditions. This allowed us to unify the work of [62], [63], and [36] for special cases of these problems as well as improve many of the bounds.

The main motivations for our improved scaling analyses were the Paulsen problem in frame theory and the tensor normal model in statistics. For the Paulsen problem, we were able to follow and refine the smoothed analysis approach of [62] by randomly perturbing the input frame, and then showing fast convergence for the frame scaling problem with high probability. We believe this kind of regularization technique to fast convergence is a general approach to scaling problems that is of independent interest.

For the tensor normal model, we generalized our matrix scaling analysis to higher order tensors, and gave strong bounds on the scaling solution for inputs satisfying strong convexity or pseudorandom conditions. We then showed that computing the maximum likelihood estimator for the tensor normal model could be reduced to solving a tensor scaling problem on random instances which satisfied these fast convergence conditions with high probability. Therefore, our main results in this section were to show sample complexity and error bounds for the general tensor normal model that nearly matched the known lower bounds for the much simpler Gaussian model.

We were also able to leverage the analysis of Franks and Moitra [35] for geodesic convex optimization algorithms to prove exponential convergence guarantees for the scaling problems above. Therefore, we were able to prove that natural iterative algorithms can be applied to make our results for the Paulsen problem and the tensor normal model constructive.

The set of scaling problems studied in this thesis are just a small sampling of the general framework for scaling presented in [20]. We believe that this perspective of scaling and geodesic convex optimization will have many more applications throughout theoretical computer science, and we hope the techniques presented in this thesis will be of use for analyzing more difficult problems in this area. Below, we present some general directions for future research.

- The previous approaches of [16] and [23] for the Paulsen problem involved fixing one of the two balanced conditions (Parseval or equal-norm according to Definition 4.1.2), and then applying some procedure to fix the other condition. It turns out that both of these procedures remain within a natural group orbit of the input frame, and therefore can be profitably understood using the perspective of group scaling. This allows us to give a principled derivation of the results in [16] and [23], and hopefully will suggest new algorithms for the general scaling framework.

- We can generalize the Paulsen problem to any problem in the scaling framework. Given any group scaling setting and a nearly balanced input, measure the distance to the nearest balanced input. The operator version of the Paulsen problem is a natural next step for applying our smoothed analysis and scaling techniques. Here, we present a very similar problem which is motivated by fundamental problems in numerical linear algebra. Recall that a matrix $A \in \mathbb{C}^{n \times n}$ is diagonalizable iff it can be written $A = VDV^{-1}$ for some invertible V and diagonal D. In this case, the columns of V, V^{-1} are the eigenvectors of the matrix acting on the left and right, respectively, and are non-unique in general. Diagonalization is a fundamental operation used as a subroutine in innumerable applications, and in practice it is only performed to some required precision. The condition number, defined below, gives a natural measure of the robustness of the diagonalization procedure to error in the input.

  $$\kappa(A) := \inf_{A = VDV^{-1}} \|V\|_{\text{op}} \|V^{-1}\|_{\text{op}}.$$

  It is therefore natural to ask whether a given input A can be made robust to noise by a small perturbation.

  Question 10.0.1. *For any $A \in \mathbb{C}^{n \times n}$ with $\|A\|_{\text{op}} \leq 1$, is there a nearby $B := A + E$ such that $\kappa(B)$ is small?*

The recent works of [8], [9], [55] show that a small perturbation suffices in order to achieve a polynomial condition number. This leaves open the regime where the condition number of the output B is close to 1. We observe that the quantity κ can be seen as a bound on the solution to a scaling problem, specifically the action $g \in \mathrm{SL}(n) \to gAg^{-1}$. Therefore, we can hope to use the techniques developed for the Paulsen problem to give new regularization theorems and a refined analysis of Question 10.0.1 when the input is close to normal.

Conjecture 10.0.2. *Given matrix $A \in \mathrm{Mat}_{\mathbb{C}}(n)$ such that $\|A\|_F^2 = 1$ and $\|AA^* - A^*A\|_{\mathrm{op}} \leq \frac{\varepsilon}{n}$, for any $\delta \gtrsim \varepsilon$, there is a perturbation $B = A + E$ such that*

$$\|E\|_F^2 \lesssim \delta \qquad \text{and} \qquad \kappa(B) \lesssim \frac{\varepsilon}{\delta}.$$

- The constructive procedures given in Section 8.5 for the Paulsen problem in the worst case relied on a random perturbation in order to guarantee fast convergence. There are many applications of scaling (including the diagonalization procedure above), for which deterministic algorithms are required. It is therefore an interesting questions whether the smoothed analysis strategy proven in Chapter 5 can be replaced with a deterministic one while maintaining the strong convergence properties of the output.

- While the tensor normal model in Definition 9.2.1 is a natural assumption for tensor valued data that is used in pratice, it is merely the simplest such assumption that can be profitably used to reduce sample complexity. In Chapter 9, we were able to analyze the MLE for this model only because of its strong connection to the tensor scaling problem (Definition 6.2.5), and specifically the Kempf-Ness function in Definition 6.2.9 which gave a geodesically convex formulation for this problem. It would be interesting to see whether these techniques can be generalized to even slightly more complex statistical models. For example, the covariance matrix can be assumed to be of the form $\Theta = \sum_{k=1}^{K} \Theta^k$ where each $\Theta^k = \otimes_{a \in [m]} \Theta_a^k$ respects the tensor product structure. Similarly, the input distribution could be assumed to be a mixture of a small number of distributions from the tensor normal model.

- The works of Bürgisser et al. [19], [20] are especially amazing because they simultaneously give a foundation to analyze a huge swath of problems from the scaling framework. Of these, tensor scaling is one of the simplest settings where efficient algorithms are not known in the worst case. The algorithms of [19] and [20] actually apply to the much more general problem of moment polytope membership testing, albeit with much worse parameters of convergence. It would be interesting to see how far the techniques in this thesis can be pushed, and for which problems it is possible to give natural sufficient conditions for beyond worst-case results.

