Algorithmic Results for Clustering and Refined Physarum Analysis

A dissertation submitted towards the degree

Doctor of Natural Science

of the Faculty of Mathematics and Computer Science of

Saarland University

by

Pavel Kolev

Saarbrücken / October 2018

Colloquium Details

Date:	27 November 2018
Dean of the Faculty:	Prof. Dr. rer. nat. Sebastian Hack
Chairman:	Prof. Dr. Markus Bläser
Examiners:	Prof. Dr. Dr. h.c. mult. Kurt Mehlhorn
	Dr. Karl Bringmann
	Prof. Dr. Raimund Seidel
Scientific Assistant:	Dr. Marvin Künnemann

Abstract

In the first part of this thesis, we study the Binary ℓ_0-Rank-k problem which given a binary matrix A and a positive integer k, seeks to find a rank-k binary matrix B minimizing the number of non-zero entries of $A - B$. A central open question is whether this problem admits a polynomial time approximation scheme. We give an affirmative answer to this question by designing the first randomized almost-linear time approximation scheme for constant k over the reals, $\mathbb{F}_2$, and the Boolean semiring. In addition, we give novel algorithms for important variants of ℓ_0-low rank approximation.

The second part of this dissertation, studies a popular and successful heuristic, known as Approximate Spectral Clustering (ASC), for partitioning the nodes of a graph G into clusters with small conductance. We give a comprehensive analysis, showing that ASC runs efficiently and yields a good approximation of an optimal k-way node partition of G.

In the final part of this thesis, we present two results on slime mold computations: i) the continuous undirected Physarum dynamics converges for undirected linear programs with a non-negative cost vector; and ii) for the discrete directed Physarum dynamics, we give a refined analysis that yields strengthened and close to optimal convergence rate bounds, and shows that the model can be initialized with any strongly dominating point.

Zusammenfassung

Im ersten Teil dieser Arbeit untersuchen wir das Binary ℓ_0-Rank-k Problem. Hier sind eine binäre Matrix A und eine positive ganze Zahl k gegeben und gesucht wird eine binäre Matrix B mit Rang k, welche die Anzahl von nicht null Einträgen in $A - B$ minimiert. Wir stellen das erste randomisierte, nahezu lineare Aproximationsschema vor konstantes k über die reellen Zahlen, $\mathbb{F}_2$ und den Booleschen Semiring. Zusätzlich erzielen wir neue Algorithmen für wichtige Varianten der ℓ_0-low rank Approximation.

Der zweite Teil dieser Dissertation beschäftigt sich mit einer beliebten und erfolgreichen Heuristik, die unter dem Namen Approximate Spectral Cluster (ASC) bekannt ist. ASC partitioniert die Knoten eines gegeben Graphen G in Cluster kleiner Conductance. Wir geben eine umfassende Analyse von ASC, die zeigt, dass ASC eine effiziente Laufzeit besitzt und eine gute Approximation einer optimale k-Weg-Knoten Partition für G berechnet.

Im letzten Teil dieser Dissertation präsentieren wir zwei Ergebnisse über Berechnungen mit Hilfe von Schleimpilzen: i) die kontinuierliche ungerichtete Physarum Dynamik konvergiert für ungerichtete lineare Programme mit einem nicht negativen Kostenvektor; und ii) für die diskrete gerichtete Physikum Dynamik geben wir eine verfeinerte Analyse, die stärkere und beinahe optimale Schranken für ihre Konvergenzraten liefert und zeigt, dass das Model mit einem beliebigen stark dominierender Punkt initialisiert werden kann.

Acknowledgments

This research thesis was conducted under the supervision of Prof. Kurt Mehlhorn and Dr. Karl Bringmann in the Department of Algorithms and Complexity at Max Planck Institute for Informatics.

First and foremost, I thank my advisor Kurt Mehlhorn for making my doctoral studies possible, for sharing his wisdom with me, and for showing me how to conduct research in a positive and sustainable manner. I am grateful to Kurt Mehlhorn for the impeccable lectures and the numerous inspiring discussions that helped me understand the significance of clarity of thought and expression, which ultimately led me to a deeper understanding of the studied problems. I am also thankful to Kurt Mehlhorn for creating and maintaining in his department a unique scientific environment that builds upon the principles of freedom, openness and collaboration. In hindsight, I have greatly benefited from it.

Secondly, I thank my advisor Karl Bringmann for his visionary, guidance and support, for the numerous inspiring discussions, for the impeccable lectures, for the invaluable advices, and especially for helping me develop the determination necessary for achieving long-term scientific goals. I thank Karl Bringmann and Kurt Mehlhorn for sharing their heroic optimism and extensive mathematical background, as without them some of the strongest theorems in this dissertation would have had still been at the drawing desk.

Thirdly, I am grateful to David P. Woodruff for initiating the study of ℓ_0-low rank approximation, for the numerous inspiring discussions, and for the exciting collaborations thereafter. I thank Richard Peng for guiding me through the process of designing and analyzing elegant divide and conquer algorithms, for the numerous inspiring discussions and the invaluable advices, and for the exciting collaboration on spectral sparsification. Further, I am thankful to Markus Bläser, Parinya Chalermsook and Geevarghese Philip for introducing me, in their fascinating lectures, to some of the most beautiful results in the areas of complexity theory and hardness of approximation. I am also grateful to Benjamin Doerr and Thomas Sauerwald for revealing in their exciting courses the vast and rich area of randomized algorithms.

I am profoundly indebted to my undergraduate mentors Krassimir Manev and Ivan Soskov, for their fascinating lectures that lay the foundations and simultaneously revealed to me the beauty of the theoretical computer science (TCS), and for encouraging and supporting me to pursue a doctoral degree in TCS. I am grateful to Kerope Chakaryan, Chavdar Lozanov and Nedyu Popivanov for their their impeccable lectures that introduced me to the beauty and the depth of mathematics. I also thank Peter Armianov, Nikolay Bujukliev, Stefan Gerdgikov and Ivan Tonov, for the numerous inspiring discussions and for guiding me in my early days through the realm of mathematics and computer science.

I thank all my collaborators from whom I learned numerous novel techniques and carefully-crated tools, and most importantly positive mental attitude towards solving challenging problems. Namely, I thank Frank Ban, Ruben Becker, Vijay Bhattiprolu, Vincenzo Bonifaci, Karl Bringmann, Gorav Jindal, Andreas Karrenbauer, Euiwoong Lee, Kurt Mehlhorn, Richard Peng, Saurabh Sawlani and David P. Woodruff.

I am grateful to my friends Andi, Eig, Gorav, Sasho and Shay for all the inspiring discussions, for sharing their pure scientific curiosity in unraveling seemingly simple yet challenging puzzles, for making this journey pleasant and joyful, and especially for the boisterous jokes highlighting our cultural differences.

I thank André for being my guide in the crypto-world; Attila for making me a Swiss chocolate fan and for the late night discussions; Ben for the invigorating ”Rick and Morty” type of jokes; Bhaskar for sharing his always seeking spirit and desire to excel; Davis for the inspiring discussions and for the incredibly addictive Indian banana chips; Daniel for reviving my interest in playing piano and guitar; Dushyant for keeping me up to date with the current state-of-the-art machine learning models; Ilya, Michael, Mohamed, and Victor for the awesome guitar parties; Marvin for guiding me through world of randomized algorithms and for suggesting splendid places for hiking; Michael for the inspiring discussions, for sharing his pure scientific curiosity, and for the vibrant and brilliant insights; and Srinath for inspiring in me the interest in building quadcopters and for the exciting night-sky sessions at the MPI-INF observatory.

Last but not least, I thank my family for their constant support and encouragement, for their love, for giving me a happy childhood, and for raising me right. Especially, I thank my wife Elena for being by my side, for lifting my spirit when I needed it most, and for showing me the magic of life.

The generous financial support by the Cluster of Excellence “Multimodal Computing and Interaction” within the Excellence Initiative of the German Federal Government is gratefully acknowledged.

Preface

In the first part of this thesis, we study NP-Hard variants of low rank approximation that are natural for problems with no underlying metric. We consider the Binary ℓ_0-Rank-k problem which given a binary matrix $A \in \{0,1\}^{m\times n}$ with $m \geqslant n$ and a positive integer k, seeks to find a rank-k binary matrix $B \in \{0,1\}^{m\times n}$ minimizing $\|A-B\|_0$, where $\|\cdot\|_0$ denotes the number of non-zero entries. This problem is known under many different names in different areas of computer science: ℓ_0-low rank approximation in computational linear algebra, constrained binary matrix factorization in data mining, Boolean factor analysis in machine learning, and matrix rigidity in computational complexity theory. A central open question is whether the Binary ℓ_0-Rank-k problem admits a polynomial time approximation scheme (PTAS).

In a joint work [BBB+18, BBB+19] with Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Euiwoong Lee and David P. Woodruff, we give an affirmative answer to this question by designing the first PTAS for the more general problem of ℓ_p-Rank-k for any $p \in [0,2)$. Approximately, Frank Ban and David P. Woodruff contributed the PTAS for $p \in (0,2)$, Vijay Bhattiprolu and Euiwoong Lee contributed the hardness results for $p \in (1,2)$ and finite fields of constant size, and Karl Bringmann and I contributed the following PTAS for what we call the Generalized Binary ℓ_0-Rank-k problem.

In particular, we give the first randomized almost-linear time approximation scheme for the Generalized Binary ℓ_0-Rank-k problem. Our algorithm finds a $(1+\varepsilon)$-approximation in time $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$, where $o(1)$ hides a factor $(\log\log n)^{1.1}/\log n$. This yields the first PTAS for the Binary ℓ_0-Rank-k problem for constant k over the reals, $\mathbb{F}_2$, and the Boolean semiring. Even for the special case of rank $k=1$ no PTAS was known before. Our algorithmic techniques crucially rely on the fact that the Generalized Binary ℓ_0-Rank-k problem is equivalent to a variant of clustering problem with constrained centers, and the crux of our work is to extend existing algorithmic techniques for unconstrained clustering to the more general setting of clustering with constrained centers.

Further, in a joint work [BKW17a, BKW17b] with Karl Bringmann and David P. Woodruff, we give novel algorithms for the following important variants of ℓ_0-low rank approximation:

(a) a polynomial time algorithm for the Reals ℓ_0-Rank-k problem, if we allow for a bicriteria solution, that outputs a matrix with larger rank $O(k\log(n/k))$ and approximation factor $O(k^2\log(n/k))$;

(b) a sublinear time $(2+\varepsilon)$-approximation algorithm for the Reals ℓ_0-Rank-1 problem;

(c) a sublinear time $(1+O(\psi))$-approximation algorithm for the Binary ℓ_0-Rank-1 problem, where $\psi = \|A\|_0/\mathrm{OPT}$, and a matching sample complexity lower bound;

(d) an exact algorithm running in time $2^{O(\mathrm{OPT}/\sqrt{\|A\|_0})}\,\mathrm{poly}(mn)$ for the Binary ℓ_0-Rank-1 problem.

The second part of this dissertation, studies a popular and successful heuristic, known as Approximate Spectral Clustering (ASC), for partitioning the nodes of a graph G into clusters for which the ratio of outside connections compared to the volume (sum of degrees) is small. ASC consists of the following two subroutines: i) compute an approximate Spectral Embedding via the Power method; and ii) partition the resulting vector set with an approximate k-means clustering algorithm. The resulting k-means partition naturally induces a k-way node partition of G.

In a joint work [KM16, KM18] with Kurt Mehlhorn, we give a comprehensive analysis of ASC building on the work of Peng et al. (2017), Boutsidis et al. (2015) and Ostrovsky et al. (2013). We show that ASC runs efficiently and yields a good approximation of an optimal k-way node partition of G. Moreover, we strengthen the quality guarantees of a structural result of Peng et al. by a factor of k, and simultaneously weaken the eigenvalue gap assumption. Further, we demonstrate that ASC finds a k-way node partition of G with the strengthened quality guarantees.

The final part of this thesis is a joint work [BBK+17, BBK+18] with Ruben Becker, Vincenzo Bonifaci, Andreas Karrenbauer and Kurt Mehlhorn. We present two results on slime mold computations:

(1) Bonifaci, Mehlhorn and Varma (2012) showed that the continuous undirected Physarum dynamics, a system of differential equations, converges for the shortest path problems. We demonstrate that the dynamics actually converges for a much wider class of problem, namely undirected linear programs with a non-negative cost vector.

(2) Combinatorial optimization researchers took the dynamics describing slime mold behavior as an inspiration for an optimization method, and in a research line culminating with the work of Straszak and Vishnoi (2016) showed that its discretization initialized with a feasible point can approximately solve linear programs with positive cost vector. We give a refined analysis that yields strengthened and close to optimal convergence rate bounds and shows that the discrete directed Physarum dynamics can be initialized with any strongly dominating point.

During my doctoral studies, I was also involved in a project on sparsifying Random Walk Laplacian (RWL) matrices. [JK15] gives density-independent algorithms for special families of RWL matrices, and this is a joint work with Gorav Jindal. [JKPS17a, JKPS17b] establishes a density-independent algorithm for general RWL matrices, and this is a joint work with Gorav Jindal, Richard Peng and Saurabh Sawlani.

The following list consists of all publications during my doctoral studies:

Journal Versions:

[BBK^{+}18] Ruben Becker, Vincenzo Bonifaci, Andreas Karrenbauer, Pavel Kolev, and Kurt Mehlhorn. Two results on slime mold computations. *Theoretical Computer Science*, 2018

[KM18] Pavel Kolev and Kurt Mehlhorn. Approximate spectral clustering: Efficiency and guarantees. *CoRR*, abs/1509.09188, 2018. Submitted. A preliminary version of this paper was presented at the 24th Annual European Symposium on Algorithms (ESA 2016)

Conference Versions:

[BBB^{+}19] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P. Woodruff. A PTAS for l_p-low rank approximation. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, Westin San Diego, San Diego, California, USA, January 6-9*, 2019. To appear

[BKW17a] Karl Bringmann, Pavel Kolev, and David P. Woodruff. Approximation algorithms for l_0-low rank approximation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NIPS 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6651–6662, 2017

[KM16] Pavel Kolev and Kurt Mehlhorn. A note on spectral clustering. In *24th Annual European Symposium on Algorithms, ESA 2016, August 22-24, 2016, Aarhus, Denmark*, pages 57:1–57:14, 2016

[JKPS17a] Gorav Jindal, Pavel Kolev, Richard Peng, and Saurabh Sawlani. Density Independent Algorithms for Sparsifying k-Step Random Walks. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*, volume 81 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik

ArXiv Versions:

[BBB^{+}18] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P. Woodruff. A PTAS for l_p-low rank approximation. *CoRR*, abs/1807.06101, 2018

[BKW17b] Karl Bringmann, Pavel Kolev, and David P. Woodruff. Approximation algorithms for l_0-low rank approximation. *CoRR*, abs/1710.11253, 2017. Full version of [BKW17a]

[BBK^{+}17] Ruben Becker, Vincenzo Bonifaci, Andreas Karrenbauer, Pavel Kolev, and Kurt Mehlhorn. Two results on slime mold computations. *CoRR*, abs/1707.06631, 2017

[JKPS17b] Gorav Jindal, Pavel Kolev, Richard Peng, and Saurabh Sawlani. Density independent algorithms for sparsifying k-step random walks. *CoRR*, abs/1702.06110, 2017

[JK15] Gorav Jindal and Pavel Kolev. An efficient parallel algorithm for spectral sparsification of laplacian and sddm matrix polynomials. *CoRR*, abs/1507.07497, 2015

Contents

Part I

Low Rank Matrix Approximation

Proceedings

This part is based on an article [BBB+19] that is accepted at *SODA'19* and on an article [BKW17a] that was presented at *NIPS'17*. The full ArXiv versions are available online [BKW17b, BBB+18].

Acknowledgements

I would like to thank Karl Bringmann for his guidance, support and fruitful collaboration, David P. Woodruff for inspiring this study and for fruitful collaboration, and Uriel Feige for helpful discussions and for pointing me to the work of Alon and Sudakov [AS99].

Chapter 1

Introduction

Low rank approximation of an $m \times n$ matrix A is an extremely well-studied problem, where the goal is to replace the matrix A with a rank-k matrix A' which well-approximates A, in the sense that $\|A - A'\|$ is small under some measure $\|\cdot\|$. Since any rank-k matrix A' can be written as $U \cdot V$, where U is $m \times k$ and V is $k \times n$, it suffices to store the $k(m+n)$ entries of U and V, which is a significant reduction compared to the mn entries of A. Furthermore, computing $A'x = U(Vx)$ takes time $O(k(m+n))$, which is much less than the time $O(mn)$ for computing Ax. We refer the reader to several surveys [KV09, Mah11, Woo14] for references to the many results on low rank approximation.

ℓ_0-low rank approximation When the measure $\|A - A'\|_0$ is the number of non-zero entries, we seek a rank-k matrix A' for which the number of entries (i,j) with $A'_{i,j} \neq A_{i,j}$ is as small as possible. This can be seen as the Hamming distance between a matrix A and its best rank k approximation A'. The ℓ_0-low rank approximation is natural for problems with no underlying metric, since the $\|\cdot\|_0$ measure directly answers the following question: *if we are allowed to ignore outliers (or anomalies), what is the best low-rank model we can get?*

A well-studied case is when A is binary, but A' and its factors U and V need not necessarily be binary. This is called *unconstrained* Binary Matrix Factorization [JPHY14]. There is also a large body of work on the *constrained* version, in which not only the input matrix A is binary, but we also have the natural restriction that the factors U, V are binary. That is, we study the following problem.

Binary ℓ_0-Rank-k Given a matrix $A \in \{0,1\}^{m\times n}$ and an integer k, compute matrices $U \in \{0,1\}^{m\times k}$ and $V \in \{0,1\}^{k\times n}$ minimizing $\|A - U \cdot V\|_0$.

Note that we did not yet specify the ground field (or, more generally, semiring) that we are working over, which affects the type of matrix multiplication used in $U \cdot V$. Specifically, for $A' = U \cdot V$ we can write the entry $A'_{i,j}$ as the inner product of the i-th row of U with the j-column of V – and the specific inner product function $\langle .,. \rangle$ depends on the ground field.

Typical choices for the ground field are as follows. Let $x, y \in \{0,1\}^k$.

(i) *Reals:* For the ground field $\mathbb{R}$, we get the standard inner product $\langle x, y\rangle = \sum_{i=1}^k x_i \cdot y_i \in \{0, 1, \ldots, k\}$. This is rather unnatural, since no number $\geqslant 2$ can match any entry of A.

(ii) *Modulo 2:* For the ground field $\mathbb{F}_2$, the inner product is $\langle x, y\rangle = \bigoplus_{i=1}^k x_i \cdot y_i \in \{0,1\}$. This is used, e.g., in [Yer11, GGYT12, DAJ+15, PRF16].

(iii) *Boolean:* Using the Boolean semiring $\{0, 1, \wedge, \vee\}$, the inner product becomes $\langle x, y\rangle = \bigvee_{i=1}^k x_i \wedge y_i = 1 - \prod_{i=1}^k (1 - x_i \cdot y_i) \in \{0,1\}$. This is used, e.g., in [BV10, DAJ+15, MMG+08, SBM03, SH06, VAG07].

Note that for $k = 1$ all three inner products are the same.

Clustering with Constrained Centers We now demonstrate that the Binary ℓ_0-Rank-k problem is equivalent to a clustering problem with constrained centers. Consider arbitrary matrices $U \in \{0,1\}^{m\times k}$ and $V \in \{0,1\}^{k\times n}$. Grouping the identical columns of V gives rise to a column partitioning, and let $C_y = \{j \in [n] : V_{:,j} = y\}$ be the set of column indices corresponding to a vector y, for every $y \in \{0,1\}^k$. Then, the expression $\|A - U \cdot V\|_0$ is equivalent to a clustering formulation whose centers are restricted to the set $S_U \stackrel{\text{def}}{=} \{U \cdot y \in \{0,1\}^m : y \in \{0,1\}^k\}$, and it reads

$$\|A - U \cdot V\|_0 = \sum_{j=1}^{n} \|A_{:,j} - U \cdot V_{:,j}\|_0 = \sum_{y \in \{0,1\}^k} \sum_{j \in C_y} \|A_{:,j} - U \cdot y\|_0.$$

Observe that any column of $U \cdot V$ is in S_U, and thus the choice of columns in V can be seen as selecting constrained centers in S_U. Formally, we rephrase the Binary ℓ_0-Rank-k problem as follows

$$\begin{aligned}\min_{U\in\{0,1\}^{m\times k}, V\in\{0,1\}^{k\times n}} \|A - U\cdot V\|_0 &= \min_{U\in\{0,1\}^{m\times k}} \sum_{j=1}^{n} \min_{V_{:,j}\in\{0,1\}^k} \|A_{:,j} - U\cdot V_{:,j}\|_0 \\ &= \min_{U\in\{0,1\}^{m\times k}} \sum_{j=1}^{n} \min_{s\in S_U} \|A_{:,j} - s\|_0.\end{aligned}$$

This is a clustering problem, where the task is to choose a set of constrained centers S_U, in order to minimize the total ℓ_0-distance of all columns of A to their closest center in S_U. The main difference to unconstrained clustering problems is that the constrained centers in S_U cannot be chosen independently, since these centers satisfy a certain system of linear equations. Because of this dependence, the techniques used for designing approximation schemes for unconstrained clustering problems are not directly applicable.

Applications The Binary ℓ_0-Rank-k problem arises in many fields of computer science, with different ground fields being important in different areas. In *computational linear algebra* the problem is known under the umbrella term "low rank matrix approximation" and used for compressing matrices, as described above [KV09, Mah11, Woo14, BKW17a].

In *complexity theory* the problem is known as "matrix rigidity", introduced by Grigoriev [Gri80] and Valiant [Val77], who showed that constructions of rigid matrices would imply circuit lower bounds, see also [AW17]. The decision problem of matrix rigidity has been studied from the viewpoint of FPT [FLM$^+$17].

In *data mining* and *machine learning*, the problem is known as "(Constrained) Binary Matrix Factorization" or "Boolean Factor Analysis" [Mie, JPHY14, DAJ$^+$15, FGP18]. In these areas, the columns of U could correspond to latent topics learned from the database A, and, more generally, the problem has numerous applications including latent variable analysis, topic models, association rule mining, clustering, and database tiling [SBM03, SH06, VAG07, MV14, GGYT12]. The usage of the problem to mine discrete patterns has also been applied in bioinformatics to analyze gene expression data [MGT15, SJY09, ZLD$^+$10]. In these situations it is natural to consider binary matrices whenever the data is categorical, which is often the case for text data [DAJ$^+$15, MV14, RPG16]. The special and important case in which A is binary and $k = 1$ was studied in [KG03, SJY09, JPHY14], as their algorithm for $k = 1$ forms the basis for their successful heuristic for general k, e.g. the PROXIMUS technique [KG03].

Another special case of Binary ℓ_0-Rank-k is the Biclique Partition problem, see, e.g., [CIK16, FGP18]. Binary ℓ_0-Rank-k over $\mathbb{F}_2$ has been applied to Independent Component Analysis of string data in the area of information theory [Yer11, GGYT12, PRF16]. If we drop the requirement that U and V are binary, and use the Frobenius norm as distance measure, we obtain the unconstrained Binary Matrix Factorization [JPHY14], which has applications to association rule mining [KG03], biclustering structure identification [ZLD$^+$10, ZLDZ07], pattern discovery for gene expression [SJY09], digits reconstruction [MGNR06], mining high-dimensional discrete-attribute data [KGR05, KGR06], market based clustering [Li05], and document clustering [ZLDZ07].

An important related problem is robust PCA [CLMW11], in which there is an underlying matrix A that can be written as a low rank matrix L plus a sparse matrix S [CLMW11]. Candès et al. [CLMW11] argue that both the components of L and S are of arbitrary magnitude, and we do not know the locations of the non-zeros in S nor how many there are. Moreover, grossly corrupted observations are common in image processing, web data analysis, and bioinformatics where some measurements are *arbitrarily* corrupted due to occlusions, malicious tampering, or sensor failures. Specific scenarios include video surveillance, face recognition, latent semantic indexing, and ranking of movies, books, etc. [CLMW11]. These problems have the common theme of being an arbitrary magnitude sparse perturbation to a low rank matrix with no natural underlying metric, and thus the ℓ_0-error measure is appropriate. In order to solve the robust PCA in practice, Candès et al. [CLMW11] relaxed the ℓ_0-error measure to the ℓ_1-norm.

It is of a fundamental importance for theory to understand the algorithmic guarantees and limits for solving the ℓ_0-low rank approximation problem.

Algorithms for Binary ℓ_0-Rank-k Let us first discuss rank $k = 1$, in which case the real, $\mathbb{F}_2$, and Boolean variants are all equal. A polynomial-time 2-approximation was designed by Shen et al. [SJY09], and simplified by Jiang et al. [JPHY14]. For $k > 1$, Dan et al. [DAJ$^+$15] presented an $n^{O(k)}$-time $O(k)$-approximation over $\mathbb{F}_2$, and an $n^{O(k)}$-time $O(2^k)$-approximation over the Boolean semiring. The work of

Chierichetti et al. [CGK+17] for ℓ_p-low rank approximation, for $p \geqslant 1$, fails to give any approximation factor for $p = 0$. Indeed, critical to their analysis is the scale-invariance property of a norm, which does not hold for $p = 0$ since ℓ_0 is not a norm. To the best of our knowledge, these are the only known approximation algorithms for this problem.

Hardness of Binary ℓ_0-Rank-k It is well-known that Binary ℓ_0-Rank-k is NP-hard, even for rank $k = 1$ [GV15, DAJ+15]. Further, if k is unbounded then even deciding whether matrix A has rank k is NP-hard over the Boolean semiring [Mie]. This suggests that the running time of any approximation algorithm must depend at least exponentially on k.

Despite the high interest in the problem in linear algebra, data mining, machine learning, and complexity theory, the approximability of the problem is wide open. In particular, this leaves the following question open. *Does the Binary ℓ_0-Rank-k problem have a polynomial time approximation scheme (PTAS) for any constant k?*

1.1 Generalized Binary ℓ_0-Rank-k

The results in this Section are proven in Chapter 2. In order to study the real, $\mathbb{F}_2$, and Boolean setting in a unified way, we introduce the following more general problem.

Generalized Binary ℓ_0-Rank-k Given a matrix $A \in \{0,1\}^{m\times n}$ with $m \geqslant n$, an integer k, and an inner product function $\langle .,. \rangle \colon \{0,1\}^k \times \{0,1\}^k \to \mathbb{R}$, compute matrices $U \in \{0,1\}^{m\times k}$ and $V \in \{0,1\}^{k\times n}$ minimizing $\|A - U\cdot V\|_0$, where the product $U \cdot V$ uses $\langle .,. \rangle$. An algorithm for the Generalized Binary ℓ_0-Rank-k problem is an α-approximation, if it outputs matrices $U \in \{0,1\}^{m\times k}$ and $V \in \{0,1\}^{k\times n}$ satisfying $\|A - U \cdot V\|_0 \leqslant \alpha \cdot \min_{U'\in\{0,1\}^{m\times k}, V'\in\{0,1\}^{k\times n}} \|A - U' \cdot V'\|_0$.

As shown above, by choosing an appropriate inner product function $\langle .,. \rangle$ which also runs in time $O(k)$, we obtain the Binary ℓ_0-Rank-k problem over the reals, $\mathbb{F}_2$, and the Boolean semiring. We assume that the function $\langle .,. \rangle$ can be evaluated in time $2^{O(k)}$, in order to simplify our running time bounds.

Our main result is a randomized almost-linear time approximation scheme for the Generalized Binary ℓ_0-Rank-k problem for any constant k. In particular, this yields the first PTAS for the Binary ℓ_0-Rank-k problem for constant k over the reals, $\mathbb{F}_2$, and the Boolean semiring. Even for the special case of rank $k = 1$ no PTAS was known before.

Theorem 1.1. *(PTAS) For any error $\varepsilon \in (0, 1/2)$, there is a $(1+\varepsilon)$-approximation algorithm for the Generalized Binary ℓ_0-Rank-k problem that runs in time $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$ and succeeds with constant probability* [1]*, where $o(1)$ hides a factor $(\log\log n)^{1.1} / \log n$.*

Moreover, we show that our PTAS has a close to optimal running time, in the sense that the runtime of any PTAS for the Generalized Binary ℓ_0-Rank-k problem must depend exponentially on $1/\varepsilon$ and doubly exponentially on k, assuming the Exponential Time Hypothesis (ETH).

Theorem 1.2. *(Hardness for Generalized Binary ℓ_0-Rank-k) Assuming the Exponential Time Hypothesis, Generalized Binary ℓ_0-Rank-k has no $(1+\varepsilon)$-approximation algorithm in time $2^{1/\varepsilon^{o(1)}} \cdot 2^{m^{o(1)}}$. Further, for any $\varepsilon \geqslant 0$, Generalized Binary ℓ_0-Rank-k has no $(1+\varepsilon)$-approximation algorithm in time $2^{2^{o(k)}} \cdot 2^{m^{o(1)}}$.*

Further, we give a faster algorithm for the Binary ℓ_0-Rank-1 problem with standard inner product. In the following, we assume [2] that we can access any entry $A_{i,j}$ in constant time, and we can also enumerate all non-zero entries in time $O(\|A\|_0)$.

Theorem 1.3. *(PTAS for the Binary ℓ_0-Rank-1 problem with standard inner product) For any $\varepsilon \in (0, 1/2)$ there is an algorithm that runs in time $(1/\varepsilon)^{O(1/\varepsilon^2)} \cdot (\|A\|_0 + m + n) \cdot \log^3(mn)$, and outputs vectors $\widetilde{u} \in \{0,1\}^m$, $\widetilde{v} \in \{0,1\}^n$ such that w.h.p.* [3] $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1+\varepsilon) \min_{u\in\{0,1\}^m, v\in\{0,1\}^n} \|A - u \cdot v^T\|_0$.

[1] The success probability can be further amplified to $1 - \delta$ for any $\delta > 0$ by running $O(\log(1/\delta))$ independent trials of the preceding algorithm.

[2] Depending on the specific storage scheme additional running time factors may be necessary. For instance, if we store A by adjacency arrays, then enumerating all non-zero entries in time $O(\|A\|_0)$ is straightforward, while accessing any entry $A_{i,j}$ can be performed in time $O(\log n)$ by a binary search over the adjacency array for row i, so in this case the stated running time has to be multiplied by a factor $O(\log n)$.

[3] An event happens *with high probability* (w.h.p.) if it has probability at least $1 - 1/n^c$ for some $c > 0$.

Note: Our results in Theorem 1.1 and Theorem 1.2 are in submission as of April 2018. Shortly after posting our manuscript [BBB+18] to arXiv on 16 July 2018, we became aware that in an unpublished work Fomin et al. have independently obtained a very similar PTAS for Binary ℓ_0-Rank-k problem. Their manuscript [FGL+18] was posted to arXiv on 18 July 2018. Interestingly, [BBB+18, FGL+18] have independently discovered i) a reduction between the Binary ℓ_0-Rank-k problem and a clustering problem with constrained centers; ii) a structural sampling theorem extending [AS99] which yields a simple but inefficient deterministic PTAS; and iii) an efficient sampling procedure, building on ideas from [KSS04, ABH+05, ABS10], which gives an efficient randomized PTAS. Notably, by establishing an additional structural result, Fomin et al. [FGL+18] design a faster sampling procedure which yields a randomized PTAS for the Binary ℓ_0-Rank-k problem that runs in linear time $(1/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn$.

Further Related Work

Suppose $A \in \{0,1\}^{m\times n}$ and the inner product $\langle .,. \rangle$ maps to $\{0,1\}$. Then, matrix $U \cdot V$ has entries in $\{0,1\}$, and hence matrix $B \overset{\text{def}}{=} A - U \cdot V$ has entries in $\{-1,0,1\}$. In this case, the ℓ_0-distance, the ℓ_1-norm of all entries, and the Frobenius norm $\|B\|_F = (\sum_{i=1}^m \sum_{j=1}^m B_{i,j}^2)^{1/2}$ are all equivalent:

$$\|B\|_0 = \sum_{i=1}^{m}\sum_{j=1}^{m} |B_{i,j}| = \|B\|_F^2.$$

Here, we also allow inner product functions mapping to numbers other than $\{0,1\}$ (note that no such number can match any entry in A).

The following problem is closely related to the Binary ℓ_0-Rank-k problem.

Hypercube Segmentation: Given a matrix $A \in \{0,1\}^{m\times n}$ and an integer k, we seek to compute vectors $u_1,\ldots,u_k \in \{0,1\}^m$ maximizing the value $\sum_{j=1}^n \min_{1\leqslant \ell \leqslant k}(m - \|A_{:,j} - u_\ell\|_0)$ (or equivalently, minimizing the expression $\sum_{j=1}^n \min_{1\leqslant \ell \leqslant k} \|A_{:,j} - u_\ell\|_0$).

The problem is NP-hard even for $k = 2$ [Fei14]. Note that $m - \|A_{:,j} - u_\ell\|_0$ is the number of entries that column $A_{:,j}$ and vector u_ℓ have in common. It can be checked that the optimum is always at least $mn/2$, and thus approximating the maximization version is easier, in the sense that any α-approximation algorithm for minimization implies an α-approximation for maximization. For the maximization version, Kleinberg et al. [KPR04] designed a polynomial-time 0.878-approximation for any fixed k, which was improved to an efficient PTAS for any fixed k by Alon and Sudakov [AS99]. For the minimization version, a PTAS was designed by Ostrovsky and Rabani [OR00]. This was extended to several clustering problems that are natural generalizations of Hypercube Segmentation [OR00, ABS10]. An efficient PTAS is known for a Euclidean variant of this clustering problem [BHI02].

Our results extend this line of work, since (the minimization version of) *Hypercube Segmentation is a special case of the Generalized Binary ℓ_0-Rank-k problem.* Indeed, let $\pi\colon \{0,1\}^k \to \{1,\ldots,k\}$ be any onto function, e.g., π maps the i-th vector in $\{0,1\}^k$ (w.r.t. any fixed ordering) to the number $\min\{i,k\}$. We define an inner product function $\langle x,y \rangle \overset{\text{def}}{=} x_{\pi(y)}$, i.e., y specifies a coordinate of the vector x. Consider the matrix product $A' = U \cdot V$ with respect to this inner product. Writing $\ell_j \overset{\text{def}}{=} \pi(V_{:,j})$, we have $A'_{i,j} = U_{i,\ell_j}$. In other words, the j-th column of A' is equal to one of the columns of U, and we can choose which one. Writing the columns of U as $u_1,\ldots,u_k$ yields the equivalence with Hypercube Segmentation.

1.2 Binary ℓ_0-rank-1 With Small Optimal Value

The results in this Section are proven in Chapter 3. Given a matrix $A \in \{0,1\}^{m \times n}$, our goal is to compute an approximate solution of the Binary ℓ_0-Rank-1 problem, and let us denote the optimal value by

$$\mathrm{OPT} \overset{\mathrm{def}}{=} \min_{u \in \{0,1\}^m,\, v \in \{0,1\}^n} \|A - u \cdot v^T\|_0. \tag{1.1}$$

In practice, approximating a matrix A by a rank-1 matrix uv^T makes most sense if A is close to being rank-1. Hence, the above optimization problem is most relevant in the case $\mathrm{OPT} \ll \|A\|_0$. For this reason, we focus in this section on the case $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, for sufficiently small $\phi > 0$.

We first give an algorithm that requires as an input a parameter $\phi \geqslant \mathrm{OPT}/\|A\|_0$.

Theorem 1.4. *Given $A \in \{0,1\}^{m \times n}$ with row and column sums, and given $\phi \in (0, \frac{1}{80}]$ with $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, we can compute in time $O(\min\{\|A\|_0 + m + n, \phi^{-1}(m+n)\log(mn)\})$ vectors $\widetilde{u} \in \{0,1\}^m$ and $\widetilde{v} \in \{0,1\}^n$ such that w.h.p. $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1 + 5\phi)\mathrm{OPT} + 37\phi^2\|A\|_0$.*

Then, we use a $(2+\varepsilon)$-approximation algorithm for the Reals ℓ_0-Rank-1 problem that captures as a special case the Binary ℓ_0-Rank-1 problem. We prove the following result in Chapter 4, see Section 4.2.

Theorem 1.5. *Given $A \in \{0,1\}^{m \times n}$ with column adjacency arrays and $\mathrm{OPT} \geqslant 1$, and given $\varepsilon \in (0, 0.1]$, we can compute w.h.p. in time*

$$O\left(\left(\frac{n \log m}{\varepsilon^2} + \min\left\{\|A\|_0,\ n + \psi^{-1}\frac{\log n}{\varepsilon^2}\right\}\right)\frac{\log^2 n}{\varepsilon^2}\right)$$

a column $A_{:,j}$ and a vector $z \in \{0,1\}^n$ such that w.h.p. $\|A - A_{:,j} \cdot z^T\|_0 \leqslant (2+\varepsilon)\mathrm{OPT}$. Further, we can compute an estimate Y such that w.h.p. $(1-\varepsilon)\mathrm{OPT} \leqslant Y \leqslant (2+2\varepsilon)\mathrm{OPT}$.

Using Theorem 1.4 in combination with Theorem 1.5, we obtain a $(1+500\psi)$-approximation algorithm for the Binary ℓ_0-Rank-1 problem that does not need the parameter ϕ as an input.

Theorem 1.6. *Given $A \in \{0,1\}^{m \times n}$ with column adjacency arrays and with row and column sums, for $\psi = \mathrm{OPT}/\|A\|_0$ we can compute w.h.p. in time $O(\min\{\|A\|_0 + m + n, \psi^{-1}(m+n)\} \cdot \log^3(mn))$ vectors $\widetilde{u} \in \{0,1\}^m$ and $\widetilde{v} \in \{0,1\}^n$ such that w.h.p. $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1 + 500\psi)\mathrm{OPT}$.*

Our algorithm simultaneously improves the approximation factor and the running time. Further, it even runs in sublinear $o(\|A\|_0)$ time, unlike all algorithms in previous works [SJY09, JPHY14]. Notably, when matrix A is very well approximated by a low rank matrix, i.e. ψ is a sub-constant, we obtain a $(1 + o(1))$-approximation which is significantly better than the previous best known 2-approximations.

Moreover, we also show that the running time of our algorithm is optimal up to a $\mathrm{poly}\log(mn)$ factor, by proving that any $(1 + O(\psi))$-approximation algorithm succeeding with constant probability must read $\Omega(\psi^{-1}(m+n))$ entries of A in the worst case.

Theorem 1.7. *Let $C \geqslant 1$. Given an $n \times n$ binary matrix A with column adjacency arrays and with row and column sums, and given $\sqrt{\log(n)/n} \ll \phi \leqslant 1/100C$ such that $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, computing a $(1 + C\phi)$-approximation of OPT requires to read $\Omega(n/\phi)$ entries of A (in the worst case over A).*

Furthermore, a variant of the algorithm from Theorem 1.6 can also be used to solve exactly the Binary ℓ_0-Rank-1 problem. This yields the following theorem, which in particular shows that the problem is in polynomial time when $\mathrm{OPT} \leqslant O\big(\sqrt{\|A\|_0}\log(mn)\big)$.

Theorem 1.8. *Given a matrix $A \in \{0,1\}^{m \times n}$, if $\mathrm{OPT}/\|A\|_0 \leqslant 1/240$ then we can solve exactly the Binary ℓ_0-Rank-1 problem in time $2^{O(\mathrm{OPT}/\sqrt{\|A\|_0})} \cdot \mathrm{poly}(mn)$.*

1.3 Algorithms for Reals ℓ_0-Rank-k Problem

The results in this Section are proven in Chapter 4. We establish approximation algorithms for several important variants of ℓ_0-low rank approximation, which significantly improve the running time and the approximation factor of previous works. In some cases our algorithms even run in sublinear time, i.e., faster than reading all non-zero entries of the matrix. This is provably impossible for other measures such as the Frobenius norm and more generally, any ℓ_p-norm for $p > 0$. For $k > 1$, our approximation algorithms are, to the best of our knowledge, the first with provable guarantees for these problems.

1.3.1 Preliminaries

For a matrix $A \in \mathbb{A}^{m\times n}$ with $m \geqslant n$ and entries $A_{i,j}$, let $A_{i,:}$ be its i-th row and $A_{:,j}$ be its j-th column. An algorithm that on input a sufficiently dense matrix A, runs in time $o(\|A\|_0)$ is called sublinear. In particular, we say that time $\widetilde{O}(m+n)$ is sublinear, since in general $m+n \ll \|A\|_0$. In this section, we study variants of the following problem.

Reals ℓ_0-Rank-k Given a matrix $A \in \mathbb{R}^{m\times n}$ and an integer k, compute matrices $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{k\times n}$ minimizing $\|A - U\cdot V\|_0$. An algorithm for the Reals ℓ_0-Rank-k problem is an α-approximation, if it outputs matrices $\widetilde{U} \in \mathbb{R}^{m\times k}$ and $\widetilde{V} \in \mathbb{R}^{k\times n}$ satisfying

$$\|A - \widetilde{U}\cdot\widetilde{V}\|_0 \leqslant \alpha \min_{U\in\mathbb{R}^{m\times k}, V\in\mathbb{R}^{k\times n}} \|A - U\cdot V\|_0.$$

Input Formats We always assume that we have random access to the entries of the given matrix A, i.e. we can read any entry $A_{i,j}$ in constant time. For our sublinear time algorithms we need more efficient access to the matrix, specifically the following two variants:

(1) We say that we are given A *with column adjacency arrays* if we are given arrays $B_1,\ldots,B_n$ and lengths $\ell_1,\ldots,\ell_n$ such that for any $k \in \{1,\ldots,\ell_j\}$ the pair $B_j[k] = (i, A_{i,j})$ stores the row i containing the k-th nonzero entry in column j as well as that entry $A_{i,j}$. This is a standard representation of matrices used in many applications. Note that given only these adjacency arrays $B_1,\ldots,B_n$, in order to access any entry $A_{i,j}$ we can perform a binary search over B_j, and hence random access to any matrix entry is in time $O(\log n)$. Moreover, we assume to have random access to matrix entries in constant time, and note that this is optimistic by at most a factor $O(\log n)$.

(2) We say that we are given matrix A *with row and column sums* if we can access the numbers $\sum_j A_{i,j}$ for $i \in [m]$ and $\sum_i A_{i,j}$ for $j \in [n]$ in constant time (and, as always, access any entry $A_{i,j}$ in constant time). Notice that storing the row and column sums takes $O(m+n)$ space, and thus while this might not be standard information it is very cheap to store.

In Subsection 4.2.5, we show that the first access type even allows to sample from the set of nonzero entries uniformly in constant time.

Lemma 1.9. *Given a matrix $A \in \mathbb{R}^{m\times n}$ with column adjacency arrays, after $O(n)$ time preprocessing we can sample a uniformly random nonzero entry (i,j) from A in time $O(1)$.*

1.3.2 Reals ℓ_0-rank-k

Given a matrix $A \in \mathbb{R}^{m\times n}$, our goal is to compute an approximate solution of the Reals ℓ_0-rank-k problem, and let us denote the optimal value by

$$\mathrm{OPT} \stackrel{\text{def}}{=} \min_{U\in\mathbb{R}^{m\times k},\, V\in\mathbb{R}^{k\times n}} \|A - U\cdot V\|_0. \tag{1.2}$$

We first give an impractical algorithm that runs in time $n^{O(k)}$ and achieves $O(k^2)$ factor approximation. To the best of our knowledge this is the first approximation algorithm for the Reals ℓ_0-Rank-k problem with non-trivial approximation guarantees.

Theorem 1.10. *Given a matrix $A \in \mathbb{R}^{m\times n}$ and an integer k, we can compute in time $O(n^{k+1}m^2k^{\omega+1})$ a set of k indices $J^{(k)} \subset [n]$ and a matrix $Z \in \mathbb{R}^{k\times n}$ such that $\|A - A_{:,J^{(k)}}\cdot Z\|_0 \leqslant O(k^2)\cdot\mathrm{OPT}$.*

To make our algorithm practical, we reduce the running time to $\mathrm{poly}(mn)$, with an exponent independent of k, if we allow for a bicriteria solution. In particular, we allow the algorithm to output a matrix A' of larger rank $O(k\log(n/k))$ and approximation factor $O(k^2\log(n/k))$.

Theorem 1.11. *Given a matrix $A \in \mathbb{R}^{m\times n}$ and an integer k, there is an algorithm that in expected time $\mathrm{poly}(m,n)$ outputs a subset of indices $J \subset [n]$ with $|J| = O(k\log(n/k))$ and a matrix $Z \in \mathbb{R}^{|J|\times n}$ such that $\|A - A_{:,J}\cdot Z\|_0 \leqslant O(k^2\log(n/k))\cdot\mathrm{OPT}$.*

Although, we do not obtain exactly rank k, many of the motivations for finding a low rank approximation, such as reducing the number of parameters and fast matrix-vector multiplication, still hold if the output rank is $O(k\log(n/k))$. We are not aware of any alternative algorithms that achieve $\mathrm{poly}(mn)$ time and any provable approximation factor, even for bicriteria solutions.

1.3.3 Reals ℓ_0-rank-1

Given a rank-1 matrix $A \in \mathbb{R}^{m \times n}$, there is an algorithm that runs in time $O(\|A\|_0)$ and finds the exact rank-1 decomposition uv^T of A. Here, we focus on the case when A is not a rank-1 matrix.

The previous best algorithm due to Jiang et al. [JPHY14] was based on the observation that there exists a column u of A spanning a 2-approximation. Hence, solving the problem $\min_v \|A - u \cdot v^T\|_0$ for each column u of matrix A yields a 2-approximation. The problem $\min_{v \in \mathbb{R}^n} \|A - u \cdot v^T\|_0$ decomposes into $\sum_i \min_i \|A_{:,i} - v_i u\|_0$, where $A_{:,i}$ is the i-th column of A, and v_i the i-th entry of vector v. The optimal v_i is the mode of the ratios $A_{i,j}/u_j$, where j ranges over indices in $\{1, 2, \ldots, m\}$ with $u_j \neq 0$. As a result, one can find a rank-1 matrix uv^T providing a 2-approximation in time $O(n\|A\|_0)$, which was the best known running time.

We design randomized algorithms for solving this problem. Let the optimal value be

$$\mathrm{OPT} \overset{\text{def}}{=} \min_{u \in \mathbb{R}^m,\, v \in \mathbb{R}^n} \|A - u \cdot v^T\|_0. \tag{1.3}$$

Our algorithm yields a $(2+\varepsilon)$-approximation and runs in nearly linear time in $\|A\|_0$, for any constant $\varepsilon > 0$. Moreover, a variant of our algorithm even runs in sublinear time when $\mathrm{OPT} \geqslant (\varepsilon^{-1} \log(mn))^4$ and $\|A\|_0 \geqslant n(\varepsilon^{-1} \log(mn))^4$.

Theorem 1.12. *Given $A \in \mathbb{R}^{m \times n}$ with column adjacency arrays and $\mathrm{OPT} \geqslant 1$, and given $\varepsilon \in (0, 0.1]$, there is an algorithm that runs w.h.p. in time*

$$O\left(\left(\frac{n \log m}{\varepsilon^2} + \min\left\{\|A\|_0,\ n + \psi^{-1} \frac{\log n}{\varepsilon^2}\right\}\right) \frac{\log^2 n}{\varepsilon^2}\right) \quad \textit{where} \quad \psi = \frac{\mathrm{OPT}}{\|A\|_0},$$

and outputs a column $A_{:,j}$ and a vector $z \in \mathbb{R}^n$ such that w.h.p. $\|A - A_{:,j} \cdot z^T\|_0 \leqslant (2+\varepsilon)\mathrm{OPT}$. The algorithm also computes an estimate Y satisfying w.h.p. $(1-\varepsilon)\mathrm{OPT} \leqslant Y \leqslant (2+2\varepsilon)\mathrm{OPT}$.

This significantly improves upon the earlier $O(n\|A\|_0)$ time algorithm for not too small ε and ψ. Our result should be contrasted to Frobenius norm low rank approximation, for which $\Omega(\|A\|_0)$ time is required even for $k = 1$, as otherwise one might miss a very large entry in A. Since ℓ_0-low rank approximation is insensitive to the magnitude of entries of A, we bypass this general impossibility result.

Chapter 2

A PTAS For Generalized Binary ℓ_0-Rank-k

Given a matrix $A \in \{0,1\}^{m \times n}$ with $m \geqslant n$, an integer k, and an inner product function $\langle .,. \rangle \colon \{0,1\}^k \times \{0,1\}^k \to \mathbb{R}$, the Generalized Binary ℓ_0-Rank-k problem asks to find matrices $U \in \{0,1\}^{m \times k}$ and $V \in \{0,1\}^{k \times n}$ minimizing $\|A - U \cdot V\|_0$, where the product $U \cdot V$ is the $m \times n$ matrix B with $B_{i,j} = \langle U_{i,:}, V_{:,j} \rangle$. An algorithm for the Generalized Binary ℓ_0-Rank-k problem is an α-approximation, if it outputs matrices $U \in \{0,1\}^{m \times k}$ and $V \in \{0,1\}^{k \times n}$ satisfying

$$\|A - U \cdot V\|_0 \leqslant \alpha \cdot \min_{U' \in \{0,1\}^{m \times k}, V' \in \{0,1\}^{k \times n}} \|A - U' \cdot V'\|_0.$$

As shown in Chapter 1, by choosing an appropriate inner product function $\langle .,. \rangle$ which also runs in time $O(k)$, we obtain the Binary ℓ_0-Rank-k problem over the reals, $\mathbb{F}_2$, and the Boolean semiring. We assume that the function $\langle .,. \rangle$ can be evaluated in time $2^{O(k)}$, in order to simplify our running time bounds.