- Even more generally, the scaling framework has recently given much motivation for geodesic convex optimization. This is a natural optimization setting in its own right which can be seen as an extension of the known tractability results for convex optimization. We believe it is an interesting future direction to develop general-purpose algorithms for the geodesic setting.

References

[1] Genevera I Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764, 2010.

[2] Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson. Much faster algorithms for matrix scaling. In *2017 IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017.

[3] Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2018.

[4] Carlos Améndola, Kathlén Kohn, Philipp Reichenbach, and Anna Seigal. Invariant theory and scaling algorithms for maximum likelihood estimation. *arXiv preprint arXiv:2003.13662*, 2020.

[5] M. Appleby, S. Flammia, G. McConnell, and J. Yard. SICs and algebraic number theory. *Foundations of Physics*, 47, 2017.

[6] Michael Atiyah. Convexity and commuting Hamiltonians. *Bulletin of the London Mathematical Society*, 14, 1982.

[7] Sheldon Axler. *Linear Algebra Done Right*. Springer, 1997.

[8] Jess Banks, Archit Kulkarni, Satyaki Mukherjee, and Nikhil Srivastava. Gaussian regularization of the pseudospectrum and Davies' conjecture. *arXiv preprint arXiv:1906.11819*, 2020.

[9] Jess Banks, Jorge Garza Vargas, Archit Kulkarni, and Nikhil Srivastava. Overlaps, eigenvalue gaps, and pseudospectrum under real Ginibre and absolutely continuous perturbations. *arXiv preprint arXiv:2005.08930*, 2020.

[10] Frank Barthe. On a reverse form of the Brascamp-Lieb inequality. *Inventiones mathematicae*, 134(2), 1998.

[11] N. Berline and M. Vergne. *Hamiltonian manifolds and moment map.* 2011.

[12] Rajendra Bhatia. *Matrix Analysis.* Springer, 1997.

[13] Rajendra Bhatia. *Positive Definite Matrices.* Princeton University Press, 2007.

[14] E. Bierstone and P. Milman. Semianalytic and subanalytic sets. *Inst. Hautes Etudes Sci. Publ. Math.*, 67, 1988.

[15] Yonatan Bilu and Nathan Linial. Lifts, discrepancy and nearly optimal spectral gap. *Combinatorica*, 26, 2006.

[16] Bernhard G. Bodmann and Peter G. Casazza. The road to equal-norm Parseval frames. *Journal of Functional Analysis*, 258(2):397–420, 2010.

[17] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[18] Peter Bürgisser, Matthias Christandl, Ketan Mulmuley, and Michael Walter. Membership in moment polytopes is in NP and coNP. *SIAM Journal on Computing*, 46, 2017.

[19] Peter Bürgisser, Cole Franks, Ankit Garg, Rafael Oliveira, Michael Walter, and Avi Wigderson. Efficient algorithms for tensor scaling, quantum marginals, and moment polytopes. In *2018 IEEE Symposium on Foundations of Computer Science (FOCS)*, 2018.

[20] Peter Bürgisser, Cole Franks, Ankit Garg, Rafael Oliveira, Michael Walter, and Avi Wigderson. Towards a theory of non-commutative optimization: geodesic 1st and 2nd order methods for moment maps and polytopes. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 845–861. IEEE, 2019.

[21] Peter Bürgisser, Yinan Li, Harold Nieuwboer, and Michael Walter. Interior-point methods for unconstrained geometric programming and scaling problems. *arXiv preprint arXiv:2008.12110*, 2020.

[22] Jameson Cahill and Peter Casazza. The Paulsen problem in operator theory. *Operators and Matrices*, 2013.

[23] Peter G. Casazza, Matthew Fickus, and Dustin G. Mixon. Auto-tuning unit norm frames. *Applied and Computational Harmonic Analysis*, 32(1):1–15, 2012.

[24] Peter G. Casazza and Gitta Kutyniok, editors. *Finite Frames: Theory and Applications*. Birkhauser Basel, 2013.

[25] P.G. Casazza. The Kadison-Singer and Paulsen problems in finite frame theory. *Finite frames: theory and applications*, 2013.

[26] Michael B. Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *2017 IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017.

[27] M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *26th International Conference on Neural Information Processing Systems (NIPS)*, 2013.

[28] E. B. Davies. Approximate diagonalization. *Journal of Matrix Analysis Applications*, 2007.

[29] Harm Derksen and Visu Makam. Maximum likelihood estimation for matrix normal models via quiver representations. *arXiv preprint arXiv:2007.10206*, 2020.

[30] Harm Derksen, Visu Makam, and Michael Walter. Maximum likelihood estimation for tensor normal models via castling transforms. *arXiv preprint arXiv:2011.03849*, 2020.

[31] Pierre Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

[32] K. Dykema and N. Strawn. Manifold structure of spaces of spherical tight frames. *International Journal of Pure and Applied Mathematics*, 28, 2006.

[33] Hamza Fawzi and James Saunderson. Lieb's concavity theorem, matrix geometric means, and semidefinite optimization. *arXiv preprint arXiv:1512.03401*, 2016.

[34] Jürgen Förster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65, 2002.

[35] Cole Franks and Ankur Moitra. Rigorous guarantees for Tyler's M-estimator via quantum expansion. *arXiv preprint arXiv:2002.00071*, 2020.

[36] Cole Franks, Rafael Oliveira, Akshay Ramachandran, and Michael Walter. Logarithmic sample complexity for dense matrix and tensor normal models. *arXiv preprint arXiv:2110.07583*, 2021.

[37] Cole Franks and Philipp Reichenbach. Barriers for recent methods in geodesic optimization. In *36th Computational Complexity Conference (CCC)*, 2021.

[38] Ankit Garg, Leonid Gurvits, Rafael Mendes de Oliveira, and Avi Wigderson. A deterministic polynomial time algorithm for non-commutative rational identity testing. *arXiv preprint arXiv:1511.03730*, 2015.

[39] Ankit Garg, Leonid Gurvits, Rafael Mendes de Oliveira, and Avi Wigderson. Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via operator scaling. *Geometric and Functional Analysis*, 28, 2018.