2.1 Technical Overview

2.1.1 PTAS For Generalized Binary ℓ_0-Rank-k

The Generalized Binary ℓ_0-Rank-k problem can be rephrased as a clustering problem with constrained centers, whose goal is to choose a set of centers satisfying a certain system of linear equations, in order to minimize the total ℓ_0-distance of all columns of A to their closest center. The main difference to usual clustering problems is that the centers cannot be chosen independently.

We view the choice of matrix U as picking a set of "cluster centers" $S_U \stackrel{\text{def}}{=} \{U \cdot y \mid y \in \{0,1\}^k\}$. Observe that any column of $U \cdot V$ is in S_U, and thus we view the choice of column $V_{:,j}$ as picking one of the constrained centers in S_U. Formally, we rephrase the Generalized Binary ℓ_0-Rank-k problem as

$$\begin{aligned}\min_{U \in \{0,1\}^{m \times k}, V \in \{0,1\}^{k \times n}} \|A - U \cdot V\|_0 &= \min_{U \in \{0,1\}^{m \times k}} \sum_{j=1}^{n} \min_{V_{:,j} \in \{0,1\}^k} \|A_{:,j} - U \cdot V_{:,j}\|_0 \\ &= \min_{U \in \{0,1\}^{m \times k}} \sum_{j=1}^{n} \min_{s \in S_U} \|A_{:,j} - s\|_0. \end{aligned} \tag{2.1}$$

Any matrix V gives rise to a "clustering" as partitioning $C_V = (C_y)_{y \in \{0,1\}^k}$ of the columns of V with $C_y = \{j \in [n] \mid V_{:,j} = y\}$. If we knew an optimal clustering $C = C_V$, for some optimal matrix V, we could compute an optimal matrix U as the best response to V. Note that

$$\min_{U \in \{0,1\}^{m \times k}} \sum_{y \in \{0,1\}^k} \sum_{j \in C_y} \|A_{:,j} - U \cdot y\|_0 = \sum_{i=1}^{m} \min_{U_{i,:} \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{j \in C_y} \|A_{i,j} - U_{i,:} \cdot y\|_0.$$

Therefore, given C we can compute independently for each $i \in [m]$ the optimal row $U_{i,:} \in \{0,1\}^k$, by enumerating over all possible binary vectors of dimension k and selecting the one that minimizes the summation $\sum_{y \in \{0,1\}^k} \sum_{j \in C_y} \|A_{i,j} - U_{i,:} \cdot y\|_0$.

What if instead we could only *sample from* C? That is, suppose that we are allowed to draw a constant number $t = \text{poly}(2^k/\varepsilon)$ of samples from each of the optimal clusters C_y uniformly at random. Denote by $\widetilde{C}_y$ the samples drawn from C_y. A natural approach is to replace the exact cost above by the following unbiased estimator:

$$\widetilde{E} \stackrel{\text{def}}{=} \sum_{y \in \{0,1\}^k} \frac{|C_y|}{|\widetilde{C}_y|} \cdot \sum_{j \in \widetilde{C}_y} \|A_{:,j} - U \cdot y\|_0.$$

We show that with good probability any matrix $U = U(\widetilde{C})$ minimizing the estimated cost $\widetilde{E}$ is close to an optimal solution. In particular, we prove for any matrix $V \in \{0,1\}^{k \times n}$ that

$$\mathbb{E}_{\widetilde{C}}[\|A - U(\widetilde{C}) \cdot V\|_0] \leqslant (1+\varepsilon) \cdot \min_{U \in \{0,1\}^{m \times k}} \|A - U \cdot V\|_0. \tag{2.2}$$

The biggest issue in proving statement (2.2) is that the number of samples $t = \operatorname{poly}(2^k/\varepsilon)$ is independent of the ambient space dimension n. A key prior probabilistic result, established by Alon and Sudakov [AS99], gives an additive $\pm\varepsilon mn$ approximation for the maximization version of a clustering problem with unconstrained centers, known as Hypercube Segmentation. Since the optimum value of this maximization problem is always at least $mn/2$, a multiplicative factor $(1+\varepsilon)$-approximation is obtained. Our contribution is twofold. First, we generalize their analysis to clustering problems with constrained centers, and second we prove a multiplicative factor $(1+\varepsilon)$-approximation for the minimization version. The proof of (2.2) takes a significant fraction of this chapter.

We combine the sampling result (2.2) with the following observations to obtain a deterministic polynomial time approximation scheme (PTAS) in time $m \cdot n^{\operatorname{poly}(2^k/\varepsilon)}$. We later discuss how to further improve this running time. Let U, V be an optimal solution to the Generalized Binary ℓ_0-Rank-k problem.

(1) To evaluate the estimated cost $\widetilde{E}$, we need the sizes $|C_y|$ of an optimal clustering C. We can guess these sizes with an n^{2^k} overhead in the running time. In fact, it suffices to know these cardinalities approximately, see Lemma 2.5, and thus this overhead [1] can be reduced to $(t + \varepsilon^{-1} \cdot \log n)^{2^k}$.

(2) Using the (approximate) size $|C_y|$ and the samples $\widetilde{C}_y$ drawn u.a.r. from C_y, for all $y \in \{0,1\}^k$, we can compute in time $2^{O(k)}mn$ a matrix $U(\widetilde{C})$ minimizing the estimated cost $\widetilde{E}$, since the estimator $\widetilde{E}$ can be split into a sum over the rows of $U(\widetilde{C})$ and each row is chosen independently as a minimizer among all possible binary vectors of dimension k.

(3) Given $U(\widetilde{C})$, we can compute a best response matrix $V(\widetilde{C})$ which has cost $\|A - U(\widetilde{C}) \cdot V(\widetilde{C})\|_0 \leqslant \|A - U(\widetilde{C}) \cdot V\|_0$, and thus by (2.2) the expected cost at most $(1+\varepsilon)\mathrm{OPT}$.

(4) The only remaining step is to draw samples $\widetilde{C}_y$ from the optimal clustering. However, in time $O(n^{2^k t}) = n^{\operatorname{poly}(2^k/\varepsilon)}$ we can enumerate all possible families $(\widetilde{C}_y)_{y \in \{0,1\}^k}$, and the best such family yields a solution that is at least as good as a random sample. In total, we obtain a PTAS in time $m \cdot n^{\operatorname{poly}(2^k/\varepsilon)}$.

The largest part of this chapter is devoted to make the above PTAS efficient, i.e., to reduce the running time from $m \cdot n^{\operatorname{poly}(2^k/\varepsilon)}$ to $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$, where $o(1)$ hides a factor $(\log \log n)^{1.1} / \log n$. By the preceding outline, it suffices to speed up Steps (1) and (4), i.e., to design a fast algorithm that guesses approximate cluster sizes and samples from the optimal clusters.

The standard sampling approach for clustering problems such as k-means [KSS04] is as follows. At least one of the clusters of the optimal solution is "large", say $|C_y| \geqslant n/2^k$. Sample t columns uniformly at random from the set $[n]$ of all columns. Then with probability at least $(1/2^k)^t$ all samples lie in C_y, and in this case they form a uniform sample from this cluster. In the usual situation without restrictions on the cluster centers, the samples from C_y allow us to determine an approximate cluster center $\widetilde{s}^{(y)}$. Do this as long as large clusters exist (recall that we have guessed approximate cluster sizes in Step (1), so we know which clusters are large). When all remaining clusters are small, remove the $n/2$ columns that are closest to the approximate cluster centers $\widetilde{s}^{(y)}$ determined so far, and estimate the cost of these columns using the centers $\widetilde{s}^{(y)}$. As there are no restrictions on the cluster centers, this yields a good cost estimation of the removed columns, and since the ℓ_0-distance is additive the algorithm recurses on the remaining columns, i.e. on an instance of twice smaller size. We continue this process until each cluster is sampled. This approach has been used to obtain linear time approximation schemes for k-means and k-median in a variety of ambient spaces [KSS04, KSS05, ABS10].

The issue in our situation is that we cannot fix a cluster center $\widetilde{s}^{(y)}$ by looking only at the samples $\widetilde{C}_y$, since we have dependencies among cluster centers. We nevertheless make this approach work, by showing that a uniformly random column $r^{(y)} \in [n]$ is a good "representative" of the cluster C_y with not-too-small probability. In the case when all remaining clusters are small, we then simply remove the $n/2$ columns that are closest to the representatives $r^{(y)}$ of the clusters that we already sampled from. Although these

[1] In Section 2.3, we establish an efficient sampling procedure, see Algorithm 2, that further reduces the total overhead for guessing the sizes $|C_y|$ of an optimal clustering to $(2^k/\varepsilon)^{2^{O(k)}} \cdot (\log n)^{(\log \log n)^{0.1}}$.

representatives can be far from the optimal cluster centers due to the linear restrictions on the latter, we show in Section 2.3 that nevertheless this algorithm yields samples from the optimal clusters.

We prove that the preceding algorithm succeeds with probability at least $(\varepsilon/t)^{2^{O(k)\cdot t}}$. Further, we show that the approximate cluster sizes $|\widetilde{C}_y|$ of an optimal clustering can be guessed with an overhead of $(2^k/\varepsilon)^{2^{O(k)}} \cdot (\log n)^{(\log\log n)^{0.1}}$. In contrast to the standard clustering approach, the representatives $r^{(y)}$ do not yield a good cost estimation of the removed columns. We overcome this issue by first collecting all samples $\widetilde{C}$ from the optimal clusters, and then computing approximate cluster centers that satisfy certain linear constraints, i.e. a matrix $U(\widetilde{C})$ and its best response matrix $V(\widetilde{C})$. The latter computation runs in linear time $2^{O(k)} \cdot mn$ in the size of the *original* instance, and this in combination with the guessing overhead, yields the total running time of $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$. For further details, we refer the reader to Algorithm 2 in Subsection 2.3.2.

Our algorithm achieves a substantial generalization of the standard clustering approach and applies to the situation with constrained centers. This yields the first randomized almost-linear time approximation scheme (PTAS) for the Generalized Binary ℓ_0-Rank-k problem.

Theorem 1.1 (from page 5)**.** *(PTAS) For any error $\varepsilon \in (0, 1/2)$, there is a $(1+\varepsilon)$-approximation algorithm for the Generalized Binary ℓ_0-Rank-k problem that runs in time $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$ and succeeds with constant probability* [2]*, where $o(1)$ hides a factor $(\log\log n)^{1.1} / \log n$.*

Our running time is close to optimal, in the sense that the running time of any PTAS for the Generalized Binary ℓ_0-Rank-k problem must depend exponentially on $1/\varepsilon$ and doubly exponentially on k, assuming the Exponential Time Hypothesis (ETH). We show this in the following.

Theorem 1.2 (from page 5)**.** *(Hardness for Generalized Binary ℓ_0-Rank-k) Assuming the Exponential Time Hypothesis, the Generalized Binary ℓ_0-Rank-k problem has no $(1+\varepsilon)$-approximation algorithm in time $2^{1/\varepsilon^{o(1)}} \cdot 2^{m^{o(1)}}$. Further, for any $\varepsilon \geqslant 0$, the Generalized Binary ℓ_0-Rank-k problem has no $(1+\varepsilon)$-approximation algorithm in time $2^{2^{o(k)}} \cdot 2^{m^{o(1)}}$.*

Regarding the dependence on ε, assume w.l.o.g. that $m \geqslant n$, and thus the input size is $O(m)$. Even for $k = 1$ the problem is known to be NP-hard [GV15, DAJ+15]. Under ETH, no NP-hard problem has an algorithm [3] in time $2^{m^{o(1)}}$. We can restrict to $\varepsilon \geqslant 1/m^2$, since any better approximation already yields an optimal solution. It follows for $k = 1$ that the Generalized Binary ℓ_0-Rank-k problem has no $(1+\varepsilon)$-approximation algorithm in time $2^{1/\varepsilon^{o(1)}} \cdot 2^{m^{o(1)}}$. In other words, in order to improve our exponential dependence on $1/\varepsilon$ to subexponential, we would need to pay an exponential factor in m.

Regarding the dependence on k, note that for any ε a $(1+\varepsilon)$-approximation algorithm for our problem decides whether the answer is 0 or larger. In particular, over the Boolean semiring it solves the problem whether a given bipartite graph can be covered with k bicliques. For this problem, Chandran et al. [CIK16] proved that even for $k = O(\log m)$ there is no algorithm in time $2^{2^{o(k)}}$, unless ETH fails. It follows that for any $\varepsilon \geqslant 0$, Generalized Binary ℓ_0-Rank-k has no $(1+\varepsilon)$-approximation algorithm in time $2^{2^{o(k)}} \cdot 2^{o(m)}$. In other words, in order to improve our doubly exponential dependence on k, we would need to pay an exponential factor in m. Together, this shows that our running time is close to optimal.

Organization In Subsection 2.2.2, we state our core sampling result. In Subsection 2.2.3, we give a simple but inefficient deterministic PTAS for the Generalized Binary ℓ_0-Rank-k problem, which serves as a blueprint for our efficient randomized PTAS. We present first the deterministic PTAS as it is conceptually simple and exhibits the main algorithmic challenge, namely, to design an efficient sampling procedure. In Subsection 2.2.4, we prove our core sampling result by extending the analysis of Alon and Sudakov [AS99] to clustering problems with constrained centers, and by further strengthening an additive $\pm\varepsilon mn$ approximation guarantee to a multiplicative factor $(1+\varepsilon)$-approximation. In Subsection 2.3, we design an efficient sampling procedure, and this yields our efficient randomized PTAS. Our approach uses ideas from clustering algorithms pioneered by Kumar et al. [KSS04] and refined in [KSS05, ABS10].

In Section 2.4, we give a faster randomized PTAS for the Binary ℓ_0-Rank-1 problem which improves the algorithm in Theorem 1.1 and runs in time $O((1/\varepsilon)^{1/\varepsilon^2} \cdot (\|A\|_0 + m + n) \cdot \log^3(mn))$.

[2] The success probability can be further amplified to $1-\delta$ for any $\delta > 0$ by running $O(\log(1/\delta))$ independent trials of the preceding algorithm.

[3] ETH postulates that 3-SAT is not in time $2^{o(m)}$. Here we only need the weaker hypothesis that 3-SAT is not in time $2^{m^{o(1)}}$.

2.2 Constant Size Sampling Suffices

2.2.1 Preliminaries

Chebyshev's inequality We now give some basic facts. Let $Z_1, \ldots, Z_n$ be independent Bernoulli random variables, with $Z_i \sim \mathrm{Ber}(p_i)$. Let $Z \overset{\text{def}}{=} Z_1 + \ldots + Z_n$ and $\mu \overset{\text{def}}{=} \mathbb{E}[Z]$.

Lemma 2.1. *For any* $\Delta > 0$*, we have* $\Pr[|Z - \mu| > \Delta] \leqslant \mu/\Delta^2$.

Proof. By independence, we have

$$\mathrm{Var}(Z) = \sum_{i=1}^{n} \mathrm{Var}(Z_i) = \sum_{i=1}^{n} p_i(1-p_i) \leqslant \sum_{i=1}^{n} p_i = \mu.$$

By Chebyshev's inequality, for any $\Delta > 0$ it holds that $\Pr[|Z - \mu| > \Delta] \leqslant \mathrm{Var}(Z)/\Delta^2$. The claim follows by $\mathrm{Var}(Z) \leqslant \mu$. □

Lemma 2.2. *For any* $\Delta > 0$*, we have* $\Pr[|Z - \mu| > \Delta] \leqslant \sqrt{n}/\Delta$.

Proof. As in the previous lemma's proof, we have $\Pr[|Z - \mu| > \Delta] \leqslant \mathrm{Var}(Z)/\Delta^2$, where $\mathrm{Var}(Z) \leqslant \mu \leqslant n$, and thus $\Pr[|Z - \mu| > \Delta] \leqslant n/\Delta^2$. It also follows that $\Pr[|Z - \mu| > \Delta] \leqslant \sqrt{n}/\Delta$, since if $\sqrt{n}/\Delta < 1$ we have $n/\Delta^2 \leqslant \sqrt{n}/\Delta$, and otherwise the inequality is trivial. □

2.2.2 Sampling Theorem

We denote the optimal value of Generalized Binary ℓ_0-Rank-k by

$$\mathrm{OPT} = \mathrm{OPT}_k \overset{\text{def}}{=} \min_{U \in \{0,1\}^{m \times k},\, V \in \{0,1\}^{k \times n}} \|A - U \cdot V\|_0.$$

Further, for a fixed matrix $V \in \{0,1\}^{k \times n}$ we let

$$\mathrm{OPT}_k^V \overset{\text{def}}{=} \min_{U \in \{0,1\}^{m \times k}} \|A - U \cdot V\|_0,$$

and we say that a matrix $U \in \{0,1\}^{m \times k}$ is a *best response* to V, if $\|A - U \cdot V\|_0 = \mathrm{OPT}_k^V$.

Given $A \in \{0,1\}^{m \times n}$, an integer k, and an inner product $\langle .,. \rangle \colon \{0,1\}^k \times \{0,1\}^k \to \mathbb{R}$, let $V \in \{0,1\}^{k \times n}$ be arbitrary and $U \in \{0,1\}^{m \times k}$ be a best response to V, i.e.,

$$\mathrm{OPT}_k^V \overset{\text{def}}{=} \|A - U \cdot V\|_0 = \min_{U' \in \{0,1\}^{m \times k}} \|A - U' \cdot V\|_0.$$

Partition the columns of V (equivalently the columns of A) into sets

$$C_y^V \overset{\text{def}}{=} \{j \,:\, V_{:,j} = y\},$$

for all $y \in \{0,1\}^k$. For any row i, vector $y \in \{0,1\}^k$, and $c \in \{0,1\}$ we define by

$$Z_{i,y,c} \overset{\text{def}}{=} |\{j \in C_y^V \mid A_{ij} = c\}| \quad \text{and} \quad Z_{i,y,\neq c} \overset{\text{def}}{=} |\{j \in C_y^V \mid A_{ij} \neq c\}|.$$

Then, the *exact cost* of a row i for any vector $x \in \{0,1\}^k$ is given by

$$E_{i,x} \overset{\text{def}}{=} \|A_{i,:} - x^T \cdot V\|_0 = \sum_{y \in \{0,1\}^k} Z_{i,y,\neq\langle x,y \rangle}. \tag{2.3}$$

Observe that $U_{i,:} \in \{0,1\}^k$ is a vector x minimizing $E_{i,x}$ (this follows from U being a best response to V), and let $E_i \overset{\text{def}}{=} E_{i,U_{i,:}}$.

We do not know the partitioning $\{C_y^V\}_{y \in \{0,1\}^k}$, however, as we will see later we can assume that (1) we can sample elements from each C_y^V and (2) we have good approximations of the sizes $|C_y^V|$ for all y.

For (1), to set up notation let $\widetilde{C} = (\widetilde{C}_y)_{y \in \{0,1\}^k}$ be a family, where $\widetilde{C}_y$ is a random multiset with elements from C_y^V. Specifically, we will work with the following *distribution* $\mathcal{D}_{V,t}$ for some $t \in \mathbb{N}$: For any

$y \in \{0,1\}^k$, if $|C_y^V| < t$ let $\widetilde{C}_y = C_y^V$, otherwise sample t elements from C_y^V with replacement and let the resulting multiset be $\widetilde{C}_y$.

For (2), we say that a sequence $\alpha = (\alpha_y)_{y \in \{0,1\}^k}$ is a sequence of *δ-approximate cluster sizes* if for any $y \in \{0,1\}^k$ with $|C_y^V| < t$ we have $\alpha_y = |C_y^V|$, and for the remaining $y \in \{0,1\}^k$ we have

$$|C_y^V| \leqslant \alpha_y \leqslant (1+\delta)|C_y^V|.$$

Then corresponding to $Z_{i,y,c}$ and $Z_{i,y,\neq c}$ we have random variables

$$\widetilde{Z}_{i,y,c} \overset{\text{def}}{=} |\{j \in \widetilde{C}_y \,:\, A_{i,j} = c\}| \quad \text{and} \quad \widetilde{Z}_{i,y,\neq c} \overset{\text{def}}{=} |\{j \in \widetilde{C}_y \,:\, A_{i,j} \neq c\}|.$$

Given $\widetilde{C}$ and α, we define the *estimated cost* of row i and vector $x \in \{0,1\}^k$ as

$$\widetilde{E}_{i,x} \overset{\text{def}}{=} \sum_{y \in \{0,1\}^k} \frac{\alpha_y}{|\widetilde{C}_y|} \widetilde{Z}_{i,y,\neq\langle x,y\rangle}. \tag{2.4}$$

If $C_y^V = \emptyset$ for some $y \in \{0,1\}^k$, then $\widetilde{Z}_{i,y,\neq\langle x,y\rangle} = 0$ and we define the corresponding summand in (2.4) to be 0. Observe that if the approximation α_y is exact, i.e., $\alpha_y = |C_y^V|$, then $\widetilde{E}_{i,x}$ is an unbiased estimator for the exact cost $E_{i,x}$.

We now simplify the problem to optimizing the estimated cost instead of the exact cost. Specifically, we construct a matrix $\widetilde{U} \in \{0,1\}^{m \times k}$ by picking for each row i any

$$\widetilde{U}_{i,:} \in \operatorname{argmin}\{\widetilde{E}_{i,x} \,:\, x \in \{0,1\}^k\}.$$

Note that matrix $\widetilde{U}$ depends on the input $(A, k, \langle .,. \rangle)$, on the sequence α, and on the sampled multisets $\widetilde{C} = (\widetilde{C}_y)_{y \in \{0,1\}^k}$. When it is clear from the context, we suppress the dependence on $A, k, \langle .,. \rangle$, and write $\widetilde{U} = \widetilde{U}(\widetilde{C}, \alpha)$. We generalize now (2.2), and show that this matrix yields a good approximation of the optimal cost.

Theorem 2.3. *For any matrix $V \in \{0,1\}^{k \times n}$, let α be a sequence of $\frac{\varepsilon}{6}$-approximate cluster sizes and draw $\widetilde{C}$ according to distribution $\mathcal{D}_{V,t}$ for $t = t(k,\varepsilon) \overset{\text{def}}{=} 2^{4k+14}/\varepsilon^2$. Then we have*

$$\mathbb{E}_{\widetilde{C}}\big[\|A - \widetilde{U}(\widetilde{C},\alpha) \cdot V\|_0\big] \leqslant (1+\varepsilon)\mathrm{OPT}_k^V.$$

We defer the proof of Theorem 2.3 to Section 2.2.4, and first show how it yields a simple but inefficient deterministic PTAS for the Generalized Binary ℓ_0-Rank-k problem running in time $m \cdot n^{\mathrm{poly}(2^k/\varepsilon)}$, see Section 2.2.3. Then, in Section 2.3, we design a sampling procedure that improves the running time to $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$, where $o(1)$ hides a factor $(\log\log n)^{1.1} / \log n$.

2.2.3 Simple PTAS

In this subsection, we show how Theorem 2.3 leads to a simple but inefficient deterministic PTAS, see Algorithm 1, for the Generalized Binary ℓ_0-Rank-k problem.

A basic, but crucial property used in our analysis is that given a matrix $A \in \{0,1\}^{m \times n}$, an integer k and a matrix V, we can compute a *best response* matrix U minimizing $\|A - U \cdot V\|_0$ in time $2^{O(k)}mn$. Indeed, we can split $\|A - U \cdot V\|_0 = \sum_{i=1}^n \|A_{i,:} - U_{i,:} \cdot V\|_0$ and brute-force the optimal solution $U_{i,:} \in \{0,1\}^k$ minimizing the i-th summand $\|A_{i,:} - U_{i,:} \cdot V\|_0$. Symmetrically, given U we can compute a best response V in time $2^{O(k)}mn$. In particular, if (U,V) is an optimal solution then U is a best response for V, and V is a best response for U.

We now present the pseudocode of Algorithm 1.

Algorithm 1 Simple deterministic PTAS for the Generalized Binary ℓ_0-Rank-k problem

Input: A matrix $A \in \{0,1\}^{m \times n}$, an integer k, an inner product $\langle ., . \rangle$, and $\varepsilon \in (0,1)$.
Output: Matrices $\widetilde{U} \in \{0,1\}^{m \times k}$, $\widetilde{V} \in \{0,1\}^{k \times n}$ such that $\|A - \widetilde{U} \cdot \widetilde{V}\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}_k$.

1. *(Guess column set sizes)* Let U, V be an optimal solution. Exhaustively guess all sizes $|C_y^V| =: \alpha_y$ for $y \in \{0,1\}^k$. There are n^{2^k} possibilities.

2. *(Guess column multisets)* Theorem 2.3 implies existence of a family $\widetilde{C} = (\widetilde{C}_y)_{y \in \{0,1\}^k}$ such that $\|A - \widetilde{U}(\widetilde{C}, \alpha) \cdot V\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}_k$, where each $\widetilde{C}_y$ is a multiset consisting of at most t indices in $\{1, \ldots, n\}$. Exhaustively guess such a family $\widetilde{C}$. There are $n^{O(t \cdot 2^k)}$ possibilities.

3. *(Compute $\widetilde{U}$)* Now we know $A, k, \langle ., . \rangle, |C_y^V|$ for all $y \in \{0,1\}^k$, and $\widetilde{C}$, thus we can compute the matrix $\widetilde{U} = \widetilde{U}(\widetilde{C}, \alpha)$, where row $\widetilde{U}_{i,:}$ is any vector x minimizing the estimated cost $\widetilde{E}_{i,x}$. Since each row $\widetilde{U}_{i,:} \in \{0,1\}^k$ can be optimized independently, this takes time $2^{O(k)}mn$. If we guessed correctly, we have $\|A - \widetilde{U} \cdot V\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}_k$.

4. *(Compute $\widetilde{V}$)* Compute $\widetilde{V}$ as a best response to $\widetilde{U}$. This takes time $2^{O(k)}mn$. If we guessed correctly, by best-response and Step 3, we have
$$\|A - \widetilde{U} \cdot \widetilde{V}\|_0 \leqslant \|A - \widetilde{U} \cdot V\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}_k.$$

5. **Return** the pair $(\widetilde{U}, \widetilde{V})$ minimizing $\|A - \widetilde{U} \cdot \widetilde{V}\|_0$ over all exhaustive guesses.

The correctness of Algorithm 1 immediately follows from Theorem 2.3. The running time is dominated by the exhaustive guessing in Step 2, so we obtain time $m \cdot n^{\mathrm{poly}(2^k/\varepsilon)}$.

2.2.4 Proof of the Sampling Theorem 2.3

We follow the notation in Section 2.2.2, in particular $V \in \{0,1\}^{k \times n}$ is an arbitrary matrix and $U \in \{0,1\}^{m \times k}$ is a best response to V. We define $D_{i,x}$ as the difference of the cost of row i w.r.t. a vector x and the cost of row i w.r.t. the optimal vector $U_{i,:}$, i.e.,

$$\begin{aligned} D_{i,x} \stackrel{\mathrm{def}}{=} E_{i,x} - E_i &= \|A_{i,:} - x^T \cdot V\|_0 - \|A_{i,:} - U_{i,:} \cdot V\|_0 \\ &= \sum_{y \in \{0,1\}^k} Z_{i,y,\neq\langle x,y \rangle} - Z_{i,y,\neq\langle U_{i,:},y \rangle}. \end{aligned} \tag{2.5}$$

Note that a vector x is suboptimal for a row i if and only if $D_{i,x} > 0$. By a straightforward splitting of the expectation, we obtain the following.

Claim 2.4. *For every $V \in \{0,1\}^{k \times n}$, we have*

$$\mathbb{E}_{\widetilde{C}}\left[\|A - \widetilde{U} \cdot V\|_0\right] = \mathrm{OPT}_k^V + \sum_{i=1}^m \sum_{\substack{x \in \{0,1\}^k \\ D_{i,x} > 0}} \Pr\left[\widetilde{U}_{i,:} = x\right] \cdot D_{i,x}.$$

Proof. We split $\|A - \widetilde{U} \cdot V\|_0 = \sum_{i=1}^m \|A_{i,:} - \widetilde{U}_{i,:} \cdot V\|_0$. This yields

$$\begin{aligned} \mathbb{E}_{\widetilde{C}}\left[\|A - \widetilde{U} \cdot V\|_0\right] &= \sum_{i=1}^m \mathbb{E}_{\widetilde{C}}\left[\|A_{i,:} - \widetilde{U}_{i,:} \cdot V\|_0\right] \\ &= \sum_{i=1}^m \sum_{x \in \{0,1\}^k} \Pr\left[\widetilde{U}_{i,:} = x\right] \cdot \|A_{i,:} - x^T \cdot V\|_0. \end{aligned}$$

By definition of $D_{i,x}$, we have

$$\begin{aligned}\mathbb{E}_{\widetilde{C}}\left[\|A-\widetilde{U}\cdot V\|_0\right] &= \sum_{i=1}^{m}\sum_{x\in\{0,1\}^k}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot(\|A_{i,:}-U_{i,:}\cdot V\|_0+D_{i,x})\\ &= \sum_{i=1}^{m}\Big(\|A_{i,:}-U_{i,:}\cdot V\|_0+\sum_{x\in\{0,1\}^k}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot D_{i,x}\Big)\\ &= \mathrm{OPT}_k^V+\sum_{i=1}^{m}\sum_{x\in\{0,1\}^k}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot D_{i,x}\\ &= \mathrm{OPT}_k^V+\sum_{i=1}^{m}\sum_{\substack{x\in\{0,1\}^k\\ D_{i,x}>0}}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot D_{i,x}. \qquad \square\end{aligned}$$

Similarly to $D_{i,x}$, we define an estimator

$$\widetilde{D}_{i,x}\overset{\text{def}}{=}\widetilde{E}_{i,x}-\widetilde{E}_{i,U_{i,:}}=\sum_{y\in\{0,1\}^k}\frac{\alpha_y}{|\widetilde{C}_y|}\cdot\Big(\widetilde{Z}_{i,y,\neq\langle x,y\rangle}-\widetilde{Z}_{i,y,\neq\langle U_{i,:},y\rangle}\Big). \tag{2.6}$$

Note that $\widetilde{U}_{i,:}$ is chosen among the vectors $x\in\{0,1\}^k$ minimizing $\widetilde{D}_{i,x}$. Hence, our goal is to show that significantly suboptimal vectors (with $D_{i,x}>\frac{\varepsilon}{3}\cdot E_i$) satisfy $\widetilde{D}_{i,x}>0$ with good probability, and thus these vectors are not picked in $\widetilde{U}$.

To this end, we split the rows i and suboptimal vectors x into:

$$\begin{aligned}L_0&\overset{\text{def}}{=}\{(i,x)\,:\,0<D_{i,x}\leqslant\frac{\varepsilon}{3}\cdot E_i\},\\ L_1&\overset{\text{def}}{=}\{(i,x)\,:\,\frac{\varepsilon}{3}\cdot E_i<D_{i,x}\leqslant E_i\},\\ L_2&\overset{\text{def}}{=}\{(i,x)\,:\,D_{i,x}>E_i\}.\end{aligned}$$

Observe that $\sum_{(x,i)\in L_0}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot D_{i,x}\leqslant\frac{\varepsilon}{3}\cdot\mathrm{OPT}_k^V$. By Claim 2.4, we can ignore all tuples $(i,x)\in L_0$, since

$$\mathbb{E}_{\widetilde{C}}\left[\|A-\widetilde{U}\cdot V\|_0\right]\leqslant(1+\frac{\varepsilon}{3})\mathrm{OPT}_k^V+\sum_{(i,x)\in L_1\cup L_2}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot D_{i,x}. \tag{2.7}$$

Hence, our goal is to upper bound the summation $\sum_{(i,x)\in L_1\cup L_2}\Pr\left[\widetilde{U}_{i,:}=x\right]\cdot D_{i,x}$.

We next establish a sufficient condition for $\widetilde{U}_{i,:}\neq x$, for any suboptimal vector x. Note that by definition of $D_{i,x}$ we have

$$D_{i,x}=\sum_{y\in\{0,1\}^k}Z_{i,y,\neq\langle x,y\rangle}-Z_{i,y,\neq\langle U_{i,:},y\rangle}=\sum_{y\in\hat{Y}_{i,x}}Z_{i,y,\neq\langle x,y\rangle}-Z_{i,y,\neq\langle U_{i,:},y\rangle}, \tag{2.8}$$

where $\hat{Y}_{i,x}\overset{\text{def}}{=}\{y\in\{0,1\}^k\,:\,\langle x,y\rangle\neq\langle U_{i,:},y\rangle\}$. Similarly, for the estimator we have

$$\widetilde{D}_{i,x}=\sum_{y\in\{0,1\}^k}\frac{\alpha_y}{|\widetilde{C}_y|}\cdot\Big(\widetilde{Z}_{i,y,\neq\langle x,y\rangle}-\widetilde{Z}_{i,y,\neq\langle U_{i,:},y\rangle}\Big)=\sum_{y\in\hat{Y}_{i,x}}\frac{\alpha_y}{|\widetilde{C}_y|}\cdot\Big(\widetilde{Z}_{i,y,\neq\langle x,y\rangle}-\widetilde{Z}_{i,y,\neq\langle U_{i,:},y\rangle}\Big). \tag{2.9}$$

Let $\mathcal{W}_{i,x}$ be the event that for every $y\in Y_{i,x}\overset{\text{def}}{=}\{y\in\hat{Y}_{i,x}\,:\,|\widetilde{C}_y|=t\}$ and every $c\in\{0,1\}$, we have

$$\left|\widetilde{Z}_{i,y,c}-\frac{|\widetilde{C}_y|}{|C_y^V|}\cdot Z_{i,y,c}\right|\leqslant\Delta_y,\qquad\text{where}\qquad\Delta_y\overset{\text{def}}{=}\frac{t\cdot D_{i,x}}{2^{k+2}\cdot\alpha_y}.$$

We now show that conditioned on the event $\mathcal{W}_{i,x}$, we have $\widetilde{D}_{i,x}>0$ for any $(i,x)\in L_1\cup L_2$, and thus $\widetilde{U}_{i,:}\neq x$.

Lemma 2.5. *For any vector $x\in\{0,1\}^k$ and row $i\in[m]$, if event $\mathcal{W}_{i,x}$ occurs then it follows that $\widetilde{D}_{i,x}\geqslant\frac{1}{2}\cdot D_{i,x}-\frac{\varepsilon}{6}\cdot E_i$. In particular, if additionally $D_{i,x}>\frac{\varepsilon}{3}\cdot E_i$ then $\widetilde{D}_{i,x}>0$.*

Proof. Observe that $\widetilde{Z}_{i,y,\neq c} \in \{\widetilde{Z}_{i,y,0},\ \widetilde{Z}_{i,y,1},\ \widetilde{Z}_{i,y,0}+\widetilde{Z}_{i,y,1}\}$ for any i, y, c. Since

$$\mathbb{E}[\widetilde{Z}_{i,y,0}+\widetilde{Z}_{i,y,1}] = |\widetilde{C}_y| = \widetilde{Z}_{i,y,0}+\widetilde{Z}_{i,y,1},$$

conditioned on the event $\mathcal{W}_{i,x}$, for any $y \in Y_{i,x}$, all three random variables $\widetilde{Z}_{i,y,0}$, $\widetilde{Z}_{i,y,1}$ and $\widetilde{Z}_{i,y,0}+\widetilde{Z}_{i,y,1}$ differ from their expectation by at most Δ_y. Hence, we have

$$\left| \frac{\alpha_y}{|\widetilde{C}_y|}\widetilde{Z}_{i,y,\neq\langle x,y\rangle} - \frac{\alpha_y}{|C_y^V|}Z_{i,y,\neq\langle x,y\rangle} \right| \leqslant \frac{D_{i,x}}{2^{k+2}}.$$

The same inequality also holds for $y \in \hat{Y}_{i,x} \setminus Y_{i,x}$, since then $\widetilde{Z}_{i,y,\neq\langle x,y\rangle} = Z_{i,y,\neq\langle x,y\rangle}$ and $|\widetilde{C}_y| = |C_y^V|$ (by definition of the distribution $\mathcal{D}_{V,t}$). In combination with (2.9) we obtain

$$\widetilde{D}_{i,x} \geqslant -\frac{D_{i,x}}{2} + \sum_{y\in\hat{Y}_{i,x}} \frac{\alpha_y}{|C_y^V|}\cdot\Big(Z_{i,y,\neq\langle x,y\rangle} - Z_{i,y,\neq\langle U_{i,:},y\rangle}\Big). \tag{2.10}$$

Let $\alpha_y = (1+\gamma_y)|C_y^V|$ with $0 \leqslant \gamma_y \leqslant \frac{\varepsilon}{6}$ for any $y \in \{0,1\}^k$. By (2.8), and since $\alpha_y = |C_y^V| = |\widetilde{C}_y|$ for every $y \in \hat{Y}_{i,x} \setminus Y_{i,x}$ (by definition of distribution $\mathcal{D}_{V,t}$), we have

$$\begin{aligned}\sum_{y\in\hat{Y}_{i,x}} \frac{\alpha_y}{|C_y^V|}\left(Z_{i,y,\neq\langle x,y\rangle} - Z_{i,y,\neq\langle U_{i,:},y\rangle}\right) &= D_{i,x} + \sum_{y\in Y_{i,x}} \gamma_y\left(Z_{i,y,\neq\langle x,y\rangle} - Z_{i,y,\neq\langle U_{i,:},y\rangle}\right)\\ &\geqslant D_{i,x} - \sum_{y\in Y_{i,x}} \gamma_y Z_{i,y,\neq\langle U_{i,:},y\rangle}\\ &\geqslant D_{i,x} - \frac{\varepsilon}{6}\sum_{y\in\{0,1\}^k} Z_{i,y,\neq\langle U_{i,:},y\rangle} = D_{i,x} - \frac{\varepsilon}{6}E_i.\end{aligned}$$

Together with (2.10), we have $\widetilde{D}_{i,x} \geqslant \frac{1}{2}D_{i,x} - \frac{\varepsilon}{6}E_i$. □

We next upper bound the probability of picking a suboptimal vector x.

Claim 2.6. *For any $x \in \{0,1\}^k$ with $D_{i,x} > \frac{\varepsilon}{3}\cdot E_i$, we have*

$$\Pr[\widetilde{U}_{i,:} = x] \leqslant \sum_{y\in Y_{i,x}} \min_{c\in\{0,1\}} \Pr\left[|\widetilde{Z}_{i,y,c} - \mathbb{E}[\widetilde{Z}_{i,y,c}]| > \Delta_y\right].$$

Proof. For any $y \in \{0,1\}^k$, we have

$$\widetilde{Z}_{i,y,0}+\widetilde{Z}_{i,y,1} = |\widetilde{C}_y| = \mathbb{E}[\widetilde{Z}_{i,y,0}] + \mathbb{E}[\widetilde{Z}_{i,y,1}].$$

Further, it holds that

$$|\widetilde{Z}_{i,y,0} - \mathbb{E}[\widetilde{Z}_{i,y,0}]| = |\widetilde{Z}_{i,y,1} - \mathbb{E}[\widetilde{Z}_{i,y,1}]|,$$

and thus

$$\begin{aligned}&\Pr\left[|\widetilde{Z}_{i,y,0} - \mathbb{E}[\widetilde{Z}_{i,y,0}]| \leqslant \Delta_y\right] = \Pr\left[|\widetilde{Z}_{i,y,1} - \mathbb{E}[\widetilde{Z}_{i,y,1}]| \leqslant \Delta_y\right]\\ &= \Pr\left[|\widetilde{Z}_{i,y,0} - \mathbb{E}[\widetilde{Z}_{i,y,0}]| \leqslant \Delta_y \text{ and } |\widetilde{Z}_{i,y,1} - \mathbb{E}[\widetilde{Z}_{i,y,1}]| \leqslant \Delta_y\right].\end{aligned}$$

Since $\widetilde{U}_{i,:} = x$ can only hold if $\widetilde{D}_{i,x} \leqslant 0$, the claim follows by Lemma 2.5 and a union bound over $y \in Y_{i,x}$. □

In the following subsections, we bound the summation in (2.7) over the sets L_1 and L_2.

Case 1: Small Difference

We show first that $|L_1|$ is small (see Claim 2.7). Then, we use a simple bound for $\Pr[\widetilde{U}_{i,:} = x]$ which is based on Lemma 2.2 (see Claim 2.8).

Claim 2.7. *It holds that*

$$\sum_{(i,x)\in L_1} \sum_{y\in Y_{i,x}} |C_y^V| \leqslant 2^{k+2} \cdot \mathrm{OPT}_k^V.$$

Proof. Fix $(i,x) \in L_1$ and let $y \in Y_{i,x}$. Note that since $\langle x,y\rangle \neq \langle U_{i,:}, y\rangle$ we have

$$\{j \in C_y^V \,:\, A_{i,j} \neq \langle x,y\rangle\} \cup \{j \in C_y^V \,:\, A_{i,j} \neq \langle U_{i,:}, y\rangle\} = C_y^V.$$

Note that this union is not necessarily disjoint, e.g., if $\langle x,y\rangle \notin \{0,1\}$. Since $E_{i,x} = D_{i,x} + E_i$ (by (2.5)) and $D_{i,x} \leqslant E_i$ (by definition of L_1), we have

$$\sum_{y\in Y_{i,x}} |C_y^V| \leqslant \sum_{y\in Y_{i,x}} Z_{i,y,\neq\langle x,y\rangle} + Z_{i,y,\neq\langle U_{i,:},y\rangle} \leqslant E_{i,x} + E_i \leqslant 3E_i. \tag{2.11}$$

Fixing x and summing over all i with $(i,x) \in L_1$, the term E_i sums to at most OPT_k^V. Also summing over all $x \in \{0,1\}^k$ yields another factor 2^k. Therefore, the claim follows. □

Claim 2.8. *It holds that*

$$\sum_{(i,x)\in L_1} \Pr[\widetilde{U}_{i,:} = x] \cdot D_{i,x} \leqslant \tfrac{\varepsilon}{3} \cdot \mathrm{OPT}_k^V.$$

Proof. Note that for any row i, vector $y \in Y_{i,x}$, and $c \in \{0,1\}$, the random variable $\widetilde{Z}_{i,y,c}$ is a sum of independent Bernoulli random variables, since the t samples from C_y^V forming $\widetilde{C}_y$ are independent, and each sample contributes either 0 or 1 to $\widetilde{Z}_{i,y,c}$. Hence, our instantiations of Chebyshev's inequality, Lemmas 2.1 and 2.2, are applicable. We use Lemma 2.2 to bound

$$\Pr\left[|\widetilde{Z}_{i,y,c} - \mathbb{E}[\widetilde{Z}_{i,y,c}]| > \Delta_y\right] \leqslant \frac{\sqrt{t}}{\Delta_y}.$$

Since $\Delta_y = t \cdot D_{i,x}/(2^{k+2} \cdot \alpha_y)$ and $\alpha_y \leqslant (1+\frac{\varepsilon}{6})|C_y^V| < 2|C_y^V|$, we have

$$\Pr\left[|\widetilde{Z}_{i,y,c} - \mathbb{E}[\widetilde{Z}_{i,y,c}]| > \Delta_y\right] \leqslant \frac{2^{k+3}|C_y^V|}{\sqrt{t}\cdot D_{i,x}},$$

and thus by Claim 2.6, we obtain

$$\Pr[\widetilde{U}_{i,:} = x] \leqslant \frac{2^{k+3}}{\sqrt{t}\cdot D_{i,x}} \sum_{y\in Y_{i,x}} |C_y^V|.$$

Claim 2.7 now yields

$$\sum_{(i,x)\in L_1} \Pr[\widetilde{U}_{i,:} = x] \cdot D_{i,x} \leqslant \frac{2^{k+3}}{\sqrt{t}} \sum_{(i,x)\in L_1} \sum_{y\in Y_{i,x}} |C_y^V| \leqslant \frac{2^{2k+5}}{\sqrt{t}} \mathrm{OPT}_k^V.$$

Since we chose $t \geqslant 2^{4k+14}/\varepsilon^2$, see Theorem 2.3, we obtain the upper bound $\frac{\varepsilon}{3}\mathrm{OPT}_k^V$. □

Case 2: Large Difference

We use here the stronger instantiation of Chebyshev's inequality, Lemma 2.1, and charge $\mu = \mathbb{E}[\widetilde{Z}_{i,y,c}]$ against OPT_k^V.

Claim 2.9. *It holds that*

$$\sum_{(i,x)\in L_2} \Pr[\widetilde{U}_{i,:} = x] \cdot D_{i,x} \leqslant \frac{\varepsilon}{3} \cdot \mathrm{OPT}_k^V.$$

Proof. Fix $(i,x) \in L_2$ and let $y \in Y_{i,x}$. As in the proof of Claim 2.8, we see that our instantiation of Chebyshev's inequality, Lemma 2.1, is applicable to $\widetilde{Z}_{i,y,c}$ for any $c \in \{0,1\}$. We obtain

$$\Pr\left[|\widetilde{Z}_{i,y,c} - \mathbb{E}[\widetilde{Z}_{i,y,c}]| > \Delta_y\right] \leqslant \frac{\mathbb{E}[\widetilde{Z}_{i,y,c}]}{\Delta_y^2}.$$

Note that $\mathbb{E}[\widetilde{Z}_{i,y,c}] = Z_{i,y,c} \cdot t/|C_y^V|$, since $|\widetilde{C}_y| = t$. Using $\min_{c\in\{0,1\}} Z_{i,y,c} \leqslant Z_{i,y,\neq\langle U_{i,:},y\rangle}$, we have

$$\min_{c\in\{0,1\}} \Pr\left[|\widetilde{Z}_{i,y,c} - \mathbb{E}[\widetilde{Z}_{i,y,c}]| > \Delta_y\right] \leqslant \frac{t}{|C_y^V|\Delta_y^2} \cdot Z_{i,y,\neq\langle U_{i,:},y\rangle}.$$

Since $\Delta_y = t \cdot D_{i,x}/(2^{k+2} \cdot \alpha_y)$ and $\alpha_y \leqslant (1+\varepsilon/6)|C_y^V| < 2|C_y^V|$, we have

$$\min_{c\in\{0,1\}} \Pr\left[|\widetilde{Z}_{i,y,c} - \mathbb{E}[\widetilde{Z}_{i,y,c}]| > \Delta_y\right] \leqslant \frac{2^{2k+6} \cdot |C_y^V|}{t \cdot (D_{i,x})^2} \cdot Z_{i,y,\neq\langle U_{i,:},y\rangle}.$$

Summing over all $y \in Y_{i,x}$, Claim 2.6 yields

$$\Pr[\widetilde{U}_{i,:} = x] \leqslant \sum_{y\in Y_{i,x}} \frac{2^{2k+6} \cdot |C_y^V|}{t \cdot (D_{i,x})^2} \cdot Z_{i,y,\neq\langle U_{i,:},y\rangle}. \tag{2.12}$$

We again use inequality (2.11), i.e., $\sum_{y\in Y_{i,x}} |C_y^V| \leqslant E_{i,x} + E_i$. Since $E_{i,x} = D_{i,x} + E_i$ (by (2.5)) and $E_i < D_{i,x}$ (by definition of L_2), we have $|C_y^V| \leqslant 3D_{i,x}$ for any $y \in Y_{i,x}$. Together with (2.12), and then using the definition of E_i, we have

$$\Pr[\widetilde{U}_{i,:} = x] \cdot D_{i,x} \leqslant \frac{2^{2k+8}}{t} \sum_{y\in Y_{i,x}} Z_{i,y,\neq\langle U_{i,:},y\rangle} \leqslant \frac{2^{2k+8}}{t} E_i,$$

Fixing x and summing over all i with $(i,x) \in L_2$, the term E_i sums to at most OPT_k^V. Also summing over all $x \in \{0,1\}^k$ yields another factor 2^k. Thus, it follows that

$$\sum_{(i,x)\in L_2} \Pr[\widetilde{U}_{i,:} = x] \cdot D_{i,x} \leqslant \frac{2^{3k+8}}{t} \mathrm{OPT}_k^V.$$

Since we chose $t \geqslant 2^{3k+10}/\varepsilon$, see Theorem 2.3, we obtain the upper bound $\frac{\varepsilon}{3} \cdot \mathrm{OPT}_k^V$. □

2.2.5 Finishing the Proof

Taken together, the above claims prove Theorem 2.3.

Proof of Theorem 2.3. By Claim 2.4, splitting into L_0, L_1 and L_2, and using Claims 2.8 and 2.9, it follows for any $\varepsilon \in (0,1)$ and

$$t = 2^{4k+12}/\varepsilon^2 \tag{2.13}$$

that the expected approximate solution satisfies

$$\begin{aligned}\mathbb{E}_{\widetilde{V}}\left[\|A - \widetilde{U} \cdot V\|_0\right] &\leqslant (1+\frac{\varepsilon}{3})\mathrm{OPT}_k^V + \sum_{(i,x)\in L_1} \Pr\left[\widetilde{U}_{i,:} = x\right] \cdot D_{i,x} + \sum_{(i,x)\in L_2} \Pr\left[\widetilde{U}_{i,:} = x\right] \cdot D_{i,x} \\ &\leqslant (1+\varepsilon)\mathrm{OPT}_k^V.\end{aligned}$$

□

2.3 Efficient Sampling Algorithm

The conceptually simple PTAS in Section 2.2.3 has two running time bottlenecks, due to the exhaustive enumeration in Step 1 and Step 2. Namely, Step 1 guesses exactly the sizes $|C_y^V|$ for each $y \in \{0,1\}^k$, and there are $n^{O(2^k)}$ possibilities; and Step 2 guesses among all columns of matrix A the multiset family $\widetilde{C}$, guaranteed to exist by Theorem 2.3, and there are $n^{O(t \cdot 2^k)}$ possibilities.

Since Theorem 2.3 needs only approximate cluster sizes, it suffices in Step 1 to guess numbers α_y with $|C_y^V| \leqslant \alpha_y \leqslant (1+\frac{\varepsilon}{6})|C_y^V|$ if $|C_y^V| \geqslant t$, and $\alpha_y = |C_y^V|$ otherwise, where $t = 2^{4k+12}/\varepsilon^2$. Hence, the runtime overhead for Step 1 can be easily improved to $(t + \varepsilon^{-1}\log n)^{2^k}$.