[40] V. Georgoulas, J. Robbin, and D. Salamon. *The moment-weight inequality and the Hilbert-Mumford criterion.* arXiv preprint arXiv:1311.0410, 2018.

[41] V. Guillemin and R. Sjamaar. *Convexity Properties of Hamiltonian Group Actions.* AMS, 2005.

[42] V. Guillemin and S. Sternberg. Convexity properties of the moment mapping. *Inventiones mathematicae*, 67, 1982.

[43] V. Guillemin and S. Sternberg. Convexity properties of the moment mapping. ii. *Inventiones mathematicae*, 77, 1984.

[44] L. Gurvits and A. Samorodnitsky. A deterministic polynomial-time algorithm for approximating mixed discriminant and mixed volume. In *32nd Annual ACM Symposium on Theory of Computing (STOC)*, 2000.

[45] Leonid Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 2004.

[46] Linus Hamilton and Ankur Moitra. The Paulsen problem made simple. In *Innovations in Theoretical Computer Science (ITCS)*, 2019.

[47] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *26th Annual Conference on Learning Theory (COLT)*, 2013.

[48] G. H. Hardy, J. E. Littlewood, and G. Pòlya. *Inequalities.* Cambridge University Press, 1988.

[49] R.B. Holmes and V.I. Paulsen. Optimal frames for erasures. *Linear Algebra and its Applications*, 2004.

[50] Max Hopkins, Daniel Kane, Shachar Lovett, and Gaurav Mahajan. Point location and active learning: Learning halfspaces almost optimally. In *e 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2020.

[51] Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76, 1954.

[52] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 2013.

[53] James Humphreys. *Introduction to Lie Algebras and Representation Theory.* Springer-Verlag, 1972.

[54] M. Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

[55] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. On the real Davies' conjecture. *arXiv preprint arXiv:abs/2005.08908*, 2020.

[56] Ilya Kachkovskiy and Yuri Safarov. Distance to normal elements in C^*-algebras of real rank zero. *Journal of the American Mathematical Society*, 29, 2016.

[57] B. Kalantari and L. Khachiyan. On the complexity of nonnegative-matrix scaling. *SIAM Journal on Matrix Analysis and Applications*, 18, 1997.

[58] George Kempf and Linda Ness. The length of vectors in representation spaces. In Knud Lønsted, editor, *Algebraic Geometry*, pages 233–243, Berlin, Heidelberg, 1979. Springer Berlin Heidelberg.

[59] Frances Kirwan. Convexity properties of the moment mapping. iii. *Inventiones mathematicae*, 77:547–552, 1984.

[60] V. M. Kravtsov. Combinatorial properties of noninteger vertices of a polytope in a threeindex axial assignment problem. *Cybernetics and Systems Analysis*, 43, 2007.

[61] R. A. Kunze. L_p Fourier transforms on locally compact unimodular groups. *Transactions of the American Mathematical Society*, 89, 1958.

[62] T.C. Kwok, L.C. Lau, Y.T. Lee, and A. Ramachandran. The Paulsen problem, continuous operator scaling, and smoothed analysis. In *Symposium on Theory of Computing (STOC)*. ACM, 2017.

[63] T.C. Kwok, L.C. Lau, Y.T. Lee, and A. Ramachandran. Spectral analysis of matrix scaling and operator scaling. *SIAM Journal of Computing*, 50, 2021.

[64] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000.

[65] Eugene Lerman. Gradient flow of the norm squared of a moment map. *L'Enseignement Mathématique*, 51, 2004.

[66] Nathan Linial, Alex Samorodnitsky, and Avi Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. *Combinatorica*, 20(4):545–568, 2000.

[67] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[68] Ameur M Manceur and Pierre Dutilleul. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49, 2013.

[69] Kanti V Mardia and Colin R Goodall. Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate environmental statistics*, 6(76):347–385, 1993.

[70] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70, 2010.

[71] Gary L. Miller, Noel J. Walkington, and Alex L. Wang. Hardy-Muckenhoupt bounds for Laplacian eigenvalues. *arXiv preprint arXiv:1812.02841*, 2018.

[72] Ketan Mulmuley. Geometric complexity theory v: Efficient algorithms for Noether normalization. *Journal of the American Mathematical Society*, 20, 2017.

[73] David Mumford, John Fogarty, and Frances Kirwan. *Geometric Invariant Theory.* Springer, 1994.

[74] Tom Needham and Clayton Shonkwiler. Symplectic geometry and connectivity of spaces of frames. *Advances in Computational Mathematics*, 2021.

[75] Yuri Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Publishing Company, 2014.

[76] E. E. Osborne. On pre-conditioning of matrices. *Journal of ACM*, 7, 1960.

[77] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. *Letta G., Pratelli M. (eds) Probability and Analysis*, 1206, 1986.

[78] Gilles Pisier. *The volume of convex bodies and Banach space geometry.* Cambridge University Press, 1989.

[79] Gilles Pisier. Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society*, 49(2):237–323, 2012.

[80] Gilles Pisier. Quantum expanders and geometry of operator spaces. *Journal of the European Mathematical Society*, 16(6):1183–1219, 2014.

[81] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18, 2013.

[82] Jean-Pierre Serre. *Lie Algebras and Lie Groups.* Springer, 1964.

[83] Richard Sinkhorn. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics*, 35(2):876 – 879, 1964.

[84] Daniel Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 2011.

[85] Daniel Spielman, Nikhil Srivastava, and Adam Marcus. Interlacing families i: Bipartite Ramanujan graphs of all degrees. *Annals of Mathematics*, 182, 2015.

[86] Daniel Spielman, Nikhil Srivastava, and Adam Marcus. Interlacing families iv: Bipartite Ramanujan graphs of all sizes. *SIAM Journal on Computing*, 47, 2015.

[87] Daniel Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal of Matrix Analysis and Applications*, 2014.

[88] Damian Straszak and Nisheeth Vishnoi. Maximum entropy distributions: Bit complexity and stability. *arXiv preprint arXiv:1711.02036*, 2019.