To reduce the exhaustive enumeration in Step 2, we design an efficient sampling procedure, see Algorithm 2, that uses ideas from clustering algorithms pioneered by Kumar et al. [KSS04] and refined in [KSS05, ABS10]. Our algorithm reduces the total exhaustive enumeration in Step 2 and the guessing overhead for the approximate cluster sizes in Step 1 to $(2^k/\varepsilon)^{2^{O(k)}} \cdot (\log n)^{(\log\log n)^{0.1}}$ possibilities.

This section is structured as follows. We first replace an optimal solution (U,V) by a "well-clusterable" solution (U,W), which will help in our correctness proof. In Subsection 2.3.2 we present pseudocode for our sampling algorithm. We then prove its correctness in Subsection 2.3.3 and analyze its running time in Subsection 2.3.4. Finally, we show how to use the sampling algorithm designed in Subsection 2.3.2 together with the ideas of the simple PTAS from Subsection 2.2.3 to prove Theorem 1.1, see Subsection 2.3.5.

2.3.1 Existence of a (U,V,ε)-Clusterable Solution

For a matrix $B \in \{0,1\}^{m\times n}$ we denote by $\mathrm{ColSupp}(B)$ the set of unique columns of B. Note that if the columns of U are linearly independent then $U \cdot \mathrm{ColSupp}(V)$ denotes the set of distinct columns of $U \cdot V$. In the clustering formulation of the Generalized Binary ℓ_0-Rank-k problem, as discussed in the introduction (2.1), the set $U \cdot \mathrm{ColSupp}(V)$ corresponds to the set of cluster centers.

Given matrices U, V, we will first replace V by a related matrix W in a way that makes all centers of $U \cdot \mathrm{ColSupp}(W)$ sufficiently different without increasing the cost too much, as formalized in the following.

Lemma 2.10. *For any $U \in \{0,1\}^{m\times k}$, $V \in \{0,1\}^{k\times n}$ and $\varepsilon \in (0,1)$, there exists a matrix $W \in \{0,1\}^{k\times n}$ such that $\|A - U\cdot W\|_0 \leqslant (1+\varepsilon)\|A - U \cdot V\|_0$ and for any distinct $y,z \in \mathrm{ColSupp}(W)$ we have*
(i) $\|U\cdot y - U\cdot z\|_0 > \varepsilon \cdot 2^{-k} \cdot \|A - U\cdot V\|_0 / \min\{|C_y^W|, |C_z^W|\}$, and
(ii) $\|A_{:,j} - U\cdot y\|_0 \leqslant \|A_{:,j} - U \cdot z\|_0$ for every $j \in C_y^W$.
We say that such a matrix W is (U,V,ε)-clusterable.

Proof. The proof is by construction of W. We initialize $W \overset{\text{def}}{=} V$ and then iteratively resolve violations of *(i)* and *(ii)*. In each step, resulting in a matrix W', we ensure that $\mathrm{ColSupp}(W') \subseteq \mathrm{ColSupp}(W)$. We call this *support-monotonicity*.

We can resolve all violations of *(ii)* at once by iterating over all columns $j \in [n]$ and replacing $W_{:,j}$ by the vector $z \in \mathrm{ColSupp}(W)$ minimizing $\|A_{:,j} - U\cdot z\|_0$. This does not increase the cost $\|A - U \cdot W\|_0$ and results in a matrix W' without any violations of *(ii)*.

So assume that there is a violation of *(i)*. That is, for distinct $y, z \in \mathrm{ColSupp}(W)$, where we can assume without loss of generality that $|C_y^W| \leqslant |C_z^W|$, we have $\|U\cdot y - U\cdot z\|_0 \leqslant \varepsilon\cdot 2^{-k}\cdot\|A - U\cdot V\|_0/|C_y^W|$. We change the matrix W by replacing for every $j \in C_y^W$ the column $W_{:,j} = y$ by z. Call the resulting matrix W'. Note that the cost of any replaced column j changes to

$$\begin{aligned}\|A_{:,j} - U\cdot W'_{:,j}\|_0 = \|A_{:,j} - U\cdot z\|_0 &\leqslant \|A_{:,j} - U\cdot y\|_0 + \|U\cdot y - U\cdot z\|_0 \\ &\leqslant \|A_{:,j} - U\cdot W_{:,j}\|_0 + \varepsilon\cdot 2^{-k}\cdot\|A - U\cdot V\|_0/|C_y^W|.\end{aligned}$$

Since the number of replaced columns is $|C_y^W|$, the overall cost increase is at most $\varepsilon\cdot 2^{-k}\cdot\|A - U\cdot V\|_0$. Note that after this step the size of $\mathrm{ColSupp}(W)$ is reduced by 1, since we removed any occurrence of column y. By support-monotonicity, the number of such steps is bounded by 2^k. Since resolving violations of *(ii)* does not increase the cost, the final cost is bounded by $(1+\varepsilon)\|A - U\cdot V\|_0$.

After at most 2^k times resolving a violation of *(i)* and then all violations of *(ii)*, we end up with a matrix W without violations and the claimed cost bound. □

2.3.2 The Algorithm Sample

Given $A \in \{0,1\}^{m\times n}$, $k \in \mathbb{N}$, $\varepsilon \in (0,1)$, and $t \in \mathbb{N}$, fix any optimal solution U, V, i.e., $\|A - U\cdot V\|_0 = \mathrm{OPT}_k$. Our proof will use the additional structure provided by well-clusterable solutions. Therefore, fix any (U, V, ε)-clusterable matrix W as in Lemma 2.10. Since $\|A - U \cdot W\|_0 \leqslant (1+\varepsilon)\|A - U\cdot V\|_0$, we can restrict to matrix W. Specifically, we fix the optimal partitioning C^W of $[n]$ for the purpose of the analysis and for the guessing steps of the algorithm. Our goal is to sample from the distribution $\mathcal{D}_{W,t}$.

Pseudocode of our sampling algorithm $\mathbf{Sample}_{A,k,\varepsilon,t}(M, \mathcal{N}, \widetilde{R}, \widetilde{C}, \alpha)$ is given below. The arguments of this procedure are as follows. Matrix M is the current submatrix of A (initialized as the full matrix A). Set $\mathcal{N} \subseteq \{0,1\}^k$ is the set of clusters that we did not yet sample from (initialized to $\{0,1\}^k$). The sequence $\widetilde{R}$ stores "representatives" of the clusters that we already sampled from (initialized to undefined entries $(\perp, \ldots, \perp)$). The sequence $\widetilde{C}$ contains our samples, so in the end we want $\widetilde{C}$ to be drawn according to $\mathcal{D}_{W,t}$ ($\widetilde{C}$ is initialized such that $\widetilde{C}_y = \emptyset$ for all $y \in \{0,1\}^k$). Finally, α contains guesses for the sizes of the clusters that we already sampled from, so in the end we want it to be a sequence of $\frac{\varepsilon}{6}$-approximate cluster sizes (α is initialized such that $\alpha_y = 0$ for all $y \in \{0,1\}^k$). This algorithm is closely related to algorithm "Irred-k-means" by Kumar et al. [KSS04], see the introduction for a discussion.

In this algorithm, at the base case we call $\mathbf{EstimateBestResponse}_{A,k}(\widetilde{C}, \alpha)$, which computes matrix $\widetilde{U} = \widetilde{U}(\widetilde{C}, \alpha)$ and a best response $\widetilde{V}$ to $\widetilde{U}$. Apart from the base case, there are three phases of algorithm **Sample**. In the *sampling* phase, we first guess some $y \in \mathcal{N}$ and an approximation α_y of $|C_y^W|$. Then we sample $\min\{t, \alpha_y\}$ columns from M to form a multiset $\widetilde{C}_y$, and we sample one column from M to form $\widetilde{R}_y$. We make a recursive call with y removed from $\mathcal{N}$ and updated $\widetilde{R}, \widetilde{C}, \alpha$ by the values $\widetilde{R}_y, \widetilde{C}_y, \alpha_y$. As an intermediate solution, we let $U^{(1)}, V^{(1)}$ be the best solution returned by the recursive calls over all exhaustive guesses. In the *pruning* phase, we delete the $n_M/2$ columns of M that are closest to $\widetilde{R}$, and we make a recursive call with the resulting matrix M', not changing the remaining arguments. Denote the returned solution by $U^{(2)}, V^{(2)}$. Finally, in the *decision* phase we return the better solution between $U^{(1)}, V^{(1)}$ and $U^{(2)}, V^{(2)}$.

Algorithm 2 Efficient Sampling

```
**Sample**$_{A,k,\varepsilon,t}(M,\ \mathcal{N},\ \widetilde{R},\ \widetilde{C},\ \alpha)$
    let $n_M$ be the number of columns of $M$
    set $\nu \overset{\mathrm{def}}{=} (\varepsilon/2^{k+4})^{2^k+2-|\mathcal{N}|}$

  **If** $\mathcal{N} = \emptyset$ **or** $n_M = 0$: **Return** $(\widetilde{U}, \widetilde{V}) = $ **EstimateBestResponse**$_{A,k}(\widetilde{C}, \alpha)$

    **\* Sampling phase \***
  **Guess** $y \in \mathcal{N}$
  **Guess** whether $|C_y^W| < t$:
      **If** $|C_y^W| < t$:  **Guess** $\alpha_y \overset{\mathrm{def}}{=} |C_y^W|$ exactly, i.e. $\alpha_y \in \{0, 1, \ldots, t-1\}$
      **Otherwise:**  **Guess** $\nu \cdot n_M \leqslant \alpha_y \leqslant n_M$ such that $|C_y^W| \leqslant \alpha_y \leqslant (1+\frac{\varepsilon}{6})|C_y^W|$
  **If** $\alpha_y = 0$: $(U^{(y,\alpha_y)}, V^{(y,\alpha_y)}) = $ **EstimateBestResponse**$_{A,k}(\widetilde{C}, \alpha)$
  **Else**
      Sample u.a.r. $\min\{t, \alpha_y\}$ columns from $M$; let $\widetilde{C}_y$ be the resulting multiset [4]
      Sample u.a.r. one column from $M$; call it $\widetilde{R}_y$
      $(U^{(y,\alpha_y)}, V^{(y,\alpha_y)}) = $ **Sample**$_{A,k,\varepsilon,t}(M,\ \mathcal{N}\backslash\{y\},\ \widetilde{R}\cup\{\widetilde{R}_y\},\ \widetilde{C}\cup\{\widetilde{C}_y\},\ \alpha\cup\{\alpha_y\})$
  Let $(U^{(1)}, V^{(1)})$ be the pair minimizing $\|A - U^{(y,\alpha_y)} \cdot V^{(y,\alpha_y)}\|_0$ over all guesses $y$ and $\alpha_y$

    **\* Pruning phase \***
  Let $M'$ be matrix $M$ after the deleting $n_M/2$ closest columns to $\widetilde{R}$,
     i.e., the $n_M/2$ columns $M_{:,j}$ with smallest values $\min_{y\in\{0,1\}^k\backslash\mathcal{N}}\|M_{:,j} - \widetilde{R}_y\|_0$
  $(U^{(2)},\ V^{(2)}) = $ **Sample**$_{A,k,\varepsilon,t}(M',\ \mathcal{N},\ \widetilde{R},\ \widetilde{C},\ \alpha)$

    **\* Decision \***
  **Return** $(U^{(\ell)}, V^{(\ell)})$ with the minimal value $\|A - U^{(\ell)} \cdot V^{(\ell)}\|_0$ over $\ell \in \{1, 2\}$.
```

[4] Given a submatrix M of A, and t columns sampled u.a.r. from M, we denote by $\widetilde{C}_y$ the resulting multiset of column indices with respect to the original matrix A.

Algorithm 3 Estimating Best Response

```
**EstimateBestResponse**$_{A,k}(\widetilde{C}, \alpha)$
1. *(Compute $\widetilde{U}$)* Compute a matrix $\widetilde{U} = \widetilde{U}(\widetilde{C}, \alpha)$, where row $\widetilde{U}_{i,:}$ is any vector $x$ minimizing the estimated cost $\widetilde{E}_{i,x}$. Note that each row $\widetilde{U}_{i,:} \in \{0,1\}^k$ can be optimized independently.
2. *(Compute $\widetilde{V}$)* Compute $\widetilde{V}$ as a best response to $\widetilde{U}$.
3. **Return** $(\widetilde{U}, \widetilde{V})$
```

2.3.3 Correctness of Algorithm Sample

With notation as above, we now prove correctness of algorithm **Sample**.

Theorem 2.11. *Algorithm* **Sample**$_{A,k,\varepsilon,t}$ *generates a recursion tree which with probability at least* $(\frac{\varepsilon}{2t})^{2^{O(k)} \cdot t}$ *has a leaf calling* **EstimateBestResponse**$_{A,k}(\widetilde{C}, \alpha)$ *such that*

(i) α *is a sequence of* $\frac{\varepsilon}{6}$*-approximate cluster sizes (w.r.t. the fixed matrix* W*), and*

(ii) $\widetilde{C}$ *is drawn according to distribution* $\mathcal{D}_{W,t}$.

The rest of this section is devoted to proving Theorem 2.11. Similarly as in the algorithm, we define parameters

$$\gamma \overset{\text{def}}{=} \varepsilon/2^{k+4} \qquad \text{and} \qquad \nu_i \overset{\text{def}}{=} \gamma^{2^k+2-i}.$$

Sort $\{0,1\}^k = \{y_1, \ldots, y_{2^k}\}$ such that $|C^W_{y_1}| \leqslant \ldots \leqslant |C^W_{y_{2^k}}|$. We construct the leaf guaranteed by the theorem inductively. In each depth $d = 0, 1, \ldots$ we focus on one recursive call, see Algorithm 2,

$$\textbf{Sample}_{A,k,\varepsilon,t}(M^{(d)},\ \mathcal{N}^{(d)},\ \widetilde{R}^{(d)},\ \widetilde{C}^{(d)},\ \alpha^{(d)}).$$

We consider the partitioning $P^{(d)} \overset{\text{def}}{=} \{P^{(d)}_y\}_{y \in \{0,1\}^k}$ induced by the partitioning C^W on $M^{(d)}$, i.e., $P^{(d)}_y$ is the set C^W_y restricted to the columns of A that appear in the submatrix $M^{(d)}$. We claim that we can find a root-to-leaf path such that the following inductive invariants hold with probability at least $(\nu_0/t)^{(2^k - |\mathcal{N}^{(d)}|)(t+1)}$:

I1. $P^{(d)}_y = C^W_y$ for all $y \in \mathcal{N}^{(d)}$, i.e., no column of an unsampled cluster has been removed,

I2. $\mathcal{N}^{(d)} = \{y_1, \ldots, y_{|\mathcal{N}^{(d)}|}\}$, i.e., the remaining clusters are the $|\mathcal{N}^{(d)}|$ smallest clusters,

I3. For any $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$ the value $\alpha^{(d)}_y$ is an $\frac{\varepsilon}{6}$-approximate cluster size, i.e., if $|C^W_y| < t$ we have $\alpha^{(d)}_y = |C^W_y|$, and otherwise $|C^W_y| \leqslant \alpha^{(d)}_y \leqslant (1 + \frac{\varepsilon}{6})|C^W_y|$,

I4. For any $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$ the multiset $\widetilde{C}^{(d)}_y$ is sampled according to distribution $\mathcal{D}_{W,t}$, i.e., if $|C^W_y| < t$ then $\widetilde{C}^{(d)}_y = C^W_y$ and otherwise $\widetilde{C}^{(d)}_y$ consists of t uniformly random samples from C^W_y with replacement, and

I5. For any $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$ the vector $\widetilde{R}^{(d)}_y$ satisfies $\|\widetilde{R}^{(d)}_y - U \cdot y\|_0 \leqslant 2\|A - U \cdot W\|_0 / |C^W_y|$.

For shorthand, we let $n^{(d)} \overset{\text{def}}{=} n_{M^{(d)}}$ and $\nu^{(d)} \overset{\text{def}}{=} \nu_{|\mathcal{N}^{(d)}|}$.

Base Case: Note that the recursion may stop in Step 1 with $\mathcal{N}^{(d)} = \emptyset$ or $n^{(d)} = 0$, or in Step 6 with $\alpha^{(d)}_y = 0$ for some guessed $y \in \mathcal{N}$. Since we only want to show existence of a leaf of the recursion tree, in the latter case we can assume that we guessed $y = y_{|\mathcal{N}^{(d)}|}$ and $\alpha^{(d)}_y = |C^W_y|$, and thus we have $|C^W_y| = 0$. Hence, in all three cases we have $|C^W_y| = 0$ for all $y \in \mathcal{N}^{(d)}$, by invariant I2 and sortedness of $y_1, \ldots, y_{2^k}$. Since we initialize $\widetilde{C}^{(0)}_y = \emptyset$ and $\alpha^{(0)}_y = 0$, we are done for all $y \in \mathcal{N}^{(d)}$. By invariants I3 and I4, we are also done for all $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$. The total success probability is at least

$$\left(\frac{\nu_0}{t}\right)^{2^k(t+1)} = \left(\frac{\varepsilon}{2^{k+4}t}\right)^{2^k(2^k+2)(t+1)} = \left(\frac{\varepsilon}{2t}\right)^{2^{O(k)} \cdot t}.$$

The proof of the inductive step proceeds by case distinction.

Case 1 (Sampling): Suppose $|P_y^{(d)}| \geqslant \nu^{(d)} \cdot n^{(d)}$ for some $y \in \mathcal{N}^{(d)}$. Since $P_y^{(d)} = C_y^W$ (by invariant I1) and sortedness, we have $|C_y^W| \geqslant \nu^{(d)} n^{(d)}$ for $y \overset{\text{def}}{=} y_{|\mathcal{N}^{(d)}|}$. We may assume that we guess $y = y_{|\mathcal{N}^{(d)}|}$ in Step 2, since we only want to prove existence of a leaf of the recursion tree. Note that there is a number $\nu^{(d)} n^{(d)} \leqslant \alpha_y \leqslant n^{(d)}$ with $|C_y^W| \leqslant \alpha_y \leqslant (1 + \frac{\varepsilon}{6})|C_y^W|$ (in particular $\alpha_y = |C_y^W|$ would work), so we can guess such a number in Step 5. Together with Steps 3 and 4, we can assume that $\alpha^{(d+1)}$ satisfies invariant I3.

In Step 8 we sample a multiset $\widetilde{C}_y$ of $\min\{t, \alpha_y\}$ columns from M. If $|C_y^W| \geqslant t$, we condition on the event that all these columns lie in C_y^W. Then $\widetilde{C}_y$ forms a uniform sample from C_y^W of size t. Since $|C_y^W| \geqslant \nu^{(d)} n^{(d)}$, this event has probability at least $(\nu^{(d)})^t$. Otherwise, if $|C_y^W| = \alpha_y < t$, we condition on the event that all α_y samples lie in C_y^W and are distinct. Then $\widetilde{C}_y = C_y^W$. The probability of this event is at least

$$\left(\frac{1}{n^{(d)}}\right)^{\alpha_y} \geqslant \left(\frac{\nu^{(d)}}{\alpha_y}\right)^{\alpha_y} \geqslant \left(\frac{\nu^{(d)}}{t}\right)^{t}.$$

In total, $\widetilde{C}^{(d+1)}$ satisfies invariant I4 with probability at least $(\nu^{(d)}/t)^t$.

In Step 9 we sample one column $\widetilde{R}_y$ uniformly at random from M. With probability at least $\nu^{(d)}$, $\widetilde{R}_y$ belongs to C_y^W, and conditioned on this event $\mathcal{E}_y$ we have

$$\mathbb{E}_{\widetilde{R}_y}\left[\|\widetilde{R}_y - U \cdot y\|_0 \;\middle|\; \mathcal{E}_y\right] = \frac{1}{|C_y^W|} \sum_{j \in C_y^W} \|A_{:,j} - U \cdot y\|_0 \leqslant \frac{\|A - U \cdot W\|_0}{|C_y^W|}.$$

By Markov's inequality, with probability at least $\nu^{(d)}/2$ we have $\|\widetilde{R}_y - U \cdot y\|_0 \leqslant 2\|A - U \cdot W\|_0/|C_y^W|$, and thus invariant I5 holds for $\widetilde{R}^{(d+1)}$.

Finally, since we did not change $M^{(d)}$, invariant I1 is maintained. We conditioned on events that hold with combined probability at least $(\nu^{(d)}/t)^t \cdot \nu^{(d)}/2 \geqslant (\nu_0/t)^{t+1}$. Since we decrement $|\mathcal{N}^{(d)}|$ by removing $y = y_{|\mathcal{N}^{(d)}|}$ from $\mathcal{N}^{(d)}$, we maintain invariant I2, and we obtain total probability at least $(\nu_0/t)^{(2^k - |\mathcal{N}^{(d+1)}|)(t+1)}$.

Case 2 (Pruning): Suppose $|P_y^{(d)}| < \nu^{(d)} \cdot n^{(d)}$ for every $y \in \mathcal{N}^{(d)}$. (Note that cases 1 and 2 are complete.) In this case, we remove the $n^{(d)}/2$ columns of $M^{(d)}$ that are closest to $\widetilde{R}^{(d)}$, resulting in a matrix $M^{(d+1)}$, and then start a recursive call on $M^{(d+1)}$. Since we do not change $\mathcal{N}^{(d)}, \widetilde{R}^{(d)}, \widetilde{C}^{(d)}$, and $\alpha^{(d)}$, invariants I2-I5 are maintained.

Invariant I1 is much more difficult to verify, as we need to check that the $n^{(d)}/2$ deleted columns do not contain any column from an unsampled cluster. We first show that *some* column of a cluster we already sampled from *survives* to depth $d+1$ and has *small* distance to $\widetilde{R}^{(d)}$ (see Claim 2.12). Then we show that *every* column of a cluster that we did not yet sample from has *large* distance to $\widetilde{R}^{(d)}$ (see Claim 2.14). Since we delete the $n^{(d)}/2$ closest columns to $\widetilde{R}^{(d)}$, it follows that every column of a cluster that we did not yet sample from *survives.*

Claim 2.12. *There exists $x \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$ and column $j \in P_x^{(d+1)}$ with*

$$\|A_{:,j} - \widetilde{R}_x^{(d)}\|_0 \leqslant 2^{k+4}\|A - U \cdot W\|_0/n^{(d)}.$$

Proof. By Case 2, we have $|P_y^{(d)}| < \nu^{(d)} \cdot n^{(d)}$ for every $y \in \mathcal{N}^{(d)}$, and since $\nu^{(d)} \leqslant \nu_{2^k} \leqslant 2^{-k-2}$ it follows that

$$\sum_{y \in \mathcal{N}^{(d)}} |P_y^{(d)}| < 2^k \nu^{(d)} n^{(d)} \leqslant n^{(d)}/4.$$

Since $|P_y^{(d)}| \geqslant |P_y^{(d+1)}|$ and $\sum_{y \in \{0,1\}^k} |P_y^{(d+1)}| = n^{(d)}/2$, we obtain $\sum_{y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}} |P_y^{(d+1)}| \geqslant n^{(d)}/4$. Hence, there is $x \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$ such that

$$|P_x^{(d+1)}| \geqslant 2^{-k-2} n^{(d)}. \tag{2.14}$$

By the minimum-arithmetic-mean inequality, there is $j \in P_x^{(d+1)}$ such that

$$\begin{aligned}\|A_{:,j} - \widetilde{R}_x^{(d)}\|_0 &\leqslant \frac{1}{|P_x^{(d+1)}|} \cdot \sum_{j' \in P_x^{(d+1)}} \|A_{:,j'} - \widetilde{R}_x^{(d)}\|_0 \\ &\leqslant \|\widetilde{R}_x^{(d)} - U \cdot x\|_0 + \frac{1}{|P_x^{(d+1)}|} \cdot \sum_{j' \in P_x^{(d+1)}} \|A_{:,j'} - U \cdot x\|_0,\end{aligned}$$

where the last step uses the triangle inequality. For the first summand we use invariant I5, and for the second we use that $P_x^{(d+1)}$ is by definition part of an induced partitioning of C^W on a smaller matrix, and thus the summation is bounded by $\|A - UW\|_0$. This yields

$$\|A_{:,j} - \widetilde{R}_x^{(d)}\|_0 \leqslant \left(\frac{2}{|C_x^W|} + \frac{1}{|P_x^{(d+1)}|} \right) \cdot \|A - U \cdot W\|_0.$$

By $P_x^{(d+1)} \subseteq C_x^W$ and by (2.14), we obtain the claimed bound. □

Claim 2.13. *For any $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$ we have $|C_y^W| \geqslant \nu^{(d)} n^{(d)}/\gamma$.*

Proof. Since $y \notin \mathcal{N}$, we sampled from this cluster in some depth $d' < d$. In the call corresponding to d', we had $\mathcal{N}^{(d')} \supseteq \mathcal{N}^{(d)} \cup \{y\}$ and thus $|\mathcal{N}^{(d')}| \geqslant |\mathcal{N}^{(d)}| + 1$, and we had $n^{(d')} \geqslant n^{(d)}$. Since we sampled from C_y^W in depth d', Case 1 was applicable, and thus

$$\begin{aligned}|C_y^W| &\overset{(\mathrm{I1})}{=} |P_y^{(d')}| \geqslant \nu^{(d')} \cdot n^{(d')} = \nu_{|\mathcal{N}^{(d')}|} \cdot n^{(d')} \\ &\geqslant \nu_{|\mathcal{N}^d|+1} \cdot n^{(d)} = \nu_{|\mathcal{N}^d|} \cdot n^{(d)}/\gamma = \nu^{(d)} \cdot n^{(d)}/\gamma.\end{aligned}$$ □

Claim 2.14. *For any $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$, $z \in \mathcal{N}^{(d)}$, and $j \in P_z^{(d)}$ we have*

$$\|A_{:,j} - \widetilde{R}_y^{(d)}\|_0 > 2^{k+4} \|A - U \cdot W\|_0 / n^{(d)}.$$

Proof. By triangle inequality, we have $\|U \cdot y - U \cdot z\|_0 \leqslant \|A_{:,j} - U \cdot y\|_0 + \|A_{:,j} - U \cdot z\|_0$. Since $j \in P_z^{(d)} = C_z^W$ and by property *(ii)* of (U, V, ε)-clustered (see Lemma 2.10), the first summand is at least as large as the second, and we obtain

$$\|U \cdot y - U \cdot z\|_0 \leqslant 2\|A_{:,j} - U \cdot y\|_0.$$

We use this and the triangle inequality to obtain

$$\begin{aligned}\|A_{:,j} - \widetilde{R}_y^{(d)}\|_0 &\geqslant \|A_{:,j} - U \cdot y\|_0 - \|\widetilde{R}_y^{(d)} - U \cdot y\|_0 \\ &\geqslant \frac{1}{2} \|U \cdot y - U \cdot z\|_0 - \|\widetilde{R}_y^{(d)} - U \cdot y\|_0.\end{aligned}$$

For the first summand we use property *(i)* of (U, V, ε)-clustered (see Lemma 2.10), for the second we use invariant I5. This yields

$$\|A_{:,j} - \widetilde{R}_y^{(d)}\|_0 > \frac{\varepsilon}{2^{k+1}} \cdot \frac{\|A - U \cdot V\|_0}{|C_z^W|} - \frac{2\|A - U \cdot W\|_0}{|C_y^W|}.$$

Since $y \in \{0,1\}^k \setminus \mathcal{N}^{(d)}$, Claim 2.13 yields $|C_y^W| \geqslant \nu^{(d)} n^{(d)}/\gamma$. Since $z \in \mathcal{N}^{(d)}$, by invariant I1, and since we are in Case 2, we have $|C_z^W| = |P_z^{(d)}| < \nu^{(d)} \cdot n^{(d)}$. Moreover, by the properties of (U, V, ε)-clustered (see Lemma 2.10), we have $\|A - UW\|_0 \leqslant (1 + \varepsilon)\|A - U \cdot V\|_0$ and thus $\|A - U \cdot V\|_0 \geqslant \frac{1}{2}\|A - U \cdot W\|_0$. Together, this yields

$$\begin{aligned}\|A_{:,j} - \widetilde{R}_y^{(d)}\|_0 &> \left(\frac{\varepsilon}{2^{k+2}} - 2\gamma \right) \cdot \frac{\|A - U \cdot W\|_0}{\nu^{(d)} n^{(d)}} \\ &= \frac{\varepsilon}{2^{k+3} \nu^{(d)}} \cdot \frac{\|A - U \cdot W\|_0}{n^{(d)}} \\ &\geqslant 2^{k+4} \cdot \frac{\|A - U \cdot W\|_0}{n^{(d)}},\end{aligned}$$

since $\gamma = \varepsilon/2^{k+4}$ and $\nu^{(d)} \leqslant \nu_{2^k} = \gamma^2 \leqslant \varepsilon/2^{2k+7}$. □

Together, Claims 2.12 and 2.14 prove that no column $j \in P_y^{(d)}$ with $y \in \mathcal{N}^{(d)}$ is removed. Indeed, we remove the columns with smallest distance to $\widetilde{R}^{(d)}$, some column in distance $2^{k+4}\|A - UW\|_0/n^{(d)}$ survives, and any column $j \in P_y^{(d)}$ with $y \in \mathcal{N}$ has larger distance to $\widetilde{R}^{(d)}$. It follows that invariant I1 is maintained, completing our proof of correctness.

2.3.4 Running Time Analysis of Algorithm Sample

We now analyze the running time of Algorithm 3.

Lemma 2.15. *Algorithm* **EstimateBestResponse** *runs in time* $2^{O(k)}mn$.

Proof. Note that if $\widetilde{C}$ is drawn according to distribution $\mathcal{D}_{W,t}$, then its total size $\sum_{y\in\{0,1\}^k}|\widetilde{C}_y|$ is at most n. Hence, we can ignore all calls violating this inequality. We can thus evaluate the estimated cost $\widetilde{E}_{i,x}$ in time $2^{O(k)}n$. Optimizing over all $x \in \{0,1\}^k$ costs another factor 2^k, and iterating over all rows i adds a factor m. Thus, Step 1 runs in time $2^{O(k)}mn$. Further, Step 2 finds a best response matrix, which can be computed in the same running time. □

We proceed by analyzing the time complexity of Algorithm 2.

Lemma 2.16. *For any* $t = \mathrm{poly}(2^k/\varepsilon)$*, Algorithm* $\textbf{Sample}_{A,k,\varepsilon,t}$ *runs in time* $(2/\varepsilon)^{2^{O(k)}} \cdot mn^{1+o(1)}$*, where* $o(1)$ *hides a factor* $(\log\log n)^{1.1} / \log n$.

Proof. Consider any recursive call $\textbf{Sample}_{A,k,\varepsilon,t}(M,\ \mathcal{N},\ \widetilde{R},\ \widetilde{C},\ \alpha)$. We express its running time as $T(a,b)$ where $a \overset{\text{def}}{=} |\mathcal{N}|$ and $b \overset{\text{def}}{=} \log(n_M)$. For notational convenience, we let $\log(0) =: -1$ and assume that n_M is a power of 2.

If we make a call to algorithm **EstimateBestResponse** then this takes time $2^{O(k)} \cdot mn$ by the preceding lemma. Note that here we indeed have the size n of the original matrix and not the size n_M of the current submatrix, since we need to determine the cost with respect to the original matrix.

In the sampling phase, in Step 2 we guess y, with $|\mathcal{N}| \leqslant 2^k$ possibilities. Moreover, in Steps 3-5 we guess either $\alpha_y \in \{0,1,\dots,t-1\}$ or $\nu_{|\mathcal{N}|} \cdot n_M \leqslant \alpha_y \leqslant n_M$ such that $|C_y^W| \leqslant \alpha_y \leqslant (1+\frac{\varepsilon}{6})|C_y^W|$. Note that there are $O(\log(1/\nu_{|\mathcal{N}|})/\log(1+\varepsilon/6)) = \mathrm{poly}(2^k/\varepsilon)$ possibilities for the latter, and thus $\mathrm{poly}(2^k/\varepsilon)$ possibilities in total. For each such guess we make one recursive call with a decremented a and we evaluate the cost of the returned solution in time $2^{O(k)} \cdot mn$.

In the pruning phase, we delete the $n_M/2$ columns that are closest to $\widetilde{R}$, which can be performed in time $2^{O(k)} \cdot m \cdot n_M$ (using median-finding in linear time). We then make one recursive call with a decremented b.

Together, we obtain the recursion

$$T(a,b) \leqslant \mathrm{poly}(2^k/\varepsilon)mn + \mathrm{poly}(2^k/\varepsilon) \cdot T(a-1,b) + T(a,b-1),$$

with base cases $T(0,b) = T(a,-1) = 2^{O(k)}mn$. The goal is to upper bound $T(2^k, \log n)$.

Let $Y = \mathrm{poly}(2^k/\varepsilon)$ and $X = Y \cdot mn$ such that

$$T(a,b) \leqslant X + Y \cdot T(a-1,b) + T(a,b-1),$$

and $T(0,b), T(a,-1) \leqslant X$. We prove by induction that $T(a,b) \leqslant X \cdot (2Y(b+2))^a$. This works in the base cases where $a = 0$ or $b = -1$. Inductively, for $a > 0$ and $b \geqslant 0$ we bound [5]

$$\begin{aligned} T(a,b) &\leqslant X + Y \cdot X \cdot (2Y(b+2))^{a-1} + X \cdot (2Y(b+1))^a \\ &= X \cdot (2Y(b+2))^a \cdot \left(\frac{1}{(2Y(b+2))^a} + \frac{1}{2(b+2)} + \left(\frac{b+1}{b+2}\right)^a\right) \\ &\leqslant X \cdot (2Y(b+2))^a \cdot \left(\frac{1}{2(b+2)} + \frac{1}{2(b+2)} + \frac{b+1}{b+2}\right) \\ &= X \cdot (2Y(b+2))^a. \end{aligned} \tag{2.15}$$

Let C be a constant to be determined soon. Using (2.15), the total running time is bounded by

$$T(2^k, \log n) \leqslant X \cdot (2Y(\log(n)+2))^{2^k} \leqslant (2/\varepsilon)^{2^{(C+1)\cdot k}} \cdot mn \cdot \log^{2^k} n \leqslant (2/\varepsilon)^{2^{(C+1)\cdot k}} \cdot mn^{1+o(1)},$$

where the last inequality follows by noting that $\log^{2^k} n > (2/\varepsilon)^{2^{(C+1)\cdot k}}$ iff $k < \log\left(\frac{\log\log n}{\log(2/\varepsilon)}\right)^{1/C}$ and in this case

$$(\log n)^{2^k} \leqslant (\log n)^{\left(\frac{\log\log n}{\log(2/\varepsilon)}\right)^{1/C}} \leqslant n^{o(1)},$$

where $o(1)$ hides a factor $(\log\log n)^{1+1/C} / \log n$. The statement follows for any $C \geqslant 10$. □

[5] Using similar arguments, for any $\alpha \in [0,1]$ the recurrence $T(a,b) \leqslant (1+\alpha)^b \cdot X + Y \cdot T(a-1,b) + T(a,b-1)$ is upper bounded by $X \cdot (2Y)^a \cdot (b^{1-\alpha}+2)^a \cdot (1+\alpha)^b$. In particular, this implies that $T(a,b) \leqslant X \cdot (2Y)^a \cdot \min\left\{(b+2)^a, 2^{a+b}\right\}$ and thus $T(2^k, \log n) \leqslant (2/\varepsilon)^{2^{O(k)}} \cdot mn \cdot \min\left\{(\log n)^{2^k}, n\right\}$.

2.3.5 The Complete PTAS: Correctness and Efficiency

Finally, we use Algorithm **Sample** to obtain an efficient PTAS for the Generalized Binary ℓ_0-Rank-k problem. Given A, k and ε, we call $\textbf{Sample}_{A,k,\varepsilon/4,t}$ with

$$t = t(k, \varepsilon/4) \overset{\text{def}}{=} 2^{4k+16}/\varepsilon^2.$$

(This means that we replace all occurrences of ε by $\varepsilon/4$, in particular we also assume that W is $(U, V, \frac{\varepsilon}{4})$-clusterable.) By Theorem 2.11, with probability at least

$$\left(\frac{\varepsilon}{2t}\right)^{2^{O(k)}\cdot t} = \left(\frac{\varepsilon}{2}\right)^{2^{O(k)}/\varepsilon^2}$$

at least one leaf of the recursion tree calls $\textbf{EstimateBestResponse}_{A,k}(\widetilde{C}, \alpha)$ with proper $\widetilde{C}$ and α such that the Sampling Theorem 2.3 is applicable. By choice of $t = t(k, \frac{\varepsilon}{4})$, this yields

$$\mathbb{E}\big[\|A - \widetilde{U}(\widetilde{C}, \alpha) \cdot W\|_0\big] \leqslant \left(1 + \frac{\varepsilon}{4}\right) \mathrm{OPT}_k^W \leqslant \left(1 + \frac{\varepsilon}{4}\right)^2 \mathrm{OPT}_k,$$

where we used that W is $(U, V, \frac{\varepsilon}{4})$-clusterable in the second step (see Lemma 2.10). The Algorithm **EstimateBestResponse** computes a matrix $\widetilde{U} = \widetilde{U}(\widetilde{C}, \alpha)$ and its best response matrix $\widetilde{V}$. This yields

$$\mathbb{E}\big[\|A - \widetilde{U} \cdot \widetilde{V}\|_0\big] \leqslant \mathbb{E}\big[\|A - \widetilde{U} \cdot W\|_0\big] \leqslant \left(1 + \frac{\varepsilon}{4}\right)^2 \mathrm{OPT}_k.$$

By Markov's inequality, with probability at least $1 - 1/(1 + \varepsilon/4) \geqslant \frac{\varepsilon}{5}$ we have

$$\|A - \widetilde{U} \cdot \widetilde{V}\|_0 \leqslant \left(1 + \frac{\varepsilon}{4}\right) \cdot \mathbb{E}\big[\|A - \widetilde{U} \cdot \widetilde{V}\|_0\big] \leqslant \left(1 + \frac{\varepsilon}{4}\right)^3 \mathrm{OPT}_k \leqslant (1 + \varepsilon)\mathrm{OPT}_k.$$

Hence, with probability at least $p = (\varepsilon/2)^{2^{O(k)}/\varepsilon^2}$ at least one solution $\widetilde{U}, \widetilde{V}$ generated by our algorithm is a $(1 + \varepsilon)$-approximation. Since we return the best of the generated solutions, we obtain a PTAS, but its success probability p is very low.

The success probability can be boosted to a constant by running $O(1/p) = (2/\varepsilon)^{2^{O(k)}/\varepsilon^2}$ independent trials of Algorithm **Sample**. By Lemma 2.16, each call runs in time $(2/\varepsilon)^{2^{O(k)}} \cdot mn^{1+o(1)}$, where $o(1)$ hides a factor $(\log \log n)^{1.1} / \log n$, yielding a total running time of $(2/\varepsilon)^{2^{O(k)}/\varepsilon^2} \cdot mn^{1+o(1)}$. This finishes the proof of Theorem 1.1. The success probability can be further amplified to $1 - \delta$ for any $\delta > 0$, by running $O(\log(1/\delta))$ independent trials of the preceding algorithm.

2.4 Faster Binary ℓ_0-Rank-1

In this section, we consider the Binary ℓ_0-Rank-1 problem with standard inner product. Given a binary matrix $A \in \{0, 1\}^{m \times n}$ and an error $\varepsilon \in (0, 1/2)$, our goal is to find binary vectors $\widetilde{u} \in \{0, 1\}^m$, $\widetilde{v} \in \{0, 1\}^n$ such that $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1 + \varepsilon)\mathrm{OPT}$, where the optimal value is defined by

$$\mathrm{OPT} = \min_{u \in \{0,1\}^m,\, v \in \{0,1\}^n} \|A - u \cdot v^T\|_0.$$

We now give a faster PTAS for the Binary ℓ_0-Rank-1 problem, which improves upon Theorem 1.1.

Theorem 1.3 (from page 5). *(PTAS for the Binary ℓ_0-Rank-1 problem with standard inner product) For any $\varepsilon \in (0, 1/2)$ there is an algorithm that runs in time $(1/\varepsilon)^{O(1/\varepsilon^2)} \cdot (\|A\|_0 + m + n) \cdot \log^3(mn)$, and outputs vectors $\widetilde{u} \in \{0, 1\}^m$, $\widetilde{v} \in \{0, 1\}^n$ such that w.h.p.* [6] $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1 + \varepsilon) \min_{u \in \{0,1\}^m, v \in \{0,1\}^n} \|A - u \cdot v^T\|_0$.

The remainder of this section is devoted to proving Theorem 1.3.

[6] An event happens *with high probability* (w.h.p.) if it has probability at least $1 - 1/n^c$ for some $c > 0$.

Structural Properties Let $R_1 = \{i \in [m] : u_i = 1\}$ and $C_1 = \{j \in [n] : v_j = 1\}$. Then, estimated cost simplifies as follows: for any row i and binary scalar $x \in \{0,1\}$, we set

$$E'_{i,x} \overset{\text{def}}{=} |\{j \in \widetilde{C}_1 : A_{i,j} \neq x\}|.$$

Lemma 2.17. *For any row i and $x, x' \in \{0,1\}$ we have $\widetilde{E}_{i,x} < \widetilde{E}_{i,x'}$ if and only if $E'_{i,x} < E'_{i,x'}$.*

Proof. By definition of $\widetilde{E}_{i,x}$ and by $k = 1$, we have $\widetilde{E}_{i,x} < \widetilde{E}_{i,x'}$ if and only if

$$\sum_{y \in \{0,1\}} \frac{|C_y|}{|\widetilde{C}_y|} \cdot |\{j \in \widetilde{C}_y : A_{i,j} \neq \langle x, y\rangle\}| < \sum_{y \in \{0,1\}} \frac{|C_y|}{|\widetilde{C}_y|} \cdot |\{j \in \widetilde{C}_y : A_{i,j} \neq \langle x', y\rangle\}|.$$

For $y = 0$ we have $\langle x, y\rangle = 0 = \langle x', y\rangle$, and thus the contribution of the corresponding summand to both sides is equal. Hence, we have, equivalently,

$$\frac{|C_1|}{|\widetilde{C}_1|} \cdot |\{j \in \widetilde{C}_1 : A_{i,j} \neq x\}| < \frac{|C_1|}{|\widetilde{C}_1|} \cdot |\{j \in \widetilde{C}_1 : A_{i,j} \neq x'\}|.$$

Removing the common factor $|C_1|/|\widetilde{C}_1|$, we equivalently arrive at $E'_{i,x} < E'_{i,x'}$. This also holds in the special case $|\widetilde{C}_1| = 0$, since then $\widetilde{E}_{i,x} = \widetilde{E}_{i,x'}$ and $E'_{i,x} = E'_{i,x'} = 0$. □

Sampling Theorem Recall that we choose the matrix $\widetilde{U}$ by picking for each row i a vector $\widetilde{U}_{i,:} = x$ minimizing $\widetilde{E}_{i,x}$. The above lemma shows that we can equivalently minimize $E'_{i,x}$. Let us formulate the resulting sampling theorem.

Corollary 2.18. *For any $\varepsilon \in (0, 1/2)$ set $t \overset{\text{def}}{=} t(1, \varepsilon/2) = 2^{16}/\varepsilon^2$. If $|C_1| < t$ then set $\widetilde{C}_1 \overset{\text{def}}{=} C_1$, otherwise sample t elements from C_1 with replacement and let the resulting multiset be $\widetilde{C}_1$. Construct $\widetilde{u} \in \{0,1\}^m$ by picking $\widetilde{u}_i = x \in \{0,1\}$ minimizing $E'_{i,x} = |\{j \in \widetilde{C}_1 : A_{i,j} \neq x\}|$. Then we have*

$$\mathbb{E}_{\widetilde{C}_1}\left[\|A - \widetilde{u} \cdot v^T\|_0\right] \leqslant \left(1 + \frac{\varepsilon}{2}\right)\mathrm{OPT}.$$

Proof. We obtain this construction of $\widetilde{u}$ as follows: Specialize the construction in Section 2.2.2 to $k = 1$, replace $\widetilde{E}_{i,x}$ by $E'_{i,x}$, which does not change the result by Lemma 2.17, and finally remove redundant steps like the sampling of $\widetilde{C}_0$, since it is no longer used in the construction. The conclusion thus follows from Theorem 2.3. For the purpose of this section, we reduced the approximation ratio from $1 + \varepsilon$ to $1 + \varepsilon/2$, thereby increasing t by a factor 4, see (2.13). □

2.4.1 The Algorithm

We present now our efficient randomized PTAS for the Binary ℓ_0-Rank-1 problem. The algorithm succeeds with probability at least $\left(\frac{\varepsilon}{5t}\right)^{t+1}$.

Algorithm 4 Faster randomized PTAS for the Binary ℓ_0-Rank-1 problem

Input: A matrix $A \in \{0,1\}^{m \times n}$ and error $\varepsilon \in (0,1)$.
Output: Vectors $\widetilde{u} \in \{0,1\}^m$ and $\widetilde{v} \in \{0,1\}^n$ satisfying $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}$.

1. *(Basic solution)* Initialize $S = \{(0_m, 0_n)\}$, where $0_d = (0, \dots, 0) \in \{0,1\}^d$.

2. *(Guess approximate column set size)* Exhaustively guess $\lceil |C_1| \rceil_2$, where $\lceil r \rceil_2$ denotes the smallest power of 2 that is at least r. There are $O(\log n)$ possibilities. If we guessed $\lceil |C_1| \rceil_2 \leqslant \lceil t \rceil_2$, then exhaustively guess $|C_1|$ exactly. There are $O(t)$ possibilities. In particular, if we guessed correctly then we know the number $\min\{t, |C_1|\}$.

3. *(Guess approximate row set size)* Exhaustively guess $\lceil |R_1| \rceil_2$ where $R_1 \overset{\text{def}}{=} \{i : u_i = 1\}$. There are $O(\log m)$ possibilities.

4. *(Ignore sparse columns)* Let $W \overset{\text{def}}{=} \{j : \|A_{:,j}\|_0 \geqslant \lceil |R_1| \rceil_2 / 4\}$ be the set of columns of A containing at least $\lceil |R_1| \rceil_2/4$ ones. Computing W takes time $O(\|A\|_0)$.

5. *(Sample columns)* Sample a multiset C_1' consisting of $\min\{t, |C_1|\}$ columns chosen independently and uniformly at random from W. This takes time $O(t)$.

6. *(Compute $\widetilde{u}$)* We plug C_1' as $\widetilde{C}_1$ into the estimated cost $E'_{i,x} = |\{j \in \widetilde{C}_1 : A_{i,j} \neq x\}|$. Compute a vector $\widetilde{u}$ by picking $\widetilde{u}_i = x \in \{0,1\}$ minimizing the cost $E'_{i,x}$, for all i. This takes time $O(mt)$ since $|C_1'| \leqslant t$.

7. *(Compute $\widetilde{v}$)* Compute $\widetilde{v}$ as a best response to $\widetilde{u}$. Add $(\widetilde{u}, \widetilde{v})$ to S.

8. **Return** the pair $(\widetilde{u}, \widetilde{v}) \in S$ minimizing $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0$ over all exhaustive guesses and the basic solution.

Note that the claimed success probability of Algorithm 4 is quite low. In order to prove Theorem 1.3 we will boost this probability by repeating the algorithm sufficiently often. In the remainder of this section we analyze the running time and the success probability of Algorithm 4.

2.4.2 Running Time

The only steps without immediate time bounds are Step 7 and Step 8.

For Step 7, we now argue that a best response $\widetilde{v}$ to $\widetilde{u}$ can be computed in time $O(\|A\|_0 + m + n)$. Observe that the cost corresponding to column $A_{:,j}$ has a fixed cost term $|\{i : A_{i,j} = 1 \text{ and } \widetilde{u}_i = 0\}|$ which is independent of the choice of $\widetilde{v}_j \in \{0,1\}$. Further, by enumerating all non-zero entries of A, we can determine for each j the number $r_j = |\{i : A_{i,j} = 1 \text{ and } \widetilde{u}_i = 1\}|$. If $r_j > \|\widetilde{u}\|_0/2$ we set $\widetilde{v}_j \overset{\text{def}}{=} 1$, and $\widetilde{v}_j \overset{\text{def}}{=} 0$ otherwise. Straightforward checking shows that this yields a best response.

In Step 8 we need to calculate $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0$ for each $(\widetilde{u}, \widetilde{v}) \in S$, in order to pick the best pair. For each $(\widetilde{u}, \widetilde{v})$, this value can be computed in time $O(\|A\|_0 + m + n)$ as follows. Enumerate all non-zero entries of A to determine the numbers $p \overset{\text{def}}{=} |\{(i,j) : A_{i,j} = 1 \text{ and } \widetilde{u}_i \cdot \widetilde{v}_j = 1\}|$ and $q \overset{\text{def}}{=} |\{(i,j) : A_{i,j} = 1 \text{ and } \widetilde{u}_i \cdot \widetilde{v}_j = 0\}|$. Then, we can infer $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 = q + \|\widetilde{u}\|_0 \cdot \|\widetilde{v}\|_0 - p$.

Note that there are $O(t \log(m) \log(n))$ possibilities for all guesses (see Steps 2-3). Considering Steps 4-6 as well as the running times for Steps 7-8 shown above, we spend time $O(\|A\|_0 + m + nt) = O((\|A\|_0 + m + n)t)$ for each guess. Hence, we obtain the following.

Lemma 2.19. *The running time of Algorithm 4 is $O(t^2(\|A\|_0 + m + n)\log(n)\log(m))$.*

2.4.3 Correctness and Success Probability

We establish now the correctness of Algorithm 4. More precisely, we will prove a lower bound of $\left(\frac{\varepsilon}{5t}\right)^{t+1}$ on its success probability.

Claim 2.20 (Step 4)**.** *We have $C_1 \subseteq W$.*

Proof. Let $R_x \overset{\text{def}}{=} \{i : u_i = x\}$ for $x \in \{0,1\}$. We split

$$\mathrm{OPT} = \|A - u \cdot v^T\|_0 = \sum_{j=1}^{n} \|A_{:,j} - u \cdot v_j\|_0,$$

where $\|A_{:,j} - u \cdot v_j\|_0$ is the cost of column j. If $v_j = 0$ then this cost is $\|A_{:,j}\|_0$, i.e. the number of 1's of $A_{:,j}$ in R_1 plus the number of 1's of $A_{:,j}$ in R_0. If $v_j = 1$ then the cost is $\|A_{:,j} - u\|_0$, i.e., the number of 0's of $A_{:,j}$ in R_1 plus the number of 1's of $A_{:,j}$ in R_0.

Suppose $\|A_{:,j}\|_0 < |R_1|/2$. Then, the number of 1's of $A_{:,j}$ in R_1 is less than $|R_1|/2$, and the number of 0's of $A_{:,j}$ in R_1 is larger than $|R_1|/2$. Thus, $v_j = 0$ has less cost, and $j \in C_0$. Hence,

$$C_1 \subseteq \Big\{ j \,:\, \|A_{:,j}\|_0 \geqslant |R_1|/2 \Big\} \subseteq \Big\{ j \,:\, \|A_{:,j}\|_0 \geqslant \lceil |R_1| \rceil_2 /4 \Big\} = W. \qquad \square$$

Assume that all guesses in Steps 2-3 are correct, so that the multiset C_1' constructed in Step 5 consists of $\min\{t, |C_1|\}$ columns chosen independently and uniformly at random from W.

We define an event $\mathcal{E}$ as follows. If $|C_1| \geqslant t$, then event $\mathcal{E}$ asserts that $C_1' \subseteq C_1$, i.e., all t sampled columns forming C_1' hit C_1. If $|C_1| < t$, then event $\mathcal{E}$ asserts that $C_1' = C_1$, i.e., all $|C_1|$ sampled columns forming C_1' hit C_1 and are distinct.

Claim 2.21 (Step 5)**.** *Let $\widetilde{C}_1$ be distributed as in Corollary 2.18. Conditioned on event $\mathcal{E}$, C_1' and $\widetilde{C}_1$ follow the same distribution.*

Proof. If $|C_1| < t$ then we set $\widetilde{C}_1 = C_1$, and conditioned on $\mathcal{E}$ we also have $C_1' = C_1$. If $|C_1| \geqslant t$, then $\widetilde{C}_1$ is formed by sampling t elements from C_1 with replacement. On the other hand, C_1' is formed by sampling t elements from W with replacement, which conditioned on $\mathcal{E}$ implies that all t elements hit C_1. Since $C_1 \subseteq W$ by Claim 2.20, the sampled elements are uniformly random in C_1. These processes describe the same distribution. □

It follows that, conditioned on event $\mathcal{E}$, Corollary 2.18 is applicable. We use it to show:

Claim 2.22. *Assuming correct guesses in Steps 2-3, the vectors $\widetilde{u}, \widetilde{v}$ computed in Steps 6-7 satisfy*

$$\Pr\Big[\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}\Big] \geqslant \frac{\varepsilon}{4} \cdot \Pr[\mathcal{E}].$$

Proof. Since Step 6 follows the construction in Corollary 2.18, except that we replaced $\widetilde{C}_1$ by C_1', and using Claim 2.21, Corollary 2.18 yields

$$\mathbb{E}_{C_1'}\Big[\|A - \widetilde{u} \cdot v^T\|_0 \;\Big|\; \mathcal{E}\Big] \leqslant \Big(1 + \frac{\varepsilon}{2}\Big)\mathrm{OPT}.$$

Using Markov's inequality with $\lambda \overset{\text{def}}{=} (1+\varepsilon)/(1+\varepsilon/2)$ yields

$$\Pr\Big[\|A - \widetilde{u} \cdot v^T\|_0 > \lambda \cdot \Big(1 + \frac{\varepsilon}{2}\Big)\mathrm{OPT} \;\Big|\; \mathcal{E}\Big] \leqslant \frac{1}{\lambda} = 1 - \frac{\varepsilon/2}{1+\varepsilon} \leqslant 1 - \frac{\varepsilon}{4},$$

or equivalently,

$$\Pr\Big[\|A - \widetilde{u} \cdot v^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT} \;\Big|\; \mathcal{E}\Big] \geqslant \frac{\varepsilon}{4}.$$

Hence, we have

$$\Pr\Big[\|A - \widetilde{u} \cdot v^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}\Big] \geqslant \Pr\Big[\|A - \widetilde{u} \cdot v^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT} \;\Big|\; \mathcal{E}\Big] \cdot \Pr[\mathcal{E}] \geqslant \frac{\varepsilon}{4} \cdot \Pr[\mathcal{E}].$$

Since in Step 7 $\widetilde{v}$ is computed as a best response to $\widetilde{u}$, we have $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant \|A - \widetilde{u} \cdot v^T\|_0$, and thus

$$\Pr\Big[\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}\Big] \geqslant \frac{\varepsilon}{4} \cdot \Pr[\mathcal{E}]. \qquad \square$$

In order to give a lower bound on the success probability, it thus suffices to bound $\Pr[\mathcal{E}]$. To this end, we first bound $|W|$ in terms of $|C_1|$.

Claim 2.23. *If* $\mathrm{OPT} < \|A\|_0/(1+\varepsilon)$*, then we have* $|W| < 5|C_1|/\varepsilon$.

Proof. As in the proof of Claim 2.20, the cost of any column $j \in C_0$ is $\|A_{:,j}\|_0$. Since any column $j \in W$ contains at least $\lceil |R_1| \rceil_2/4 \geqslant |R_1|/4$ 1's, the cost of any column $j \in C_0 \cap W$ is at least $|R_1|/4$. It follows that

$$|C_0 \cap W| \leqslant 5 \cdot \mathrm{OPT}/|R_1|. \tag{2.16}$$

Since $u \cdot v^T$ contains $|R_1| \cdot |C_1|$ 1's, we can lower bound $\mathrm{OPT} = \|A - u \cdot v^T\|_0$ by $\|A\|_0 - |R_1| \cdot |C_1|$. Together with the assumption, we obtain that

$$\mathrm{OPT} \geqslant \|A\|_0 - |R_1| \cdot |C_1| > (1+\varepsilon)\mathrm{OPT} - |R_1| \cdot |C_1|,$$

or equivalently, $\varepsilon\mathrm{OPT} < |R_1| \cdot |C_1|$. This implies $\mathrm{OPT}/|R_1| < |C_1|/\varepsilon$ and thus by inequality (2.16) we have $|C_0 \cap W| \leqslant 4|C_1|/\varepsilon$. Further, since $|C_1 \cap W| \leqslant |C_1| \leqslant |C_1|/\varepsilon$ and $(C_0 \cap W) \cup (C_1 \cap W) = W$, we obtain the claimed upper bound of $|W| \leqslant 5|C_1|/\varepsilon$. □

Claim 2.24. *If* $\mathrm{OPT} < \|A\|_0/(1+\varepsilon)$*, then we have* $\Pr[\mathcal{E}] \geqslant \left(\frac{\varepsilon}{5t}\right)^t$.

Proof. If $|C_1| \geqslant t$ then event $\mathcal{E}$ asserts that $C_1' \subseteq C_1$. Since $C_1 \subseteq W$ (Claim 2.20) and $|W| \leqslant 5|C_1|/\varepsilon$ (Claim 2.23), a random element from W hits C_1 with probability at least $\varepsilon/5$. It follows that all t samples forming C_1' hit C_1 with probability at least $(\varepsilon/5)^t$.

If $|C_1| < t$ then event $\mathcal{E}$ asserts that $C_1' = C_1$. Note that the i-th sample from W is equal to the i-th element of C_1 with probability $1/|W| \geqslant \varepsilon/(5|C_1|) > \varepsilon/(5t)$. It follows that $C_1' = C_1$ with probability at least $(\varepsilon/(5t))^{|C_1|} \geqslant (\varepsilon/(5t))^t$. □

Lemma 2.25. *Algorithm 4 computes vectors* $\widetilde{u} \in \{0,1\}^m$ *and* $\widetilde{v} \in \{0,1\}^n$ *such that with probability at least* $\left(\frac{\varepsilon}{5t}\right)^{t+1}$ *it holds that* $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}$.

Proof. When $\mathrm{OPT} \geqslant \|A\|_0/(1+\varepsilon)$, for the basic solution $(0_m, 0_n)$ from Step 1 we have $\|A - 0_m \cdot 0_n^T\|_0 = \|A\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}$. Otherwise, if $\mathrm{OPT} < \|A\|_0/(1+\varepsilon)$, the statement follows by combining Claim 2.22 and Claim 2.24. □

2.4.4 Proof of Theorem 1.3

By Lemma 2.25, Algorithm 4 computes vectors $\widetilde{u}, \widetilde{v}$ satisfying $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}$ with probability at least $\left(\frac{\varepsilon}{5t}\right)^{t+1}$. By Lemma 2.19, the algorithm runs in time

$$O(t^2(\|A\|_0 + m + n)\log(m)\log(n)).$$

Repeating the algorithm $O\big((5t/\varepsilon)^{t+1}\log n\big)$ times and returning the best solution found, yields a success probability of $1 - n^{-\Omega(1)}$. Since we set $t = 2^{16}/\varepsilon^2$, the total running time is

$$O\Big((5t/\varepsilon)^{t+3}(\|A\|_0 + m + n)\log^3(mn)\Big) = (1/\varepsilon)^{O(1/\varepsilon^2)}(\|A\|_0 + m + n)\log^3(mn).$$

□

Chapter 3

Algorithmic Results For Binary ℓ_0-Rank-1 With Small Optimal Value

Given a binary matrix $A \in \{0,1\}^{m\times n}$ with $m \geqslant n$, our goal is to compute an approximate solution of the Binary ℓ_0-Rank-1 problem, and let us denote the optimal value by

$$\mathrm{OPT} \stackrel{\text{def}}{=} \min_{u\in\{0,1\}^m,\, v\in\{0,1\}^n} \|A - u\cdot v^T\|_0. \tag{3.1}$$

In practice, approximating a matrix A by a rank-1 matrix uv^T makes most sense if A is close to being rank-1. Hence, the above optimization problem is most relevant in the case when $\mathrm{OPT} \ll \|A\|_0$. For this reason, we focus our studies on the setting when the ratio $\mathrm{OPT}/\|A\|_0 \leqslant \phi$ for sufficiently small $\phi > 0$.

Organization In Section 3.1, we give a simple $(1+O(\phi))$-approximation algorithm running in time $O(\min\{\|A\|_0 + m + n,\, \phi^{-1}(m+n)\log(mn)\})$. Then, in Section 3.2, we establish a sample complexity lower bound of $\Omega((m+n)/\phi)$ for any $(1+O(\phi))$-approximation algorithm, showing that our algorithm has an optimal runtime up to a $\operatorname{poly}\log(mn)$ factor. In Section 3.3, we give an algorithm that runs in time $2^{O(\mathrm{OPT}/\sqrt{\|A\|_0})} \cdot \operatorname{poly}(mn)$ and solves exactly the Binary ℓ_0-Rank-1 problem.

3.1 A Simple Approximation Algorithm

We start by stating our core algorithmic result, which requires as an input a parameter $\phi \geqslant \mathrm{OPT}/\|A\|_0$.

Theorem 1.4 (from page 7)**.** *Given $A \in \{0,1\}^{m\times n}$ with row and column sums, and given $\phi \in (0, \frac{1}{80}]$ with $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, we can compute in time $O(\min\{\|A\|_0+m+n, \phi^{-1}(m+n)\log(mn)\})$ vectors $\widetilde{u} \in \{0,1\}^m$ and $\widetilde{v} \in \{0,1\}^n$ such that w.h.p. $\|A - \widetilde{u}\cdot\widetilde{v}^T\|_0 \leqslant (1+5\phi)\mathrm{OPT} + 37\phi^2\|A\|_0$.*

In Chapter 4, see Section 4.2, we give a $(2+\varepsilon)$-approximation algorithm for the Reals ℓ_0-Rank-1 problem, which captures as a special case the Binary ℓ_0-Rank-1 problem. In particular, we obtain a $(2+\varepsilon)$-approximation of OPT, and thus a $(2+\varepsilon)$-approximation of the ratio $\mathrm{OPT}/\|A\|_0$.

Theorem 1.5 (from page 7)**.** *Given $A \in \{0,1\}^{m\times n}$ with column adjacency arrays and $\mathrm{OPT} \geqslant 1$, and given $\varepsilon \in (0, 0.1]$, we can compute w.h.p. in time*

$$O\left(\left(\frac{n\log m}{\varepsilon^2} + \min\left\{\|A\|_0,\ n + \psi^{-1}\frac{\log n}{\varepsilon^2}\right\}\right)\frac{\log^2 n}{\varepsilon^2}\right)$$

a column $A_{:,j}$ and a vector $z \in \{0,1\}^n$ such that w.h.p. $\|A - A_{:,j}\cdot z^T\|_0 \leqslant (2+\varepsilon)\mathrm{OPT}$. Further, we can compute an estimate Y such that w.h.p. $(1-\varepsilon)\mathrm{OPT} \leqslant Y \leqslant (2+2\varepsilon)\mathrm{OPT}$.

Combining Theorem 1.4 and Theorem 1.5, yields an algorithm that does not need ϕ as an input and computes a $(1+500\psi)$-approximate solution of the Binary ℓ_0-Rank-1 problem.

Theorem 1.6 (from page 7)**.** *Given $A \in \{0,1\}^{m\times n}$ with column adjacency arrays and with row and column sums, for $\psi = \mathrm{OPT}/\|A\|_0$ we can compute w.h.p. in time $O(\min\{\|A\|_0 + m + n, \psi^{-1}(m+n)\}\cdot\log^3(mn))$ vectors $\widetilde{u} \in \{0,1\}^m$ and $\widetilde{v} \in \{0,1\}^n$ such that w.h.p. $\|A - \widetilde{u}\cdot\widetilde{v}^T\|_0 \leqslant (1+500\psi)\mathrm{OPT}$.*

Proof of Theorem 1.6. We compute a 3-approximation of OPT by applying Theorem 1.5 with $\varepsilon = 0.1$. This yields a value ϕ satisfying $\psi \leqslant \phi \leqslant 3\psi$. If $\phi > 1/80$, then the 3-approximation is already good enough, since $\psi > 1/240$ and $1 + 500\psi > 3$. Otherwise, we run Theorem 1.4 with ϕ. Further, using $\phi^2\|A\|_0 \leqslant 9\psi^2\|A\|_0 = 9\psi\mathrm{OPT}$, the total error is at most

$$(1+5\phi)\mathrm{OPT} + 37\phi^2\|A\|_0 \leqslant (1+15\psi)\mathrm{OPT} + 37\cdot 9\psi\mathrm{OPT} \leqslant (1+500\psi)\mathrm{OPT}.$$

A rough upper bound on the running time is $O(\min\{\|A\|_0 + m + n, \psi^{-1}(m+n)\}\cdot\log^3(mn))$. □

The remainder of this section is devoted to proving Theorem 1.4.

3.1.1 Preparations

Given a matrix $A \in \{0,1\}^{m\times n}$, let $u \in \{0,1\}^m$ and $v \in \{0,1\}^n$ be an optimal solution to the Binary ℓ_0-Rank-1 problem, realizing $\mathrm{OPT} = \|A - u \cdot v^T\|_0$. Moreover, set $\alpha \overset{\text{def}}{=} \|u\|_0$ and $\beta \overset{\text{def}}{=} \|v\|_0$. We start with the following technical preparations.

Lemma 3.1. *For any row $i \in \{1,\dots,m\}$ let x_i be the number of 0's in columns selected by v, i.e., $x_i \overset{\text{def}}{=} \{j \in [n] : A_{i,j} = 0,\, v_j = 1\}$, and let y_i be the number of 1's in columns not selected by v, i.e., $y_i \overset{\text{def}}{=} \{j \in [n] : A_{i,j} = 1,\, v_j = 0\}$. Let $R = \{i \in [m] : u_i = 1\}$ be the rows selected by u, and let $\bar{R} \overset{\text{def}}{=} [m] \setminus R$. Symmetrically, let C be the columns selected by v. Then we have*

1. *$\|A_{i,:}\|_0 = \beta - x_i + y_i$, for any $i \in [m]$;*
2. *$\mathrm{OPT} = \sum_{i\in R}(x_i + y_i) + \sum_{i\in\bar{R}}(\beta - x_i + y_i)$;*
3. *$\mathrm{OPT} \geqslant \sum_{i\in R} |x_i - y_i|$;*
4. *$x_i \leqslant \beta/2$ for any $i \in R$, and $x_i \geqslant \beta/2$, for any $i \in \bar{R}$;*
5. *$\mathrm{OPT} \geqslant \sum_{i=1}^m \min\{\|A_{i,:}\|_0,\, |\|A_{i,:}\|_0 - \beta|\}$;*
6. *$|\|A\|_0 - \alpha\beta| \leqslant \mathrm{OPT}$;*
7. *If $\mathrm{OPT} \leqslant \phi\|A\|_0$, then $(1-\phi)\|A\|_0 \leqslant \alpha\beta \leqslant (1+\phi)\|A\|_0$.*

Proof. For (1), note that in the β columns C selected by v, row i has $\beta - x_i$ 1's, and in the remaining $n - |C|$ columns row i has y_i 1's. Hence, the total number of 1's in row i is $\|A_{i,:}\|_0 = \beta - x_i + y_i$.

(2) We split $\mathrm{OPT} = \|A - uv^T\|_0$ into a sum over all rows, so that $\mathrm{OPT} = \sum_{i=1}^m \|A_{i,:} - u_i v^T\|_0$. For $i \in \bar{R}$, the i-th term of this sum is simply $\|A_{i,:}\|_0 = \beta - x_i + y_i$. For $i \in R$, the i-th term is $\|A_{i,:} - v^T\|_0 = x_i + y_i$.

(3) follows immediately from (2).

(4) follows from (2), since for $x_i > \beta/2$ and $i \in R$ we can change u_i from 1 to 0, reducing the contribution of row i from $x_i + y_i$ to $\beta - x_i + y_i$, which contradicts optimality of OPT.

For (5), we use that $x_i + y_i \geqslant |x_i - y_i| = |\|A_{i,:}\|_0 - \beta|$ by (1), and

$$\mathrm{OPT} = \sum_{i\in R}(x_i + y_i) + \sum_{i\in\bar{R}}(\beta - x_i + y_i) = \sum_{i\in R}(x_i + y_i) + \sum_{i\in\bar{R}}\|A_{i,:}\|_0.$$

(6) is shown similarly to (5) by noting that

$$\begin{aligned}\mathrm{OPT} &= \sum_{i\in R}(x_i + y_i) + \sum_{i\in\bar{R}}(\beta - x_i + y_i) \geqslant \sum_{i\in R}|\|A_{i,:}\|_0 - \beta| + \sum_{i\in\bar{R}}\|A_{i,:}\|_0 \\ &\geqslant \sum_{i\in R}(\|A_{i,:}\|_0 - \beta) + \sum_{i\in\bar{R}}\|A_{i,:}\|_0 = \|A\|_0 - \alpha\beta,\end{aligned}$$

and similarly

$$\mathrm{OPT} \geqslant \sum_{i\in R}|\|A_{i,:}\|_0 - \beta| + \sum_{i\in\bar{R}}\|A_{i,:}\|_0 \geqslant \sum_{i\in R}(\beta - \|A_{i,:}\|_0) - \sum_{i\in\bar{R}}\|A_{i,:}\|_0 = \alpha\beta - \|A\|_0.$$

Finally, (7) follows immediately from (6) by plugging in the upper bound $\mathrm{OPT} \leqslant \phi\|A\|_0$. □

3.1.2 Approximating α and β

We now show how to approximate $\alpha = \|u\|_0$ and $\beta = \|v\|_0$, where (u,v) is an optimal solution.

Lemma 3.2. *Given $A \in \{0,1\}^{m\times n}$ and $\phi \in (0, 1/30]$ with $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, we can compute in time $O(\|A\|_0 + m + n)$ an integer $\widetilde{\beta} \in [m]$ with*

$$\frac{1-3\phi}{1-\phi}\beta \leqslant \widetilde{\beta} \leqslant \frac{1+\phi}{1-\phi}\beta.$$

Symmetrically, we can approximate α by $\widetilde{\alpha}$. If we are additionally given the number of 1's in each row and column, then the running time becomes $O(m+n)$.

Proof. Let

$$\Lambda \overset{\text{def}}{=} \min_{\beta' \in [n]} \sum_{i=1}^{m} \min \left\{ \|A_{i,:}\|_0, \left| \|A_{i,:}\|_0 - \beta' \right| \right\},$$

and let $\widetilde{\beta}$ be the value of β' realizing Λ.

We first verify the approximation guarantee. Consider the set of rows R selected by u. Let x_i, y_i for $i \in R$ be as in Lemma 3.1. Then we have

$$\Lambda \geqslant \sum_{i \in R} \min \left\{ \|A_{i,:}\|_0, \left| \|A_{i,:}\|_0 - \widetilde{\beta} \right| \right\} = \sum_{i \in R} \min \left\{ \beta + y_i - x_i, |\beta - \widetilde{\beta} + y_i - x_i| \right\},$$

where we used Lemma 3.1.1. Assume for the sake of contradiction that $|\beta - \widetilde{\beta}| > \frac{2\phi}{1-\phi}\beta$. Since $|x - y| \geqslant |x| - |y|$ for any numbers x, y, we obtain

$$|\beta - \widetilde{\beta} + y_i - x_i| \geqslant |\beta - \widetilde{\beta}| - |x_i - y_i| > \frac{2\phi}{1-\phi}\beta - |x_i - y_i|.$$

Similarly, we have $\beta + y_i - x_i > \frac{2\phi}{1-\phi}\beta - |x_i - y_i|$. Hence,

$$\Lambda > \sum_{i \in R} \left(\frac{2\phi\beta}{1-\phi} - |x_i - y_i| \right) \geqslant \frac{2\phi\beta}{1-\phi}|R| - \mathrm{OPT},$$

where we used Lemma 3.1.3. Since R is the set of rows selected by u, we have $|R| = \alpha$. By Lemma 3.1.7, we have $\mathrm{OPT} \leqslant \phi\|A\|_0 \leqslant \frac{\phi}{1-\phi}\alpha\beta$. Together, this yields $\Lambda > \mathrm{OPT}$, contradicting

$$\Lambda \leqslant \sum_{i=1}^{m} \min\{\|A_{i,:}\|_0, |\|A_{i,:}\|_0 - \beta|\} \leqslant \mathrm{OPT}$$

by Lemma 3.1.5. Hence, $|\beta - \widetilde{\beta}| \leqslant \frac{2\phi}{1-\phi}\beta$.

It remains to design a fast algorithm. We first compute all numbers $\|A_{i,:}\|_0$ in time $O(\|A\|_0)$ (this step can be skipped if we are given these numbers as input). We sort these numbers, obtaining a sorted order $\|A_{\pi(1),:}\|_0 \leqslant \ldots \leqslant \|A_{\pi(m),:}\|_0$. Using counting sort, this takes time $O(m+n)$. We precompute prefix sums $P(k) \overset{\text{def}}{=} \sum_{\ell=1}^{k} \|A_{\pi(\ell),:}\|_0$, which allows us to evaluate in constant time any sum

$$\sum_{\ell=x}^{y} \|A_{\pi(\ell),:}\|_0 = P(y) - P(x-1).$$

Finally, we precompute the inverse

$$\ell(\beta') \overset{\text{def}}{=} \max\{\ell : \|A_{\pi(\ell),:}\|_0 \leqslant \beta'\},$$

or $\ell(\beta') = 0$ if there is no ℓ with $\|A_{\pi(\ell),:}\|_0 \leqslant \beta'$. By a simple sweep, all values $\ell(\beta')$ can be computed in total time $O(m+n)$.

Note that for any fixed β' and row i, the term realizing $\min\{\|A_{i,:}\|_0, |\|A_{i,:}\|_0 - \beta'|\}$ is equal to: (a) $\|A_{i,:}\|_0$ if $\|A_{i,:}\|_0 \leqslant \beta'/2$; (b) $\beta' - \|A_{i,:}\|_0$, if $\beta'/2 < \|A_{i,:}\|_0 \leqslant \beta'$; and (c) $\|A_{i,:}\|_0 - \beta'$, if $\|A_{i,:}\|_0 > \beta'$. Hence, we obtain

$$\begin{aligned}
&\sum_{i=1}^{n} \min \left\{ \|A_{i,:}\|_0, \left| \|A_{i,:}\|_0 - \beta' \right| \right\} \\
&= \left(\sum_{i=1}^{\ell(\beta'/2)} \|A_{\pi(i),:}\|_0 \right) + \left(\sum_{i=\ell(\beta'/2)+1}^{\ell(\beta')} \beta' - \|A_{\pi(i),:}\|_0 \right) + \left(\sum_{i=\ell(\beta')+1}^{n} \|A_{\pi(i),:}\|_0 - \beta' \right) \\
&= P(n) - 2\left[P(\ell(\beta')) - P(\ell(\beta'/2))\right] - \left[n + \ell(\beta'/2) - 2 \cdot \ell(\beta')\right]\beta'.
\end{aligned}$$

This shows that after the above precomputation the sum $\sum_{j=1}^{n} \min\{\|A_{i,:}\|_0, |\|A_{:j}\|_0 - \beta'|\}$ can be evaluated in time $O(1)$ for any β'. Minimizing over all $\beta' \in [m]$ yields $\widetilde{\beta}$. This finishes our algorithm, which runs in total time $O(\|A\|_0 + m + n)$, or $O(m+n)$ if we are given the number of 1's in each row and column. □

3.1.3 The Algorithm

We now design an approximation algorithm for the Binary ℓ_0-Rank-1 problem that will yield Theorem 1.4. We present the pseudocode of this Algorithm 5 below.

Algorithm 5 Binary ℓ_0-Rank-1 With Small Optimal Value

```
Input: $A \in \{0,1\}^{m\times n}$ and $\phi \in (0, 1/80]$ such that $\mathrm{OPT}/\|A\|_0 \leqslant \phi$.
Output: Vectors $\widetilde{u} \in \{0,1\}^m$, $\widetilde{v} \in \{0,1\}^n$ such that $\|A - \widetilde{u}\cdot\widetilde{v}^T\|_0 \leqslant (1+\varepsilon)\mathrm{OPT}$.

1. Compute approximations $\frac{1-3\phi}{1-\phi}\alpha \leqslant \widetilde{\alpha} \leqslant \frac{1+\phi}{1-\phi}\alpha$ and $\frac{1-3\phi}{1-\phi}\beta \leqslant \widetilde{\beta} \leqslant \frac{1+\phi}{1-\phi}\beta$ using Lemma 3.2.
Initialize $R^R \overset{\text{def}}{=} [m]$, $C^R \overset{\text{def}}{=} [n]$, $R^S \overset{\text{def}}{=} \emptyset$, $C^S \overset{\text{def}}{=} \emptyset$.
2. For any row $i$, if $\|A_{i,:}\|_0 < \frac{1-\phi}{1+\phi}\cdot\frac{\widetilde{\beta}}{2}$ then set $\widetilde{u}_i = 0$ and remove $i$ from $R^R$.
   For any column $j$, if $\|A_{:,j}\|_0 < \frac{1-\phi}{1+\phi}\cdot\frac{\widetilde{\alpha}}{2}$ then set $\widetilde{v}_j = 0$ and remove $j$ from $C^R$.
3. For any $i \in R^R$ compute an estimate $X_i$ with $\big|X_i - \|A_{i,C^R}\|_0\big| \leqslant \frac{1}{9}|C^R|$.
   For any $j \in C^R$ compute an estimate $Y_j$ with $\big|Y_j - \|A_{R^R,j}\|_0\big| \leqslant \frac{1}{9}|R^R|$.
4. For any $i \in R^R$, if $X_i > \frac{2}{3}\widetilde{\beta}$ then set $\widetilde{u}_i = 1$ and add $i$ to $R^S$.
   For any $j \in C^R$, if $Y_j > \frac{2}{3}\widetilde{\alpha}$ then set $\widetilde{v}_j = 1$ and add $j$ to $C^S$.
5. For any $i \in R^R \setminus R^S$, compute an estimate $X'_i$ with $|X'_i - \|A_{i,C^S}\|_0| \leqslant \phi|C^S|$,
   For any $j \in C^R \setminus C^S$, compute an estimate $Y'_j$ with $|Y'_j - \|A_{R^S,j}\|_0| \leqslant \phi|R^S|$.
6. For any $i \in R^R \setminus R^S$, set $\widetilde{u}_i = 1$ if $X'_i \geqslant |C^S|/2$ and 0 otherwise,
   For any $j \in C^R \setminus C^S$, set $\widetilde{v}_j = 1$ if $Y'_j \geqslant |R^S|/2$ and 0 otherwise.
7. Return $(\widetilde{u}, \widetilde{v})$.
```

Running Time By Lemma 3.2, Step 1 runs in time $O(\|A\|_0 + m + n)$, or in time $O(m+n)$ if we are given the number of 1's in each row and column. Steps 2, 4, and 6 clearly run in time $O(m+n)$. For steps 3 and 5, there are two ways to implement them.

Variant (1) is an exact algorithm. We enumerate all nonzero entries of A and count how many contribute to the required numbers $\|A_{i,C^R}\|_0, \|A_{R^R,j}\|_0$ etc. This takes total time $O(\|A\|_0)$, and hence the total running time of the algorithm is $O(\|A\|_0 + m + n)$.

Variant (2) uses random sampling. In order to estimate $\|A_{i,C^R}\|_0$, consider a random variable Z that draws a uniformly random column $j \in C^R$ and returns 1 if $A_{i,j} \neq 0$ and 0 otherwise. Then $\mathbb{E}[Z] = \|A_{i,C^R}\|_0/|C^R|$. Taking independent copies $Z_1, \ldots, Z_\ell$ of Z, where $\ell = \Theta(\log(mn)/\delta^2)$ with sufficiently large hidden constant, a standard Chernoff bound argument shows that w.h.p.

$$\left|(Z_1 + \ldots + Z_\ell)\cdot\frac{|C^R|}{\ell} - \|A_{i,C^R}\|_0\right| \leqslant \delta\cdot|C^R|,$$

which yields the required approximation. For Step 3 we use this procedure with $\delta = \frac{1}{9}$ and obtain running time $O(\log(mn))$ per row and column, or $O((m+n)\log(mn))$ in total. For Step 5 we use $\delta = \phi$, resulting in time $O(\phi^{-2}\log(mn))$ for computing one estimate X'_i or Y'_j. By Claim 3.6 below there are only $O(\phi(m+n))$ rows and columns in $R^R \setminus R^S$ and $C^R \setminus C^S$, and hence the total running time for Step 5 is $O(\phi^{-1}(m+n)\log(mn))$. This dominates the total running time.

Combining both variants, we obtain the claimed running time of

$$O(\min\{\|A\|_0 + m + n, \phi^{-1}(m+n)\log(mn)\}).$$

Correctness In the following we analyze the correctness of the above algorithm.

Claim 3.3. *For any row i deleted in Step 2 we have $\widetilde{u}_i = u_i$. Symmetrically, for any column j deleted in Step 2 we have $\widetilde{v}_j = v_j$.*

Proof. If row i is deleted, then by the approximation guarantee of $\widetilde{\beta}$ we have

$$\|A_{i,:}\|_0 < \frac{1-\phi}{1+\phi}\cdot\frac{\widetilde{\beta}}{2} \leqslant \frac{\beta}{2}$$

Note that for x_i (the number of 0's in row i in columns selected by v) we have $x_i \geqslant \beta - \|A_{i,:}\|_0$. Together, we obtain $x_i > \beta/2$, and thus row i cannot be selected by u, by Lemma 3.1.4. Hence, we have $u_i = 0 = \widetilde{u}_i$. The statement for the columns is symmetric. □

Claim 3.4. *After Step 2, it holds for the remaining rows R^R and columns C^R that*

$$|R^R| \leqslant \left(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\alpha \quad \textit{and} \quad |C^R| \leqslant \left(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\beta.$$

Proof. By Claim 3.3 the α rows R selected by u remain. We split the rows R^R remaining after Step 2 into $R \cup R'$, and bound $|R'|$ from above. Since any $i \in R'$ is not selected by u, it contributes $\|A_{i,:}\|_0$ to OPT. Note that

$$\|A_{i,:}\|_0 \geqslant \frac{1-\phi}{1+\phi} \cdot \frac{\widetilde{\beta}}{2} \geqslant \frac{1-\phi}{1+\phi} \cdot \frac{1-3\phi}{1-\phi} \cdot \frac{\beta}{2} = \frac{1-3\phi}{1+\phi} \cdot \frac{\beta}{2},$$

and thus $|R'| \leqslant \mathrm{OPT} \cdot \frac{1+\phi}{1-3\phi} \cdot \frac{2}{\beta}$. Since

$$\mathrm{OPT} \leqslant \phi\|A\|_0 \leqslant \frac{\phi}{1-\phi} \cdot \alpha\beta$$

by Lemma 3.1.7, we obtain $|R'| \leqslant \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi} \cdot \alpha$. Thus, we have in total

$$|R^R| = |R| + |R'| \leqslant \left(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\alpha.$$

The statement for the columns is symmetric. □

Claim 3.5. *The rows and columns selected in Step 4 are also selected by the optimal solution u, v, i.e., for any $i \in R^S$ we have $u_i = 1$ and for any $j \in C^S$ we have $v_j = 1$.*

Proof. If row i is selected in Step 4, then we have by the approximation guarantee of X_i, definition of Step 4, Claim 3.4, and Lemma 3.2

$$\begin{aligned}\|A_{i,C^R}\|_0 &\geqslant X_i - \frac{1}{9}|C^R| > \frac{2}{3}\widetilde{\beta} - \frac{1}{9}\left(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\beta \\ &\geqslant \frac{2}{3} \cdot \frac{1-3\phi}{1-\phi}\beta - \frac{1}{9}\left(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\beta.\end{aligned}$$

It is easy to see that for sufficiently small $\phi \geqslant 0$ this yields

$$\|A_{i,C^R}\|_0 > \frac{\beta}{2} + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\beta.$$

One can check that $0 \leqslant \phi \leqslant 1/80$ is sufficient. Since there are $|C^R| \leqslant \left(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\beta$ columns remaining, in particular the β columns $C \subseteq C^R$ which are selected by v, we obtain

$$\|A_{i,C}\|_0 \geqslant \|A_{i,C^R}\|_0 - (|C^R| - \beta) > \beta/2.$$

By Lemma 3.1.4, we thus obtain that row i is selected by the optimal u. The statement for the columns is symmetric. □

Claim 3.6. *After Step 4 there are $|R^R \setminus R^S| \leqslant 6\phi\alpha$ remaining rows and $|C^R \setminus C^S| \leqslant 6\phi\beta$ remaining columns.*

Proof. After Step 4, every remaining row i, for any $0 \leqslant \phi \leqslant 1/80$, satisfies

$$\|A_{i,:}\|_0 \geqslant \frac{1-\phi}{1+\phi} \cdot \frac{\widetilde{\beta}}{2} \geqslant \frac{1-\phi}{1+\phi} \cdot \frac{1-3\phi}{1-\phi} \cdot \frac{\beta}{2} \geqslant \frac{2}{5}\beta,$$

Moreover, each such row satisfies

$$\|A_{i,C^R}\|_0 \leqslant X_i + \frac{1}{9}|C^R| \leqslant \frac{2}{3}\widetilde{\beta} + \frac{1}{9}|C^R|,$$

which together with $\widetilde{\beta} \leqslant \frac{1+\phi}{1-\phi}\beta$ (Lemma 3.2) and $|C^R| \leqslant \big(1 + \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\big)\beta$ (Claim 3.4) yields

$$\|A_{i,C^R}\|_0 \leqslant \left(\frac{2}{3} \cdot \frac{1+\phi}{1-\phi} + \frac{1}{9} + \frac{1}{9} \cdot \frac{1+\phi}{1-3\phi} \cdot \frac{2\phi}{1-\phi}\right)\beta.$$

It is easy to see that for sufficiently small $\phi \geqslant 0$ we have $\|A_{i,C^R}\|_0 \leqslant \frac{4}{5}\beta$, and it can be checked that $0 \leqslant \phi \leqslant 1/80$ is sufficient.

If i is not selected by u, then its contribution to OPT is $\|A_{i,:}\|_0 \geqslant \frac{2}{5}\beta$. If i is selected by u, then since $C \subseteq C^R$ its contribution to OPT is at least

$$\beta - \|A_{i,C}\|_0 \geqslant \beta - \|A_{i,C^R}\|_0 \geqslant \beta - \frac{4}{5}\beta = \frac{1}{5}\beta.$$

Thus, the number of remaining rows is at most

$$\frac{\mathrm{OPT}}{\beta/5} \leqslant \frac{5\phi\alpha\beta}{(1-\phi)\beta} \leqslant 6\phi\alpha,$$

where we used Lemma 3.1.7. The statement for the columns is symmetric. □

We are now ready to prove correctness of Algorithm 5.

Proof of Theorem 1.4. The rows and columns removed in Step 2 are also not picked by the optimal solution, by Claim 3.3. Hence, in the region $([m] \setminus R^R) \times [n]$ and $[m] \times ([n] \setminus C^R)$ we incur the same error as the optimal solution. The rows and columns chosen in Step 4 are also picked by the optimal solution, by Claim 3.5. Hence, in the region $R^S \times C^S$ we incur the same error as the optimal solution. We split the remaining matrix into three regions: $(R^R \setminus R^S) \times C^S$, $R^S \times (C^R \setminus C^S)$, and $(R^R \setminus R^S) \times (C^R \setminus C^S)$.

In the region $(R^R \setminus R^S) \times C^S$ we compute for any row $i \in R^R \setminus R^S$ an additive $\phi|C|$-approximation X_i' of $\|A_{i,C}\|_0$, and we pick row i iff $X_i' \geqslant |C|/2$. In case $\big|\|A_{i,C}\|_0 - |C|/2\big| > \phi|C|$, we have $X_i' \geqslant |C|/2$ if and only if $\|A_{i,C}\|_0 \geqslant |C|/2$, and thus our choice for row i is optimal, restricted to region $(R^R \setminus R^S) \times C^S$. Otherwise, if $\big|\|A_{i,C}\|_0 - |C|/2\big| \leqslant \phi|C|$, then no matter whether we choose row i or not, we obtain approximation ratio

$$\frac{|C|/2 + \phi|C|}{|C|/2 - \phi|C|} = \frac{1+2\phi}{1-2\phi} \leqslant 1 + 5\phi,$$

restricted to region $(R^R \setminus R^S) \times C^S$. The region $R^S \times (C^R \setminus C^S)$ is symmetric.

Finally, in region $(R^R \setminus R^S) \times (C^R \setminus C^S)$ we pessimistically assume that every entry is an error. By Claim 3.6 and Lemma 3.1.7, this submatrix has size at most

$$6\phi\alpha \cdot 6\phi\beta \leqslant 36\phi^2(1+\phi)\|A\|_0 \leqslant 37\phi^2\|A\|_0.$$

In total, over all regions, we computed vectors $\widetilde{u}, \widetilde{v}$ such that

$$\|A - \widetilde{u}\widetilde{v}^T\|_0 \leqslant (1+5\phi)\mathrm{OPT} + 37\phi^2\|A\|_0.$$

This completes the correctness prove of Algorithm 5. □

3.2 Sample Complexity Lower Bound

We give now a lower bound of $\Omega(n/\phi)$ on the number of samples of any $1+O(\phi)$-approximation algorithm for the Binary ℓ_0-Rank-1 problem, where $\phi \geqslant \mathrm{OPT}/\|A\|_0$ as before.

Theorem 1.7 (from page 7)**.** *Let $C \geqslant 1$. Given an $n \times n$ binary matrix A with column adjacency arrays and with row and column sums, and given $\sqrt{\log(n)/n} \ll \phi \leqslant 1/100C$ such that $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, computing a $(1+C\phi)$-approximation of OPT requires to read $\Omega(n/\phi)$ entries of A (in the worst case over A).*

The technical core of our argument is the following lemma.

Lemma 3.7. *Let $\phi \in (0, 1/2)$. Let $X_1, \ldots, X_k$ be binary random variables with expectations $p_1, \ldots, p_k$, where $p_i \in \{1/2 - \phi, 1/2 + \phi\}$ for each i. Let $\mathcal{A}$ be an algorithm which can adaptively obtain any number of samples of each random variable, and which outputs bits b_i for every $i \in [1:k]$. Suppose that with probability at least 0.95 over the joint probability space of $\mathcal{A}$ and the random samples, $\mathcal{A}$ outputs for at least a 0.95 fraction of all i that $b_i = 1$ if $p_i = 1/2 + \phi$ and $b_i = 0$ otherwise. Then, with probability at least 0.05, $\mathcal{A}$ makes $\Omega(k/\phi^2)$ samples in total, asymptotically in k.*

Proof. Consider the following problem P: let X be a binary random variable with expectation p drawn uniformly in $\{1/2 - \phi, 1/2 + \phi\}$. It is well-known that any algorithm which, with probability at least 0.6, obtains samples from X and outputs 0 if $p = 1/2 - \phi$ and outputs 1 if $p = 1/2 + \phi$, requires $\Omega(1/\phi^2)$ samples; see, e.g., Theorem 4.32 of [BY02]. Let $c > 0$ be such that c/ϕ^2 is a lower bound on the number of samples for this problem P.

Let $\mathcal{A}$ be an algorithm solving the problem in the lemma statement. Since $\mathcal{A}$ succeeds with probability at least 0.95 in obtaining the guarantees of the lemma for given sequence $p_1, \ldots, p_k$, it also succeeds with this probability when $(p_1, \ldots, p_k)$ is drawn from the uniform distribution on $\{1/2 - \phi, 1/2 + \phi\}^k$.

Suppose, towards a contradiction, that $\mathcal{A}$ takes less than $0.05 \cdot ck/\phi^2$ samples with probability at least 0.95. By stopping $\mathcal{A}$ before taking $0.05 \cdot ck/\phi^2$ samples, we obtain an algorithm A' that always takes less than $0.05 \cdot ck/\phi^2$ samples. By the union bound, A' obtains the guarantees of the lemma for the output bits b_i with probability at least 0.9, over the joint probability space of A' and the random samples.

Note that the expected number of samples A' takes from a given X_i is less than $0.05 \cdot c/\phi^2$. By Markov's inequality, for a 0.95 fraction of indices i, A' takes less than c/ϕ^2 samples from X_i. We say that i is *good* if A' takes less than c/ϕ^2 samples from X_i and the output bit b_i is correct. By union bound, at least a $1 - (1 - 0.9) - (1 - 0.95) = 0.85$ fraction of indices i is good.

Since $(p_1, \ldots, p_k)$ is drawn from the uniform distribution on $\{1/2 - \phi, 1/2 + \phi\}^k$, with probability at least 0.95 the number $k_+ = |\{i : p_i = 1/2 + \phi\}|$ satisfies $0.45k \leqslant k_+ \leqslant 0.55k$ (for sufficiently large k). This implies that a 0.65 fraction of indices $\{i : p_i = 1/2 + \phi\}$ is good, as otherwise the number of bad i's is at least $(1 - 0.65) \cdot 0.45k > 0.15k$. Similarly, a 0.65 fraction of indices $\{i : p_i = 1/2 - \phi\}$ is good.

Given an instance of problem P with random variable X and expectation p, we choose a uniformly random $i \in [k]$, and set $X_i = X$. For $j \neq i$, we independently and uniformly at random choose $p_j \in \{1/2 - \phi, 1/2 + \phi\}$. We then run algorithm A'. Whenever A' samples from X_i, we sample a new value of X as in problem P. Whenever A' samples from X_j for $j \neq i$, we flip a coin with probability p_j and report the output to A'. If A' takes c/ϕ^2 samples from X_i, then we abort, thus ensuring that A' always takes less than c/ϕ^2 samples from $X_i = X$. Observe that the input to A' is a sequence of random variables $X_1, \ldots, X_k$ with expectations $p_1, \ldots, p_k$ which are independent and uniformly distributed in $\{1/2 - \phi, 1/2 + \phi\}$. In particular, except for their expectation these random variables are indistinguishable.

We now condition on $0.45k \leqslant k_+ \leqslant 0.55k$, which has success probability at least 0.95 for sufficiently large k. Then no matter whether $p_i = 1/2 + \phi$ or $p_i = 1/2 - \phi$, at least a 0.65 fraction of indices j with $p_j = p_i$ is good. Since i was chosen to be a uniformly random position independently of the randomness of the sampling and the algorithm A', and the X_j with $p_j = p_i$ are indistinguishable, with probability at least 0.65 index i is good. In this case, A' takes less than c/ϕ^2 samples from $X_i = X$ and correctly determines the output bit b_i, i.e., whether $p_i = 1/2 + \phi$. As by union bound the total success probability is $1 - (1 - 0.65) - (1 - 0.95) = 0.6$, this contradicts the requirement of c/ϕ^2 samples mentioned above for solving P. Hence, the assumption was wrong, and $\mathcal{A}$ takes $\Omega(k/\phi^2)$ samples with probability at least 0.05. □

We start with a simplified version of our result, where we only have random access to the matrix entries. Below we extend this lower bound to the situation where we even have random access to the adjacency lists of all rows and columns.

Theorem 3.8. *Let $C \geqslant 1$. Given an $n \times n$ binary matrix A by random access to its entries, and given $\sqrt{\log(n)/n} \ll \phi \leqslant 1/100C$ such that $\mathrm{OPT}/\|A\|_0 \leqslant \phi$, computing a $(1 + C\phi)$-approximation of* OPT *requires to read $\Omega(n/\phi)$ entries of A (in the worst case over A).*

Proof. Set $\phi' \overset{\text{def}}{=} 25C\phi$ and $k \overset{\text{def}}{=} \phi n/2$. As in Lemma 3.7, consider binary random variables $X_1, \ldots, X_k$ with expectations $p_1, \ldots, p_k$, where $p_i \in \{1/2 - \phi', 1/2 + \phi'\}$ for each i. We (implicitly) construct an $n \times n$ matrix A as follows. For ever $k < i \leqslant n$, $1 \leqslant j \leqslant n$ we set $A_{i,j} \overset{\text{def}}{=} 1$. For any $1 \leqslant i \leqslant k$, $1 \leqslant j \leqslant n$ we sample a bit $b_{i,j}$ from X_i and set $A_{i,j} \overset{\text{def}}{=} b_{i,j}$. Note that we can run any Binary ℓ_0-Rank-1 algorithm implicitly on A: whenever the algorithm reads an entry $A_{i,j}$ we sample a bit from X_i to determine the entry (and we remember the entry for possible further accesses).

Let us determine the optimal solution for A. Note that for each $i > k$, since the row $A_{i,:}$ is all-ones, it is always better to pick this row than not to pick it, and thus without loss of generality any solution u, v has $u_i = 1$. Similarly, for any j, since the column $A_{:,j}$ has $n - k > n/2$ 1's in rows picked by u, it is always better to pick the column than not to pick it, and thus $v_j = 1$, i.e., v is the all-ones vector. Hence, the only choice is for any $1 \leqslant i \leqslant k$ to pick or not to pick row i. Note that no matter whether we pick these rows or not, the total error is at most $\phi n^2/2$, since these rows in total have $kn = \phi n^2/2$ entries, and all remaining entries of A are correctly recovered by the product uv^T by the already chosen entries of u and v. Hence, $\mathrm{OPT} \leqslant \phi n^2/2$, and since $\|A\|_0 \geqslant (n-k)n \geqslant n^2/2$, we obtain, as required, $\mathrm{OPT}/\|A\|_0 \leqslant \phi$.

Now consider the rows $1 \leqslant i \leqslant k$ more closely. Since v is the all-ones vector, not picking row i incurs cost for each 1 in the row, which is cost $\|A_{i,:}\|_0$, while picking row i incurs cost for each 0 in the row, which is cost $n - \|A_{i,:}\|_0$. Note that the expected number of 1's in row $1 \leqslant i \leqslant k$ is $p_i n$. The Chernoff bound yields concentration: We have w.h.p. $|\|A_{i,:}\|_0 - p_i n| \leqslant 0.01 \cdot \phi' n$, where we used $\phi' \gg \sqrt{\log(n)/n}$. In the following we condition on this event and thus drop "w.h.p." from our statements. In particular, for any i with $p_i = 1/2 + \phi'$ we have $\|A_{i,:}\|_0 \geqslant (1/2 + 0.99\phi')n$, and for any i with $p_i = 1/2 - \phi'$ we have $\|A_{i,:}\|_0 \leqslant (1/2 - 0.99\phi')n$.