[89] Wei Sun, Zhaoran Wang, Han Liu, and Guang Cheng. Non-convex statistical optimization for sparse tensor graphical model. *Advances in Neural Information Processing Systems*, 28, 2015.

[90] Terence Tao. *Topics in random matrix theory.* American Mathematical Society, 2012.

[91] J.A. Tropp, I.S. Dhillon, R.W. Heath Jr., and T. Strohmer. Designing structured tight frames via an alternating projection method. *IEEE Transactions on Information Theory*, 51, 2005.

[92] Theodoros Tsiligkaridis, Alfred O III Hero, and Shuheng Zhou. On convergence of Kronecker graphical lasso algorithms. *IEEE Transactions on Signal Processing*, 61(7), 2013.

[93] Joran van Apeldoorn, Sander Gribling, Yinan Li, Harold Nieuwboer, Michael Walter, and Ronald de Wolf. Quantum algorithms for matrix scaling and matrix balancing. *arXiv preprint arXiv:2011.12823*, 2021.

[94] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[95] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[96] Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press, 2019.

[97] N. R. Wallach. *Geometric Invariant Theory: Over the Real and Complex Numbers.* Springer, 2017.

[98] John Watrous. *The Theory of Quantum Information.* Cambridge University Press, 2018.

[99] Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, 2008.

[100] Ami Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

[101] Shuheng Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

[102] A. Zygmund. *Trigonometric series*. Cambridge University Press, 2002.

Appendix A

Supplementary Proofs

A.1 Lower Bound Example for the Paulsen Problem

The exponent of ε is best possible due to the following example.

Example A.1.1 (Example 9 in [23])**.** *For a fixed $\theta \in [0, \pi]$ (small enough) consider the following frame $U \in \mathrm{Mat}(2, 4)$:*

$$U := \begin{pmatrix} \cos(2\theta) & \cos(2\theta) & 0 & 0 \\ \sin(2\theta) & -\sin(2\theta) & 1 & 1 \end{pmatrix}.$$

It can be shown that the nearest doubly balanced frame to U is

$$V := \begin{pmatrix} \cos(\theta) & \cos(-\theta) & \sin(-\theta) & \sin(\theta) \\ \sin(\theta) & \sin(-\theta) & \cos(-\theta) & \cos(\theta) \end{pmatrix}.$$

Therefore, U is an equal-norm ε-Parseval frame with $\varepsilon \lesssim \sin^2 \theta$ such that the minimum distance from U to a doubly balanced frame is $\|V - U\|_F^2 \gtrsim \sin^2 \theta \gtrsim \varepsilon$.

We will show the bounds for the example via the following three claims.

Claim A.1.2. *U is an equal-norm ε-Parseval frame with $\varepsilon \lesssim \sin^2 \theta$.*

Proof. U is clearly equal-norm by construction. To show the Parseval condition, we calculate outer products to show

$$\sum_{j=1}^{4} u_j u_j^* = 2 \begin{pmatrix} \cos^2(2\theta) & 0 \\ 0 & 1 + \sin^2(2\theta) \end{pmatrix} \implies \|UU^* - 2I_2\|_{op} = 2\sin^2(2\theta) \approx 8\sin^2(\theta),$$

where the last approximation is $\sin(2\theta) = 2\sin(\theta)\cos(\theta) \approx 2\sin(\theta)$, when $\theta \leq \frac{1}{4}$ is small. □

Claim A.1.3. *Any doubly balanced frame $V \in \mathrm{Mat}(d, n)$ for $d = 2, n = 4$ must be the union of two orthonormal bases.*

Proof. The claim follows by some straightforward (tedious) calculations. We assume, by a change of basis if necessary, that V is of the following form

$$(v_1, v_2, v_3, v_4) = \begin{pmatrix} 1 & \alpha & \beta & \gamma \\ 0 & \sqrt{1-\alpha^2} & \sqrt{1-\beta^2} & \sqrt{1-\gamma^2} \end{pmatrix},$$

for some $\alpha, \beta, \gamma \in [-1, 1]$. By assumption $VV^* = 2I_2$, so the top left entry of this equation gives the constraint

$$2 = 1 + \alpha^2 + \beta^2 + \gamma^2 \implies \gamma^2 = 1 - \alpha^2 - \beta^2,$$

so we can write $v_4 = (\pm\sqrt{1-\alpha^2-\beta^2}, \sqrt{\alpha^2+\beta^2})$. By the bottom right entry, we get

$$2 = (1-\alpha^2) + (1-\beta^2) + (1-\gamma^2) = 2 - \alpha^2 - \beta^2 + \alpha^2 + \beta^2,$$

which is a redundant equation. The off diagonal entries give the equation

$$\begin{aligned} 0 &= \alpha\sqrt{1-\alpha^2} + \beta\sqrt{1-\beta^2} + \gamma\sqrt{1-\gamma^2} \\ \pm\sqrt{(1-\alpha^2-\beta^2)(\alpha^2+\beta^2)} &= \alpha\sqrt{1-\alpha^2} + \beta\sqrt{1-\beta^2} \\ (1-\alpha^2-\beta^2)(\alpha^2+\beta^2) &= \alpha^2(1-\alpha^2) + \beta^2(1-\beta^2) + 2\alpha\beta\sqrt{(1-\alpha^2)(1-\beta^2)} \\ -2\alpha^2\beta^2 &= 2\alpha\beta\sqrt{(1-\alpha^2)(1-\beta^2)}. \end{aligned}$$

Note that if either $\alpha = 0$ or $\beta = 0$, we are done since (v_1, v_2) or (v_1, v_3) is an orthonormal basis, so in order for the frame to be Parseval, the other pair must be as well. So we can continue by canceling $2\alpha\beta$ from both sides to show

$$-\alpha\beta = \sqrt{(1-\alpha^2)(1-\beta^2)} \iff \alpha^2\beta^2 = 1 - \alpha^2 - \beta^2 + \alpha^2\beta^2,$$

which shows $\gamma^2 = 1 - \alpha^2 - \beta^2 = 0$. So (v_1, v_4) is an orthonormal basis, and therefore so must the other pair be. □

Claim A.1.4. *The nearest doubly balanced frame to U is*

$$V := \begin{pmatrix} \cos(\theta) & \cos(-\theta) & \sin(-\theta) & \sin(\theta) \\ \sin(\theta) & \sin(-\theta) & \cos(-\theta) & \cos(\theta) \end{pmatrix},$$

and the distance is $\|V - U\|_F^2 \gtrsim \sin^2\theta$.