By picking all rows $i \leqslant k$ with $p_i = 1/2 + \phi'$ and not pick the rows with $p_i = 1/2 - \phi'$, we see that $\mathrm{OPT} \leqslant (1/2 - 0.99\phi')kn$. Now consider a solution u that among the rows $1 \leqslant i \leqslant k$ with $p_i = 1/2 + \phi'$ picks g_+ many and does not pick b_+ many. Similarly, among the rows with $p_i = 1/2 - \phi'$ it picks g_- and does not pick b_-. Note that each of the g_+ "good" rows incurs cost

$$n - \|A_{i:}\|_0 \geqslant n - (1/2 + 1.01\phi')n = (1/2 - 1.01\phi')n.$$

Each of the b_+ "bad" rows incurs a cost of $\|A_{i:}\|_0 \geqslant (1/2 + 0.99\phi')n$. Similar statements hold for g_- and b_-, and thus for $g \overset{\text{def}}{=} g_+ + g_-$ and $b \overset{\text{def}}{=} b_+ + b_-$, with $g + b = k$, we obtain a total cost of

$$\begin{aligned} \|A - uv^T\|_0 &\geqslant g \cdot (1/2 - 1.01\phi')n + b \cdot (1/2 + 0.99\phi')n \\ &= k(1/2 - 0.99\phi')n + 2b\phi' n - 0.02k\phi' n \\ &\geqslant \mathrm{OPT} + 2b\phi' n - 0.02\phi' kn. \end{aligned}$$

If $b \geqslant 0.02k$, then

$$\|A - uv^T\|_0 \geqslant \mathrm{OPT} + 0.02\phi' kn \geqslant (1 + 0.04\phi')\mathrm{OPT}.$$

By contraposition, if we compute a $(1 + 0.04\phi' = 1 + C\phi)$-approximation on A, then $b \leqslant 0.02k$, and thus the vector u correctly identifies for at least a 0.98 fraction of the random variables X_i whether $p_i = 1/2 + \phi'$ or $p_i = 1/2 - \phi'$. Since this holds w.h.p., by Lemma 3.7 we need $\Omega(k/\phi'^2) = \Omega(n/(\phi C^2))$ samples from the variables X_i, and thus $\Omega(n/(\phi C^2))$ reads in A. Since $C \geqslant 1$ is constant, we obtain a lower bound of $\Omega(n/\phi)$. This lower bound holds in expectation over the constructed distribution of A-matrices, and thus also in the worst case over A. □

The construction of the above theorem does not work in case when we have random access to the adjacency lists of the rows, since this allows us to quickly determine the numbers of 1's per row, which is all we need to determine whether we want to pick a particular row in the matrix constructed above. To treat this issue, we adapt the construction as follows.

Proof of Theorem 1.7. We assume that n is even. Let $\phi', k, X_1, \ldots, X_k, p_1, \ldots, p_k$ be as in the proof of Theorem 3.8. We adapt the construction of the matrix A as follows. For any $2k < i \leqslant n$, $1 \leqslant j \leqslant n/2$ we set $A_{i,2j} \overset{\text{def}}{=} 1$ and $A_{i,2j-1} \overset{\text{def}}{=} 0$. For any $1 \leqslant i \leqslant k$, $1 \leqslant j \leqslant n/2$ we sample a bit $b_{i,j}$ from X_i and set $A_{2i,2j} \overset{\text{def}}{=} A_{2i-1,2j-1} \overset{\text{def}}{=} b_{i,j}$ and $A_{2i-1,2j} \overset{\text{def}}{=} A_{2i,2j-1} \overset{\text{def}}{=} 1 - b_{i,j}$.

As before, when running any Binary ℓ_0-Rank-1 algorithm on A we can easily support random accesses to entries $A_{i,j}$, by sampling from $X_{\lceil i/2 \rceil}$ to determine the entry (and remembering the sampled bit for possible further accesses). Furthermore, we can now allow random accesses to the *adjacency arrays* of rows and columns. Specifically, if we want to determine the ℓ-th 1 in row $i \leqslant 2k$, we know that among the entries $A_{i,1}, \ldots, A_{i,2\ell}$ there are exactly ℓ 1's, since by construction $A_{i,2j-1} + A_{i,2j} = 1$. Hence, the ℓ-th 1 in row i is at position $A_{i,2\ell-1}$ or $A_{i,2\ell}$, depending only on the sample $b_{\lceil i/2 \rceil, \ell}$ from $X_{\lceil i/2 \rceil}$. For rows $i > 2k$, the ℓ-th 1 is simply at position $A_{i,2\ell}$. Thus, accessing the ℓ-th 1 in any row takes at most one sample, so we can simulate any algorithm on A with random access to the adjacency lists of rows. The situation for columns is essentially symmetric. Similarly, we can allow constant time access to the row and column sums.

In the remainder we show that the constructed matrix A has essentially the same properties as the construction in Theorem 3.8. We first argue that any 2-approximation u, v for the Binary ℓ_0-Rank-1 problem on A picks all rows $i > 2k$ and picks all even columns and does not pick any odd column. Thus, the only remaining choice is which rows $i \leqslant 2k$ to pick. To prove this claim, first note that any solution following this pattern has error at most $2kn = \phi n^2$, since the $2k$ undecided rows have $2kn$ entries, and all other entries are correctly recovered by the already chosen parts of uv^T. Hence, we have $\mathrm{OPT} \leqslant \phi n^2$. Now consider any 2-approximation u, v, which must have cost at most $2\phi n^2$. Note that u picks at least

$(1-5\phi)n$ of the rows $\{2k+1,\dots,n\}$, since each such row contains $n/2$ 1's that can only be recovered if we pick the row, so we can afford to ignore at most $8k = 4\phi n$ of these $n-2k = (1-\phi)n$ rows. Now, each even column contains at least $(1-5\phi)n > n/2$ 1's in picked rows, and thus it is always better to pick the even columns. Similarly, each odd column contains at least $n/2$ 0's in picked rows, and thus it is always better not to pick the odd columns. Hence, we obtain without loss of generality $v_{2j} = 1$ and $v_{2j-1} = 0$. Finally, each row $i > 2k$ contains $n/2$ 1's in columns picked by v and $n/2$ 0's in columns not picked by v, and thus it is always better to pick row i. Hence, we obtain without loss of generality $u_i = 1$ for $i > 2k$.

Our goal now is to lower bound $\|A - uv^T\|_0$ in terms of OPT and the error term $b\phi' n$, similarly to the proof in Theorem 3.8. Notice that we may ignore the odd columns, as they are not picked by v. Restricted to the even columns, row $2i$ is exactly as row i in the construction in Theorem 3.8, while row $2i-1$ is row $2i$ negated. Thus, analogously as in the proof of Theorem 3.8, we obtain w.h.p. $\mathrm{OPT} \leqslant (1/2 - 0.99\phi')2kn$ and

$$\|A - uv^T\|_0 \geqslant \mathrm{OPT} + 2b\phi' n - 0.04k\phi' n \geqslant (1 + 0.04\phi')\mathrm{OPT},$$

where $b \geqslant 0.04k$ is the number of "bad" rows $i \leqslant 2k$. Again analogously, if we compute a $(1+0.04\phi') = 1 + C\phi)$-approximation on A, then $b \leqslant 0.04k$, and thus w.h.p. we correctly identify for at least a 0.96 fraction of the random variables X_i whether $p_i = 1/2 + \phi'$ or $p_i = 1/2 - \phi'$. As before, this yields a lower bound of $\Omega(n/\phi)$ samples. □

3.3 Exact Algorithm

A variant of the algorithm from Theorem 1.4 can also be used to solve the Binary ℓ_0-Rank-1 problem exactly. This yields the following theorem, which in particular shows that the problem is in polynomial time when $\mathrm{OPT} \leqslant O\big(\sqrt{\|A\|_0}\log(mn)\big)$.

Theorem 1.8 (from page 7)**.** *Given a matrix $A \in \{0,1\}^{m\times n}$, if $\mathrm{OPT}/\|A\|_0 \leqslant 1/240$ then we can solve exactly the Binary ℓ_0-Rank-1 problem in time $2^{O(\mathrm{OPT}/\sqrt{\|A\|_0})} \cdot \mathrm{poly}(mn)$.*

Proof. This algorithm builds upon the algorithmic results established in Theorem 1.6 and Theorem 1.12, and it consists of the following three phases:

1. Run the algorithm in Theorem 1.12 to compute a 3-approximation of $\psi = \mathrm{OPT}/\|A\|_0$, i.e. a number $\phi \in [\psi, 3\psi]$.

2. Run Steps 1-4 of Algorithm 5, resulting in selected rows R^S and columns C^S, and undecided rows $R' = R^R \setminus R^S$ and columns $C' = C^R \setminus C^S$. As shown above, the choices made by these steps are optimal.

3. For the remaining rows R' and columns C' we use brute force to find the optimum solution. Specifically, assume without loss of generality that $|R'| \leqslant |C'|$. Enumerate all binary vectors $u' \in \{0,1\}^{R'}$. For each u', set $\widetilde{u}_i = u'_i$ for all $i \in R'$ to complete the specification of a vector $\widetilde{u} \in \{0,1\}^m$. We can now find the optimal choice of vector $\widetilde{v}$ in polynomial time, since the optimal choice is to set $\widetilde{v}_j = 1$ iff column $A_{:,j}$ has more 1's than 0's in the support of $\widetilde{u}$. Since some u' gives rise to the optimal vector $\widetilde{u} = u$, we solve the Binary ℓ_0-Rank-1 problem exactly.

To analyze the running time, note that by Claim 3.6 we have

$$\min\{|R'|, |C'|\} \leqslant 6\phi \min\{\alpha, \beta\} \leqslant 6\phi\sqrt{\alpha\beta}.$$

By Lemma 3.1.7 and $\phi \leqslant 3\psi$, we obtain $\min\{|R'|, |C'|\} = O(\psi\sqrt{\|A\|_0})$. Hence, we enumerate $2^{O(\psi\sqrt{\|A\|_0})} = 2^{O(\mathrm{OPT}/\sqrt{\|A\|_0})}$ vectors u', and the total running time is $2^{O(\mathrm{OPT}/\sqrt{\|A\|_0})} \cdot \mathrm{poly}(mn)$. This completes the proof of Theorem 1.8. □

Chapter 4

Algorithms For Reals ℓ_0-Rank-k

Given a matrix $A \in \mathbb{R}^{m\times n}$ with $m \geqslant n$, an integer k and an error $\varepsilon \in (0, 1/2)$, our goal is to find an approximate solution $\widetilde{u} \in \mathbb{R}^m$, $\widetilde{v} \in \mathbb{R}^n$ of the Reals ℓ_0-Rank-k problem such that $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (1 + \varepsilon)\mathrm{OPT}_k$, where the optimal value is defined by

$$\mathrm{OPT}_k \overset{\text{def}}{=} \min_{U\in\mathbb{R}^{m\times k},\, V\in\mathbb{R}^{k\times n}} \|A - U \cdot V\|_0. \tag{4.1}$$

Organization In Section 4.1, we design a poly$(k, \log n)$ bicriteria algorithm for the Reals ℓ_0-Rank-k problem, which runs in polynomial time. In Section 4.2, we give an efficient $(2 + \varepsilon)$-approximation algorithm for the Reals ℓ_0-Rank-1 problem.

4.1 Polytime Bicriteria Algorithm For Reals ℓ_0-Rank-k

This section is organized as follows. In Subsection 4.1.1, we prove a structural lemma that guarantees the existence of k columns that yield a $(k+1)$-approximation of OPT_k, and we also give an $\Omega(k)$-approximation lower bound for any algorithm that selects k columns from the input matrix A. In Subsection 4.1.2, we give an approximation algorithm that runs in time poly(n^k, m) and achieves $O(k^2)$-approximation. To the best of our knowledge, this is the first algorithm with provable non-trivial approximation guarantees. In Subsection 4.1.3, we design a practical algorithm that runs in time poly(m, n) with an exponent independent of k, if we allow for a bicriteria solution.

4.1.1 Structural Results

We give a novel structural result showing that any matrix A contains k columns which provide a $(k + 1)$-approximation for the Reals ℓ_0-Rank-k problem (4.1).

Lemma 4.1. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix and k be an integer. There is a subset $J^{(k)} \subset [n]$ of size k and a matrix $Z \in \mathbb{R}^{k\times n}$ such that $\|A - A_{:,J^{(k)}} \cdot Z\|_0 \leqslant (k + 1)\mathrm{OPT}_k$.*

Proof. Let $Q^{(0)}$ be the set of columns j with $UV_{:,j} = 0$, and let $R^{(0)} \overset{\text{def}}{=} [n] \setminus Q^{(0)}$. Let $S^{(0)} \overset{\text{def}}{=} [n]$, $T^{(0)} \overset{\text{def}}{=} \emptyset$. We split the value OPT_k into $\mathrm{OPT}(S^{(0)}, R^{(0)}) \overset{\text{def}}{=} \|A_{S^{(0)},R^{(0)}} - UV_{S^{(0)},R^{(0)}}\|_0$ and

$$\mathrm{OPT}(S^{(0)}, Q^{(0)}) \overset{\text{def}}{=} \|A_{S^{(0)},Q^{(0)}} - UV_{S^{(0)},Q^{(0)}}\|_0 = \|A_{S^{(0)},Q^{(0)}}\|_0.$$

Suppose $\mathrm{OPT}(S^{(0)}, R^{(0)}) \geqslant |S^{(0)}||R^{(0)}|/(k + 1)$. Then, for any subset $J^{(k)}$ it follows that

$$\min_Z \|A - A_{S^{(0)},J^{(k)}} Z\|_0 \leqslant |S^{(0)}||R^{(0)}| + \|A_{S^{(0)},Q^{(0)}}\|_0 \leqslant (k + 1)\mathrm{OPT}_k.$$

Otherwise, there is a column $i^{(1)}$ such that

$$\left\|A_{S^{(0)},i^{(1)}} - (UV)_{S^{(0)},i^{(1)}}\right\|_0 \leqslant \mathrm{OPT}(S^{(0)}, R^{(0)})/|R^{(0)}| \leqslant \mathrm{OPT}_k/|R^{(0)}|.$$

Let $T^{(1)}$ be the set of indices on which $(UV)_{S^{(0)},i^{(1)}}$ and $A_{S^{(0)},i^{(1)}}$ disagree, and similarly $S^{(1)} \overset{\text{def}}{=} S^{(0)} \backslash T^{(1)}$ on which they agree. Then we have $|T^{(1)}| \leqslant \mathrm{OPT}_k/|R^{(0)}|$. Hence, in the submatrix $T^{(1)} \times R^{(0)}$ the total error is at most $|T^{(1)}| \cdot |R^{(0)}| \leqslant \mathrm{OPT}_k$. Let $R^{(1)}, D^{(1)}$ be a partitioning of $R^{(0)}$ such that $A_{S^{(1)},j}$ is linearly dependent on $A_{S^{(1)},i^{(1)}}$ iff $j \in D^{(1)}$. Then by selecting column $A_{:,i^{(1)}}$ the incurred cost on matrix $S^{(1)} \times D^{(1)}$ is zero. For the remaining submatrix $S^{(\ell)} \times R^{(\ell)}$, we perform a recursive call of the algorithm.

We make at most k recursive calls, on instances $S^{(\ell)} \times R^{(\ell)}$ for $\ell \in \{0, \ldots, k-1\}$. In the ℓ^{th} iteration, either $\mathrm{OPT}(S^{(\ell)}, R^{(\ell)}) \geqslant |S^{(\ell)}||R^{(\ell)}|/(k+1-\ell)$ and we are done, or there is a column $i^{(\ell+1)}$ which partitions $S^{(\ell)}$ into $S^{(\ell+1)}, T^{(\ell+1)}$ and $R^{(\ell)}$ into $R^{(\ell+1)}, D^{(\ell+1)}$ such that

$$|S^{(\ell+1)}| \geqslant m \cdot \prod_{i=0}^{\ell} \left(1 - \frac{1}{k+1-i}\right) = \frac{k-\ell}{k+1} \cdot m$$

and for every $j \in D^{(\ell)}$ the column $A_{S^{(\ell+1)},j}$ belongs to the span of $\{A_{S^{(\ell+1)},i^{(t)}}\}_{t=1}^{\ell+1}$.

Suppose we performed k recursive calls. We show now that the incurred cost in submatrix $S^{(k)} \times R^{(k)}$ is at most $\mathrm{OPT}(S^{(k)}, R^{(k)}) \leqslant \mathrm{OPT}_k$. By construction, $|S^{(k)}| \geqslant m/(k+1)$ and the sub-columns $\{A_{S^{(k)},i}\}_{i \in I^{(k)}}$ are linearly independent, where $I^{(k)} = \{i^{(1)}, \ldots, i^{(k)}\}$ is the set of the selected columns, and $A_{S^{(k)},I^{(k)}} = (UV)_{S^{(k)},I^{(k)}}$. Since $\mathrm{rank}(A_{S^{(k)},I^{(k)}}) = k$, it follows that $\mathrm{rank}(U_{S^{(k)},:}) = k$, $\mathrm{rank}(V_{:,I^{(k)}}) = k$ and the matrix $V_{:,I^{(k)}} \in \mathbb{R}^{k \times k}$ is invertible. Hence, for matrix $Z = (V_{:,I^{(k)}})^{-1} V_{:,R^k}$ we have

$$\mathrm{OPT}(S^{(k)}, R^{(k)}) = \|A_{S^k,R^k} - A_{S^k,I^k} Z\|_0.$$

The statement follows by noting that the recursive calls accumulate a total cost of at most $k \cdot \mathrm{OPT}_k$ in the submatrices $T^{(\ell+1)} \times R^{(\ell)}$ for $\ell \in \{0, 1, \ldots, k-1\}$, as well as cost at most OPT_k in submatrix $S^{(k)} \times R^{(k)}$. □

We now show that any algorithm that selects k columns of a matrix A incurs at least an $\Omega(k)$-approximation for the Reals ℓ_0-Rank-k problem.

Lemma 4.2. *Let $k \leqslant n/2$. Suppose $A = (G_{k \times n};\ I_{n \times n}) \in \mathbb{R}^{(n+k) \times n}$ is a matrix composed of a Gaussian random matrix $G \in \mathbb{R}^{k \times n}$ with $G_{i,j} \sim N(0,1)$ and identity matrix $I_{n \times n}$. Then for any subset $J^{(k)} \subset [n]$ of size k, we have $\min_{Z \in \mathbb{R}^{k \times n}} \|A - A_{:,J^{(k)}} \cdot Z\|_0 = \Omega(k) \cdot \mathrm{OPT}_k$.*

Proof. Notice that the optimum cost is at most n, achieved by selecting $U = (I_{k \times k};\ 0_{n \times k})$ and $V = G_{k \times n}$. It is well known that Gaussian matrices are invertible with probability 1, see e.g. [SST06, Thm 3.3]. Hence, $G_{:,J^{(k)}}$ is a nonsingular matrix for every subset $J^{(k)} \subset [n]$ of size k.

We will show next that for any subset $J^{(k)}$ of k columns the incurred cost is at least $(n-k)k \geqslant nk/2$. Without loss of generality, the chosen columns $J^{(k)} = [k]$ are the first k columns of A. Let $R = [2k]$ be the first $2k$ rows and $C = [n] \setminus J$ be the last $n-k$ columns. We bound

$$\begin{aligned} \min_Z \|A - A_{:,[k]} Z\|_0 &\geqslant \min_Z \|A_{R,C} - A_{R,[k]} Z\|_0 \\ &= \sum_{j \in C} \min_{z^{(j)}} \|A_{R,j} - A_{R,[k]} z^{(j)}\|_0, \end{aligned}$$

i.e. we ignore all rows and columns except R and C. Consider any column $j \in C$. Since $A_{R,j} = (G_{:,j}, 0_k)$ and $A_{R,[k]} = (G_{:,[k]}, I_{k \times k})$, for any vector $z \in \mathbb{R}^k$ we have

$$\begin{aligned} \|A_{R,j} - A_{R,J^{(k)}} z\|_0 &= \|G_{:,j} - G_{:,[k]} z\|_0 + \|I_{k \times k} z\|_0 \\ &= \|G_{:,j} - G_{:,[k]} z\|_0 + \|z\|_0. \end{aligned}$$

Let $\ell \overset{\text{def}}{=} \|z\|_0$. By symmetry, without loss of generality we can assume that the first ℓ entries of z are non-zero and the remaining entries are 0. Let $x \in \mathbb{R}^\ell$ be the vector containing the first ℓ entries of z. Then we have

$$\|A_{R,j} - A_{R,J^{(k)}} z\|_0 = \|G_{:,j} - G_{:,[\ell]} x\|_0 + \ell.$$

We consider w.l.o.g. the first k columns of A, and we construct the optimum matrix Z that minimizes $\|A_{:,1:k} Z - A\|_0$. Observe that it is optimal to set the first k columns of Z to $I_{k \times k}$, and since $A_{2k+1:n,1:k} = 0$ we can focus only on the submatrix $A_{1:2k,k+1:n} = (G_{1:k,k+1:n}; 0_{k \times n-k})$.

Consider a column $A_{1:2k,j}$ for $j \in [k+1, n]$. Our goal is to find a vector $v \in \mathbb{R}^k$ minimizing the objective function $\Psi = \min_v \{\|v\|_0 + \|G^{(k)} v - g\|_0\}$, where $G^{(k)} \overset{\text{def}}{=} \{G_{1:k,1:k}\}$ and $g \overset{\text{def}}{=} G_{1:k,j}$. It holds with probability 1 that $G^{(k)}$ and g do not have an entry equal to zero. Moreover, since $G^{(k)}$ is invertible every row in $[G^{(k)}]^{-1}$ is non-zero, and thus with probability 1 a vector $v = [G^{(k)}]^{-1} g$ has entry equal to zero.

Let $v = (x;\ 0)$ be an arbitrary vector with $\|x\|_1 = \ell$. Let $G^{(\ell)}$ be a submatrix of $G^{(k)}$ induced by the first ℓ columns. For every subset $S \subset [m]$ of ℓ rows the corresponding submatrix $G_{S,:}^{(\ell)}$ has a full rank.

Suppose there is a subset S such that for $G^{(\ell)}_{S,:}$ and g_S there is a vector $x \in \mathbb{R}^k$ satisfying $G^{(\ell)}_{S,:}x = g_S$. Since $G^{(\ell)}_{S,:}$ is invertible, the existence of x implies its uniqueness. On the other hand, for any row $i \in [m] \backslash S$ the probability of the event $G^{(\ell)}_{i,:}x = g_i$ is equals to 0. Since $G^{(k)}v = G^{(\ell)}x$ and there are finitely many possible subsets S as above, i.e. $\binom{m}{\ell} \leqslant m^\ell$, by union bound it follows that $\|G^{(k)}v - g\|_0 \geqslant k - \ell$. Therefore, it holds that $\phi \geqslant k$.

The statement follows by noting that the total cost incurred by $A_{:,1:k}$ and any Z is lower bounded by $(n-k)k + (n-k) = (1 - k/n)(k+1)n$. □

4.1.2 Basic Algorithm

We give an impractical algorithm that runs in time $\text{poly}(n^k, sm)$ and achieves $O(k^2)$-approximation. To the best of our knowledge this is the first approximation algorithm for the ℓ_0-Rank-k problem with non-trivial approximation guarantees.

Theorem 1.10 (from page 8)**.** *Given $A \in \mathbb{R}^{m\times n}$ and $k \in [n]$ we can compute in $O(n^{k+1}m^2k^{\omega+1})$ time a set of k indices $J^{(k)} \subset [n]$ and a matrix $Z \in \mathbb{R}^{k\times n}$ such that $\|A - A_{:,J^{(k)}} \cdot Z\|_0 \leqslant O(k^2) \cdot \text{OPT}_k$.*

We use as a subroutine the algorithm of Berman and Karpinski [BK02] (attributed also to Kannan in that paper) which given a matrix U and a vector b approximates $\min_x \|Ux - b\|_0$ in polynomial time. Specifically, we invoke in our algorithm the following variant of this result established by Alon, Panigrahy, and Yekhanin [APY09].

Theorem 4.3. *[APY09] There is an algorithm that given a matrix $A \in \mathbb{R}^{m\times k}$ and a vector $b \in \mathbb{R}^m$, outputs in time $O(m^2k^{\omega+1})$ a vector $z \in \mathbb{R}^k$ such that w.h.p. $\|Az - b\|_0 \leqslant k \cdot \min_x \|Ax - b\|_0$.*

Proof of Theorem 1.10. The existence of a subset J^* of k columns of A and matrix $Z^* \in \mathbb{R}^{k\times n}$ with $\|A - A_{:,J^*}Z^*\|_0 \leqslant (k+1)\text{OPT}_k$ follows by Lemma 4.1. We enumerate all $\binom{n}{k}$ subsets $J^{(k)}$ of k columns. For each $J^{(k)}$, we split $\min_Z \|A_{:,J^{(k)}}Z - A\|_0 = \sum_{i=1}^n \min_{z^{(i)}} \|A_{:,J^{(k)}}z^{(i)} - A_{:,i}\|_0$, and we run the algorithm from Theorem 4.3 for each column $A_{:,i}$, obtaining approximate solutions $\widetilde{z}^{(1)}, \ldots, \widetilde{z}^{(n)}$ that form a matrix $\widetilde{Z}$. Then, we return the best solution $(A_{:,J^{(k)}}, \widetilde{Z})$. To verify that this yields a $k(k+1)$-approximation, note that for $J^{(k)} = J^*$ we have

$$\begin{aligned}\|A_{:,J^*}\widetilde{Z} - A\|_0 &= \sum_{i=1}^n \|A_{:,J^*}\widetilde{z}^{(i)} - A_{:,i}\|_0 \leqslant k \sum_{i=1}^n \min_{z^{(i)}} \|A_{:,J^*}z^{(i)} - A_{:,i}\|_0 \\ &= k \cdot \min_Z \|A_{:,J^*}Z - A\|_0 \leqslant k(k+1) \cdot \text{OPT}_k.\end{aligned}$$

The time bound $O(n^{k+1}m^2k^{\omega+1})$ is immediate from Theorem 4.3. This proves the statement. □

4.1.3 Polynomial Time Bicriteria Algorithm

Our main contribution in this section is to design a practical algorithm that runs in time $\text{poly}(n, m)$ with an exponent independent of k, if we allow for a bicriteria solution.

Theorem 1.11 (from page 8)**.** *Given a matrix $A \in \mathbb{R}^{m\times n}$ and an integer k, there is an algorithm that in expected time $\text{poly}(m, n)$ outputs a subset of indices $J \subset [n]$ with $|J| = O(k \log(n/k))$ and a matrix $Z \in \mathbb{R}^{|J|\times n}$ such that $\|A - A_{:,J} \cdot Z\|_0 \leqslant O(k^2 \log(n/k)) \cdot \text{OPT}_k$.*

The structure of the proof follows a recent approximation algorithm [CGK+17, Algorithm 3] for the ℓ_p-low rank approximation problem, for any $p \geqslant 1$. We note that the analysis of [CGK+17, Theorem 7] is missing an $O(\log^{1/p} n)$ approximation factor, and naïvely provides an $O(k \log^{1/p} n)$-approximation rather than the stated $O(k)$-approximation. Further, it might be possible to obtain an efficient algorithm yielding an $O(k^2 \log k)$-approximation for Theorem 1.11 using unpublished techniques in [SWZ18]; we leave the study of obtaining the optimal approximation factor to future work.

There are two critical differences with the proof of [CGK+17, Theorem 7]. We cannot use the earlier [CGK+17, Theorem 3] which shows that any matrix A contains k columns which provide an $O(k)$-approximation for the ℓ_p-low rank approximation problem, since that proof requires $p \geqslant 1$ and critically uses scale-invariance, which does not hold for $p = 0$. Our combinatorial argument in Lemma 4.1 seems fundamentally different than the maximum volume submatrix argument in [CGK+17] for $p \geqslant 1$.

Second, unlike for ℓ_p-regression for $p \geqslant 1$, the ℓ_0-regression problem $\min_x \|Ux - b\|_0$ given a matrix U and vector b is not efficiently solvable, since it corresponds to a nearest codeword problem which is

Algorithm 6 Bicriteria Algorithm: Selecting $O(k\log(n/k))$ Columns

```
ApproximatelySelectColumns($A, k$)
  ensure $A$ has at least $k\log(n/k))$ columns
1. If the number of columns of matrix $A$ is less than or equal to $2k$
2.     Return all the columns of $A$
3. Else
4.     Repeat
5.         Let $R$ be a set of $2k$ uniformly random columns of $A$
6.     Until at least $1/10$ fraction of columns of $A$ are nearly approximately covered
7.     Let $A_{\overline{R}}$ be the columns of $A$ not nearly approximately covered by $R$
8.     Return $R \cup$ ApproximatelySelectColumns($A_{\overline{R}}, k$)
```

NP-hard [Ale11]. Thus, we resort to an approximation algorithm for ℓ_0-regression, based on ideas for solving the nearest codeword problem in [APY09, BK02].

Note that $\mathrm{OPT}_k \leqslant \|A\|_0$. Since there are only $mn+1$ possibilities of OPT_k, we can assume we know OPT_k and we can run the Algorithm 6 for each such possibility, obtaining a rank-$O(k\log n)$ solution, and then outputting the solution found with the smallest cost.

This can be further optimized by forming instead $O(\log(mn))$ guesses of OPT_k. One of these guesses is within a factor of 2 from the true value of OPT_k, and we note that the following argument only needs to know OPT_k up to a factor of 2.

We start by defining the notion of approximate coverage, which is different than the corresponding notion in [CGK$^+$17] for $p \geqslant 1$, due to the fact that ℓ_0-regression cannot be efficiently solved. Consequently, approximate coverage for $p = 0$ cannot be efficiently tested. Let $Q \subseteq [n]$ and $M = A_{:,Q}$ be an $m \times |Q|$ submatrix of A. We say that a column $M_{:,i}$ is (S, Q)-*approximately covered* by a submatrix $M_{:,S}$ of M, if $|S| = 2k$ and

$$\min_x \|M_{:,S} \cdot x - M_{:,i}\|_0 \leqslant \frac{100(k+1)\mathrm{OPT}_k}{|Q|}. \tag{4.2}$$

Lemma 4.4. *(Similar to [CGK$^+$17, Lemma 6], but using Lemma 4.1) Let $Q \subseteq [n]$ and $M = A_{:,Q}$ be a submatrix of A. Suppose we select a subset R of $2k$ uniformly random columns of M. Then with probability at least $1/3$, at least a $1/10$ fraction of the columns of M are (R, Q)-approximately covered.*

Proof. To show this, as in [CGK$^+$17], consider a uniformly random column index i not in the set R. Let $T \overset{\text{def}}{=} R \cup \{i\}$, $\eta \overset{\text{def}}{=} \min_{\mathrm{rank}(B)=k} \|M_{:,T} - B\|_0$, and $B^\star \overset{\text{def}}{=} \arg\min_{\mathrm{rank}(B)=k} \|M - B\|_0$. Since T is a uniformly random subset of $2k+1$ columns of M, we have

$$\begin{aligned} \mathbb{E}_T \eta &\leqslant \mathbb{E}_T \|M_{:,T} - B^\star_{:,T}\|_0 = \sum_{T \in \binom{|Q|}{2k+1}} \sum_{i \in T} \|M_{:,i} - B^\star_{:,i}\|_0 Pr\,[T] \\ &= \sum_{i \in Q} \frac{\binom{|Q|-1}{|T-1|}}{\binom{|Q|}{|T|}} \|M_{:,i} - B^\star_{:,i}\|_0 = \frac{(2k+1)\mathrm{OPT}_k^M}{|Q|} \leqslant \frac{(2k+1)\mathrm{OPT}_k}{|Q|}. \end{aligned}$$

Then, by a Markov bound, we have $\Pr[\eta \leqslant \frac{10(2k+1)\mathrm{OPT}_k}{|Q|}] \geqslant 9/10$. Let $\mathcal{E}_1$ denotes this event.

Fix a configuration $T = R \cup \{i\}$ and let $L(T) \subset T$ be the subset guaranteed by Lemma 4.1 such that $|L(T)| = k$ and

$$\min_X \|M_{:,L(T)} X - M_{:,T}\|_0 \leqslant (k+1) \min_{\mathrm{rank}(B)=k} \|M_{:,T} - B\|_0.$$

Notice that

$$\mathbb{E}_i \left[\min_x \|M_{:,L(T)} x - M_{:,i}\|_0 \mid T \right] = \frac{1}{2k+1} \min_X \|M_{:,L(T)} X - M_{:,T}\|_0,$$

and thus by the law of total probability we have

$$\mathbb{E}_T \left[\min_x \|M_{:,L(T)} x - M_{:,i}\|_0 \right] \leqslant \frac{(k+1)\eta}{2k+1}.$$

Let $\mathcal{E}_2$ denote the event that $\min_x \|M_{:,L} x - M_{:,i}\|_0 \leqslant \frac{10(k+1)\eta}{2k+1}$. By a Markov bound, $\Pr[\mathcal{E}_2] \geqslant 9/10$.

Further, as in [CGK+17], let $\mathcal{E}_3$ be the event that $i \notin L$. Observe that there are $\binom{k+1}{k}$ ways to choose a subset $R' \subset T$ such that $|R'| = 2k$ and $L \subset R'$. Since there are $\binom{2k+1}{2k}$ ways to choose R', it follows that $\Pr[L \subset R \mid T] = \frac{k+1}{2k+1} > 1/2$. Hence, by the law of total probability, we have $\Pr[\mathcal{E}_3] > 1/2$.

As in [CGK+17], $\Pr[\mathcal{E}_1 \wedge \mathcal{E}_2 \wedge \mathcal{E}_3] > 2/5$, and conditioned on $\mathcal{E}_1 \wedge \mathcal{E}_2 \wedge \mathcal{E}_3$,

$$\min_x \|M_{:,R}x - M_{:,i}\|_0 \leqslant \min_x \|M_{:,L}x - M_{:,i}\|_0 \leqslant \frac{10(k+1)\eta}{2k+1} \leqslant \frac{100(k+1)\mathrm{OPT}_k}{|Q|}, \tag{4.3}$$

where the first inequality uses that L is a subset of R given $\mathcal{E}_3$, and so the regression cost cannot decrease, while the second inequality uses the occurrence of $\mathcal{E}_2$ and the final inequality uses the occurrence of $\mathcal{E}_1$.

As in [CGK+17], if Z_i is an indicator random variable indicating whether i is approximately covered by R, and $Z = \sum_{i \in Q} Z_i$, then $\mathbb{E}_R[Z] \geqslant \frac{2|Q|}{5}$ and $\mathbb{E}_R[|Q| - Z] \leqslant \frac{3|Q|}{5}$. By a Markov bound, it follows that $\Pr[|Q| - Z \geqslant \frac{9|Q|}{10}] \leqslant \frac{2}{3}$. Thus, probability at least $1/3$, at least a $1/10$ fraction of the columns of M are (R, Q)-approximately covered. □

Given Lemma 4.4, we are ready to prove Theorem 1.11. As noted above, a key difference with the corresponding [CGK+17, Algorithm 3] for ℓ_p and $p \geqslant 1$, is that we cannot efficiently test if the i-th column is approximately covered by the set R. We will instead again make use of Theorem 4.3.

Proof of Theorem 1.11. The computation of matrix Z force us to relax the notion of (R, Q)-approximately covered to the notion of (R, Q)-nearly-approximately covered as follows: we say that a column $M_{:,i}$ is (R, Q)-*nearly-approximately covered* if, the algorithm in Theorem 4.3 returns a vector z such that

$$\|M_{:,R}z - M_{:,i}\|_0 \leqslant \frac{100(k+1)^2\mathrm{OPT}_k}{|Q|}. \tag{4.4}$$

By the guarantee of Theorem 4.3, if $M_{:,i}$ is (R, Q)-approximately covered then it is also with probability at least $1 - 1/\operatorname{poly}(mn)$ (R, Q)-nearly-approximately covered.

Suppose Algorithm 6 makes t iterations and let $A_{:,\cup_{i=1}^t R_i}$ and Z be the resulting solution. We bound now its cost. Let $B_0 = [n]$, and consider the i-th iteration of Algorithm 6. We denote by R_i a set of $2k$ uniformly random columns of B_{i-1}, by G_i a set of columns that is (R_i, B_{i-1})-nearly-approximately covered, and by $B_i = B_{i-1} \setminus \{G_i \cup R_i\}$ a set of the remaining columns. By construction, $|G_i| \geqslant |B_{i-1}|/10$ and

$$|B_i| \leqslant \frac{9}{10}|B_{i-1}| - 2k < \frac{9}{10}|B_{i-1}|.$$

Since Algorithm 6 terminates when $B_{t+1} \leqslant 2k$, we have

$$2k < |B_t| < \left(1 - \frac{1}{10}\right)^t |B_0| = \left(1 - \frac{1}{10}\right)^t n,$$

and thus the number of iterations $t \leqslant 10 \log(n/2k)$. By construction, $|G_i| = (1 - \alpha_i)|B_{i-1}|$ for some $\alpha_i \leqslant 9/10$, and hence

$$\sum_{i=1}^t \frac{|G_i|}{|B_{i-1}|} \leqslant t \leqslant 10 \log \frac{n}{2k}. \tag{4.5}$$

Therefore, the solution cost is bounded by

$$\begin{aligned} \|A_{:,\cup_{i=1}^t R_i} Z - A\|_0 &= \sum_{i=1}^t \sum_{j \in G_i} \|A_{:,R_i} z^{(j)} - A_{:,j}\|_0 \\ &\overset{\text{Lem.4.4}}{\leqslant} \sum_{i=1}^t \sum_{j \in G_i} k \cdot \min_{x^{(j)}} \|A_{:,R_i} x^{(j)} - A_{:,j}\|_0 \overset{(4.4)}{\leqslant} \sum_{i=1}^t \sum_{j \in G_i} \frac{100\,(k+1)^2\,\mathrm{OPT}_k}{|B_{i-1}|} \\ &= 100\,(k+1)^2\,\mathrm{OPT}_k \cdot \sum_{i=1}^t \frac{|G_i|}{|B_{i-1}|} \overset{(4.5)}{\leqslant} O\left(k^2 \cdot \log \frac{n}{2k}\right) \cdot \mathrm{OPT}_k. \end{aligned}$$

By Lemma 4.4, the expected number of iterations of selecting a set R_i such that $|G_i| \geqslant 1/10|B_{i-1}|$ is $O(1)$. Since the number of recursive calls t is bounded by $O(\log(n/k))$, it follows by a Markov bound that Algorithm 6 chooses $O(k \log(n/k))$ columns in total. Since the approximation algorithm of Theorem 4.3 runs in polynomial time, our entire algorithm has expected polynomial time. □

4.2 Approximation Algorithm for Reals ℓ_0-Rank-1

Given a matrix $A \in \mathbb{R}^{m\times n}$ with $m \geqslant n$, and an error $\varepsilon \in (0, 1/2)$, our goal is to find an approximate solution $\widetilde{u} \in \mathbb{R}^m$, $\widetilde{v} \in \mathbb{R}^n$ of the Reals ℓ_0-Rank-1 problem such that $\|A - \widetilde{u} \cdot \widetilde{v}^T\|_0 \leqslant (2 + \varepsilon)\mathrm{OPT}_1$, where the optimal value is defined by

$$\mathrm{OPT}_1 \overset{\mathrm{def}}{=} \min_{u\in\mathbb{R}^m,\ v\in\mathbb{R}^n} \|A - u \cdot v^T\|_0. \tag{4.6}$$

In the trivial case when $\mathrm{OPT}_1 = 0$, there is an optimal algorithm that runs in time $O(\|A\|_0)$ and finds the exact rank-1 decomposition uv^T of a matrix A. Here, we focus on the case when $\mathrm{OPT}_1 \geqslant 1$. We will show that Algorithm 7 yields a $(2 + \varepsilon)$-approximation and runs in nearly linear time in $\|A\|_0$, for any constant $\varepsilon > 0$. Further, a variant of our algorithm even runs in sublinear time, if $\|A\|_0$ is large and

$$\psi \overset{\mathrm{def}}{=} \mathrm{OPT}_1/\|A\|_0 \tag{4.7}$$

is not too small. In particular, we obtain sublinear time $o(\|A\|_0)$ when $\mathrm{OPT}_1 \geqslant (\varepsilon^{-1}\log(mn))^4$ and $\|A\|_0 \geqslant n(\varepsilon^{-1}\log(mn))^4$.

Theorem 1.12 (from page 9)**.** *There is an algorithm that, given $A \in \mathbb{R}^{m\times n}$ with column adjacency arrays and $\mathrm{OPT}_1 \geqslant 1$, and given $\varepsilon \in (0, 0.1]$, runs w.h.p. in time*

$$O\left(\left(\frac{n\log m}{\varepsilon^2} + \min\left\{\|A\|_0,\ n + \psi^{-1}\frac{\log n}{\varepsilon^2}\right\}\right)\frac{\log^2 n}{\varepsilon^2}\right)$$

and outputs a column $A_{:,j}$ and a vector $z \in \mathbb{R}^n$ such that w.h.p. $\|A - A_{:,j} \cdot z^T\|_0 \leqslant (2 + \varepsilon)\mathrm{OPT}_1$. The algorithm also computes an estimate Y satisfying w.h.p. $(1 - \varepsilon)\mathrm{OPT}_1 \leqslant Y \leqslant (2 + 2\varepsilon)\mathrm{OPT}_1$.

In fact, our analysis of Theorem 1.12 directly applies to the Binary ℓ_0-Rank-1 problem, and yields as a special case the following result (which is used to prove Theorem 1.6 in Section 1.2).

Theorem 1.5 (from page 7)**.** *Let $\mathrm{OPT} = \min_{u\in\{0,1\}^m,\, v\in\{0,1\}^n} \|A - u \cdot v^T\|_0$. Given a binary matrix $A \in \{0, 1\}^{m\times n}$ with column adjacency arrays and $\mathrm{OPT} \geqslant 1$, and given $\varepsilon \in (0, 0.1]$, we can compute w.h.p. in time*

$$O\left(\left(\frac{n\log m}{\varepsilon^2} + \min\left\{\|A\|_0,\ n + \psi^{-1}\frac{\log n}{\varepsilon^2}\right\}\right)\frac{\log^2 n}{\varepsilon^2}\right)$$

a column $A_{:,j}$ and a binary vector $z \in \{0, 1\}^n$ such that w.h.p. $\|A - A_{:,j} \cdot z^T\|_0 \leqslant (2 + \varepsilon)\mathrm{OPT}$. Further, we can compute an estimate Y such that w.h.p. $(1 - \varepsilon)\mathrm{OPT} \leqslant Y \leqslant (2 + 2\varepsilon)\mathrm{OPT}$.

The rest of this section is devoted to proving Theorem 1.12. We start by presenting the pseudocode of Algorithm 7.

Algorithm 7 Reals ℓ_0-Rank-1: Approximation Scheme

```
Input: $A \in \mathbb{R}^{m\times n}$ and $\varepsilon \in (0, 0.1)$.

1. Partition the columns of $A$ into weight-classes $\mathcal{S} = \{S^{(0)}, \dots, S^{(1+\log n)}\}$ such that
    i) $S^{(0)}$ contains all columns $j$ with $\|A_{:,j}\|_0 = 0$, and
    ii) $S^{(i)}$ contains all columns $j$ with $2^{i-1} \leqslant \|A_{:,j}\|_0 < 2^i$.
2. For each weight-class $S^{(i)}$ do:
    2.1 Sample a set $C^{(i)}$ of $\Theta(\varepsilon^{-2}\log n)$ elements uniformly at random from $S^{(i)}$.
    2.2 Find a $(1 + \frac{\varepsilon}{15})$-approximate solution $z^{(j)} \in \mathbb{R}^n$ for each column $A_{:,j} \in C^{(i)}$, i.e.

        $$\left\|A - A_{:,j} \cdot [z^{(j)}]^T\right\|_0 \leqslant \left(1 + \frac{\varepsilon}{15}\right)\min_v \left\|A - A_{:,j} \cdot v^T\right\|_0. \tag{4.8}$$

3. Compute a $(1 + \frac{\varepsilon}{15})$-approximation $Y_j$ of $\|A - A_{:,j} \cdot [z^{(j)}]^T\|_0$ for every $j \in \bigcup_{i\in[|\mathcal{S}|]} C^{(i)}$.
4. Return the pair $(A_{:,j}, z^{(j)})$ corresponding to the minimal value $Y_j$.
```

The only steps for which the implementation details are not immediate are Steps 2.2 and 3. We will discuss them in Sections 4.2.2 and 4.2.3, respectively.

Note that the algorithm from Theorem 1.12 selects a column $A_{:,j}$ and then finds a good vector z such that the product $A_{:,j} \cdot z^T$ approximates A. We show that the $(2 + \varepsilon)$-approximation guarantee is essentially tight for algorithms following this pattern.

Lemma 4.5. *There exist a matrix $A \in \mathbb{R}^{n\times n}$ such that $\min_z \|A - A_{:,j} \cdot z^T\|_0 \geqslant 2(1-1/n)\mathrm{OPT}_1$, for every column $A_{:,j}$.*

Proof of Lemma 4.5. Let $A = I + J \in \mathbb{R}^{n\times n}$, where I is an identity matrix and $J = \mathbf{1}\mathbf{1}^T$ is an all-ones matrix. Note that $\mathrm{OPT}_1 \leqslant n$ is achieved by approximating A with the rank-1 matrix J. On the other hand, when we choose $u = A_{:,i}$ for any $i \in [n]$, the incurred cost on any column $A_{:,j}$, $j \neq i$, is $\min_x \|A_{:,j} - xA_{:,i}\|_0 = 2$, since there are two entries where $A_{:,i}$ and $A_{:,j}$ disagree. Hence, the total cost is at least $2n - 2 \geqslant (2 - 2/n)\mathrm{OPT}_1$. □

4.2.1 Correctness

We first prove the following structural result, capturing Steps 1-2.2 of Algorithm 7.

Lemma 4.6. *Let $C^{(0)}, \ldots, C^{(\log n+1)}$ be the sets constructed in Step 2.1 of Algorithm 7, and let $C = C^{(0)} \cup \ldots \cup C^{(\log n+1)}$. Then w.h.p. C contains an index j such that*

$$\min_z \|A - A_{:,j} \cdot z^T\|_0 \leqslant (2 + \varepsilon/2)\mathrm{OPT}_1.$$

Proof. Let u, v be an optimum solution of (4.6). For the weight class $S^{(0)}$ containing all columns without nonzero entries, setting $z_c = 0$ for any $c \in S^{(0)}$ gives zero cost on these columns, no matter what column $A_{:,j}$ we picked. Hence, without loss of generality in the following we assume that $S^{(0)} = \emptyset$.

For any $i \geqslant 1$, we partition the weight class $S^{(i)}$ into $N^{(i)}, Z^{(i)}$ such that $v_i = 0$ for every $i \in Z^{(i)}$ and $v_i \neq 0$ for $i \in N^{(i)}$. We denote by $\mathcal{S}^+$ the set of weight-classes $S^{(i)}$ with $|N^{(i)}| \geqslant \frac{1}{3}|S^{(i)}|$. Let $\mathcal{R} = \bigcup_{i\in\mathcal{S}^+} S^{(i)}$ and $\mathcal{W} = [n]\backslash\mathcal{R}$. We partition $\mathcal{R} = \mathcal{N} \cup \mathcal{Z}$ such that $v_i = 0$ for every $i \in \mathcal{Z}$ and $v_i \neq 0$ for $i \in \mathcal{N}$. Further, using the three sets $\mathcal{N}, \mathcal{Z}$ and $\mathcal{W}$ we decompose OPT_1 into

$$\begin{aligned} \mathrm{OPT}_1 &= \mathrm{OPT}_{\mathcal{N}} + \mathrm{OPT}_{\mathcal{Z}} + \mathrm{OPT}_{\mathcal{W}} \\ &= \|A_{:,\mathcal{N}} - u \cdot v_{\mathcal{N}}^T\|_0 + \|A_{:,\mathcal{Z}}\|_0 + \|A_{:,\mathcal{W}} - u \cdot v_{\mathcal{W}}^T\|_0. \end{aligned}$$

The proof proceeds by case distinction:

The set $\mathcal{Z}$: For any column $A_{:,j}$ of A, we have

$$\min_{z_{\mathcal{Z}}} \|A_{:,\mathcal{Z}} - A_{:,j} \cdot z_{\mathcal{Z}}^T\|_0 \leqslant \|A_{:,\mathcal{Z}} - A_{:,j} \cdot 0\|_0 = \|A_{:,\mathcal{Z}}\|_0 = \mathrm{OPT}_{\mathcal{Z}}. \tag{4.9}$$

The set $\mathcal{W}$: Note that $\mathcal{W}$ consists of all weight classes $S^{(i)}$ with $|Z^{(i)}| > \frac{2}{3}|S^{(i)}|$. For any such weight class $S^{(i)}$, the optimum cost satisfies

$$\|A_{:,S^{(i)}} - uv_{S^{(i)}}^T\|_0 \geqslant \|A_{:,Z^{(i)}}\|_0 \geqslant \frac{2}{3}|S^{(i)}|2^{i-1} = \frac{1}{3}|S^{(i)}|2^i.$$

Further, for any column $A_{:,j}$ of A, we have

$$\begin{aligned} \min_z \|A_{:,S^{(i)}} - A_{:,j}z^T\|_0 &\leqslant \|A_{:,S^{(i)}}\|_0 \leqslant \|A_{:,Z^{(i)}}\|_0 + \frac{1}{3}|S^{(i)}|2^i \\ &\leqslant 2\|A_{:,Z^{(i)}}\|_0 \leqslant 2\|A_{:,S^{(i)}} - uv_{S^{(i)}}^T\|_0, \end{aligned}$$

and thus the total cost in $\mathcal{W}$ is bounded by

$$\min_{z_{\mathcal{W}}} \|A_{:,\mathcal{W}} - A_{:,j} \cdot z_{\mathcal{W}}^T\|_0 = \sum_{i\in\mathcal{W}} \min_{z_{S^{(i)}}} \|A_{:,S^{(i)}} - A_{:,j} \cdot z_{S^{(i)}}^T\|_0 \leqslant 2\|A_{:,\mathcal{W}} - uv_{\mathcal{W}}^T\|_0 = 2\,\mathrm{OPT}_{\mathcal{W}}. \tag{4.10}$$

The set $\mathcal{N}$: By an averaging argument there is a subset $G \subseteq \mathcal{N}$ of size $|G| \geqslant \frac{\varepsilon}{3}|\mathcal{N}|$ such that for every $j \in G$ we have

$$\|A_{:,j} - v_j \cdot u\|_0 \leqslant \frac{1}{1-\varepsilon/3} \cdot \frac{\mathrm{OPT}_{\mathcal{N}}}{|\mathcal{N}|} \leqslant \left(1 + \frac{\varepsilon}{2}\right) \cdot \frac{\mathrm{OPT}_{\mathcal{N}}}{|\mathcal{N}|}.$$

Let $j \in G$ be arbitrary. Furthermore, let $P^{(j)}$ be the set of all rows i with $A_{i,j} = v_j \cdot u_i$, and let $Q^{(j)} = [m] \setminus P^{(j)}$. By construction, we have $|Q^{(j)}| \leqslant (1 + \frac{\varepsilon}{2})\mathrm{OPT}_{\mathcal{N}}/|\mathcal{N}|$. Moreover, since $j \in \mathcal{N}$ we have

$v_j \neq 0$, and thus we may choose $z_{\mathcal{N}} = \frac{1}{v_j} v_{\mathcal{N}}$. This yields

$$\begin{aligned} \min_{z_{\mathcal{N}}} \|A_{:,\mathcal{N}} - A_{:,j} \cdot z_{\mathcal{N}}^T\|_0 &\leqslant \|A_{:,\mathcal{N}} - A_{:,j} \cdot \frac{1}{v_j} v_{\mathcal{N}}^T\|_0 \\ &= \|A_{P^{(j)},\mathcal{N}} - u_{P^{(j)}} \cdot v_{\mathcal{N}}^T\|_0 + \|A_{Q^{(j)},\mathcal{N}} - A_{:,j} \cdot \frac{1}{v_j} v_{\mathcal{N}}^T\|_0 \\ &\leqslant \mathrm{OPT}_{\mathcal{N}} + \left(1 + \frac{\varepsilon}{3}\right) \mathrm{OPT}_{\mathcal{N}} = \left(2 + \frac{\varepsilon}{3}\right) \mathrm{OPT}_{\mathcal{N}}. \end{aligned} \tag{4.11}$$

Hence, by combining (4.9), (4.10) and (4.11), it follows for any index $j \in G$ that

$$\begin{aligned} &\min_z \|A - A_{:,j} \cdot z^T\|_0 \\ = &\min_{z_{\mathcal{W}}} \|A_{:,\mathcal{W}} - A_{:,j} \cdot z_{\mathcal{W}}^T\|_0 + \min_{z_{\mathcal{Z}}} \|A_{:,\mathcal{Z}} - A_{:,j} \cdot z_{\mathcal{Z}}^T\|_0 + \min_{z_{\mathcal{N}}} \|A_{:,\mathcal{N}} - A_{:,j} \cdot z_{\mathcal{N}}^T\|_0 \\ \leqslant &\, 2\,\mathrm{OPT}_{\mathcal{W}} + \mathrm{OPT}_{\mathcal{Z}} + \left(2 + \frac{\varepsilon}{2}\right) \mathrm{OPT}_{\mathcal{N}} \leqslant \left(2 + \frac{\varepsilon}{2}\right) \mathrm{OPT}_1. \end{aligned}$$

This yields the desired approximation guarantee, provided that we sampled a column $A_{:,j}$ from G. We show next that whenever $\mathcal{N} \neq \emptyset$, our algorithm samples with high probability at least one column from G.

Before that let us consider the case when $\mathcal{N} = \emptyset$. Then, since the bounds (4.9) and (4.10) hold for any column $A_{:,j}$ of A, the set C contains only good columns. Thus, we may assume that $\mathcal{N} \neq \emptyset$.

We now analyze the probability of sampling a column $A_{:,j}$ from G. By construction, the set $\mathcal{N}$ is the union of all $N^{(i)}$ such that $|N^{(i)}| \geqslant \frac{1}{3}|S^{(i)}|$. As shown above, we have $|G| \geqslant \frac{\varepsilon}{3}|\mathcal{N}|$, and thus there is an index i satisfying $|G \cap S^{(i)}| \geqslant \frac{\varepsilon}{3}|N^{(i)}| \geqslant \frac{\varepsilon}{9}|S^{(i)}|$. Hence, when sampling a uniformly random element from $S^{(i)}$ we hit G with probability at least $\frac{\varepsilon}{9}$. Since we sample $\Theta(\varepsilon^{-2} \log n)$ elements from $S^{(i)}$, we hit G with high probability. This finishes the proof. □

Correctness Proof of Algorithm 7: It remains to show that the pair $(A_{:,j}, z^j)$ with minimum estimate Y_j yields a $(2+\varepsilon)$-approximation to OPT_1. By Step 3, for every column j we have

$$\left(1 + \frac{\varepsilon}{15}\right)^{-1} \cdot \|A - A_{:,j}[z^{(j)}]^T\|_0 \leqslant Y_j \leqslant \left(1 + \frac{\varepsilon}{15}\right) \cdot \|A - A_{:,j}[z^{(j)}]^T\|_0. \tag{4.12}$$

Since $Y_j \leqslant Y_{j'}$ for any other column j', (4.12) and the approximation guarantee of Steps 2.2 yield

$$\left(1 + \frac{\varepsilon}{15}\right)^{-1} \|A - A_{:,j}[z^{(j)}]^T\|_0 \leqslant \left(1 + \frac{\varepsilon}{15}\right) \|A - A_{:,j'}[z^{(j')}]^T\|_0 \leqslant \left(1 + \frac{\varepsilon}{15}\right)^2 \min_z \left\|A - A_{:,j'} z^T\right\|_0.$$

By Lemma 4.6, w.h.p. there exists a column $j' \in C$ with $\min_z \|A - A_{:,j'} z^T\|_0 \leqslant (2 + \frac{\varepsilon}{2})\mathrm{OPT}_1$. We obtain a total approximation ratio of $(1 + \frac{\varepsilon}{15})^3 (2 + \frac{\varepsilon}{2}) \leqslant 2 + \varepsilon$ for any error $0 < \varepsilon \leqslant 0.1$, i.e. we have $\|A - A_{:,j}[z^{(j)}]^T\|_0 \leqslant (2 + \varepsilon)\mathrm{OPT}_1$. Therefore, it holds that

$$(1 - \varepsilon)\mathrm{OPT}_1 \leqslant \left(1 + \frac{\varepsilon}{15}\right)^{-1} \mathrm{OPT}_1 \leqslant Y_j \leqslant \left(1 + \frac{\varepsilon}{15}\right) (2 + \varepsilon)\mathrm{OPT}_1 \leqslant (2 + 2\varepsilon)\mathrm{OPT}_1.$$

This finishes the correctness proof.

4.2.2 Implementing Step 2.2

Step 2.2 of Algorithm 7 uses the following sublinear procedure.

Algorithm 8 Reals ℓ_0-Rank-1: Objective Value Estimation

```
Input: $A \in \mathbb{R}^{m \times n}$, $u \in \mathbb{R}^m$ and $\varepsilon \in (0,1)$.
    let $t \overset{\text{def}}{=} \Theta(\varepsilon^{-2} \log m)$, $N \overset{\text{def}}{=} \text{supp}(u)$, and $p \overset{\text{def}}{=} t/|N|$.
1. Select each index $i \in N$ with probability $p$ and let $S$ be the resulting set.
2. Compute a vector $z \in \mathbb{R}^n$ such that $z_j = \arg\min_{r \in \mathbb{R}} \|A_{S,j} - r \cdot u_S\|_0$ for all $j \in [n]$.
3. Return the vector $z$.
```

We prove now the correctness of Algorithm 8 and we analyze its runtime.

Lemma 4.7. *Given $A \in \mathbb{R}^{m\times n}$, $u \in \mathbb{R}^m$ and $\varepsilon \in (0,1)$ we can compute in time $O\left(\varepsilon^{-2} n \log m\right)$ a vector $z \in \mathbb{R}^n$ such that w.h.p. for every $i \in [n]$ it holds that*

$$\|A_{:,i} - z_i u\|_0 \leqslant (1+\varepsilon) \min_{v_i \in \mathbb{R}} \|A_{:,i} - v_i u\|_0.$$

Proof. Let N, Z be a partitioning of $[m]$ such that $u_i = 0$ for $i \in Z$ and $u_i \neq 0$ for $i \in N$. Since $\|A - u \cdot z^T\|_0 = \|A_{N,:} - u_N \cdot z^T\|_0 + \|A_{Z,:}\|_0$, it suffices to find a vector z such that for every $j \in [n]$ we have

$$\|A_{N,j} - z_j \cdot u_N\|_0 \leqslant (1+\varepsilon) \cdot \min_{v_j} \|A_{N,j} - v_j \cdot u_N\|_0 . \tag{4.13}$$

Let $j \in [n]$ be arbitrary. For $r \in \mathbb{R}$ let $G(r) \overset{\text{def}}{=} \{i \in N : A_{i,j}/u_i = r\}$ be the set of entries of $A_{N,j}$ that we correctly recover by setting $z_j = r$. Note that $\|A_{N,j} - z_j \cdot u_N\|_0 = |N| - |G(z_j)|$ holds for any $z_j \in \mathbb{R}$. Hence, the optimal solution sets $z_j = r^\star \overset{\text{def}}{=} \arg\max_{r\in\mathbb{R}} |G(r)|$.

Let $X_{G(r)}$ be a random variable indicating the number of elements selected from group $G(r)$ in Step 1 of Algorithm 8. Notice that $\mathbb{E}[X_{G(r)}] = t \cdot |G(r)|/|N|$, and by Chernoff bound w.h.p. we have

$$|X_{G(r)} - \mathbb{E}[X_{G(r)}]| \leqslant (\varepsilon/8) \cdot t. \tag{4.14}$$

Let $S \subseteq N$ be the set of selected indices. Further, since $|S| = \sum_\ell X_{G(r)}$ and $\mathbb{E}[|S|] = t$, by Chernoff bound we have w.h.p. $|S| \leqslant (1+\varepsilon)|t|$. Observe that Step 2 of Algorithm 8 selects $z_j = \arg\max_{r\in\mathbb{R}} X_{G(r)}$, since $\|A_{S,j} - r \cdot u_S\|_0 = |S| - X_{G(r)}$. We now relate z_j to $r^\star$. The proof proceeds by case distinction on $\delta^\star \overset{\text{def}}{=} |G(r^\star)|/|N|$.

Case 1: Suppose $\delta^\star \leqslant \varepsilon/4$. Then $\|A_{N,j} - r \cdot u_N\|_0 \geqslant (1-\varepsilon/4)|N|$ for every $r \in \mathbb{R}$, and thus no matter which z_j is selected we obtain a $(1+\varepsilon)$-approximation, since

$$\|A_{N,j} - z_j \cdot u_N\|_0 \leqslant |N| \leqslant (1+\varepsilon) \min_r \|A_{N,j} - r \cdot u_N\|_0.$$

Case 2: Suppose $\delta^\star \geqslant 1/2 + \varepsilon$. Then, by (4.14) w.h.p. we have

$$X_{G(r^\star)} \geqslant \mathbb{E}[X_{G(r^\star)}] - (\varepsilon/8)t = (\delta^\star - \varepsilon/8)t > (1+\varepsilon)t/2 \geqslant |S|/2,$$

and thus $X_{r^\star}$ is maximal among all X_r. Hence, we select the optimal $z_j = r^\star$.

Case 3: Suppose $\varepsilon/4 < \delta^\star < 1/2 + \varepsilon$. Let $z_j = r$ be the value chosen by Algorithm 8. By (4.14), the event of making a mistake, given by $X_{G(r)} \geqslant X_{G(r^\star)}$, happens when

$$\mathbb{E}[X_{G(r)}] + (\varepsilon/8)t \geqslant \mathbb{E}[X_{G(r^\star)}] - (\varepsilon/8)t. \tag{4.15}$$

Let $\delta \overset{\text{def}}{=} |G(r)|/|N|$ and note that (4.15) implies $\delta \geqslant \delta^\star - \varepsilon/4$. Hence, for the selected $r \neq r^\star$ we have

$$\begin{aligned}\|A_{N,j} - r \cdot u_N\|_0 &= (1-\delta)|N| \leqslant (1-\delta^\star+\varepsilon/4)|N| \\ &\leqslant (1+\varepsilon)(1-\delta^\star)|N| = (1+\varepsilon)\|A_{N,j} - r^\star \cdot u_N\|_0.\end{aligned}$$

Therefore, in each of the preceding three cases, we obtain w.h.p. a $(1+\varepsilon)$-approximate solution. The statement follows by the union bound. □

4.2.3 Implementing Step 3

In Step 3 of Algorithm 7 we want to compute a $(1+\frac{\varepsilon}{15})$-approximation Y_j of $\|A - A_{:,j} \cdot [z^{(j)}]^T\|_0$ for every $j \in \bigcup_{i\in[|\mathcal{S}|]} C^{(i)}$. We present two solutions, an exact algorithm (see Lemma 4.8) and a sublinear time sampling-based algorithm (see Lemma 4.10).

Lemma 4.8. *Suppose $A, B \in \mathbb{R}^{m\times n}$ are represented by column adjacency arrays. Then, we can compute in time $O(\|A\|_0 + n)$ the measure $\|A - B\|_0$.*

Proof. We partition the entries of A into five sets:

$$\begin{aligned}
T_1 &= \{(i,j) : A_{ij} = 0 \text{ and } B_{ij} \neq 0\}\,, & T_4 &= \{(i,j) : 0 \neq A_{ij} = B_{ij} \neq 0\}\,,\\
T_2 &= \{(i,j) : A_{ij} \neq 0 \text{ and } B_{ij} = 0\}\,, & T_5 &= \{(i,j) : A_{ij} = B_{ij} = 0\}\,,\\
T_3 &= \{(i,j) : 0 \neq A_{ij} \neq B_{ij} \neq 0\}\,. & &
\end{aligned}$$

Observe that $\|A-B\|_0 = |T_1| + |T_2| + |T_3|$ and $\|B\|_0 = |T_1| + |T_3| + |T_4|$. Since $\|A-B\|_0 = \|B\|_0 + |T_2| - |T_4|$, it suffices to compute the numbers $\|B\|_0$, $|T_2|$ and $|T_4|$. We compute $|T_2|$ and $|T_4|$ in $O(\|A\|_0)$ time, by enumerating all non-zero entries of A and performing $O(1)$ checks for each. For $\|B\|_0$, we sum the column lengths of B in time $O(n)$. □

For our second, sampling-based implementation of Step 3, we make use of an algorithm by Dagum et al. [DKLR00] for estimating the expected value of a random variable. We note that the runtime of their algorithm is a random variable, the magnitude of which is bounded w.h.p. within a certain range.

Theorem 4.9. *[DKLR00] Let X be a random variable taking values in $[0,1]$ with $\mu \overset{\text{def}}{=} \mathbb{E}[X] > 0$. Let $0 < \varepsilon, \delta < 1$ and $\rho_X = \max\{\text{Var}[X], \varepsilon\mu\}$. There is an algorithm with sample access to X that computes an estimator $\widetilde{\mu}$ in time t such that for a universal constant c we have*

i) $\Pr[(1-\varepsilon)\mu \leqslant \widetilde{\mu} \leqslant (1+\varepsilon)\mu] \geqslant 1-\delta$ *and* *ii)* $\Pr[t \geqslant c \cdot \varepsilon^{-2}\mu^{-2}\rho_X \log(1/\delta)] \leqslant \delta$.

We state now our key technical insight, on which we build upon our sublinear algorithm.

Lemma 4.10. *There is an algorithm that, given $A, B \in \mathbb{R}^{m\times n}$ with column adjacency arrays and $\|A-B\|_0 \geqslant 1$, and given $\varepsilon > 0$, computes an estimator Z that satisfies w.h.p.*

$$(1-\varepsilon)\|A-B\|_0 \leqslant Z \leqslant (1+\varepsilon)\|A-B\|_0.$$

The algorithm runs w.h.p. in time $O(n + \varepsilon^{-2}\frac{\|A\|_0+\|B\|_0}{\|A-B\|_0}\log n\})$.

Proof. By Lemma 1.9, after $O(n)$ preprocessing time we can sample a uniformly random non-zero entry from A or B in time $O(1)$.

We consider the following random process:

1. Sample $C \in \{A, B\}$ such that $\Pr[C = A] = \frac{\|A\|_0}{\|A\|_0+\|B\|_0}$ and $\Pr[C = B] = \frac{\|B\|_0}{\|A\|_0+\|B\|_0}$.
2. Sample (i,j) uniformly at random from the non-zero entries of C
3. Return:

$$X = \begin{cases} 0, & \text{if } A_{ij} = B_{ij}; \\ 1/2, & \text{if } 0 \neq A_{ij} \neq B_{ij} \neq 0; \\ 1, & \text{if } A_{ij} \neq B_{ij} \text{ and either } A_{ij} \text{ or } B_{ij} \text{ equals } 0. \end{cases}$$

Observe that

$$\begin{aligned} \mathbb{E}[X] &= \sum_{(i,j)\colon\, A_{ij}\neq B_{ij}=0} \frac{\|A\|_0}{\|A\|_0+\|B\|_0}\cdot\frac{1}{\|A\|_0} + \sum_{(i,j)\colon\, 0=A_{ij}\neq B_{ij}} \frac{\|B\|_0}{\|A\|_0+\|B\|_0}\cdot\frac{1}{\|B\|_0} \\ &\quad + \sum_{(i,j)\colon\, 0\neq A_{ij}\neq B_{ij}\neq 0} \left(\frac{1}{2}\cdot\frac{\|A\|_0}{\|A\|_0+\|B\|_0}\cdot\frac{1}{\|A\|_0} + \frac{1}{2}\cdot\frac{\|B\|_0}{\|A\|_0+\|B\|_0}\cdot\frac{1}{\|B\|_0}\right) \\ &= \frac{\|A-B\|_0}{\|A\|_0+\|B\|_0}. \end{aligned}$$

Straightforward checking shows that $X \in [0,1]$ implies $\text{Var}[X] \leqslant \mathbb{E}[X]$, and thus

$$\rho_X = \max\{\text{Var}[X], \varepsilon\cdot\mathbb{E}[X]\} \leqslant \mathbb{E}[X].$$

Theorem 4.9 applied with $\delta = 1/\text{poly}(n)$, yields w.h.p. in $O(\varepsilon^{-2}\mathbb{E}[X]^{-1}\log n) = O(\varepsilon^{-2}\frac{\|A\|_0+\|B\|_0}{\|A-B\|_0}\log n)$ time an estimator $(1-\varepsilon)\mathbb{E}[X] \leqslant \widetilde{\mu} \leqslant (1+\varepsilon)\mathbb{E}[X]$. Then, w.h.p. the estimator $Z \overset{\text{def}}{=} (\|A\|_0 + \|B\|_0)\widetilde{\mu}$ satisfies the statement. □

We present now our main result in this section.

Theorem 4.11. *There is an algorithm that, given $A \in \mathbb{R}^{m\times n}$ with column adjacency arrays and $\text{OPT}_1 \geqslant 1$, and given $j \in [n]$, $v \in \mathbb{R}^m$ and $\varepsilon \in (0,1)$, outputs an estimator Y that satisfies w.h.p.*

$$(1-\varepsilon)\|A - A_{:,j}\cdot v^T\|_0 \leqslant Y \leqslant (1+\varepsilon)\|A - A_{:,j}\cdot v^T\|_0.$$

The algorithm runs w.h.p. in time $O(\min\{\|A\|_0, n + \varepsilon^{-2}\psi^{-1}\log n\})$, where $\psi = \text{OPT}_1/\|A\|_0$.

Proof. Let $B \overset{\text{def}}{=} A_{:,j}v^T$ and observe that $\|A-B\|_0 \geqslant \mathrm{OPT}_1 \geqslant 1$. Note that we implicitly have column adjacency arrays for B, since for any column c with $v_c = 0$ there are no non-zero entries in $B_{:,c}$, and for any column c with $v_c = 1$ the non-zero entries of $B_{:,c}$ are the same as for $A_{:,j}$. Hence, Lemma 4.8 and Lemma 4.10 are applicable.

We analyze the running time of Lemma 4.10. Note that if $\|B\|_0 \leqslant (1+\psi)\|A\|_0$ then $\frac{\|A\|_0+\|B\|_0}{\|A-B\|_0} \leqslant (2+\psi)/\psi$, and otherwise, i.e. $(1+\psi)\|A\|_0 \leqslant \|B\|_0$, we have

$$\|A-B\|_0 \geqslant \|B\|_0 - \|A\|_0 \geqslant \tfrac{\psi}{1+\psi}\|B\|_0$$

and thus $\frac{\|A\|_0+\|B\|_0}{\|A-B\|_0} \leqslant 2(1+\psi)/\psi$. Hence, $\frac{\|A\|_0+\|B\|_0}{\|A-B\|_0} < 4/\psi$, which yields w.h.p. time $O(n+\varepsilon^{-2}\psi^{-1}\log n)$.

We execute in parallel the algorithms from Lemma 4.8 and Lemma 4.10. Once the faster algorithm outputs a solution, we terminate the execution of the slower one. Note that this procedure runs w.h.p in time $O(\min\{\|A\|_0, n+\varepsilon^{-2}\psi^{-1}\log n\})$, and returns w.h.p. the desired estimator Y. □

To implement Step 3 of Algorithm 7, we simply apply Theorem 4.11 with A, ε and $v = z^{(j)}$ to each sampled column $j \in \bigcup_{0\leqslant i\leqslant \log n+1} C^{(i)}$.

4.2.4 Analyzing the Runtime of Algorithm 7

Consider Algorithm 7. In Steps 1, 2 and 2.1, from each of the $O(\log n)$ weight classes we sample $O(\varepsilon^{-2}\log n)$ columns. In Step 2.2, for each sampled column we use Lemma 4.7, which takes time $O(\varepsilon^{-2}n\log m)$ per column, or $O(\varepsilon^{-4}n\log m\log^2 n)$ in total. Further, in Step 3, we use Theorem 4.11 for each sampled column, which w.h.p. takes time $O(\min\{\|A\|_0, n+\varepsilon^{-2}\psi^{-1}\log n\})$ per column, or in total

$$O(\min\{\|A\|_0, n+\varepsilon^{-2}\psi^{-1}\log n\}\cdot\varepsilon^{-2}\log^2 n).$$

Then, the total runtime is bounded by

$$O(\varepsilon^{-4}n\log m\log^2 n + \min\{\|A\|_0\varepsilon^{-2}\log^2 n,\ \varepsilon^{-4}\psi^{-1}\log^3 n\}).$$

4.2.5 Proof of Lemma 1.9

Note that we are given the number of nonzero entries $\ell_j = \|A_{:,j}\|_0$ for each column. We want to first sample a column $X \in [n]$ such that $\Pr[X=j] = \ell_j/\sum_{k\in[n]}\ell_k$, then sample $Y \in [\ell_j]$ uniformly, read $B_X[Y] = (i, A_{i,X})$, and return $A_{i,X}$. Observe that this process indeed samples each nonzero entry of A with the same probability, since the probability of sampling a particular nonzero entry (i,j) is $(\ell_j/\sum_{k\in[n]}\ell_k)\cdot(1/\ell_j) = 1/\sum_{k\in[n]}\ell_k$. Sampling $Y \in [\ell_j]$ uniformly can be done in constant time by assumption. For sampling X, we use the classic Alias Method by Walker [Wal74], which is given the probabilities $\Pr[X=1],\dots,\Pr[X=n]$ as an input and computes, in time $O(n)$, a data structure that allows to sample from X in time $O(1)$. This finishes the construction. □

Part II

Approximate Spectral Clustering

Proceedings

This part is based on an article [KM16] that was presented at *ESA'16* and on its journal version which is under submission as of July 2018. The full ArXiv version is available online [KM18].

Acknowledgements

I would like to thank Kurt Mehlhorn for his guidance, support and fruitful collaboration, Michael B. Cohen for a helpful discussion that led me to the work of Boutsidis et al. [BKG15], Emanuele Natale for a helpful discussion and for pointing me to the book [Ste90], and anonymous reviewers of our submission [KM16] for their helpful remarks.

Chapter 5

Introduction

A *cluster* in an undirected graph $G = (V, E)$ is a subset S of nodes whose volume is large compared to the number of outside connections. Formally, the *conductance* of S is defined as

$$\phi(S) \overset{\text{def}}{=} \frac{|E(S, \overline{S})|}{\min\{\mu(S), \mu(\overline{S})\}}, \tag{5.1}$$

where the volume of S is given by $\mu(S) \overset{\text{def}}{=} \sum_{v \in S} \deg(v)$. We are interested in the problem of partitioning the nodes into a given number k of clusters in a way that (approximately) minimizes the *k-way conductance*

$$\widehat{\rho}(k) \overset{\text{def}}{=} \min_{\text{partition } (P_1, \ldots, P_k) \text{ of } V} \; \max_{i \in \{1, \ldots, k\}} \phi(P_i). \tag{5.2}$$

The 2-way partitioning constant is also known as the conductance of the graph and is denoted as

$$\phi_G \overset{\text{def}}{=} \min_{S \subseteq V} \phi(S). \tag{5.3}$$

The k-way partitioning problem arises in many applications, e.g., image segmentation and exploratory data analysis. We refer to the survey [vL07] for additional information. Further, the surveys [SM00, KVV04, vL07] discuss properties of graphs with small or large conductance.

Hardness and Approximation

The k-way partitioning problem is known to be NP-hard, even for $k = 2$ [MS90]. In the case when $k = 2$, the k-way partitioning problem reduces to the graph conductance problem (5.3), for which there is an approximation algorithm [Chu97] that computes a bipartition $(S, \overline{S})$ such that $\phi(S) \leqslant \sqrt{2\phi_G}$. The algorithm computes an eigenvector corresponding to the second smallest eigenvalue of a normalized Laplacian matrix, sorts the eigenvector's entries, and performs a sweep over the sorted vector. It is guaranteed that one of the resulting sweep sets satisfies the approximation bound.

The fact that the second eigenvector encodes sufficient information for computing an approximate bipartition with small conductance, motivate researchers to consider the bottom k eigenvectors in order to approximately solve the k-way partitioning problem. The resulting approach is called Spectral Clustering.

Spectral Clustering

Given an undirected graph $G = (V, E)$ and a number of clusters k, the Spectral Clustering algorithm consists of the following two steps:

(i) Compute the bottom k eigenvectors of the normalized Laplacian matrix of G, and store them as the columns of a matrix $Y \in \mathbb{R}^{n \times k}$. The i-th node of G is associated with the i-th row of Y, i.e. a vector in $\mathbb{R}^k$. This step is known as Spectral Embedding (SE).

(ii) Partition the resulting vector set into k clusters using a k-means clustering algorithm.

This meta algorithm has been successfully applied in practice for solving challenging clustering problems, in the fields of: image segmentation, pattern recognition, data mining, community detection and VLSI design [AY95, SM00, NJW01, MBLS01, BN01, LZ04, ZP04, WS05, vL07, WD12, Tas12, CKC+16].

Approximate Spectral Clustering

Exact computation of Spectral Clustering is expensive due to the following two bottlenecks:

(i) the best algorithm for computing a SE exactly requires time $\Omega(n^\omega)$, cf. [Woo14];

(ii) the k-means clustering problem is NP-hard [MNV12].

It is therefore necessary to relax the preceding two problems and to focus on designing approximation schemes for them. Several approximation techniques were developed for Spectral Clustering [Pre81, ST14, YHJ09, CCDL14, FBCM04, PP04, BHH+06, WLRB09, Nys30, WD12, Tas12, LC10, Woo14].

The Power method [LC10, Woo14] is perhaps the most popular technique for computing an approximate SE, due to its simplicity and ease of implementation. Furthermore, this technique was successfully applied for low-rank matrix approximation [Woo14], and it has a worst case convergence guarantee in terms of a principal angle between the space spanned by the approximate and the true eigenvectors [GVL96, Theorem 8.2.4].

Although, the k-means clustering problem is NP-hard [MNV12], it admits a polynomial time approximation scheme (PTAS) [KSS04, HK05, FMS07, ORSS13]. However, the best PTAS for computing a $(1+\varepsilon)$ approximation incurs a factor $2^{\text{poly}(k/\varepsilon)}$ in the runtime.

On the other hand, it is folklore that the approximate variant of Spectral Clustering which computes an approximate SE via the Power method, and applies to it an approximate k-means clustering algorithm, recovers a good approximation of an optimal k-way node partition of G and at the same time runs efficiently (in nearly-linear time).

It is an important task for theory to explain the practical success of Approximate Spectral Clustering, and in particular to resolve the following three questions. In order to state them, we need some notation. Let Y be a SE computed exactly, and $\widetilde{Y}$ be an approximate SE computed via the Power method. Further, let X $(\widetilde{X})$ be an optimal k-means clustering partition of the rows of Y $(\widetilde{Y})$. Let $\widetilde{X_\alpha}$ be a k-way row partition of $\widetilde{Y}$, computed by an α-approximate k-means clustering algorithm. The following questions arise:

Q1. Show that $\widetilde{X_\alpha}$ is a good approximation of X.

Q2. Show that the k-way node partition of G induced by $\widetilde{X_\alpha}$, yields a good approximation of an optimal k-way node partition of G.

Q3. Show that Approximate Spectral Clustering runs efficiently (in nearly-linear time).

Eigenvalue Gaps and k-Way Partitions

Let $0 = \lambda_1 \leqslant \ldots \leqslant \lambda_n \leqslant 2$ be the eigenvalues of a normalized Laplacian matrix of G. It was observed experimentally [vL07, For10] that a large gap between λ_{k+1} and λ_k guarantees a good k-way node partition of G and this was formally proven in [LGT12, GT14]. Lee, Gharan and Trevisan [LGT12] studied the k-way expansion constant defined as

$$\rho(k) \overset{\text{def}}{=} \min_{\text{disjoint } S_1,\ldots,S_k} \max_{i\in\{1,\ldots,k\}} \phi(S_i), \tag{5.4}$$

and related it to λ_k via higher-order Cheeger inequalities

$$\lambda_k/2 \leqslant \rho(k) \leqslant O(k^2)\sqrt{\lambda_k}. \tag{5.5}$$

For related works on higher-order Cheeger inequalities, we refer the reader to [LRTV12, KLL17]. Gharan and Trevisan [GT14] showed that the k-way conductance is at most a factor k away from the k-way expansion constant, i.e.,

$$\rho(k) \leqslant \widehat{\rho}(k) \leqslant k \cdot \rho(k). \tag{5.6}$$

In particular, (5.5) and (5.6) together yield that $\lambda_{k+1} \gg O(k^3)\sqrt{\lambda_k}$ implies $\widehat{\rho}(k+1) \gg \widehat{\rho}(k)$. Thus, there is a k-way node partition $(P_1,\ldots,P_k)$ of G such that $\phi(P_i) \leqslant O(k^3)\sqrt{\lambda_k}$ for all i, and simultaneously the best $(k+1)$-way partition is significantly worse.

Prior Work

Ng et al. [NJW01] reported that ASC performs very well on challenging clustering instances, and initiated the study for finding a formal explanation for the practical success of ASC. Using tools from matrix perturbation theory, they derived sufficient conditions under which the vectors of a SE form tight clusters. However, their analysis does not apply to approximate SEs, and does not give guarantees for the induced k-way node partition of G.

Peng et al. [PSZ17] showed that for all instances satisfying the eigenvalue gap assumption $\lambda_{k+1}/\widehat{\rho}(k) \geqslant \Omega(k^3)$, any $O(1)$-approximate k-means partition of a normalized SE Y' induces a good approximation of an optimal k-way node partition of G. Notably, their analysis yields the first approximation guarantees in terms of the k-way conductance . However, their analysis does not apply to approximate SE, and also computing an $O(1)$-approximation k-means partition using any known PTAS [HK05, FMS07, ORSS13] incurs an exponential factor of $2^{\Omega(k)}$ in the running time.

Boutsidis et al. [BKG15] showed that an approximate k-means partition of an approximate SE $\widetilde{Y}$ computed via the Power method, yields a k-means partition P of the exact SE Y such that the k-means cost of P yields an additive approximation to the optimum k-means cost of Y. This gives an affirmative answer to question Q1. Further, the authors stated as main open problems to resolve questions Q2 and Q3.

Besides designing a PTAS for the k-means clustering problems, Ostrovsky et al. [ORSS13] analyzed a variant of Lloyd k-means clustering algorithm. They showed that on input a set of n vectors in $\mathbb{R}^k$ satisfying a natural well-clusterable assumption, the algorithm *efficiently* computes a good approximation of an optimal k-means partition. In particular, the algorithm runs in time $O(k^2(n+k^2))$.

A natural question to ask is whether the analysis of Peng et al. [PSZ17], Boutsidis et al. [BKG15] and Ostrovsky et al. [ORSS13] can be integrated and extended to answer the questions Q2 and Q3?

Our Contribution: An Overview

We give a comprehensive analysis of ASC building on the work of Peng et al. [PSZ17], Boutsidis et al. [BKG15] and Ostrovsky et al. [ORSS13]. We show that the Approximate Spectral Clustering i) runs efficiently, and ii) yields a good approximation of an optimal k-way node partition of G. Moreover, we strengthen the quality guarantees of a structural result of Peng et al. [PSZ17] by a factor of k, and simultaneously weaken the eigenvalue gap assumption. Further, our analysis shows that the Approximate Spectral Clustering finds a k-way node partition of G with the strengthened quality guarantees. This gives an affirmative answer to questions Q2 and Q3.

5.1 Notation

k-means Clustering Problem

Let $\mathcal{X}$ be a set of vectors of the same dimension. The k-means cost of a partition $(X_1,\ldots,X_k)$ of $\mathcal{X}$ is given by

$$\mathrm{Cost}(\{X_i\}_{i=1}^k) \overset{\text{def}}{=} \sum_{i=1}^{k}\sum_{x\in X_i}\|x-c_i\|_2^2\,, \qquad \text{where} \qquad c_i \overset{\text{def}}{=} \frac{1}{|X_i|}\sum_{x\in X_i} x$$

is the gravity center of X_i, for all $i\in\{1,\ldots,k\}$. Then, the optimum k-means cost (of clustering $\mathcal{X}$ into k sets) is defined as

$$\triangle_k(\mathcal{X}) \overset{\text{def}}{=} \min_{\text{partition } (X_1,\ldots,X_k) \text{ of } \mathcal{X}} \mathrm{Cost}(\{X_i\}_{i=1}^k).$$

A k-means partition $(X_1,\ldots,X_k)$ of $\mathcal{X}$, with corresponding gravity centers $c_1,\ldots,c_k$ as above, is α-approximate if $\mathrm{Cost}(\{X_i\}_{i=1}^k)\leqslant \alpha\cdot\triangle_k(\mathcal{X})$. Given a matrix Y, we abuse notation and write $\triangle_k(Y)$ to denote the optimum k-means cost of partitioning the rows of matrix Y.

Spectral Embeddings

Given an undirected graph $G=(V,E)$ with $m=|E|$ edges and $n=|V|$ nodes, let D be the diagonal degree matrix and A be the adjacency matrix. Then, the graph Laplacian matrix is defined as $L=D-A$, and the normalized Laplacian matrix is given by $\mathcal{L}_G=I-\mathcal{A}$, where $\mathcal{A}=D^{-1/2}AD^{-1/2}$. Further, let $f_i\in\mathbb{R}^V$ be the eigenvector corresponding to the i-th smallest eigenvalue λ_i of $\mathcal{L}_G$.

The *canonical* Spectral Embedding, for short *canonical* SE, is defined as a matrix $Y\in\mathbb{R}^{n\times k}$ composed of the bottom k eigenvectors [1] of $\mathcal{L}_G$ corresponding to the k smallest eigenvalues. The *approximate* SE is computed via the Power method [2]. Namely, let $S\in\mathbb{R}^{n\times k}$ be a matrix whose entries are i.i.d. samples from

[1] The Eigendecomposition theorem guarantees that all eigenvectors are orthonormal.

[2] Given a symmetric matrix M and a number k, the Power method approximates the top k eigenvectors of M corresponding to the largest k eigenvalues. Since we seek a good approximation of the bottom k eigenvectors of $\mathcal{L}_G=I-\mathcal{A}$, associated with the smallest k eigenvalues, we initialize the Power method with $M=I+\mathcal{A}$.

the standard Gaussian distribution $N(0,1)$ and p be the number of iterations. Then, the *approximate* SE $\widetilde{Y}$ is given by:

$$1)\ M \overset{\text{def}}{=} I + \mathcal{A}; \quad 2)\ \text{Let } \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}} \text{ be the SVD }^{3} \text{ of } M^{p}S; \quad \text{and} \quad 3)\ \widetilde{Y} \overset{\text{def}}{=} \widetilde{U} \in \mathbb{R}^{n\times k}. \tag{5.7}$$

Peng et al. [PSZ17] do not apply k-means directly to the *canonical* SE, but first normalize it by dividing the row corresponding to u by $\sqrt{d(u)}$ and then put $d(u)$ copies of the resulting vector into the k-means clustering instance. This repetition of vectors is crucial for their analysis, in order to achieve approximation guarantees in terms of volume overlap and conductance. We follow their approach.

We construct a matrix $Y' \in \mathbb{R}^{2m\times k}$ such that for every node $u \in V$, we insert $d(u)$ many copies of the normalized row $Y(u,:)/\sqrt{d(u)}$ to Y'. Formally, the *normalized* SE Y' and the *approximate normalized* SE $\widetilde{Y'}$ are defined by

$$Y' \overset{\text{def}}{=} \begin{pmatrix} \mathbf{1}_{d(1)} \frac{Y(1,:)}{\sqrt{d(1)}} \\ \cdots \\ \mathbf{1}_{d(n)} \frac{Y(n,:)}{\sqrt{d(n)}} \end{pmatrix}_{2m\times k} \quad \text{and} \quad \widetilde{Y'} \overset{\text{def}}{=} \begin{pmatrix} \mathbf{1}_{d(1)} \frac{\widetilde{Y}(1,:)}{\sqrt{d(1)}} \\ \cdots \\ \mathbf{1}_{d(n)} \frac{\widetilde{Y}(n,:)}{\sqrt{d(n)}} \end{pmatrix}_{2m\times k}, \tag{5.8}$$

where $\mathbf{1}_{d(i)}$ is the all-one column vector with dimension $d(i)$.

We can assume w.l.o.g. that a k-means clustering algorithm applied on Y' ($\widetilde{Y'}$), outputs a k-means partition such that all copies of row $Y(v,:)/\sqrt{d(v)}$ ($\widetilde{Y}(v,:)/\sqrt{d(v)}$) belong to the same cluster, for all nodes v. Thus, the algorithm induces a k-way node partition of G.

5.1.1 Prior Structural Result

A key prior structural result, established by Peng et al. [PSZ17], connects the normalized SE Y', α-approximate k-means clustering, the k-way conductance $\widehat{\rho}(k)$, as given in Equation (5.2), and the $(k+1)$-st eigenvalue λ_{k+1} of the normalized Laplacian matrix $\mathcal{L}_G$. In particular, they proved the following statement under a gap assumption defined in terms of

$$\Upsilon \overset{\text{def}}{=} \frac{\lambda_{k+1}}{\widehat{\rho}(k)}.$$

Theorem 5.1. *[PSZ17, Theorem 1.2] Let $k \geqslant 3$ and G be a graph satisfying the gap assumption* [4]

$$\delta \overset{\text{def}}{=} 2\cdot 10^{5}\cdot k^{3}/\Upsilon \leqslant 1/2. \tag{5.9}$$

Let $(P_1,\ldots,P_k)$ be a k-way node partition of G achieving $\widehat{\rho}(k)$, and let $(A_1,\ldots,A_k)$ be the k-way node partition of G induced by an α-approximate k-means partition of the normalized SE Y'. Then, for every $i \in \{1,\ldots,k\}$ it hold (after suitable renumbering of one of the partitions) that

$$1)\ \mu(A_i \triangle P_i) \leqslant \alpha\delta\cdot\mu(P_i) \quad \text{and} \quad 2)\ \phi(A_i) \leqslant (1+2\alpha\delta)\cdot\phi(P_i) + 2\alpha\delta.$$

Under a stronger eigenvalue gap assumption $2\cdot 10^{5}\cdot k^{5}/\Upsilon \leqslant 1/2$, Peng et al. [PSZ17] gave an algorithm that finds in time $O(m\cdot \operatorname{poly}\log(n))$ a k-way node partition of G with essentially the guarantees stated in Theorem 5.1. However, their algorithmic result substitutes normalized SE with Heat Kernel Embedding and k-means clustering with locality sensitive hashing. Thus, the algorithmic part of their paper does not explain the success of Approximate Spectral Clustering.

5.1.2 Our Contribution

We give affirmative answer to the questions Q2 and Q3. On the way, we also strengthen the approximation guarantees in Theorem 5.1 by a factor of k and simultaneously weaken the eigenvalue gap assumption.

Let $\mathcal{P}$ be the set of all k-way partitions $(P_1,\ldots,P_k)$ achieving the k-way conductance $\widehat{\rho}(k)$. Let

$$\widehat{\rho}_{\text{avr}}(k) \overset{\text{def}}{=} \min_{(P_1,\ldots,P_k)\in\mathcal{P}} \frac{1}{k}\sum_{i=1}^{k}\phi(P_i)$$

[3] SVD abbreviates Singular Value Decomposition, see [Woo14].

[4] Note that $\lambda_k/2 \leqslant \widehat{\rho}(k)$, see (5.8). Thus, the assumption implies $\lambda_k/2 \leqslant \widehat{\rho}(k) \leqslant \delta\lambda_{k+1}/(2\cdot 10^{5}\cdot k^{3})$, i.e., there is a substantial gap between the $(k+1)$-th and the k-th eigenvalue.

be the *minimum average conductance* over all k-way partitions in $\mathcal{P}$. Note that $\widehat{\rho}_{\mathrm{avr}}(k) \leqslant \widehat{\rho}(k)$. Our gap assumption is defined in terms of

$$\Psi \overset{\mathrm{def}}{=} \frac{\lambda_{k+1}}{\widehat{\rho}_{\mathrm{avr}}(k)}.$$

In the remainder, we denote by $(P_1, \ldots, P_k)$ a k-way node partition of G achieving $\widehat{\rho}_{\mathrm{avr}}(k)$.

We now present our main result, consisting of a structural and an algorithmic statement.

Theorem 5.2. *a) (Existence of a Good Clustering) Let $k \geqslant 3$ and G be a graph satisfying*

$$\delta \overset{\mathrm{def}}{=} 20^4 \cdot k^3/\Psi \leqslant 1/2. \tag{5.10}$$

Let $(P_1, \ldots, P_k)$ be a k-way node partition of G achieving $\widehat{\rho}_{\mathrm{avr}}(k)$, and let $(A_1, \ldots, A_k)$ be the k-way node partition of G induced by an α-approximate k-means partition of the normalized SE Y'. Then, for every $i \in \{1, \ldots, k\}$ it hold (after suitable renumbering of one of the partitions) that

$$1)\ \mu(A_i \triangle P_i) \leqslant \frac{\alpha\delta}{10^3 k} \cdot \mu(P_i) \quad \textit{and} \quad 2)\ \phi(A_i) \leqslant \left(1 + \frac{2\alpha\delta}{10^3 k}\right) \cdot \phi(P_i) + \frac{2\alpha\delta}{10^3 k}.$$

b) (An Efficient Algorithm) If in addition $k/\delta \geqslant 10^9$, then the variant of Lloyd algorithm analyzed by Ostrovsky et al. [ORSS13] when applied to the approximate normalized SE $\widetilde{Y'}$, induces in time $O(m(k^2 + \frac{\ln n}{\lambda_{k+1}}))$ with constant probability a k-way node partition $(A_1, \ldots, A_k)$ of G such that for every $i \in \{1, \ldots, k\}$ it hold (after suitable renumbering of one of the partitions) that

$$3)\ \mu(A_i \triangle P_i) \leqslant \frac{2\delta}{10^3 k} \cdot \mu(P_i) \quad \textit{and} \quad 4)\ \phi(A_i) \leqslant \left(1 + \frac{4\delta}{10^3 k}\right) \cdot \phi(P_i) + \frac{4\delta}{10^3 k}.$$

Part (a) of Theorem 5.2 strengthens the quality guarantees in Theorem 5.1 by a factor of k, and simultaneously weaken the eigenvalue gap assumption. Part (b) of Theorem 5.2 gives a comprehensive analysis of Approximate Spectral Clustering, and demonstrates that the algorithm i) runs efficiently, and ii) yields a good approximation of an optimal k-way node partition of G.
Further, it shows that the Approximate Spectral Clustering finds a k-way node partition of G with the strengthened quality guarantees, and whenever $k \leqslant (\log n)^{O(1)}$ and $\lambda_{k+1} \geqslant 1/(\log n)^{O(1)}$, the algorithm runs in nearly linear time. This answers affirmatively questions Q2 and Q3.

Remarks

The variant of Lloyd k-means clustering algorithm, analyzed by Ostrovsky et al. [ORSS13], is efficient only for inputs $\mathcal{X}$ satisfying $\triangle_k(\mathcal{X}) \leqslant \varepsilon^2 \triangle_{k-1}(\mathcal{X})$ for some $\varepsilon \in (0, \varepsilon_0]$, where $\varepsilon_0 = 6/10^7$. The authors stated that their result should also hold for a larger ε_0, and mentioned that they did not attempt to maximize ε_0.

An anonymous reviewer of the conference version of this paper, suggested to include a numerical example. Consider a graph which consists of k cliques each of size n/k [5], plus k additional edges that connect the cliques in the form of a ring such that no two edges of the ring share a vertex. This graph is a trivial clustering instance, and for any constant k it holds that $\lambda_{k+1} \geqslant 1 - k/n$, see Theorem 7.17. Observe that $\widehat{\rho}_{\mathrm{avr}}(k) = \widehat{\rho}(k) \approx (k/n)^2$. For the gap assumption to hold we need $\lambda_{k+1} \geqslant 2 \cdot 20^4 \cdot k^3 \cdot \widehat{\rho}_{\mathrm{avr}}(k)$. This implies $n \geqslant \sqrt{2 \cdot 20^4 \cdot k^5/\lambda_{k+1}}$. For small k, this is a modest requirement on the size of the graph.

For the algorithmic result, we need in addition $\delta \leqslant k \cdot \varepsilon_0/600$. For the gap condition to hold, we need $\lambda_{k+1} \geqslant (600/\varepsilon_0 k) \cdot 20^4 \cdot k^3 \cdot (k^2/n^2)$ or $n \geqslant \sqrt{20^6 \cdot k^4/(\varepsilon_0 \lambda_{k+1})}$. For $\varepsilon_0 = 6/10^7$, this amounts to $n \geqslant \sqrt{2^4 \cdot 10^{13} \cdot k^4/\lambda_{k+1}}$, a quite large lower bound on n.

The statement that Part (b) of Theorem 5.2 gives a theoretical support for the practical success of Approximate Spectral Clustering, therefore has to be taken with a grain of salt. It is only an asymptotic statement and does not explain the good behavior on small graphs.

[5] A graph G has k connected components iff $\lambda_k = 0$. For any clique K_n, we have $\lambda_1 = 0$ and $\lambda_2 = \cdots = \lambda_n = 1$. Further, when G consists of k cliques $K_{n/k}$ disconnected from each other, then $\lambda_1 = \cdots = \lambda_k = 0$ and $\lambda_{k+1} = \cdots = \lambda_n = 1$.

5.2 Our Techniques

In Chapter 6, we give a refined spectral analysis of [PSZ17] which yields the improved structural result in Part (a) of Theorem 5.2. In Chapter 7, we connect Part (a) of Theorem 5.2 with the work of Ostrovsky et al. [ORSS13] and Boutsidis et al. [BKG15], yielding the algorithmic result in Part (b) of Theorem 5.2.

Ostrovsky et al. [ORSS13] analyzed a variant of Lloyd k-means clustering algorithm. We refer to this algorithm as the ORSS clustering algorithm. The ORSS-algorithm is efficient only for inputs $\mathcal{X}$ satisfying: some partition into k clusters is much better than any partition into $k-1$ clusters. Formally, it states

Theorem 5.3. *[ORSS13, Theorem 4.15] Assuming that $\triangle_k(\mathcal{X}) \leqslant \varepsilon^2 \cdot \triangle_{k-1}(\mathcal{X})$ for $\varepsilon \in (0, 6\cdot 10^{-7}]$, the* ORSS*-algorithm runs in time $O(nkd + k^3 d)$ and returns with probability at least $1 - O(\sqrt{\varepsilon})$ a k-way partition of $\mathcal{X}$ with cost at most $[(1-\varepsilon^2)/(1-37\varepsilon^2)]\triangle_k(\mathcal{X})$.*

Let $Z \in \mathbb{R}^{n\times k}$ be a matrix and $(R_1, \ldots, R_k)$ be a row partition of Z. Let $c_j = \frac{1}{|R_j|}\sum_{u\in R_j} Z_{u,:}$ be the gravity center of cluster R_j, for all $j \in \{1, \ldots, k\}$. We next express in matrix notation the k-means cost of partition $(R_1, \ldots, R_k)$. To this end, we introduce an *indicator* matrix $X \in \mathbb{R}^{n\times k}$ such that $X_{ij} = 1/\sqrt{|R_j|}$ if row $Z_{i,:}$ belongs to cluster R_j, and $X_{ij} = 0$ otherwise. Then, $(XX^T Z)_{i,:} = c_j$, where row $Z_{i,:}$ belongs to cluster R_j. Hence, the k-means cost of $(R_1, \ldots, R_k)$ becomes

$$\mathrm{Cost}(\{R_i\}_{i=1}^k) = \sum_{j=1}^k \sum_{u\in R_j} \|Z_{u,:} c_j\|_2^2 = \|Z - XX^{\mathrm{T}}Z\|_F^2. \tag{5.11}$$

Our Analytical Approach

Our main technical contribution is to prove that the approximate normalized SE $\widetilde{Y'}$ computed via the Power method is ε-separated, i.e. the assumption $\triangle_k(\widetilde{Y'}) < \varepsilon^2 \cdot \triangle_{k-1}(\widetilde{Y'})$ of Ostrovsky et al. [ORSS13] is satisfied. This implies, by Theorem 5.3, that the ORSS-algorithm runs efficiently on $\widetilde{Y'}$. Let the resulting k-way row partition of $\widetilde{Y'}$ be encoded by the indicator matrix $\widetilde{X'}$.