Proof. If $\{(v_1, v_2), (v_3, v_4)\}$ are the bases, then for small θ the distance $\|V - U\|_F^2$ will be $\Omega(1)$ since one of (u_1, v_1) or (u_2, v_2) will be nearly orthogonal. Since $u_3 = u_4$ we can assume wlog that $\{(v_1, v_3), (v_2, v_4)\}$ are the pairs of orthonormal bases, and so the nearest doubly balanced V is of the form

$$V := \begin{pmatrix} \cos(\phi) & \cos(\psi) & -\sin(\phi) & \sin(\psi) \\ \sin(\phi) & -\sin(\psi) & \cos(\phi) & \cos(\psi) \end{pmatrix}.$$

Since $u_3 = u_4$ and u_1, u_2 are symmetric across the x-axis, we can for now consider

$$\begin{aligned} \|v_1 - u_1\|_2^2 + \|v_3 - u_3\|_2^2 &= 2(1 - \langle v_1, u_1\rangle) + 2(1 - \langle v_3, u_3\rangle) \\ &= 4 - 2(\cos(\phi)\cos(2\theta) + \sin(\phi)\sin(2\theta) - 0 + \cos(\phi)), \end{aligned}$$

where in the first step we used that all vectors are unit norm. To minimize the distance, we would like to maximize the term in paranthesis, so we continue

$$\begin{aligned} \cos(\phi)(1 + \cos(2\theta)) + \sin(\phi)\sin(2\theta) &= \cos(\phi)(1 + \cos^2(\theta) - \sin^2(\theta)) + 2\sin(\phi)\sin(\theta)\cos(\theta) \\ &= 2\cos(\theta)(\cos(\phi)\cos(\theta) + \sin(\phi)\sin(\theta)), \end{aligned}$$

which is maximized when $\phi = \theta$. Arguing symmetrically for $\|v_2 - u_2\|_2^2 + \|v_4 - u_4\|_2^2$ we get that the nearest doubly balanced frame to U is the following:

$$V := \begin{pmatrix} \cos(\theta) & \cos(-\theta) & \sin(-\theta) & \sin(\theta) \\ \sin(\theta) & \sin(-\theta) & \cos(-\theta) & \cos(\theta) \end{pmatrix}.$$

We can lower bound this distance by just considering the last two vectors:

$$\|V - U\|_F^2 \geq \|v_3 - u_3\|_2^2 + \|v_4 - u_4\|_2^2 \geq 2\sin^2(\theta).$$

□

Putting these claims together, we see that U is an ε-doubly balanced frame for which the nearest doubly balanced frame V satisfies the distance bound $\|V - U\|_F^2 \gtrsim \varepsilon$. Therefore, the distance function $p(d, n, \varepsilon)$ in Conjecture 4.1.4 must depend linearly on ε, and the ε^2 results of [16] and [23] cannot hold for general d, n.

A.2 Tightness of Commutative Robustness

Here we show that the function $e^{-\|(X,Y)\|_\infty}$ in Lemma 3.2.4 cannot be improved in general.

Example A.2.1. *Consider $A := \frac{1}{\sqrt{dn}} J_{dn} \in \mathrm{Mat}(d,n)$ where J is the all-ones matrix. Then A is 1-strongly convex according to Definition 3.2.1.*

On the other hand, for any $S \in \binom{[d]}{d/2}$, consider diagonal scaling $X' := \mathrm{diag}(1_S - 1_{\overline{S}})$ with $Tr[X'] = |S| - |\overline{S}| = 0$ and $\|X\|_{\mathrm{op}} = \|1_S - 1_{\overline{S}}\|_\infty = 0$. Then the scaling $B := e^{X'}A$ is at most $e^{-2\|X'\|_{\mathrm{op}}} = e^{-2}$-strongly convex.

Proof. Recall that by Definition 3.2.1, to show strong convexity we would like to lower bound the following quadratic form for every $(X,Y) \in \mathfrak{t}$:

$$\begin{aligned}
\partial^2_{\delta=0} f_A(\delta X, \delta Y) &= \sum_{i=1}^{d} \sum_{j=1}^{n} |A_{ij}|^2 (X_i + Y_j)^2 \\
&= \frac{1}{dn} \sum_{i=1}^{d} X_i^2 \sum_{j=1}^{n} 1 + \frac{2}{dn} \sum_{i=1}^{d} \sum_{j=1}^{n} X_i Y_j + \frac{1}{dn} \sum_{j=1}^{n} Y_j^2 \sum_{i=1}^{d} 1 \\
&= \frac{1}{d} \sum_{i=1}^{d} X_i^2 + \frac{2}{dn} \left(\sum_{i=1}^{d} X_i \right) \left(\sum_{j=1}^{n} Y_j \right) + \frac{1}{n} \sum_{j=1}^{n} Y_j^2 = \|(X,Y)\|_{\mathfrak{t}}^2,
\end{aligned}$$

where the first line was by Definition 3.2.1, the second was by definition $A_{ij} = 1/\sqrt{dn}$, and the cross-term in the final step vanished because $(X,Y) \in \mathfrak{t}$ so $\sum_{i=1}^{d} X_i = \sum_{j=1}^{n} Y_j = 0$.