Then, building on the work of [BM14, BKG15], we show that $\widetilde{X'}$ is a good approximation of an optimal k-means partition of the corresponding normalized SE Y'. Further, using our strengthened structural result in Part (a) of Theorem 5.2, we show that $\widetilde{X'}$ induces a good approximation of an optimal k-way node partition of graph G, in terms of volume overlap and conductance.

First, we establish in Section 7.1 the assumption of Ostrovsky et al. [ORSS13] for the normalized SE Y'.

Theorem 5.4. *(normalized SE is ε-separated) Let G be a graph that satisfies $\Psi = 20^4 \cdot k^3/\delta$, $\delta \in (0, 1/2]$ and $k/\delta \geqslant 10^9$. Then for $\varepsilon = 6 \cdot 10^{-7}$ it holds $\triangle_k(Y') \leqslant \varepsilon^2 \cdot \triangle_{k-1}(Y')$.*

Theorem 5.4 does not suffice for proving Part (b) of Theorem 5.2, since it requires the analogous statement for the approximate normalized SE $\widetilde{Y'}$.

In Subsection 7.2.2, we show that an α-approximate k-means clustering algorithm applied to the approximate normalized SE $\widetilde{Y'}$, yields an approximate k-way row partition of the corresponding normalized SE Y'.

Theorem 5.5. *(Similar to [BKG15, Theorem 6], but analyzes the approximate normalized SE) Let $\varepsilon, \delta_p \in (0,1)$ be arbitrary. Compute the approximate normalized SE $\widetilde{Y'}$ via the Power method with $p \geqslant \ln(8nk/\varepsilon\delta_p)/\ln(1/\gamma_k)$ iterations and $\gamma_k = (2-\lambda_{k+1})/(2-\lambda_k) < 1$. Run on the rows of $\widetilde{Y'}$ an α-approximate k-means clustering algorithm with failure probability δ_α. Let the outcome be a clustering indicator matrix $\widetilde{X'_\alpha} \in \mathbb{R}^{n\times k}$. Then, with probability at least $1 - 2e^{-2n} - 3\delta_p - \delta_\alpha$, it holds that*

$$\|Y' - \widetilde{X'_\alpha}(\widetilde{X'_\alpha})^{\mathrm{T}} Y'\|_F^2 \leqslant (1+4\varepsilon)\cdot\alpha\cdot\triangle_k(Y') + 4\varepsilon^2.$$

In Subsection 7.2.3, using Theorem 5.4 and Theorem 5.5, we show that the approximate normalized SE $\widetilde{Y'}$ satisfies the assumption of Ostrovsky et al. [ORSS13].

Theorem 5.6. *(approximate normalized SE is ε-separated) Assume $\Psi = 20^4 \cdot k^3/\delta$ and $k/\delta \geqslant 10^9$ for some $\delta \in (0, 1/2]$. Compute the approximate normalized SE $\widetilde{Y'}$ via the Power method with $p \geqslant \Omega(\frac{\ln n}{\lambda_{k+1}})$ iterations. Then, for $\varepsilon = 6 \cdot 10^{-7}$ it holds with high probability that $\triangle_k(\widetilde{Y'}) < 5\varepsilon^2 \cdot \triangle_{k-1}(\widetilde{Y'})$.*

Finally, in Subsection 7.3, we prove Part (b) of Theorem 5.2 by combining Part (a) of Theorem 5.2, Theorem 5.3, Theorem 5.5 and Theorem 5.6.

Chapter 6

Improved Structural Result

6.1 Notation

We use the notation adopted in [PSZ17]. Let λ_j be the j-th eigenvalue of the normalized Laplacian matrix $\mathcal{L}_G$, and let $f_j \in \mathbb{R}^V$ be the associated eigenvector ($\mathcal{L}_G f_j = \lambda_j f_j$).

Let $\overline{g_i} = \frac{D^{1/2}\chi_{P_i}}{\left\|D^{1/2}\chi_{P_i}\right\|_2}$, where χ_{P_i} is the characteristic vector of the subset $P_i \subseteq V$. Note that $\overline{g_i}$ is the normalized characteristic vector of P_i and $\left\|D^{1/2}\chi_{P_i}\right\|_2^2 = \sum_{v\in P_i} d(v) = \mu(P_i)$. The Rayleigh quotient is defined by and satisfies

$$\mathcal{R}\left(\overline{g_i}\right) \overset{\text{def}}{=} \frac{\overline{g_i}^{\mathrm{T}}\mathcal{L}_G\overline{g_i}}{\overline{g_i}^{\mathrm{T}}\overline{g_i}} = \frac{1}{\mu(P_i)}\chi_{P_i}^{\mathrm{T}} L \chi_{P_i} = \frac{|E(S,\overline{S})|}{\mu(P_i)} = \phi(P_i),$$

where the Laplacian matrix $L = D - A$ and the normalized Laplacian matrix $\mathcal{L}_G = D^{-1/2}LD^{-1/2}$.

The eigenvectors $\{f_i\}_{i=1}^n$ form an orthonormal basis of $\mathbb{R}^n$. Thus each characteristic vector $\overline{g_i}$ can be expressed as $\overline{g_i} = \sum_{j=1}^n \alpha_j^{(i)} f_j$ for all $i \in \{1,\ldots,k\}$. We define its *projection* onto the first k eigenvectors by $\widehat{f_i} = \sum_{j=1}^k \alpha_j^{(i)} f_j$.

Peng et al. [PSZ17] proved that if the gap parameter Υ is large enough then $\mathrm{span}(\{\widehat{f_i}\}_{i=1}^k) = \mathrm{span}(\{f_i\}_{i=1}^k)$ and the first k eigenvectors can be expressed by $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f_j}$, for all $i \in \{1,\ldots,k\}$. Moreover, they demonstrated that each vector $\widehat{g_i} = \sum_{j=1}^k \beta_j^{(i)} \overline{g_j}$ approximates the eigenvector f_i, for all $i \in \{1,\ldots,k\}$. We will show that similar statements hold with weakened gap parameter Ψ.

The *estimation centers* induced by the canonical SE are given by

$$p^{(i)} = \frac{1}{\sqrt{\mu(P_i)}}\left(\beta_i^{(1)},\ldots,\beta_i^{(k)}\right)^{\mathrm{T}}. \tag{6.1}$$

Our analysis crucially relies on spectral properties of the following two matrices. Let $F, B \in \mathbb{R}^{k\times k}$ be square matrices defined by

$$F_{j,i} = \alpha_j^{(i)} \quad \text{and} \quad B_{j,i} = \beta_j^{(i)}. \tag{6.2}$$

In Figure 6.1, we show the relation among the vectors f_i, $\widehat{f_i}$, $\widehat{g_i}$ and $\overline{g_i}$.

$$\begin{tikzcd}
\widehat{f_i} = \sum_{j=1}^k \alpha_j^{(i)} f_j \arrow[r, dash, "\|\widehat{f_i} - \overline{g_i}\|_2^2 \leqslant \phi(P_i)/\lambda_{k+1}"] \arrow[d, dash] & \overline{g_i} = \frac{D^{1/2}\chi_{P_i}}{\sqrt{\mu(P_i)}} = \sum_{j=1}^n \alpha_j^{(i)} f_j \arrow[d, dash] \\
f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f_j} \arrow[r, dash, "\|f_i - \widehat{g_i}\|_2^2 \leqslant (1 + 3k/\Psi)\cdot k/\Psi"] & \widehat{g_i} = \sum_{j=1}^k \beta_j^{(i)} \overline{g_j}
\end{tikzcd}$$

Figure 6.1: The vectors $\{f_i\}_{i=1}^n$ are eigenvectors of the normalized Laplacian matrix $\mathcal{L}_G$. The vectors $\{\overline{g}_i\}_{i=1}^k$ are the normalized characteristic vectors of an optimal partition $(P_1,\ldots,P_k)$. For each $i \in \{1,\ldots,k\}$ the vector $\widehat{f_i}$ is the projection of vector $\overline{g_i}$ onto $\mathrm{span}(f_1,\ldots,f_k)$. The vectors $\widehat{f_i}$ and $\overline{g}_i$ are close for $i \in \{1,\ldots,k\}$. If $\Psi > 4\cdot k^{3/2}$, then $\mathrm{span}(f_1,\ldots,f_k) = \mathrm{span}(\widehat{f_1},\ldots,\widehat{f_k})$ and thus we can write $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f_j}$. Further, the vectors f_i and $\widehat{g_i} = \sum_{j=1}^k \beta_j^{(i)} \overline{g_j}$ are close for $i \in \{1,\ldots,k\}$.

6.2 Technical Insights

The analysis of Part (a) of Theorem 5.2 follows the proof approach in [PSZ17, Theorem 1.2], but improves upon it in essential ways.

Our first technical insight is that the matrices $B^{\mathrm{T}}B$ and BB^{T} are close to the identity matrix. We prove this in two steps. In Section 6.4, we show that the vectors $\widehat{g}_i$ and f_i are close, and then in Section 6.5 we analyze the column space and row space of matrix B.

Theorem 6.1 (Matrix BB^{T} is Close to Identity Matrix). *If $\Psi \geqslant 10^4 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0,1)$ then for all distinct $i, j \in \{1,\ldots,k\}$ it holds*

$$1-\varepsilon \leqslant \langle B_{i,:}, B_{i,:}\rangle \leqslant 1+\varepsilon \quad \textit{and} \quad |\langle B_{i,:}, B_{j,:}\rangle| \leqslant \sqrt{\varepsilon}.$$

Using Theorem 6.1, we give a strengthened version of [PSZ17, Lemma 4.2] that depends on the weaken gap parameter Ψ.

Lemma 6.2. *If $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0,1]$ then for every $i \in \{1,\ldots,k\}$ it holds that*

$$\left(1-\sqrt{\delta}/4\right)\frac{1}{\mu(P_i)} \leqslant \left\|p^{(i)}\right\|_2^2 \leqslant \left(1+\sqrt{\delta}/4\right)\frac{1}{\mu(P_i)}.$$

Proof. By definition $p^{(i)} = \frac{1}{\sqrt{\mu(P_i)}} \cdot B_{i,:}$ and Theorem 6.1 yields $\|B_{i,:}\|_2^2 \in [1 \pm \sqrt{\delta}/4]$. □

Using Theorem 6.1 and Lemma 6.2, we establish a strengthened version of [PSZ17, Lemma 4.3] that depends on the weaken gap parameter Ψ, and simultaneously shows that the ℓ_2 distance between estimation centers is larger by a factor of k.

Lemma 6.3 (Larger Distance Between Estimation Centers). *If $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, \frac{1}{2}]$ then for any distinct $i, j \in \{1,\ldots,k\}$ it holds that*

$$\left\|p^{(i)} - p^{(j)}\right\|_2^2 \geqslant \left[2 \cdot \min\left\{\mu(P_i), \mu(P_j)\right\}\right]^{-1}.$$

Proof. Since $p^{(i)}$ is a row of matrix B, Theorem 6.1 with $\varepsilon = \sqrt{\delta}/4$ yields

$$\left\langle \frac{p^{(i)}}{\|p^{(i)}\|_2}, \frac{p^{(j)}}{\|p^{(j)}\|_2} \right\rangle = \frac{\langle B_{i,:}, B_{j,:}\rangle}{\|B_{i,:}\|_2 \|B_{j,:}\|_2} \leqslant \frac{\sqrt{\varepsilon}}{1-\varepsilon} = \frac{2\delta^{1/4}}{3}.$$

W.l.o.g. assume that $\|p^{(i)}\|_2^2 \geqslant \|p^{(j)}\|_2^2$, say $\|p^{(j)}\|_2 = \alpha \cdot \|p^{(i)}\|_2$ for some $\alpha \in (0,1]$. Then by Lemma 6.2 we have $\|p^{(i)}\|_2^2 \geqslant (1-\sqrt{\delta}/4) \cdot [\min\{\mu(P_i), \mu(P_j)\}]^{-1}$, and hence

$$\begin{aligned}\left\|p^{(i)} - p^{(j)}\right\|_2^2 &= \left\|p^{(i)}\right\|_2^2 + \left\|p^{(j)}\right\|_2^2 - 2\left\langle \frac{p^{(i)}}{\|p^{(i)}\|_2}, \frac{p^{(j)}}{\|p^{(j)}\|_2} \right\rangle \left\|p^{(i)}\right\|_2 \left\|p^{(j)}\right\|_2 \\ &\geqslant \left(\alpha^2 - \frac{4\delta^{1/4}}{3} \cdot \alpha + 1\right)\left\|p^{(i)}\right\|_2^2 \geqslant \left[2 \cdot \min\left\{\mu(P_i), \mu(P_j)\right\}\right]^{-1}.\end{aligned}$$

□

Using Lemma 6.2 and Lemma 6.3, the observation that Υ can be replaced by Ψ in all statements in [PSZ17] is technically easy.

Our second technical contribution is to show that the larger ℓ_2 distance between estimation centers, in Lemma 6.3, strengthens [PSZ17, Lemma 4.5] by a factor of k. Before we state our result, we need some notation.

The normalized Spectral Embedding map $\mathcal{F} : V \to \mathbb{R}^k$ is defined by

$$\mathcal{F}(v) \overset{\text{def}}{=} \frac{1}{\sqrt{d(v)}}\left(f_1(v), \ldots, f_k(v)\right)^{\mathrm{T}} = \frac{1}{\sqrt{d(v)}} \cdot [Y(v,:)]^{\mathrm{T}},$$

for every node $v \in V$. Recall that the normalized SE Y' contains duplicate rows, namely, $d(u)$ many copies of $\mathcal{F}(u)$ for each node $u \in V$.

Suppose an α-approximate k-means clustering algorithm outputs a k-way row partition $(R_1, \ldots, R_k)$ of Y'. We can assume w.l.o.g. that all identical rows of Y' are assigned to same cluster, and thus $(R_1, \ldots, R_k)$ induces a k-way node partition $(A_1, \ldots, A_k)$ of G. For an arbitrary point set $c_1, \ldots, c_k$ in $\mathbb{R}^k$, we abuse the notation and denote the k-means cost of a tuple $\{A_i, c_i\}_{i=1}^k$ by

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) = \sum_{i=1}^{k} \sum_{u \in A_i} d(u) \left\| \mathcal{F}(u) - c_i \right\|_2^2 . \tag{6.3}$$

When each point $c_j = \frac{1}{\mu(A_j)} \sum_{u \in A_j} d(u)\mathcal{F}(u)$ is the gravity center of cluster R_j, for brevity we write $\mathrm{Cost}(\{A_i\}_{i=1}^k)$ to denote the k-means cost of tuple $\{A_i, c_i\}_{i=1}^k$.

Lemma 6.4 (Volume Overlap)**.** *Let $(P_1, \ldots, P_k)$ and $(A_1, \ldots, A_k)$ be k-way node partitions of G. Suppose for every permutation $\pi : \{1, \ldots, k\} \to \{1, \ldots, k\}$ there is an index $i \in \{1, \ldots, k\}$ such that*

$$\mu(A_i \triangle P_{\pi(i)}) \geqslant \frac{2\varepsilon}{k} \cdot \mu(P_{\pi(i)}), \tag{6.4}$$

where $\varepsilon \in (0, 1)$ is a parameter. If $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, \frac{1}{2}]$, and $\varepsilon \geqslant 64\alpha \cdot k^3/\Psi$ then

$$\mathrm{Cost}(\{A_i\}_{i=1}^k) > \frac{2\alpha k^2}{\Psi}.$$

With the above lemmas in place, the proof of Part (a) of Theorem 5.2 is then completed as in [PSZ17]. For completeness, we present the proof.

6.3 Proof of Improved Structural Result

In this Section, we prove Part (a.1) of Theorem 5.2. Crucial to our analysis is the following result, which we prove in the next Section 6.4, showing that vectors $\widehat{g}_i$ and f_i are close, c.f. Figure 6.1.

Theorem 6.5. *If $\Psi > 4 \cdot k^{3/2}$, then for every $i \in \{1, \ldots, k\}$ the vectors f_i and $\widehat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \overline{g_j}$ satisfy*

$$\| f_i - \widehat{g}_i \|^2 \leqslant \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k}{\Psi}.$$

Lemma 6.6 ($(P_1, \ldots, P_k)$ is a good k-means partition)**.** *If $\Psi > 4 \cdot k^{3/2}$, then there are vectors $\{p^{(i)}\}_{i=1}^k$ such that*

$$\mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leqslant \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi}.$$

Proof. Let $(P_1, \ldots, P_k)$ be a k-way node partition of G achieving $\widehat{\rho}_{\mathrm{avr}}(k)$. Peng et al. [PSZ17, Lemma 4.1] showed that $\mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) = \sum_{j=1}^k \| f_j - \widehat{g}_j \|_2^2$, and thus the statement follows by Theorem 6.5.

For completeness, we now prove the preceding equation. By definition, $p_j^{(i)} = \beta_i^{(j)} / \sqrt{\mu(P_i)}$ and $\widehat{g}_j = \sum_{i=1}^k \beta_i^{(j)} \cdot \frac{D^{1/2}\chi_{P_i}}{\sqrt{\mu(P_i)}}$, where χ_{P_i} is characteristic vector of the node subset P_i. Then,

$$\begin{aligned} \mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) &= \sum_{i=1}^{k} \sum_{u \in P_i} d(u) \| \mathcal{F}(u) - p^{(i)} \|_2^2 \\ &= \sum_{j=1}^{k} \sum_{i=1}^{k} \sum_{u \in P_i} \left(f_j(u) - \frac{\sqrt{d(u)}}{\sqrt{\mu(P_i)}} \beta_i^{(j)} \right)^2 = \sum_{j=1}^{k} \| f_j - \widehat{g}_j \|_2^2 . \end{aligned}$$

□

Lemma 6.7 (Only partitions close to $(P_1, \ldots, P_k)$ are good)**.** *Under the hypothesis of Theorem 5.2, the following holds. If for every permutation $\sigma : \{1, \ldots, k\} \to \{1, \ldots, k\}$ there exists an index $i \in \{1, \ldots, k\}$ such that*

$$\mu(A_i \triangle P_{\sigma(i)}) \geqslant \frac{8\alpha\delta}{10^4 k} \cdot \mu(P_{\sigma(i)}). \tag{6.5}$$

Then it holds that

$$\mathrm{Cost}(\{A_i\}_{i=1}^k) > \frac{2\alpha k^2}{\Psi}. \tag{6.6}$$

We note that Lemma 6.7 follows directly by applying Lemma 6.4 with $\varepsilon = 64 \cdot \alpha \cdot k^3/\Psi$. Since $(A_1, \ldots, A_k)$ is an α approximate solution to $\triangle_k(Y')$, we obtain a contradiction

$$\frac{2\alpha k^2}{\Psi} < \mathrm{Cost}(\{A_i\}_{i=1}^k) \leqslant \alpha \cdot \triangle_k(Y') \leqslant \alpha \cdot \mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leqslant \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{\alpha k^2}{\Psi}.$$

Therefore, there exists a permutation π (the identity after suitable renumbering of one of the partitions) such that $\mu(A_i \triangle P_i) < \frac{8\alpha\delta}{10^4 k} \cdot \mu(P_i)$ for all $i \in \{1, \ldots, k\}$.

Part (a.2) of Theorem 5.2 follows from Part (a.1). Indeed, for $\delta' = 8\delta/10^4$ we have

$$\mu(A_i) \geqslant \mu(P_i \cap A_i) = \mu(P_i) - \mu(P_i \setminus A_i) \geqslant \mu(P_i) - \mu(A_i \triangle P_i) \geqslant \left(1 - \frac{\alpha\delta'}{k}\right) \cdot \mu(P_i)$$

and $|E(A_i, \overline{A_i})| \leqslant |E(P_i, \overline{P_i})| + \mu(A_i \Delta P_i)$, since every edge that is counted in $|E(A_i, \overline{A_i})|$ but not in $|E(P_i, \overline{P_i})|$ must have an endpoint in $A_i \Delta P_i$. Thus

$$\Phi(A_i) = \frac{|E(A_i, \overline{A_i})|}{\mu(A_i)} \leqslant \frac{|E(P_i, \overline{P_i})| + \frac{\alpha\delta'}{k} \cdot \mu(P_i)}{(1 - \frac{\alpha \cdot \delta'}{k}) \cdot \mu(P_i)} \leqslant \left(1 + \frac{2\alpha\delta'}{k}\right) \cdot \phi(P_i) + \frac{2\alpha\delta'}{k}.$$

This completes the proof of Part (a) of Theorem 5.2.

6.4 Vectors $\widehat{g_i}$ and f_i are Close

In this section, we prove Theorem 6.5. We argue in a similar manner as in [PSZ17], but in contrast our results depend on the weaken gap parameter Ψ. For completeness, we show in Subsection 6.4.1 that the span of the first k eigenvectors of $\mathcal{L}_G$ equals the span of the projections of P_i's characteristic vectors onto the first k eigenvectors. Then, in Subsection 6.4.2, we conclude the proof of Theorem 6.5 by analyzing the eigenvectors $\{f_i\}_{i=1}^k$ in terms of projection vectors $\{\widehat{f_i}\}_{i=1}^k$.

6.4.1 Analyzing the Columns of Matrix F

We show now that the span of the first k eigenvectors $\{f_i\}_{i=1}^k$ equals the span of the projection vectors $\{\widehat{f_i}\}_{i=1}^k$.

Lemma 6.8. *If $\Psi > k^{3/2}$ then the $\mathrm{span}(\{\widehat{f_i}\}_{i=1}^k) = \mathrm{span}(\{f_i\}_{i=1}^k)$ and thus each eigenvector can be expressed as $f_i = \sum_{j=1}^k \beta_j^{(i)} \cdot \widehat{f_j}$ for every $i \in \{1, \ldots, k\}$.*

To prove Lemma 6.8, we build upon the following result established by Peng et al. [PSZ17].

Lemma 6.9. *[PSZ17, Theorem 1.1 Part 1] For $P_i \subset V$ let $\overline{g_i} = \frac{D^{1/2}\chi_{P_i}}{\|D^{1/2}\chi_{P_i}\|_2}$. Then any $i \in \{1, \ldots, k\}$ it holds that*

$$\left\|\overline{g_i} - \widehat{f_i}\right\|_2^2 = \sum_{j=k+1}^n \left(\alpha_j^{(i)}\right)^2 \leqslant \frac{\mathcal{R}(\overline{g_i})}{\lambda_{k+1}} = \frac{\phi(P_i)}{\lambda_{k+1}}.$$

Our analysis crucially relies on the following two technical lemmas.

Lemma 6.10. *For every $i \in \{1, \ldots, k\}$ and $p \neq q \in \{1, \ldots, k\}$ it holds that*

$$1 - \phi(P_i)/\lambda_{k+1} \leqslant \left\|\widehat{f_i}\right\|_2^2 = \left\|\alpha^{(i)}\right\|_2^2 \leqslant 1 \quad \textit{and} \quad \left|\left\langle \widehat{f_p}, \widehat{f_q}\right\rangle\right| = |\langle \alpha^p, \alpha^q \rangle| \leqslant \frac{\sqrt{\phi(P_p) \cdot \phi(P_q)}}{\lambda_{k+1}}.$$

Proof. The first part follows by Lemma 6.9 and the following chain of inequalities

$$1 - \frac{\phi(P_i)}{\lambda_{k+1}} \leqslant 1 - \sum_{j=k+1}^n \left(\alpha_j^{(i)}\right)^2 = \left\|\widehat{f_i}\right\|_2^2 = \sum_{j=1}^k \left(\alpha_j^{(i)}\right)^2 \leqslant \sum_{j=1}^n \left(\alpha_j^{(i)}\right)^2 = 1.$$

We show now the second part. Since $\{f_i\}_{i=1}^n$ are orthonormal eigenvectors we have for all $p \neq q$ that

$$\langle f_p, f_q \rangle = \sum_{l=1}^n \alpha_\ell^{(p)} \cdot \alpha_\ell^{(q)} = 0. \tag{6.7}$$

We combine (6.7) and Cauchy-Schwarz to obtain

$$\begin{aligned}\left|\left\langle \widehat{f}_p, \widehat{f}_q\right\rangle\right| &= \left|\sum_{l=1}^{k}\alpha_\ell^{(p)}\cdot\alpha_\ell^{(q)}\right| = \left|\sum_{l=k+1}^{n}\alpha_\ell^{(p)}\cdot\alpha_\ell^{(q)}\right| \\ &\leqslant \sqrt{\sum_{l=k+1}^{n}\left(\alpha_\ell^{(p)}\right)^2}\cdot\sqrt{\sum_{l=k+1}^{n}\left(\alpha_\ell^{(q)}\right)^2} \leqslant \frac{\sqrt{\phi(P_p)\cdot\phi(P_q)}}{\lambda_{k+1}}.\end{aligned}$$

□

Lemma 6.11. *If $\Psi > k^{3/2}$ then the columns $\{F_{:,i}\}_{i=1}^{k}$ are linearly independent.*

Proof. We show that the columns of matrix F are almost orthonormal. Consider the symmetric matrix $F^{\mathrm{T}}F$. It is known that $\ker\left(F^{\mathrm{T}}F\right) = ker(F)$ and that all eigenvalues of matrix $F^{\mathrm{T}}F$ are real numbers. We proceeds by showing that the smallest eigenvalue $\lambda_{\min}(F^{\mathrm{T}}F) > 0$. This would imply that $ker(F) = \emptyset$ and hence yields the statement.

By combining Gersgorin Circle Theorem, Lemma 6.10 and Cauchy-Schwarz it holds that

$$\begin{aligned}\lambda_{\min}(F^{\mathrm{T}}F) &\geqslant \min_{i\in\{1,\ldots,k\}}\left\{\left(F^{\mathrm{T}}F\right)_{ii} - \sum_{j\neq i}^{k}\left|\left(F^{\mathrm{T}}F\right)_{ij}\right|\right\} = \min_{i\in\{1,\ldots,k\}}\left\{\left\|\alpha^{(i)}\right\|_2^2 - \sum_{j\neq i}^{k}\left|\left\langle\alpha^{(j)},\alpha^{(i)}\right\rangle\right|\right\} \\ &\geqslant 1 - \sum_{j=1}^{k}\sqrt{\frac{\phi(P_j)}{\lambda_{k+1}}}\sqrt{\frac{\phi(P_{i^\star})}{\lambda_{k+1}}} \geqslant 1 - \sqrt{k}\sqrt{\sum_{j=1}^{k}\frac{\phi(P_j)}{\lambda_{k+1}}}\sqrt{\frac{\phi(P_{i^\star})}{\lambda_{k+1}}} \geqslant 1 - \frac{k^{3/2}}{\Psi} > 0,\end{aligned}$$

where $i^\star \in \{1,\ldots,k\}$ is the index that minimizes the expression above. □

We present now the proof of Lemma 6.8.

Proof of Lemma 6.8. Let $\nu \in \mathbb{R}^k$ be an arbitrary non-zero vector. Notice that

$$\sum_{i=1}^{k}\nu_i\cdot\widehat{f}_i = \sum_{i=1}^{k}\nu_i\sum_{j=1}^{k}\alpha_j^{(i)}f_j = \sum_{j=1}^{k}\left(\sum_{i=1}^{k}\nu_i\alpha_j^{(i)}\right)f_j = \sum_{j=1}^{k}\gamma_j f_j, \quad \text{where} \quad \gamma_j = \langle F_{j,:},\nu\rangle. \tag{6.8}$$

By Lemma 6.11, the columns $\{F_{:,i}\}_{i=1}^{k}$ are linearly independent and since $\gamma = F\nu$, it follows that at least one component $\gamma_j \neq 0$. Hence, the vectors $\{\widehat{f}_i\}_{i=1}^{k}$ are linearly independent, and since each vector $\widehat{f}_i$ is a projection onto the span of the first k eigenvectors $\{f_i\}_{i=1}^{k}$, it follows that $\operatorname{span}(\{\widehat{f}_i\}_{i=1}^{k}) = \operatorname{span}(\{f_i\}_{i=1}^{k})$. Thus, each eigenvector f_i can be expressed as a linear combination of the projection vectors $\{\widehat{f}_i\}_{i=1}^{k}$. □

6.4.2 Analyzing Eigenvectors f in terms of $\widehat{f}_j$

In this section, we prove Theorem 6.5. Using Lemma 6.8, we first express each eigenvector $f_i = \sum_{j=1}^{k}\beta_j^{(i)}\cdot\widehat{f}_j$ as a linear combination of the projection vectors $\{\widehat{f}_j\}_{j=1}^{k}$, and we bound the squared ℓ_2 norm of the corresponding coefficient vector $\beta^{(i)} = B_{:,i}$ for all $i \in \{1,\ldots,k\}$. Then, we conclude the proof of Theorem 6.5.

Lemma 6.12. *If $\Psi > k^{3/2}$ then for $i \in [k]$ it holds*

$$\left(1+\frac{2k}{\Psi}\right)^{-1} \leqslant \sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2 \leqslant \left(1-\frac{2k}{\Psi}\right)^{-1}.$$

Proof. We show now the upper bound. By Lemma 6.8 $f_i = \sum_{j=1}^{k}\beta_j^{(i)}\widehat{f}_j$ for all $i \in \{1,\ldots,k\}$ and thus

$$\begin{aligned}1 &= \|f_i\|_2^2 = \left\langle\sum_{a=1}^{k}\beta_a^{(i)}\widehat{f}_a, \sum_{b=1}^{k}\beta_b^{(i)}\widehat{f}_b\right\rangle \\ &= \sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2\left\|\widehat{f}_j\right\|_2^2 + \sum_{a=1}^{k}\sum_{b\neq a}^{k}\beta_a^{(i)}\beta_b^{(i)}\left\langle\widehat{f}_a,\widehat{f}_b\right\rangle \\ &\overset{(\star)}{\geqslant} \left(1-\frac{2k}{\Psi}\right)\cdot\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2.\end{aligned}$$

To prove the inequality $(\star)$ we consider the two terms separately.

By Lemma 6.10, $\left\|\widehat{f}_j\right\|_2^2 \geqslant 1-\phi(P_j)/\lambda_{k+1}$. We then apply $\sum_i a_i b_i \leqslant (\sum_i a_i)(\sum_i b_i)$ for all non-negative vectors a, b and obtain

$$\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2\left(1-\frac{\phi(P_j)}{\lambda_{k+1}}\right)=\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2-\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2\frac{\phi(P_j)}{\lambda_{k+1}}\geqslant\left(1-\frac{k}{\Psi}\right)\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2.$$

Again by Lemma 6.10, we have $\left|\left\langle\widehat{f}_a,\widehat{f}_b\right\rangle\right| \leqslant \sqrt{\phi(P_a)\phi(P_b)}/\lambda_{k+1}$, and by Cauchy-Schwarz it holds

$$\begin{aligned}\sum_{a=1}^{k}\sum_{b\neq a}^{k}\beta_a^{(i)}\beta_b^{(i)}\left\langle\widehat{f}_a,\widehat{f}_b\right\rangle &\geqslant -\sum_{a=1}^{k}\sum_{b\neq a}^{k}\left|\beta_a^{(i)}\right|\cdot\left|\beta_b^{(i)}\right|\cdot\left|\left\langle\widehat{f}_a,\widehat{f}_b\right\rangle\right|\\ &\geqslant -\frac{1}{\lambda_{k+1}}\sum_{a=1}^{k}\sum_{b\neq a}^{k}\left|\beta_a^{(i)}\right|\sqrt{\phi(P_a)}\cdot\left|\beta_b^{(i)}\right|\sqrt{\phi(P_b)}\\ &\geqslant -\frac{1}{\lambda_{k+1}}\left(\sum_{j=1}^{k}\left|\beta_j^{(i)}\right|\sqrt{\phi(P_j)}\right)^2\geqslant-\frac{k}{\Psi}\cdot\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2.\end{aligned}$$

The lower bound follows by analogous arguments. □

We are now ready to prove Theorem 6.5.

Proof of Theorem 6.5. By Lemma 6.8, we have $f_i=\sum_{j=1}^{k}\beta_j^{(i)}\widehat{f}_j$ and recall that $\widehat{g}_i=\sum_{j=1}^{k}\beta_j^{(i)}\overline{g_j}$ for all $i\in\{1,\ldots,k\}$. Further, by combining triangle inequality, Cauchy-Schwarz, Lemma 6.9 and Lemma 6.12, we obtain that

$$\begin{aligned}\|f_i-\widehat{g}_i\|_2^2&=\left\|\sum_{j=1}^{k}\beta_j^{(i)}\left(\widehat{f}_j-\overline{g_j}\right)\right\|_2^2\leqslant\left(\sum_{j=1}^{k}|\beta_j^i|\cdot\left\|\widehat{f}_j-\overline{g_j}\right\|_2\right)^2\\ \leqslant&\left(\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2\right)\cdot\left(\sum_{j=1}^{k}\left\|\widehat{f}_j-\overline{g_j}\right\|_2^2\right)\leqslant\left(1-\frac{2k}{\Psi}\right)^{-1}\left(\frac{1}{\lambda_{k+1}}\sum_{j=1}^{k}\phi(P_j)\right)\\ =&\left(1-\frac{2k}{\Psi}\right)^{-1}\cdot\frac{k}{\Psi}\leqslant\left(1+\frac{3k}{\Psi}\right)\cdot\frac{k}{\Psi},\end{aligned}$$

where the last inequality uses $\Psi > 4k$. □

6.5 Spectral Properties of Matrix B

In this Section, we prove Theorem 6.1 in two steps. In Subsection 6.5.1, we analyzes the column space of matrix B and we show that matrix $B^{\mathrm{T}}B$ is close to the identity matrix. Then, in Subsection 6.5.2, we analyze the row space of matrix B and we prove that matrix BB^{T} is close to the identity matrix.

6.5.1 Analyzing the Column Space of Matrix B

We show below that the matrix $B^{\mathrm{T}}B$ is close to the identity matrix.

Lemma 6.13. *(Columns) If $\Psi > 4\cdot k^{3/2}$ then for all distinct $i, j \in \{1,\ldots,k\}$ it holds*

$$1-\frac{3k}{\Psi}\leqslant\langle B_{:,i},B_{:,i}\rangle\leqslant1+\frac{3k}{\Psi}\quad\textit{and}\quad|\langle B_{:,i},B_{:,j}\rangle|\leqslant4\sqrt{\frac{k}{\Psi}}.$$

Proof. By Lemma 6.12 it holds that

$$1-\frac{3k}{\Psi}\leqslant\langle B_{:,i},B_{:,i}\rangle=\sum_{j=1}^{k}\left(\beta_j^{(i)}\right)^2\leqslant1+\frac{3k}{\Psi}.$$

Recall that $\widehat{g_i} = \sum_{j=1}^{k} \beta_j^{(i)} \cdot \overline{g_j}$. Moreover, since the eigenvectors $\{f_i\}_{i=1}^{k}$ and the characteristic vectors $\{\overline{g_i}\}_{i=1}^{k}$ are orthonormal by combing Cauchy-Schwarz and by Theorem 6.5 it holds

$$\begin{aligned}
|\langle B_{:,i}, B_{:,j}\rangle| &= \sum_{l=1}^{k} \beta_\ell^{(i)} \beta_\ell^{(j)} = \left\langle \sum_{a=1}^{k} \beta_a^{(i)} \cdot \overline{g_a}, \sum_{b=1}^{k} \beta_b^{(j)} \cdot \overline{g_b} \right\rangle = \langle \widehat{g_i}, \widehat{g_j}\rangle \\
&= \langle (\widehat{g_i} - f_i) + f_i, (\widehat{g_j} - f_j) + f_j \rangle \\
&= \langle \widehat{g_i} - f_i, \widehat{g_j} - f_j\rangle + \langle \widehat{g_i} - f_i, f_j\rangle + \langle f_i, \widehat{g_j} - f_j\rangle \\
&\leqslant \|\widehat{g_i} - f_i\|_2 \cdot \|\widehat{g_j} - f_j\|_2 + \|\widehat{g_i} - f_i\|_2 + \|\widehat{g_j} - f_j\|_2 \\
&\leqslant \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k}{\Psi} + 2\sqrt{\left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k}{\Psi}} \leqslant 4\sqrt{\frac{k}{\Psi}}.
\end{aligned}$$

□

We demonstrate now that the columns of matrix B are linearly independent.

Lemma 6.14. *If $\Psi > 25 \cdot k^3$ then the columns $\{B_{:,i}\}_{i=1}^{k}$ are linearly independent.*

Proof. Since $\ker(B) = \ker\left(B^{\mathrm{T}}B\right)$ and $B^{\mathrm{T}}B$ is SPSD[1] matrix, it suffices to show that the smallest eigenvalue

$$\lambda(B^{\mathrm{T}}B) = \min_{x \neq 0} \frac{x^{\mathrm{T}} B^{\mathrm{T}} B x}{x^{\mathrm{T}} x} > 0.$$

By Lemma 6.13,

$$\sum_{i=1}^{k} \sum_{j \neq i}^{k} |x_i|\,|x_j| \left| \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle \right| \leqslant 4\sqrt{\frac{k}{\Psi}} \left(\sum_{i=1}^{k} |x_i| \right)^2 \leqslant \|x\|_2^2 \cdot 4k\sqrt{\frac{k}{\Psi}},$$

and

$$\begin{aligned}
x^{\mathrm{T}} B^{\mathrm{T}} B x &= \left\langle \sum_{i=1}^{k} x_i \beta^{(i)}, \sum_{j=1}^{k} x_j \beta^{(j)} \right\rangle = \sum_{i=1}^{k} x_i^2 \left\| \beta^{(i)} \right\|_2^2 + \sum_{i=1}^{k} \sum_{j \neq i}^{k} x_i x_j \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle \\
&\geqslant \left(1 - \frac{3k}{\Psi}\right) \|x\|_2^2 - \sum_{i=1}^{k} \sum_{j \neq i}^{k} |x_i|\,|x_j| \left| \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle \right| \geqslant \left(1 - 5k\sqrt{\frac{k}{\Psi}}\right) \cdot \|x\|_2^2.
\end{aligned}$$

Hence, $\lambda(B^{\mathrm{T}}B) > 0$ and the statement follows. □

6.5.2 Analyzing the Row Space of Matrix B

In this section, we show that matrix BB^{T} is close to the identity matrix. We bound now the squared ℓ_2 norm of the rows in matrix B, i.e. the diagonal entries in matrix BB^{T}.

Lemma 6.15. *(Rows) If $\Psi \geqslant 400 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0,1)$ then for all distinct $i, j \in \{1, \ldots, k\}$ it holds*

$$1 - \varepsilon \leqslant \langle B_{i,:}, B_{i,:}\rangle \leqslant 1 + \varepsilon.$$

Proof. We show that the eigenvalues of matrix BB^{T} are concentrated around 1. This would imply that $\chi_i^{\mathrm{T}} BB^{\mathrm{T}} \chi_i = \langle B_{i,:}, B_{i,:}\rangle \approx 1$, where χ_i is a characteristic vector. By Lemma 6.13 we have

$$\left(1 - \frac{3k}{\Psi}\right)^2 \leqslant \left(\beta^{(i)}\right)^{\mathrm{T}} \cdot BB^{\mathrm{T}} \cdot \beta^{(i)} = \left\| \beta^{(i)} \right\|_2^4 + \sum_{j \neq i}^{k} \left\langle \beta^{(j)}, \beta^{(i)} \right\rangle^2 \leqslant \left(1 + \frac{3k}{\Psi}\right)^2 + \frac{16k^2}{\Psi} \leqslant 1 + \frac{23k^2}{\Psi}$$

and

$$\left| \left(\beta^{(i)}\right)^{\mathrm{T}} \cdot BB^{\mathrm{T}} \cdot \beta^{(j)} \right| \leqslant \sum_{l=1}^{k} \left| \left\langle \beta^{(i)}, \beta^{(l)} \right\rangle \right| \left| \left\langle \beta^{(l)}, \beta^{(j)} \right\rangle \right| \leqslant 8\left(1 + \frac{3k}{\Psi}\right)\sqrt{\frac{k}{\Psi}} + 16\frac{k^2}{\Psi} \leqslant 11\sqrt{\frac{k}{\Psi}}.$$

[1] We denote by SPSD the class of symmetric positive semi-definite matrices.

By Lemma 6.14 every vector $x \in \mathbb{R}^k$ can be expressed as $x = \sum_{i=1}^k \gamma_i \beta^{(i)}$.

$$\begin{aligned} x^{\mathrm{T}} B B^{\mathrm{T}} x &= \sum_{i=1}^{k} \gamma_i \left(\beta^{(i)}\right)^{\mathrm{T}} \cdot BB^{\mathrm{T}} \cdot \sum_{j=1}^{k} \gamma_j \beta^{(j)} \\ &= \sum_{i=1}^{k} \gamma_i^2 \left(\beta^{(i)}\right)^{\mathrm{T}} \cdot BB^{\mathrm{T}} \cdot \beta^{(i)} + \sum_{i=1}^{k}\sum_{j\neq i}^{k} \gamma_i\gamma_j \left(\beta^{(i)}\right)^{\mathrm{T}} \cdot BB^{\mathrm{T}} \cdot \beta^{(j)} \\ &\geqslant \left(1 - \frac{23k^2}{\Psi} - 11 \cdot k\sqrt{\frac{k}{\Psi}}\right) \|\gamma\|_2^2 \geqslant \left(1 - 14 \cdot k\sqrt{\frac{k}{\Psi}}\right) \|\gamma\|_2^2 . \end{aligned}$$

and

$$x^{\mathrm{T}} x = \sum_{i=1}^{k}\sum_{j=1}^{k} \gamma_i\gamma_j \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle = \sum_{i=1}^{k} \gamma_i^2 \left\|\beta^{(i)}\right\|_2^2 + \sum_{i=1}^{k}\sum_{j\neq i}^{k} \gamma_i\gamma_j \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle$$

By Lemma 6.13 we have $\left|\sum_{i=1}^{k}\sum_{j\neq i}^{k} \gamma_i\gamma_j \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle\right| \leqslant \|\gamma\|_2^2 \cdot 4k\sqrt{\frac{k}{\Psi}}$ and $\left\|\beta^{(i)}\right\|_2^2 \leqslant 1 + \frac{3k}{\Psi}$. Thus, it holds

$$\left(1 - 5k\sqrt{\frac{k}{\Psi}}\right) \|\gamma\|_2^2 \leqslant x^{\mathrm{T}} x \leqslant \left(1 + 5k\sqrt{\frac{k}{\Psi}}\right) \|\gamma\|_2^2 .$$

Hence, we have

$$1 - 20k\sqrt{\frac{k}{\Psi}} \leqslant \lambda(BB^{\mathrm{T}}) \leqslant 1 + 20k\sqrt{\frac{k}{\Psi}}.$$

□

This proves the first part of Theorem 6.1. We turn now to the second part and restate it in the following Lemma.

Lemma 6.16. *(Rows) If $\Psi \geqslant 10^4 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0,1)$ then for all distinct $i, j \in \{1, \ldots, k\}$ it holds*

$$|\langle B_{i,:}, B_{j,:} \rangle| \leqslant \sqrt{\varepsilon}.$$

Let $E \in \mathbb{R}^{k\times k}$ be a symmetric matrix such that $B^{\mathrm{T}} B = I + E$ and $|E_{ij}| \leqslant 4\sqrt{k/\Psi}$. Then,

$$\left(BB^{\mathrm{T}}\right)^2 = B\left(I+E\right)B^{\mathrm{T}} = BB^{\mathrm{T}} + BEB^{\mathrm{T}}. \tag{6.9}$$

We show next that the absolute value of every eigenvalue of matrix BEB^{T} is small, and further demonstrate that this implies that all entries of matrix BEB^{T} are small. Then, we conclude the proof of Lemma 6.16.

Lemma 6.17. *If $\Psi \geqslant 40^2 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0,1)$, then every eigenvalue λ of matrix BEB^{T} satisfies*

$$\left|\lambda(BEB^{\mathrm{T}})\right| \leqslant \varepsilon/5.$$

Proof. Let $z = B^{\mathrm{T}} x$. We upper bound the quadratic form

$$\left|x^{\mathrm{T}} BEB^{\mathrm{T}} x\right| = \left|z^{\mathrm{T}} E z\right| \leqslant \sum_{ij} |E_{ij}|\,|z_i|\,|z_j| \leqslant 4\sqrt{\frac{k}{\Psi}} \cdot \left(\sum_{i=1}^{k} |z_i|\right)^2 \leqslant \|z\|_2^2 \cdot 4k\sqrt{\frac{k}{\Psi}}.$$

By Lemma 6.15, we have $1 - \varepsilon \leqslant \lambda(BB^{\mathrm{T}}) \leqslant 1 + \varepsilon$ and since $\|z\|_2^2 = \frac{xBB^{\mathrm{T}}x}{x^{\mathrm{T}}x} \cdot \|x\|_2^2$, it follows that

$$\frac{\|z\|_2^2}{1+\varepsilon} \leqslant \|x\|_2^2 \leqslant \frac{\|z\|_2^2}{1-\varepsilon},$$

and hence

$$\left|\lambda(BEB^{\mathrm{T}})\right| \leqslant \max_x \frac{\left|x^{\mathrm{T}} BEB^{\mathrm{T}} x\right|}{x^{\mathrm{T}} x} \leqslant 4\left(1+\varepsilon\right) \cdot k\sqrt{\frac{k}{\Psi}} \leqslant \varepsilon/5.$$

□

Lemma 6.18. *If $\Psi \geqslant 40^2 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0,1)$, then it holds that $|(BEB^{\mathrm{T}})_{ij}| \leqslant \varepsilon/5$ for every $i,j \in \{1,\dots,k\}$.*

Proof. Since matrix $E \in \mathbb{R}^{k\times k}$ is a symmetric, by construction matrix, $BEB^{\mathrm{T}} \in \mathbb{R}^{k\times k}$ is also symmetric. Using the SVD Theorem, there is an orthonormal basis $\{u_i\}_{i=1}^k$ such that $BEB^{\mathrm{T}} = \sum_{i=1}^k \lambda_i(BEB^{\mathrm{T}})\cdot u_i u_i^{\mathrm{T}}$. Thus, it suffices to bound the expression

$$|(BEB^{\mathrm{T}})_{ij}| \leqslant \sum_{l=1}^{k} |\lambda_\ell(BEB^{\mathrm{T}})| \cdot |(u_\ell u_\ell^{\mathrm{T}})_{ij}|.$$

Let $U \in \mathbb{R}^{k\times k}$ be a square matrix whose i-th column is vector u_i. By construction, matrix U is orthogonal and satisfies $U^{\mathrm{T}}U = I = UU^{\mathrm{T}}$. In particular, it holds that $\|U_{i,:}\|_2^2 = 1$, for all i. Therefore, we have

$$\sum_{l=1}^{k} |(u_\ell)_i| \cdot |(u_\ell)_j| \leqslant \sqrt{\|U_{i,:}\|_2^2}\sqrt{\|U_{j,:}\|_2^2} = 1.$$

We apply now Lemma 6.17 to obtain

$$\sum_{l=1}^{k} |\lambda_\ell(BEB^{\mathrm{T}})| \cdot |(u_\ell u_\ell^{\mathrm{T}})_{ij}| \leqslant \frac{\varepsilon}{5} \cdot \sum_{l=1}^{k} |(u_\ell)_i| \cdot |(u_\ell)_j| \leqslant \frac{\varepsilon}{5}.$$

□

We are now ready to prove Lemma 6.16.

Proof of Lemma 6.16. By (6.9) we have $\left(BB^{\mathrm{T}}\right)^2 = BB^{\mathrm{T}} + BEB^{\mathrm{T}}$. Observe that the (i,j)-th entry of matrix BB^{T} is equal to the inner product between the i-th and j-th row of matrix B, i.e. $\left(BB^{\mathrm{T}}\right)_{ij} = \langle B_{i,:}, B_{j,:}\rangle$. Moreover, we have

$$\left[\left(BB^{\mathrm{T}}\right)^2\right]_{ij} \quad = \quad \sum_{l=1}^{k} \left(BB^{\mathrm{T}}\right)_{i,l} \left(BB^{\mathrm{T}}\right)_{l,j} = \sum_{l=1}^{k} \langle B_{i,:}, B_{l,:}\rangle \langle B_{l,:}, B_{j,:}\rangle .$$

For the entries on the main diagonal, it holds

$$\langle B_{i,:}, B_{i,:}\rangle^2 + \sum_{l\neq i}^{k} \langle B_{i,:}, B_{l,:}\rangle^2 = [(BB^{\mathrm{T}})^2]_{ii} = [BB^{\mathrm{T}} + BEB^{\mathrm{T}}]_{ii} = \langle B_{i,:}, B_{i,:}\rangle + \left(BEB^{\mathrm{T}}\right)_{ii},$$

and hence by applying Lemma 6.15 with $\varepsilon' = \varepsilon/5$ and Lemma 6.18 with $\varepsilon' = \varepsilon$ we obtain

$$\langle B_{i,:}, B_{j,:}\rangle^2 \leqslant \sum_{l\neq i} \langle B_{i,:}, B_{l,:}\rangle^2 \leqslant \left(1+\frac{\varepsilon}{5}\right) + \frac{\varepsilon}{5} - \left(1-\frac{\varepsilon}{5}\right)^2 \leqslant \varepsilon.$$

□

6.6 Volume Overlap Lemma

In this section, we prove Lemma 6.4. Our main technical contribution is to strengthen the lower bound of k-means cost in [PSZ17, Lemma 4.5] by a factor of k, under the weaken gap assumption.

We begin by stating a useful Corollary of Lemma 6.3.