To upper bound the strong convexity of $B = e^{X'}A$, consider arbitrary $(X,0) \in \mathfrak{t}$ such that $X_{i \in S} = 0$. Then we calculate

$$\begin{aligned}
\partial^2_{\delta=0} f_B(\delta X, \delta Y) &= \sum_{i \in S} \sum_{j=1}^{n} e^2 |A_{ij}|^2 (X_i + Y_j)^2 + \sum_{i \notin S} \sum_{j=1}^{n} e^{-2} |A_{ij}|^2 (X_i + Y_j)^2 \\
&= \frac{e^{-2}}{dn} \sum_{i=1}^{d} \sum_{j=1}^{n} (X_i + Y_j)^2 + \frac{e^2 - e^{-2}}{dn} \sum_{i \in S} \sum_{j=1}^{n} e^2 |A_{ij}|^2 (X_i + Y_j)^2 \\
&= \frac{e^{-2}}{d} \sum_{i=1}^{d} X_i^2 + 0 = e^{-2} \|(X,Y)\|_{\mathfrak{t}}^2,
\end{aligned}$$

where the first line was by our choice of scaling $X' = 1_S - 1_{\overline{S}}$, and in the third line the first term was calculated above and the second term vanishes by our choice $X_{i \in S} = Y_{j \in [n]} = 0$.

Note that $\|\delta\|_\infty = 1$, and the same argument can be applied by interchanging the roles of d, n. Therefore Definition 3.2.1 is tight in general. □

It is clear that A is doubly balanced. By the convex formulation for matrix scaling in Proposition 3.1.10, this implies $s(B) \geq s(A)$. This means that this also provides robustness in terms of α/s.

This also gives the same lower bound for robustness of pseudorandom property, i.e. shows that Lemma 3.3.4 is tight.

A.3 Alternate Scaling Algorithm

Here we prove the well-known properties of the alternate scaling algorithm in Eq. (4.2) for nearly doubly balanced frames.

Fact A.3.1 (Restatement of Fact 4.1.5)**.** *For any input frame $U \in \mathrm{Mat}(d,n)$ with $s(U) = 1$, the two transformations*

$$\tilde{u}_j := \frac{u_j}{\sqrt{n}\|u_j\|_2}, \qquad \tilde{u}_j := \left(d \sum_{j=1}^{n} u_j u_j^* \right)^{-\frac{1}{2}} u_j \tag{A.1}$$

produce the equal-norm and Parseval frames which are nearest to U.

Further, if U is ε-doubly balanced for $\varepsilon \leq \frac{1}{3}$, then in both cases $\tilde{U}$ is 3ε-doubly balanced and satisfies the distance bound

$$\|\tilde{U} - U\|_F^2 \leq \varepsilon^2.$$

Proof. The fact that the transformations satisfies the equal-norm and Parseval condition is easily verified:

$$\tilde{U}\tilde{U}^* = \left(d \sum_{j=1}^{n} u_j u_j^* \right)^{-\frac{1}{2}} UU^* \left(d \sum_{j=1}^{n} u_j u_j^* \right)^{-\frac{1}{2}} = \frac{1}{d} I_d,$$

$$\|\tilde{u}_j\|_2^2 = \frac{\|u_j\|_2^2}{n\|u_j\|_2^2} = \frac{1}{n},$$

where the two lines correspond to the left and right normalizations respectively.

To show that these are the nearest such frames, we use the following claim which is a simple application of the triangle inequality.

Claim A.3.2. *Let $x \in \mathbb{R}^m$ and $\tilde{x} := \frac{x}{\|x\|_2} \in S^{m-1}$. Then*

$$\inf_{y \in S^{m-1}} \|y - x\|_2 = \|\tilde{x} - x\|_2 = \Big| \|\tilde{x}\|_2 - \|x\|_2 \Big|.$$

Now consider any equal norm frame $V \in \mathrm{Mat}(d, n)$ with size $s(V) = 1$ and note

$$\|V - U\|_F^2 = \sum_{j=1}^n \|v_j - u_j\|_2^2 \geq \sum_{j=1}^n \|\tilde{u}_j - u_j\|_2^2 = \|\tilde{U} - U\|_F^2,$$

where the inequality was by the claim above applied to $x = \sqrt{n} u_j$.

To show the statement for the left normalization, assume without loss of generality that UU^* is diagonal, and let V be an arbitrary Parseval frame of size $s(V) = 1$. This implies in particular that for all $i \in [d] : \|e_i^* V\|_2^2 = \frac{1}{d}$. Therefore, we can again use the claim to bound the distance

$$\|V - U\|_F^2 = \sum_{i=1}^d \|e_i^* V - e_i^* U\|_2^2 \geq \sum_{i=1}^d \|e_i^* \tilde{U} - e_i^* U\|_2^2 = \|\tilde{U} - U\|_F^2,$$

where the inequality was again by the claim above applied to $x = \sqrt{d} e_i^* U$ and the fact that UU^* was diagonal so $\tilde{U}$ is exactly the row-normalization in the standard basis.

Now assume that $U \in \mathrm{Mat}(d, n)$ is ε-doubly balanced. Then we can bound the distance of the right normalization as

$$\|\tilde{U} - U\|_F^2 = \sum_{j=1}^n \|\tilde{u}_j - u_j\|_2^2 = \sum_{j=1}^n (\|\tilde{u}_j\|_2 - \|u_j\|_2)^2 \leq \frac{n}{n}(1 - \sqrt{1 \pm \varepsilon})^2 \leq \varepsilon^2,$$

where we used the fact that U is ε-nearly equal norm in the third step, and the final inequality was by Taylor approximation $|\sqrt{1+x} - 1| \leq |x|$ for $|x| \leq \frac{1}{2}$.

Further, $\tilde{U}$ is equal norm by definition, so we show it is nearly Parseval.

$$\tilde{U}\tilde{U}^* = \sum_{j=1}^n \frac{u_j u_j^*}{n\|u_j\|_2^2} \preceq \frac{UU^*}{1 - \varepsilon} \preceq \frac{1 + 3\varepsilon}{d} I_d,$$

where we used that U is ε-doubly balanced and Taylor approximation on $\frac{1+x}{1-x}$ for $|x| \leq \frac{1}{3}$. The lower bound is shown similarly by reversing inequalities.

The distance to the left normalization can be bounded the same way. Assume again that UU^* is diagonal without loss of generality. Then

$$\|\tilde{U} - U\|_F^2 = \sum_{i=1}^{d} \|e_i^* \tilde{U} - e_i^* U\|_2^2 \leq \frac{d}{d}(1 - \sqrt{1 \pm \varepsilon})^2 \leq \varepsilon^2,$$

where we used the fact that U is ε-Parseval in the second step, and the final inequality was by Taylor approximation.

Further, $\tilde{U}$ is Parseval by definition, so we show it is nearly equal norm.

$$\|\tilde{u}_j\|_2^2 = \langle u_j, (dUU^*)^{-1} u_j \rangle \leq \frac{\|u_j\|_2^2}{1 - \varepsilon} \leq \frac{1 + 3\varepsilon}{n},$$

where we used that U is ε-doubly balanced and Taylor approximation on $\frac{1+x}{1-x}$ for $|x| \leq \frac{1}{3}$. The lower bound is shown similarly. □

A.4 Robustness of Strong Convexity

The following example shows that there can be no multiplicative robustness bound for non-commutative scalings. This example was found during joint work with Cole Franks, Rafael Oliveira, and Michael Walter [36].

Example A.4.1. *Consider input* $V := \begin{pmatrix} \sqrt{2} & 1 \\ 1 & \sqrt{2} \end{pmatrix}$ *and let* $G = (SL(2), ST(2))$ *act by left-right scaling, along with polar* $P = (SPD(2), ST_+(2))$ *and infinitesimal vector space* $\mathfrak{p} = \mathfrak{spd}(2) \oplus \mathfrak{st}_+(2)$ *according to Definition 6.2.3. Then* V *is* $\Omega(1)$-$\mathfrak{p}$*-strongly convex, but* $V^{-1}V = I_2$ *is not* α-$\mathfrak{p}$*-strongly convex for any* $\alpha > 0$.

Before we prove these facts about $\mathfrak{p}$-strong convexity, note that V^{-1} is a bounded scaling of V which destroys $\mathfrak{p}$-strong convexity, so multiplicative robustness is impossible for strong convexity under non-commutative scalings.

Proof. To show V is strongly convex as a frame, consider arbitrary element $(X, Y) \in \mathfrak{p} = \mathfrak{spd}(2) \oplus \mathfrak{st}_+(2)$, which can be specified by orthogonal basis of eigenvectors: $u = (\cos\theta, \sin\theta), v = (-\sin\theta, \cos\theta)$, and $x, y \in \mathbb{R}$:

$$X = xuu^* - xvv^* = x \begin{pmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \sin^2\theta - \cos^2\theta \end{pmatrix} = x \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}$$

and $Y = y(E_{11} - E_{22})$. $\theta \in [0, 2\pi], x, y \in \mathbb{R}$ gives every element of $\mathfrak{p} = \mathfrak{spd}(2) \oplus \mathfrak{st}_+(2)$.

$$\begin{aligned}\partial^2_{\eta=0} f_V(e^{\eta X} \otimes e^{\eta Y}) &= \left\| x \begin{pmatrix} \sqrt{2}\cos 2\theta + \sin 2\theta & \cos 2\theta + \sqrt{2}\sin 2\theta \\ \sqrt{2}\sin 2\theta - \cos 2\theta & \sin 2\theta - \sqrt{2}\cos 2\theta \end{pmatrix} + \begin{pmatrix} \sqrt{2} & -1 \\ 1 & -\sqrt{2} \end{pmatrix} y \right\|_F^2 \\ &= 2x^2(2\cos^2 2\theta + \sin^2 2\theta + 2\sin^2 2\theta + \cos^2 2\theta) + 6y^2 + 2xy(4\cos 2\theta - 2\cos 2\theta) \\ &= 6(x^2 + y^2) + 4xy\cos 2\theta \geq 4(x^2 + y^2) = 4\Big(\frac{\|X\|_F^2}{2} + \frac{\|Y\|_F^2}{2}\Big)\end{aligned}$$

where we used Lemma 3.1.9 to calculate the second order derivative in the first step, in the second step we plugged in our definitions of X, Y, the third step was by $\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle$, the remaining steps used various trignometric identities as well as $|\cos 2\theta| \leq 1, |2xy| \leq x^2 + y^2$, and the final step was once again by our choice of X, Y with $\|X\|_F^2 = 2x^2, \|Y\|_F^2 = 2y^2$. Since $(X, Y) \in \mathfrak{p}$ was arbitrary, this verifies that V is 4-$\mathfrak{p}$-strongly convex according to Definition 4.2.11.

To show I_2 is not strongly convex, we consider direction $X = E_{11} - E_{22}$ and $Y = E_{22} - E_{11} = -X$:

$$\partial^2_{\eta=0} f_{I_2}(e^{\eta X} \otimes e^{\eta Y}) = \|XI_2 + I_2 Y\|_F^2 = 0,$$

where the first step was by the formula in Lemma 3.1.9.

Note that $V \succeq 0$ and $\det(V) = 1$ so the scaling $V^{-1} \in SPD(2)$, i.e. restricting scalings to P does not improve non-commutative robustness. □

A.5 Strong Convexity and Size

In this section, we discuss some relations between strong convexity and size.

Proposition A.5.1. *For matrix tuple $A \in \mathrm{Mat}(d, n)^K$, if A is α-strongly convex according to Definition 3.2.1, then*

$$\alpha \leq s(A).$$

Proof. Recall that according to Definition 3.2.1, strong convexity is given by the following variational formula:

$$\alpha = \inf_{(X,Y) \in \mathfrak{t}} \frac{\partial^2_{\eta=0} f_A(\eta X, \eta Y)}{\|(X, Y)\|_{\mathfrak{t}}^2}.$$

We can write out the second derivative explicitly as

$$\begin{aligned} \partial^2_{\eta=0} f_A(\eta X, \eta Y) &= \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{K} |(A_k)_{ij}|^2 (X_i + Y_j)^2 \\ &= \sum_{i=1}^{d} r_i(A) X_i^2 + \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{K} |(A_k)_{ij}|^2 (2X_i Y_j) + \sum_{j=1}^{n} c_j(A) Y_j^2, \end{aligned}$$

where the first step is by Lemma 3.1.9, and the second step was by collecting terms and using Definition 3.1.1 of row and column sums.

To bound the diagonal terms, let $i \in \arg\min_{i' \in [d]} r_i(A)$ and define $X = E_{ii} - \frac{1}{d} I_d$ so that $Tr[X] = 1 - \frac{d}{d} = 0$. Then we can bound the diagonal term

$$\sum_{i'=1}^{d} r_{i'}(A) X_i^2 = \frac{(d-1)^2 - 1}{d^2} r_i(A) + \frac{1}{d^2} \sum_{i'=1}^{d} r_{i'}(A) \le \frac{(d-1)^2 - 1}{d^2} \frac{s(A)}{d} + \frac{s(A)}{d^2} \le \frac{s(A)}{d} \sum_{i'=1}^{d} X_{i'}^2,$$

where in the first step we used the definition of $X = E_{ii} - \frac{1}{d} I_d$, in the second step we bounded the first term by $\min_{i' \in [d]} r_{i'}(A) \le \frac{1}{d} \sum_{i'=1}^{d} r_{i'}(A) = \frac{s(A)}{d}$ and the second term by Definition 3.1.1 of size, and the final step was once again by definition of X.

Similarly, let $j \in \arg\min_{j' \in [n]} c_j(A)$ and define $Y = E_{jj} - \frac{1}{n} I_n$ so that $Tr[Y] = 0$, and in total $(X, Y) \in \mathfrak{t}$. Combining with the same calculation for Y, we get

$$\sum_{i'=1}^{d} r_{i'}(A) X_{i'}^2 + \sum_{j'=1}^{n} c_{j'}(A) Y_{j'}^2 \le s(A) \Big(\frac{1}{d} \sum_{i'=1}^{d} X_{i'}^2 + \frac{1}{n} \sum_{j'=1}^{n} Y_{j'}^2 \Big) = s(A) \|(X, Y)\|_{\mathfrak{t}}^2,$$

where in the last step we used Definition 3.1.11 of $\| \cdot \|_{\mathfrak{t}}$.

Finally, assuming the cross-term is non-positive by replacing $X \to -X$ if necessary,

$$\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{K} |(A_k)_{ij}|^2 (2X_i Y_j) \le 0.$$

So combining all the terms gives the proposition:

$$\alpha \le \frac{\partial^2_{\eta=0} f_A(\eta X, \eta Y)}{\|(X, Y)\|_{\mathfrak{t}}^2} \le \frac{\sum_{i'=1}^{d} r_{i'}(A) X_{i'}^2 + \sum_{j'=1}^{n} c_{j'}(A) Y_{j'}^2 + 0}{\|(X, Y)\|_{\mathfrak{t}}^2} \le s(A).$$

□

We can simply generalize this to arbitrary tensor scaling.

Proposition A.5.2. *Let* $V = \otimes_{a\in[m]} V_a$ *be a tensor product of inner product spaces of dimension* $\dim(V_a) = d_a$ *and consider scaling group* $(G, P, \mathfrak{p})$ *according to Definition 6.2.3. Then for any input tuple* $x \in V^K$ *that is* α*-*$\mathfrak{p}$*-strongly convex according to Definition 7.1.7,*

$$\alpha \leq s(x).$$

Proof. Consider arbitrary commutative subgroup $(T_a \subseteq G_a, T_b \subseteq G_b)$ and its associated infinitesimal vector space $\mathfrak{t}$. Assume x is α-$\mathfrak{p}$-strongly convex, and we apply Proposition A.5.1 to this subspace:

$$\alpha := \inf_{Z\in\mathfrak{p}} \frac{\langle \rho_x, Z^2\rangle}{\|Z\|_{\mathfrak{p}}^2} \leq \inf_{(X,Y)\in\mathfrak{t}} \frac{\langle \rho_x^{(ab)}, (X\otimes I_b + I_a \otimes Y)^2\rangle}{\|(X,Y)\|_{\mathfrak{t}}^2} \leq s(x),$$

where the first step was by Definition 7.1.7 of $\mathfrak{p}$-strong convexity, in the second step we restricted to the subspace $\mathfrak{t} \subseteq \mathfrak{p}$ and used the fact that on this subspace $\|\cdot\|_{\mathfrak{p}} = \|\cdot\|_{\mathfrak{t}}$, and the final step was shown in the proof of Proposition A.5.1.

□

This shows that the all-ones matrix $A := \frac{1}{\sqrt{dn}} J \in \mathrm{Mat}(d,n)$ is maximally strongly convex with respect to its size according to Proposition A.5.1. Explicitly, $s(\frac{1}{\sqrt{dn}}J) = 1$, and for any $(X,Y) \in \mathfrak{t}$:

$$\partial^2_{\eta=0} f_A(\eta X, \eta Y) = \sum_{i=1}^{d}\sum_{j=1}^{n} \frac{(X_i + Y_j)^2}{dn} = \sum_{i=1}^{d} \frac{X_i^2}{d} + 0 + \sum_{j=1}^{n} \frac{Y_j^2}{n} = \|(X,Y)\|_{\mathfrak{t}}^2,$$

where the first step was by Lemma 3.1.9 of the second derivative; in the second step the cross term vanished by the calculation below:

$$\sum_{i=1}^{d}\sum_{j=1}^{n} X_i Y_j = \left(\sum_{i=1}^{d} X_i\right)\left(\sum_{j=1}^{n} Y_j\right) = 0$$

by the defining condition of $(X,Y) \in \mathfrak{t}$ according to Definition 3.1.5; and in the final step we used Definition 3.1.11 of $\|\cdot\|_{\mathfrak{t}}$. Since $(X,Y) \in \mathfrak{t}$ was arbitrary, this verifies Definition 3.2.1 of $\alpha = 1$-strong convexity. Therefore, for this $A = \frac{1}{\sqrt{dn}}J$, the inequality in Proposition A.5.1 is tight as $\alpha = s(A)$.