Corollary 6.19. *Let $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0,1/2]$. Suppose c_i is the center of a cluster A_i. If $\left\|c_i - p^{(i_1)}\right\|_2 \geqslant \left\|c_i - p^{(i_2)}\right\|_2$, then it holds that*

$$\left\|c_i - p^{(i_1)}\right\|_2^2 \geqslant \frac{1}{4}\left\|p^{(i_1)} - p^{(i_2)}\right\|_2^2 \geqslant \left[8 \cdot \min\left\{\mu(P_{i_1}), \mu(P_{i_2})\right\}\right]^{-1}.$$

We restate now [PSZ17, Lemma 4.6] whose analysis crucially relies on the following function

$$\sigma(\ell) = \arg\max_{j\in\{1,\dots,k\}} \frac{\mu(A_\ell \cap P_j)}{\mu(P_j)}, \quad \text{for all } \ell \in \{1,\dots,k\}. \tag{6.10}$$

Lemma 6.20. *[PSZ17, Lemma 4.6] Let $(P_1, \ldots, P_k)$ and $(A_1, \ldots, A_k)$ be k-way node partitions of G. Suppose for every permutation $\pi : \{1, \ldots, k\} \to \{1, \ldots, k\}$ there is an index $i \in \{1, \ldots, k\}$ such that*

$$\mu(A_i \triangle P_{\pi(i)}) \geqslant 2\varepsilon \cdot \mu(P_{\pi(i)}), \tag{6.11}$$

where $\varepsilon \in (0, 1/2)$ is a parameter. Then one of the following three statements holds:
1. If σ is a permutation and $\mu(P_{\sigma(i)} \backslash A_i) \geqslant \varepsilon \cdot \mu(P_{\sigma(i)})$, then for every index $j \neq i$ there is a real $\varepsilon_j \geqslant 0$ such that

$$\mu(A_j \cap P_{\sigma(j)}) \geqslant \mu(A_j \cap P_{\sigma(i)}) \geqslant \varepsilon_j \cdot \min\{\mu(P_{\sigma(j)}), \mu(P_{\sigma(i)})\},$$

and $\sum_{j \neq i} \varepsilon_j \geqslant \varepsilon$.
2. If σ is a permutation and $\mu(A_i \backslash P_{\sigma(i)}) \geqslant \varepsilon \cdot \mu(P_{\sigma(i)})$, then for every $j \neq i$ there is a real $\varepsilon_j \geqslant 0$ such that

$$\mu(A_i \cap P_{\sigma(i)}) \geqslant \varepsilon_j \cdot \mu(P_{\sigma(i)}), \quad \mu(A_i \cap P_{\sigma(j)}) \geqslant \varepsilon_j \cdot \mu(P_{\sigma(i)}),$$

and $\sum_{j \neq i} \varepsilon_j \geqslant \varepsilon$.
3. If σ is not a permutation, then there is an index $\ell \notin \{\sigma(1), \ldots, \sigma(k)\}$ and for every index j there is a real $\varepsilon_j \geqslant 0$ such that

$$\mu(A_j \cap P_{\sigma(j)}) \geqslant \mu(A_j \cap P_\ell) \geqslant \varepsilon_j \cdot \min\{\mu(P_{\sigma(j)}), \mu(P_\ell)\},$$

and $\sum_{j=1}^{k} \varepsilon_j = 1$.

We strengthen now the lower bound of k-means cost in [PSZ17, Lemma 4.5] by a factor of k.

Lemma 6.21. *Suppose the hypothesis of Lemma 6.20 is satisfied and $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, 1/2]$. Then it holds*

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) \geqslant \frac{\varepsilon}{16} - \frac{2k^2}{\Psi}.$$

Proof. By definition

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) = \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{u \in A_i \cap P_j} d(u) \left\| \mathcal{F}(u) - c_i \right\|_2^2 \stackrel{\text{def}}{=} \Lambda. \tag{6.12}$$

Since for every vectors $x, y, z \in \mathbb{R}^k$ it holds

$$2 \left(\|x - y\|_2^2 + \|z - y\|_2^2 \right) \geqslant (\|x - y\|_2 + \|z - y\|_2)^2 \geqslant \|x - z\|_2^2 ,$$

we have for all indices $i, j \in \{1, \ldots, k\}$ that

$$\|\mathcal{F}(u) - c_i\|_2^2 \geqslant \frac{\left\|p^{(j)} - c_i\right\|_2^2}{2} - \left\|\mathcal{F}(u) - p^{(j)}\right\|_2^2 . \tag{6.13}$$

Our proof proceeds by considering three cases. Let $i \in \{1, \ldots, k\}$ be the index from the hypothesis in Lemma 6.20.

Case 1. Suppose the first conclusion of Lemma 6.20 holds. For every index $j \neq i$ let

$$p^{\gamma(j)} = \begin{cases} p^{\sigma(j)} & \text{, if } \left\|p^{\sigma(j)} - c_j\right\|_2 \geqslant \left\|p^{\sigma(i)} - c_j\right\|_2 ; \\ p^{\sigma(i)} & \text{, otherwise.} \end{cases}$$

Then by combining (6.13), Corollary 6.19 and Lemma 6.6, we have

$$\begin{aligned} \Lambda &\geqslant \frac{1}{2} \sum_{j \neq i} \sum_{u \in A_j \cap P_{\gamma(j)}} d(u) \left\|p^{\gamma(j)} - c_j\right\|_2^2 - \sum_{j \neq i} \sum_{u \in A_j \cap P_{\gamma(j)}} \left\|\mathcal{F}(u) - p^{\gamma(j)}\right\|_2^2 \\ &\geqslant \frac{1}{16} \sum_{j \neq i} \frac{\mu(A_j \cap P_{\gamma(j)})}{\min\{\mu(P_{\sigma(i)}), \mu(P_{\sigma(j)})\}} - \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi} \geqslant \frac{\varepsilon}{16} - \frac{2k^2}{\Psi}. \end{aligned}$$

Case 2. Suppose the second conclusion of Lemma 6.20 holds. Notice that if $\mu(A_i \cap P_{\sigma(i)}) \leqslant (1 - \varepsilon) \cdot \mu(P_{\sigma(i)})$ then $\mu(P_{\sigma(i)} \backslash A_i) \geqslant \varepsilon \cdot \mu(P_{\sigma(i)})$ and thus we can argue as in Case 1. Hence, we can assume that it holds

$$\mu(A_i \cap P_{\sigma(i)}) \geqslant (1 - \varepsilon) \cdot \mu(P_{\sigma(i)}). \tag{6.14}$$

We proceed by analyzing two subcases.

a) If $\left\|p^{\sigma(j)} - c_i\right\|_2 \geqslant \left\|p^{\sigma(i)} - c_i\right\|$ holds for all $j \neq i$ then by combining (6.13), Corollary 6.19 and Lemma 6.6 it follows

$$\begin{aligned}\Lambda &\geqslant \frac{1}{2}\sum_{j\neq i}\sum_{u\in A_i\cap P_{\sigma(j)}} d(u)\left\|p^{\sigma(j)} - c_i\right\|_2^2 - \sum_{j\neq i}\sum_{u\in A_i\cap P_{\sigma(j)}} \left\|\mathcal{F}(u) - p^{\sigma(j)}\right\|_2^2 \\ &\geqslant \frac{1}{2}\sum_{j\neq i}\frac{\mu(A_i\cap P_{\sigma(j)})}{\min\{\mu(P_{\sigma(i)}),\mu(P_{\sigma(j)})\}} - \left(1+\frac{3k}{\Psi}\right)\cdot\frac{k^2}{\Psi} \geqslant \frac{\varepsilon}{16} - \frac{2k^2}{\Psi}.\end{aligned}$$

b) Suppose there is an index $j \neq i$ such that $\left\|p^{\sigma(j)} - c_i\right\|_2 < \left\|p^{\sigma(i)} - c_i\right\|$. Then by triangle inequality combined with Corollary 6.19 we have

$$\left\|p^{\sigma(i)} - c_i\right\|_2^2 \geqslant \frac{1}{4}\left\|p^{\sigma(i)} - p^{\sigma(j)}\right\|_2 \geqslant \left[8\cdot\min\{\mu(P_{\sigma(i)}),\mu(P_{\sigma(j)})\}\right]^{-1}.$$

Thus, by combining (6.13), (6.14) and Lemma 6.6 we obtain

$$\begin{aligned}\Lambda &\geqslant \frac{1}{2}\sum_{u\in A_i\cap P_{\sigma(i)}} d(u)\left\|p^{\sigma(i)} - c_i\right\|_2^2 - \sum_{u\in A_i\cap P_{\sigma(i)}} d(u)\left\|\mathcal{F}(u) - p^{\sigma(i)}\right\|_2^2 \\ &\geqslant \frac{1}{16}\cdot\frac{\mu(A_i\cap P_{\sigma(i)})}{\min\{\mu(P_{\sigma(i)}),\mu(P_{\sigma(j)})\}} - \left(1+\frac{3k}{\Psi}\right)\cdot\frac{k^2}{\Psi} \geqslant \frac{1-\varepsilon}{16} - \frac{2k^2}{\Psi}.\end{aligned}$$

Case 3. Suppose the third conclusion of Lemma 6.20 holds, i.e., σ is not a permutation. Then there is an index $\ell \in \{1,\ldots,k\}\setminus\{\sigma(1),\ldots,\sigma(k)\}$ and for every index $j\in\{1,\ldots,k\}$ let

$$p^{\gamma(j)} = \begin{cases} p^{\ell} & \text{, if } \left\|p^{\ell} - c_j\right\|_2 \geqslant \left\|p^{\sigma(j)} - c_j\right\|_2; \\ p^{\sigma(j)} & \text{, otherwise.}\end{cases}$$

By combining (6.13), Corollary 6.19 and Lemma 6.6 it follows that

$$\begin{aligned}\Lambda &\geqslant \frac{1}{2}\sum_{j=1}^{k}\sum_{u\in A_j\cap P_{\gamma(j)}} d(u)\left\|p^{\gamma(j)} - c_j\right\|_2^2 - \sum_{j=1}^{k}\sum_{u\in A_j\cap P_{\gamma(j)}} d(u)\left\|\mathcal{F}(u) - p^{\gamma(j)}\right\|_2^2 \\ &\geqslant \frac{1}{16}\sum_{j=1}^{k}\frac{\mu(A_j\cap P_{\gamma(j)})}{\min\{\mu(P_{\sigma(j)}),\mu(P_{\ell})\}} - \left(1+\frac{3k}{\Psi}\right)\cdot\frac{k^2}{\Psi} \geqslant \frac{1}{16} - \frac{2k^2}{\Psi}.\end{aligned}$$

□

We are now ready to prove Lemma 6.4.

Proof of Lemma 6.4. We apply Lemma 6.20 with $\varepsilon' = \varepsilon/k$. Then, by Lemma 6.21 we have

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^{k}) \geqslant \frac{\varepsilon}{16k} - \frac{2k^2}{\Psi},$$

and the desired result follows by setting $\varepsilon \geqslant 64\alpha\cdot k^3/\Psi$. □

Chapter 7

Analysis of Approximate Spectral Clustering

7.1 Normalized Spectral Embedding

In this section, we prove Theorem 5.4, showing that the normalized SE Y' is ε-separated. For convenience of the reader, we restate the result.

Theorem 5.4 (from page 62)**.** *Let G be a graph that satisfies $\Psi = 20^4 \cdot k^3/\delta$, $\delta \in (0, 1/2]$ and $k/\delta \geqslant 10^9$. Then for $\varepsilon = 6 \cdot 10^{-7}$ it holds*

$$\triangle_k(Y') \leqslant \varepsilon^2 \cdot \triangle_{k-1}(Y'). \tag{7.1}$$

We establish first a lower bound on $\triangle_{k-1}(Y')$.

Lemma 7.1. *Let G be a graph that satisfies $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, 1/2]$. Then for $\delta' = 2\delta/20^4$ it holds*

$$\triangle_{k-1}(Y') \geqslant \frac{1}{12} - \frac{\delta'}{k}. \tag{7.2}$$

Before we present the proof of Lemma 7.1, we show that it implies (7.1). By Lemma 6.6, we have

$$\triangle_k(Y') \leqslant \frac{2k^2}{\Psi} = \frac{\delta'}{k},$$

and thus, by applying Lemma 7.1 with $k/\delta \geqslant 10^9$ and $\varepsilon = 6 \cdot 10^{-7}$, we obtain

$$\triangle_{k-1}(Y') \geqslant \frac{1}{12} - \frac{\delta'}{k} = \frac{1}{12} - \frac{2}{20^4} \cdot \frac{\delta}{k} \geqslant \frac{10^{10}}{9 \cdot 2^5} \cdot \frac{\delta}{k} = \frac{1}{\varepsilon^2} \cdot \frac{\delta'}{k} \geqslant \frac{1}{\varepsilon^2} \cdot \triangle_k(Y').$$

7.1.1 Proof of Lemma 7.1

We argue in a similar manner as in Lemma 6.21 (c.f. Case 3). We start by giving some notation, then we establish Lemma 7.2 and apply it in the proof of Lemma 7.1.

We redefine the function σ, see (6.10), such that for any two partitions $(P_1, \ldots, P_k)$ and $(Z_1, \ldots, Z_{k-1})$ of V, we define a mapping $\sigma : \{1, \ldots, k-1\} \mapsto \{1, \ldots, k\}$ by

$$\sigma(i) = \arg \max_{j \in \{1,\ldots,k\}} \frac{\mu(Z_i \cap P_j)}{\mu(P_j)}, \quad \text{for every } i \in \{1, \ldots, k-1\}.$$

We lower bound now the overlapping of clusters between any k-way and $(k-1)$-way partitions of V in terms of volume.

Lemma 7.2. *Suppose $(P_1, \ldots, P_k)$ and $(Z_1, \ldots, Z_{k-1})$ are partitions of V. Then for any index $\ell \in \{1, \ldots, k\} \setminus \{\sigma(1), \ldots, \sigma(k-1)\}$ (there is at least one such ℓ) and for every $i \in \{1, \ldots, k-1\}$ it holds*

$$\left\{\mu(Z_i \cap P_{\sigma(i)}), \mu(Z_i \cap P_\ell)\right\} \geqslant \tau_i \cdot \min\left\{\mu(P_\ell), \mu(P_{\sigma(i)})\right\},$$

where $\sum_{i=1}^{k-1} \tau_i = 1$ and $\tau_i \geqslant 0$.

Proof. By pigeonhole principle there is an index $\ell \in \{1, \ldots, k\}$ such that $\ell \notin \{\sigma(1), \ldots, \sigma(k-1)\}$. Thus, for every $i \in \{1, \ldots, k-1\}$ we have $\sigma(i) \neq \ell$ and

$$\frac{\mu(Z_i \cap P_{\sigma(i)})}{\mu(P_{\sigma(i)})} \geqslant \frac{\mu(Z_i \cap P_\ell)}{\mu(P_\ell)} \overset{\text{def}}{=} \tau_i,$$

where $\sum_{i=1}^{k-1} \tau_i = 1$ and $\tau_i \geqslant 0$ for all i. Hence, the statement follows. □

We present now the proof of Lemma 7.1.

Proof of Lemma 7.1. Let $(Z_1, \ldots, Z_{k-1})$ be a $(k-1)$-way partition of V with centers $c'_1, \ldots, c'_{k-1}$ that achieves $\triangle_{k-1}(Y')$, and $(P_1, \ldots, P_k)$ be a k-way partition of V achieving $\widehat{\rho}_{\mathrm{avr}}(k)$. Our goal now is to lower bound the optimum $(k-1)$-means cost

$$\triangle_{k-1}(Y') = \sum_{i=1}^{k-1} \sum_{j=1}^{k} \sum_{u \in Z_i \cap P_j} d_u \left\| \mathcal{F}(u) - c'_i \right\|_2^2 . \tag{7.3}$$

By Lemma 7.2 there is an index $\ell \in \{1, \ldots, k\} \setminus \{\sigma(1), \ldots, \sigma(k-1)\}$. For $i \in \{1, \ldots, k-1\}$ let

$$p^{\gamma(i)} = \begin{cases} p^{\ell} & \text{, if } \left\|p^{\ell} - c'_i\right\|_2 \geqslant \left\|p^{\sigma(i)} - c'_i\right\|_2 ; \\ p^{\sigma(i)} & \text{, otherwise.} \end{cases}$$

Then by combining Corollary 6.19 and Lemma 7.2, we have

$$\left\|p^{\gamma(i)} - c'_i\right\|_2^2 \geqslant \left[8 \cdot \min\left\{\mu(P_\ell), \mu(P_{\sigma(i)})\right\}\right]^{-1} \text{ and } \mu(Z_i \cap P_{\gamma(i)}) \geqslant \tau_i \cdot \min\left\{\mu(P_\ell), \mu(P_{\sigma(i)})\right\}, \tag{7.4}$$

where $\sum_{i=1}^{k-1} \tau_i = 1$. We now lower bound the expression in (7.3). Since

$$\left\|\mathcal{F}(u) - c'_i\right\|_2^2 \geqslant \frac{1}{2} \left\|p^{\gamma(i)} - c'_i\right\|_2^2 - \left\|\mathcal{F}(u) - p^{\gamma(i)}\right\|_2^2,$$

it follows for $\delta' = 2\delta/20^4$ that

$$\begin{aligned} \triangle_{k-1}(\mathcal{X}_V) &= \sum_{i=1}^{k-1} \sum_{j=1}^{k} \sum_{u \in Z_i \cap P_j} d_u \left\|\mathcal{F}(u) - c'_i\right\|_2^2 \geqslant \sum_{i=1}^{k-1} \sum_{u \in Z_i \cap P_{\gamma(i)}} d_u \left\|\mathcal{F}(u) - c'_i\right\|_2^2 \\ &\geqslant \frac{1}{2} \sum_{i=1}^{k-1} \sum_{u \in Z_i \cap P_{\gamma(i)}} d_u \left\|p^{\gamma(i)} - c'_i\right\|_2^2 - \sum_{i=1}^{k-1} \sum_{u \in Z_i \cap P_{\gamma(i)}} d_u \left\|\mathcal{F}(u) - p^{\gamma(i)}\right\|_2^2 \\ &\geqslant \frac{1}{2} \sum_{i=1}^{k-1} \frac{\mu(Z_i \cap P_{\gamma(i)})}{8 \cdot \min\left\{\mu(P_{\gamma(i)}), \mu(P_{\sigma(i)})\right\}} - \sum_{i=1}^{k} \sum_{u \in P_i} d_u \left\|\mathcal{F}(u) - p^i\right\|_2^2 \\ &\geqslant \frac{1}{16} - \frac{\delta'}{k}, \end{aligned}$$

where the last inequality holds due to (7.4) and Lemma 6.6. □

7.2 Approximate Normalized Spectral Embedding

In this section, we prove Theorem 5.6, which shows that the approximate normalized SE $\widetilde{Y'}$, computed via the Power method, is ε-separated.

Before we state our results, we need some notation. Let X'_{opt} be an *indicator* matrix, see (5.11), corresponding to an optimal k-way row partition of the normalized SE Y'. Then, the optimum k-means cost of Y' in matrix notation reads

$$\triangle_k(Y') = \|Y' - X'_{\mathrm{opt}}(X'_{\mathrm{opt}})^{\mathrm{T}} Y'\|_F^2.$$

Similarly, for the approximate normalized SE $\widetilde{Y'}$, let $\widetilde{X'_{\mathrm{opt}}}$ be an indicator matrix such that

$$\triangle_k(\widetilde{Y'}) = \|\widetilde{Y'} - \widetilde{X'_{\mathrm{opt}}}(\widetilde{X'_{\mathrm{opt}}})^{\mathrm{T}} \widetilde{Y'}\|_F^2.$$

In Subsection 7.2.1, using techniques from [BKG15, Lemma 5] and [BM14, Lemma 7], we prove the following statement.

Lemma 7.3. *Let λ_k and λ_{k+1} be the k-th and $(k+1)$-st smallest eigenvalue of $\mathcal{L}_G$, Y be the canonical SE, and $S \in \mathbb{R}^{n\times k}$ be a matrix whose entries are i.i.d. samples from the standard Gaussian distribution. For any $\beta, \varepsilon \in (0,1)$ and $p \geqslant \ln(8nk/\varepsilon\beta)/\ln(1/\gamma_k)$, where $\gamma_k = \frac{2-\lambda_{k+1}}{2-\lambda_k} < 1$, compute the approximate SE $\widetilde{Y}$ via the Power method:*

$$1)\ M \overset{\text{def}}{=} I + D^{-1/2}AD^{-1/2};\quad 2)\ \text{Let } \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}} \text{ be the SVD of } M^pS;\quad \text{and}\quad 3)\ \widetilde{Y} \overset{\text{def}}{=} \widetilde{U} \in \mathbb{R}^{n\times k}.$$

Then, with probability at least $1 - 2e^{-2n} - 3\beta$, it holds that

$$\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\|_F \leqslant \varepsilon.$$

In Subsection 7.2.2, we establish technical lemmas that allows us to apply the proof technique developed in [BKG15, Theorem 6] for approximate SE $\widetilde{Y}$, and to prove a similar statement for the approximate normalized SE $\widetilde{Y'}$.

Theorem 5.5 (from page 62). *Let $\varepsilon, \delta_p \in (0,1)$ be arbitrary. Compute the approximate normalized SE $\widetilde{Y'}$ via the Power method with $p \geqslant \ln(8nk/\varepsilon\delta_p)/\ln(1/\gamma_k)$ iterations and $\gamma_k = (2-\lambda_{k+1})/(2-\lambda_k) < 1$. Run on the rows of $\widetilde{Y'}$ an α-approximate k-means clustering algorithm with failure probability δ_α. Let the outcome be a clustering indicator matrix $\widetilde{X'_\alpha} \in \mathbb{R}^{n\times k}$. Then, with probability at least $1 - 2e^{-2n} - 3\delta_p - \delta_\alpha$, it holds that*

$$\left\|Y' - \widetilde{X'_\alpha}\left(\widetilde{X'_\alpha}\right)^{\mathrm{T}} Y'\right\|_F^2 \leqslant (1+4\varepsilon)\cdot\alpha\cdot\left\|Y' - X'_{\mathrm{opt}}\left(X'_{\mathrm{opt}}\right)^{\mathrm{T}} Y'\right\|_F^2 + 4\varepsilon^2.$$

In Subsection 7.2.3, we prove Theorem 5.6 using Lemma 7.3 and Theorem 5.5, showing that the approximate normalized SE $\widetilde{Y'}$, computed via the Power method, is ε-separated.

Theorem 5.6 (from page 62)**.** *Assume $\Psi = 20^4 \cdot k^3/\delta$, $k/\delta \geqslant 10^9$ for some $\delta \in (0, 1/2]$. Compute the approximate normalized SE $\widetilde{Y'}$ via the Power method with $p \geqslant \Omega(\frac{\ln n}{\lambda_{k+1}})$. Then, for $\varepsilon = 6\cdot 10^{-7}$ it holds with high probability that*

$$\triangle_k(\widetilde{Y'}) < 5\varepsilon^2 \cdot \triangle_{k-1}(\widetilde{Y'}).$$

In Subsection 7.3, we show that Part (b) of Theorem 5.2 follows by combining Part (a) of Theorem 5.2, Theorem 5.3, Theorem 5.5 and Theorem 5.6.

7.2.1 Proof of Lemma 7.3

We argue in a similar manner as in [BM14, Lemma 7]. Our analysis uses the following two probabilistic results on Gaussian matrices.

Lemma 7.4 (Norm of a Gaussian Matrix [DS01])**.** *Let $M \in \mathbb{R}^{n\times k}$ be a matrix of i.i.d. standard Gaussian random variables, where $n \geqslant k$. Then, for $t \geqslant 4$, $\Pr\{\sigma_1(M) \geqslant t\sqrt{n}\} \geqslant \exp\{-nt^2/8\}$.*

Lemma 7.5 (Invertibility of a Gaussian Matrix [SST06])**.** *Let $M \in \mathbb{R}^{n\times n}$ be a matrix of i.i.d. standard Gaussian random variables. Then, for any $\beta \in (0,1)$, $\Pr\{\sigma_n(M) \leqslant \beta/(2.35\sqrt{n})\} \leqslant \beta$.*

Using the preceding two lemmas, we obtain the following probabilistic statement.

Lemma 7.6 (Rectangular Gaussian Matrix)**.** *Let $S \in \mathbb{R}^{n\times k}$ be a matrix of i.i.d. standard Gaussian random variables, $V \in \mathbb{R}^{n\times\rho}$ be a matrix with orthonormal columns and $n \geqslant \rho \geqslant k$. Then, with probability at least $1 - e^{-2n}$ it holds $\mathrm{rank}(V^{\mathrm{T}}S) = k$.*

Proof. Let $S' \in \mathbb{R}^{n\times\rho}$ be an extension of S such that $S' = [S\ S'']$, where $S'' \in \mathbb{R}^{n\times\rho-k}$ is a matrix of i.i.d. standard Gaussian random variables. Notice that $V^{\mathrm{T}}S' \in \mathbb{R}^{\rho\times\rho}$ is a matrix of i.i.d. standard Gaussian random variables. We apply now Lemma 7.5 with $\beta = e^{-2n}$ which yields with probability at least $1-e^{-2n}$ that $\sigma_\rho(V^{\mathrm{T}}S') > 1/(2.35\cdot e^{2n}\sqrt{\rho}) > 0$ and thus $\mathrm{rank}(V^{\mathrm{T}}S') = \rho$. In particular, $\mathrm{rank}(V^{\mathrm{T}}S) = k$ with probability at least $1 - e^{-2n}$. □

Proof of Lemma 7.3. By the Eigendecomposition theorem, $\mathcal{L}_G = U\Sigma' U^{-1}$ where $U \in \mathbb{R}^{n\times n}$ is an orthonormal matrix whose i-th column equals the eigenvector of $\mathcal{L}_G$ corresponding to the i-th smallest eigenvalue λ_i, and Σ' is a non-negative diagonal matrix such that $\Sigma'_{ii} = \lambda_i$, for all i. Since the canonical SE $Y \in \mathbb{R}^{n\times k}$ consists of the bottom k eigenvectors of $\mathcal{L}_G$, we have $U = [Y\ U_{n-k}]$ where $U_{n-k} \in \mathbb{R}^{n\times n-k}$, and similarly $\Sigma = [\Sigma_k\ 0_{k,n-k};\ 0_{n-k,k}\ \Sigma_{n-k}]$.

Further, by the Eigendecomposition theorem $M = U\Sigma U^{\mathrm{T}}$, where $\Sigma = 2I - \Sigma'$ and in particular $\Sigma_{ii} = 2 - \lambda_i \geqslant 0$ for all i. Since $M^p = U\Sigma^p U^{\mathrm{T}}$, it follows that $\ker(M^p S) = \ker(U^{\mathrm{T}} S)$. By Lemma 7.6 with probability at least $1 - e^{-2n}$ we have $\operatorname{rank}(U^{\mathrm{T}} S) = k$ and thus matrix $M^p S$ has k singular values. Further, the SVD decomposition $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}}$ of $M^p S$ satisfies: $\widetilde{U} \in \mathbb{R}^{n\times k}$ is a matrix with orthonormal columns, $\widetilde{\Sigma} \in \mathbb{R}^{k\times k}$ is a positive diagonal matrix and $\widetilde{V}^{\mathrm{T}} \in \mathbb{R}^{k\times k}$ is an orthonormal matrix. Recall that approximate SE is defined by $\widetilde{Y} = \widetilde{U}$.

Let $R \stackrel{\text{def}}{=} \widetilde{\Sigma}\widetilde{V}^{\mathrm{T}} \in \mathbb{R}^{k\times k}$ and observe that $\widetilde{Y}R = M^p S = [Y\ U_{n-k}]\Sigma^p[Y^{\mathrm{T}};\ U_{n-k}^{\mathrm{T}}]S$. We use the facts:

$$\begin{aligned}
\widetilde{Y}R &= Y\Sigma_k^p Y^{\mathrm{T}} S + U_{n-k}\Sigma_{n-k}^p U_{n-k}^{\mathrm{T}} S; && (7.5)\\
\sigma_i(\widetilde{Y}R) &\geqslant \sigma_k\left(Y\Sigma_k^p Y^{\mathrm{T}} S\right) \geqslant (2-\lambda_k)^p \cdot \sigma_k\left(Y^{\mathrm{T}} S\right); && (7.6)\\
\sigma_i(\widetilde{Y}R) &= \sigma_i\left(R\right); && (7.7)\\
\|X\widetilde{Y}\|_2 &\geqslant \|X\widetilde{Y}\|_2 \cdot \sigma_k(\widetilde{Y}), \quad \text{for any } X \in \mathbb{R}^{\ell\times k}. && (7.8)
\end{aligned}$$

(7.5) follows from the eigenvalue decomposition of M and the fact that $M^p = U\Sigma^p U^{\mathrm{T}}$; (7.6) follows by (7.5) due to Y and U_{n-k} span orthogonal spaces, and since the minimum singular value of a product is at least the product of the minimum singular values; (7.7) holds due to $\widetilde{Y}^{\mathrm{T}}\widetilde{Y} = I_k$; Recall that with probability at least $1 - e^{-2n}$ we have $\sigma_k(R) > 0$ and hence (7.8) follows by

$$\|X\|_2 = \max_{x\neq 0} \frac{\|XRx\|_2}{\|Rx\|_2} \leqslant \max_{x\neq 0} \frac{\|XRx\|_2}{\sigma_k(R)\,\|x\|_2} = \frac{\|XR\|_2}{\sigma_k(R)}.$$

[GVL12, Theorem 2.6.1] shows that for every two $m \times k$ orthonormal matrices W, Z with $m \geqslant k$ it holds

$$\left\|WW^{\mathrm{T}} - ZZ^{\mathrm{T}}\right\|_2 = \left\|Z^{\mathrm{T}} W^{\perp}\right\|_2 = \left\|W^{\mathrm{T}} Z^{\perp}\right\|_2,$$

where $[Z, Z^{\perp}] \in \mathbb{R}^{m\times m}$ is full orthonormal basis. Therefore, we have

$$\left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_2 = \left\|\widetilde{Y}^{\mathrm{T}} Y^{\perp}\right\|_2 = \left\|(Y^{\perp})^{\mathrm{T}}\widetilde{Y}\right\|_2 = \left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}\right\|_2, \tag{7.9}$$

where the last equality is due to $Y^{\perp} = U_{n-k}$.

To upper bound $\left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}\right\|_2$ we establish the following inequalities:

$$\begin{aligned}
\left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}R\right\|_2 &\geqslant \left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}\right\|_2 \cdot \sigma(R) \geqslant \left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}\right\|_2 \cdot (2-\lambda_k)^p \cdot \sigma_k\left(Y^{\mathrm{T}} S\right), && (7.10)\\
\left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}R\right\|_2 &= \left\|\Sigma_{n-k}^p U_{n-k}^{\mathrm{T}} S\right\|_2 \leqslant (2-\lambda_{k+1})^p \cdot \sigma_1\left(U_{n-k}^{\mathrm{T}} S\right), && (7.11)
\end{aligned}$$

where (7.10) follows by (7.8), (7.7) and (7.6); and (7.11) is due to (7.5) and $2 = \Sigma_{11} \geqslant \cdots \geqslant \Sigma_{nn} \geqslant 0$.

By Lemma 7.4 and Lemma 7.5, it follows by the Union bound that with probability at least $1 - e^{-2n} - 3\beta$, we have

$$\frac{\beta}{\sqrt{k}} \leqslant \sigma_k\left(Y^{\mathrm{T}} S\right) \quad \text{and} \quad \sigma_1\left(U_{n-k}^{\mathrm{T}} S\right) \leqslant 4\sqrt{n}. \tag{7.12}$$

Using (7.9), (7.10), (7.11) and (7.12) we obtain

$$\left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_2 = \left\|U_{n-k}^{\mathrm{T}}\widetilde{Y}\right\|_2 \leqslant (4/\beta) \cdot \sqrt{nk} \cdot \gamma_k^p. \tag{7.13}$$

Since $\|M\|_F \leqslant \sqrt{\operatorname{rank}(M)} \cdot \|M\|_2$ for every matrix M and $\operatorname{rank}(YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}) \leqslant 2k$, it follows

$$\left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_F \leqslant 2k \cdot \left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_2 \leqslant (8/\beta) \cdot n^{1/2} k^{3/2} \cdot \gamma_k^p \leqslant \varepsilon,$$

where the last two inequalities are due to (7.13) and the choice of γ_k. □

7.2.2 Proof of Theorem 5.5

[BKG15, Theorem 6] relates canonical SE and approximate SE, whereas our goal is to establish similar result for the normalized SE and approximate normalized SE. We present next four technical lemmas that combined with Lemma 7.3, allow us to apply the proof technique developed in [BKG15, Theorem 6].

Lemma 7.7. *Let $X', \widetilde{X'} \in \mathbb{R}^{m \times k}$ be indicator matrices returned by an α-approximate k-means clustering algorithm applied on inputs Y' and $\widetilde{Y'}$, respectively, for any $\alpha \geqslant 1$. Then, it holds that $X'(X')^{\mathrm{T}}$ and $\widetilde{X'}(\widetilde{X'})^{\mathrm{T}}$ are projection matrices.*

Proof. We prove now the first conclusion. By construction, there are $d(v)$ many copies of row $Y(v,:)/\sqrt{d(v)}$ in Y', for all $v \in V$. W.l.o.g. the indicator matrix X' has all copies of row $Y(v,:)/\sqrt{d(v)}$ assigned to the same cluster, for all $v \in V$. By definition, $X'_{ij} = 1/\sqrt{\mu(C_j)}$ if row $Y'_{i,:}$ belongs to the j-th cluster C_j and $X'_{ij} = 0$ otherwise. Hence, $(X')^{\mathrm{T}} X' = I_{k \times k}$ and thus $[X'(X')^{\mathrm{T}}]^2 = X'(X')^{\mathrm{T}}$. The second part follows similarly, since matrix $\widetilde{U}$ is orthonormal. □

Lemma 7.8. *The normalized SE Y' and the approximate normalized SE $\widetilde{Y'}$ are orthonormal matrices.*

Proof. We prove now $(Y')^{\mathrm{T}} Y' = I_{k \times k}$. The equality $\widetilde{Y'}^{\mathrm{T}} \widetilde{Y'} = I_{k \times k}$ follows similarly. Note that

$$\begin{aligned}
\left[(Y')^{\mathrm{T}} Y'\right]_{ij} &= \begin{pmatrix} \frac{Y(1,i)}{\sqrt{d(1)}} \mathbf{1}_{d(1)}^{\mathrm{T}} & \cdots & \frac{Y(n,i)}{\sqrt{d(n)}} \mathbf{1}_{d(n)}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \frac{Y(1,j)}{\sqrt{d(1)}} \mathbf{1}_{d(1)} \\ \cdots \\ \frac{Y(n,j)}{\sqrt{d(n)}} \mathbf{1}_{d(n)} \end{pmatrix} \\
&= \sum_{\ell=1}^{n} d(\ell) \frac{Y(\ell,i)}{\sqrt{d(\ell)}} \frac{Y(\ell,j)}{\sqrt{d(\ell)}} = \langle Y(:,i), Y(:,j) \rangle = \delta(i,j),
\end{aligned}$$

where $\delta(i,j)$ is the Kronecker delta function. Hence, the statement follows. □

Lemma 7.9. *It holds that $\|Y'(Y')^{\mathrm{T}} - \widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}}\|_F = \|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\|_F$.*

Proof. Let $\mathbf{1}_{d(i)} \in \{0,1\}^m$ be an indicator vector of the $d(i)$ copies of row $Y(i,:)/\sqrt{d(i)}$ in matrix Y'. By definition

$$Y'(Y')^{\mathrm{T}} = \sum_{\ell=1}^{k} Y'_{:,\ell} Y'^{T}_{:,\ell} \quad \text{where} \quad Y'_{:,\ell} = \begin{pmatrix} \frac{Y(1,\ell)}{\sqrt{d(1)}} \mathbf{1}_{d(1)} \\ \cdots \\ \frac{Y(n,\ell)}{\sqrt{d(n)}} \mathbf{1}_{d(n)} \end{pmatrix}_{m \times 1}$$

and

$$\left(Y'_{:,\ell} Y'^{T}_{:,\ell}\right)_{d(i)d(j)} = \frac{Y(i,\ell) Y(j,\ell)}{\sqrt{d(i)d(j)}} \cdot \mathbf{1}_{d(i)} \mathbf{1}_{d(j)}^{\mathrm{T}}.$$

Hence, we have

$$\begin{aligned}
\left\| Y'(Y')^{\mathrm{T}} - \widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}} \right\|_F^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \left(Y'(Y')^{\mathrm{T}} - \widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}} \right)_{d(i)d(j)} \right\|_F^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \sum_{\ell=1}^{k} \left(Y'_{:,\ell}(Y'_{:,\ell})^{\mathrm{T}} - \widetilde{Y'}_{:,\ell}(\widetilde{Y'}_{:,\ell})^{\mathrm{T}} \right)_{d(i)d(j)} \right\|_F^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \left\{ \sum_{\ell=1}^{k} \left(\frac{Y(i,\ell)Y(j,\ell)}{\sqrt{d(i)d(j)}} - \frac{\widetilde{Y}(i,\ell)\widetilde{Y}(j,\ell)}{\sqrt{d(i)d(j)}} \right) \right\} \cdot \mathbf{1}_{d(i)} \mathbf{1}_{d(j)}^{\mathrm{T}} \right\|_F^2 .
\end{aligned}$$

By definition of Frobenius norm, it holds that

$$\begin{aligned}\left\|Y'(Y')^{\mathrm{T}}-\widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}}\right\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} d(i)d(j)\left[\sum_{\ell=1}^{k}\left(\frac{Y(i,\ell)Y(j,\ell)}{\sqrt{d(i)d(j)}}-\frac{\widetilde{Y}(i,\ell)\widetilde{Y}(j,\ell)}{\sqrt{d(i)d(j)}}\right)\right]^2\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\sum_{\ell=1}^{k}\left(Y(i,\ell)Y(j,\ell)-\widetilde{Y}(i,\ell)\widetilde{Y}(j,\ell)\right)\right]^2\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(YY^{\mathrm{T}}-\widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right)_{ij}^2\\ &= \left\|YY^{\mathrm{T}}-\widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_F^2.\end{aligned}$$

□

Lemma 7.10. *For any matrix U with orthonormal columns and every matrix A it holds*

$$\left\|UU^{\mathrm{T}}-AA^{\mathrm{T}}UU^{\mathrm{T}}\right\|_F=\left\|U-AA^{\mathrm{T}}U\right\|_F. \tag{7.14}$$

Proof. The statement follows by the Frobenius norm property $\|M\|_F^2=\mathrm{Tr}[M^{\mathrm{T}}M]$, the cyclic property of trace $\mathrm{Tr}[UM^{\mathrm{T}}MU^{\mathrm{T}}]=\mathrm{Tr}[M^{\mathrm{T}}M\cdot U^{\mathrm{T}}U]$ and the orthogonality of matrix U. □

Using the preceding lemmas, we are ready to prove Theorem 5.5.

Proof of Theorem 5.5. Using Lemma 7.3 and Lemma 7.9 with probability at least $1-2e^{-2n}-3\delta_p$ we have

$$\left\|Y'(Y')^{\mathrm{T}}-\widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}}\right\|_F=\left\|YY^{\mathrm{T}}-\widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_F\leqslant\varepsilon.$$

Let $Y'(Y')^{\mathrm{T}}=\widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}}+E$ such that $\|E\|_F\leqslant\varepsilon$. By combining Lemma 7.8 and Lemma 7.10, (7.14) holds for the matrices Y' and $\widetilde{Y'}$. Thus, by Lemma 7.7 and the proof techniques in [BKG15, Theorem 6], it follows that

$$\left\|Y'-\widetilde{X'_\alpha}(\widetilde{X'_\alpha})^{\mathrm{T}}Y'\right\|_F\leqslant\sqrt{\alpha}\cdot\left(\left\|Y'-X'_{\mathrm{opt}}(X'_{\mathrm{opt}})^{\mathrm{T}}Y'\right\|_F+2\varepsilon\right). \tag{7.15}$$

The desired statement follows by simple algebraic manipulations of (7.15). □

7.2.3 Proof of Theorem 5.6

In this section, we demonstrate that the approximate normalized SE $\widetilde{Y'}$ is ε-separated, i.e. $\triangle_k(\widetilde{Y'})<5\varepsilon^2\cdot\triangle_{k-1}(\widetilde{Y'})$. Our analysis builds upon Theorem 5.4, Theorem 5.5 and the proof techniques in [BKG15, Theorem 6].

Before we present the proof of Theorem 5.6, we establish two technical Lemmas.

Lemma 7.11. *Suppose $\Psi\geqslant 20^4\cdot k^3/\delta$ for some $\delta\in(0,1/2]$. Then, it holds that*

$$\ln\left(\frac{2-\lambda_k}{2-\lambda_{k+1}}\right)\geqslant\frac{1}{2}\left(1-\frac{4\delta}{20^4k^2}\right)\lambda_{k+1}.$$

Proof. By (5.5), the following higher-order Cheeger inequalities hold

$$\lambda_k/2\leqslant\rho(k)\leqslant O(k^2)\cdot\sqrt{\lambda_k}. \tag{7.16}$$

Using the LHS of (7.16), we have

$$k^3\widehat{\rho}_{\mathrm{avr}}(k)=k^2\sum_{i=1}^{k}\phi(P_i)\geqslant k^2\max_{i\in\{1,\ldots,k\}}\phi(P_i)\geqslant k^2\cdot\rho(k)\geqslant\frac{k^2\lambda_k}{2},$$

and thus the k-th smallest eigenvalue of $\mathcal{L}_G$ satisfies $\lambda_k\leqslant 2k\cdot\widehat{\rho}_{\mathrm{avr}}(k)$. Then, the gap assumption yields

$$\lambda_{k+1}\geqslant\frac{20^4k^2}{2\delta}\cdot 2k\cdot\widehat{\rho}_{\mathrm{avr}}(k)\geqslant\frac{20^4k^2}{2\delta}\cdot\lambda_k.$$

The statement follows by

$$\frac{2-\lambda_k}{2-\lambda_{k+1}}\geqslant\frac{1-\frac{\delta}{20^4k^2}\cdot\lambda_{k+1}}{1-\frac{1}{2}\lambda_{k+1}}\geqslant\exp\left\{\frac{1}{2}\left(1-\frac{4\delta}{20^4k^2}\right)\lambda_{k+1}\right\}.$$

□

Lemma 7.12. *For any matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$, it holds that $\|AB\|_F \leqslant \|A\|_2 \cdot \|B\|_F$.*

Proof. By definition, $\|B\|_F^2 = \sum_{i=1}^k \|B_{:,i}\|_2^2$ and $\|Ax\|_2 \leqslant \|A\|_2 \|x\|_2$, and thus we have

$$\|AB\|_F^2 = \sum_{i=1}^k \|AB_{:,i}\|_2^2 \leqslant \|A\|_2^2 \sum_{i=1}^k \|B_{:,i}\|_2^2 = \|A\|_2^2 \cdot \|B\|_2^2. \qquad \square$$

In the following, we use interchangeably X'_{opt} and $X'^{(k)}_{\text{opt}}$ to denote an optimal indicator matrix for the k-means clustering problem on Y'. Similarly, we denote by $X'^{(k-1)}_{\text{opt}}$ an optimal indicator matrix for the $(k-1)$-means clustering problem on Y'.

We are now ready to prove Theorem 5.6.

Proof of Theorem 5.6. We set the approximation parameter in Theorem 5.5 to $\varepsilon' = \varepsilon/30$. By Theorem 5.4, we have

$$\begin{aligned} &\left\|Y' - X'^{(k)}_{\text{opt}}\left(X'^{(k)}_{\text{opt}}\right)^{\mathrm{T}} Y'\right\|_F = \sqrt{\triangle_k(Y')} \\ \leqslant \quad & \varepsilon\sqrt{\triangle_{k-1}(Y')} = \varepsilon \left\|Y' - X'^{(k-1)}_{\text{opt}}\left(X'^{(k-1)}_{\text{opt}}\right)^{\mathrm{T}} Y'\right\|_F . \end{aligned} \tag{7.17}$$

We compute an approximate SE $\widetilde{Y} \in \mathbb{R}^{n\times k}$, defined in (5.7), via the Power method which runs $p = O(\frac{\ln n}{\lambda_{k+1}})$ iterations.

By combining Lemma 7.3 and Lemma 7.9, for the normalized and approximate normalized SE, Y' and $\widetilde{Y'}$ respectively, we obtain w.h.p. that

$$\left\|Y'Y'^{\mathrm{T}} - \widetilde{Y'}(\widetilde{Y'})^{\mathrm{T}}\right\|_F = \left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_F \leqslant \varepsilon'.$$

Let $Y'Y'^{\mathrm{T}} = \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} + E$ such that $\|E\|_F \leqslant \varepsilon'$. By Lemma 7.8, Y' and $\widetilde{Y'}$ are orthonormal matrices. Hence, by Lemma 7.10 applied on $\widetilde{Y'}$, we obtain

$$\begin{aligned} \sqrt{\triangle_k(\widetilde{Y'})} \quad &= \quad \left\|\widetilde{Y'} - \widetilde{X'^{(k)}_{\text{opt}}}\left(\widetilde{X'^{(k)}_{\text{opt}}}\right)^{\mathrm{T}} \widetilde{Y'}\right\|_F = \left\|\widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} - \widetilde{X'^{(k)}_{\text{opt}}}\left(\widetilde{X'^{(k)}_{\text{opt}}}\right)^{\mathrm{T}} \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}}\right\|_F \\ &= \quad \left\|Y'(Y')^{\mathrm{T}} - \widetilde{X'^{(k)}_{\text{opt}}}\left(\widetilde{X'^{(k)}_{\text{opt}}}\right)^{\mathrm{T}} Y'(Y')^{\mathrm{T}} - \left[I_{m\times m} - \widetilde{X'^{(k)}_{\text{opt}}}\left(\widetilde{X'^{(k)}_{\text{opt}}}\right)^{\mathrm{T}}\right]E\right\|_F \\ &\leqslant \quad \left\|Y' - \widetilde{X'^{(k)}_{\text{opt}}}\left(\widetilde{X'^{(k)}_{\text{opt}}}\right)^{\mathrm{T}} Y'\right\|_F + \|E\|_F , \end{aligned} \tag{7.18}$$

where the last step uses triangle inequality, Lemma 7.10 applied on Y', Lemma 7.12, Lemma 7.7 and $\|I - PP^{\mathrm{T}}\|_2 \leqslant 1$ for any projection matrix P. Then, we apply Theorem 5.5 with an exact k-means clustering algorithm, i.e. $\alpha = 1$, and parameters $\delta_p = n^{-O(1)}$ and ε' (as above). By Lemma 7.11, for any $p \geqslant \Omega(\frac{\ln n}{\lambda_{k+1}})$ it holds w.h.p.

$$\left\|Y' - \widetilde{X'^{(k)}_{\text{opt}}}\left(\widetilde{X'^{(k)}_{\text{opt}}}\right)^{\mathrm{T}} Y'\right\|_F^2 \leqslant (1+4\varepsilon') \cdot \left\|Y' - X'^{(k)}_{\text{opt}}\left(X'^{(k)}_{\text{opt}}\right)^{\mathrm{T}} Y'\right\|_F^2 + 4(\varepsilon')^2. \tag{7.19}$$

The proof proceeds by case distinction.

Case 1: Suppose $\varepsilon' \leqslant \frac{1}{4}\sqrt{\triangle_k(Y')}$. Combining (7.18), (7.19) and $\|E\|_F \leqslant \varepsilon'$, yields

$$\begin{aligned} \sqrt{\triangle_k(\widetilde{Y'})} \quad &\leqslant \quad \varepsilon' + \sqrt{(1+4\varepsilon') \cdot \triangle_k(Y') + 4(\varepsilon')^2} \\ &\leqslant \quad 2 \cdot \sqrt{\triangle_k(Y')} \leqslant 2\varepsilon \cdot \sqrt{\triangle_{k-1}(Y')}, \end{aligned} \tag{7.20}$$

where the last inequality follows by (7.17). Moreover, it holds that

$$\begin{aligned}
\sqrt{\triangle_{k-1}(Y')} &= \left\| Y' - X'^{(k-1)}_{\mathrm{opt}} \left(X'^{(k-1)}_{\mathrm{opt}} \right)^{\mathrm{T}} Y' \right\|_F \leqslant \left\| Y' - \widetilde{X'^{(k-1)}_{\mathrm{opt}}} \left(\widetilde{X'^{(k-1)}_{\mathrm{opt}}} \right)^{\mathrm{T}} Y' \right\|_F \\
&= \left\| Y'(Y')^{\mathrm{T}} - \widetilde{X'^{(k-1)}_{\mathrm{opt}}} \left(\widetilde{X'^{(k-1)}_{\mathrm{opt}}} \right)^{\mathrm{T}} Y'(Y')^{\mathrm{T}} \right\|_F \\
&= \left\| \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} - \widetilde{X'^{(k-1)}_{\mathrm{opt}}} \left(\widetilde{X'^{(k-1)}_{\mathrm{opt}}} \right)^{\mathrm{T}} \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} + \left[I_{m\times m} - \widetilde{X'^{(k-1)}_{\mathrm{opt}}} \left(\widetilde{X'^{(k-1)}_{\mathrm{opt}}} \right)^{\mathrm{T}} \right] E \right\|_F \\
&\leqslant \left\| \widetilde{Y'} - \widetilde{X'^{(k-1)}_{\mathrm{opt}}} \left(\widetilde{X'^{(k-1)}_{\mathrm{opt}}} \right)^{\mathrm{T}} \widetilde{Y'} \right\|_F + \|E\|_F \qquad (7.21) \\
&\leqslant \sqrt{\triangle_{k-1}(\widetilde{Y'})} + \frac{\varepsilon}{4} \sqrt{\triangle_{k-1}(Y')},
\end{aligned}$$

where the last inequality uses the definition of $\triangle_{k-1}(\widetilde{Y'})$ and

$$\|E\|_F \leqslant \varepsilon' \leqslant \frac{1}{4}\sqrt{\triangle_k(Y')} \leqslant \frac{\varepsilon}{4}\sqrt{\triangle_{k-1}(Y')}.$$

Hence,

$$\sqrt{\triangle_{k-1}(Y')} \leqslant \left(1 + \frac{\varepsilon}{2}\right) \sqrt{\triangle_{k-1}(\widetilde{Y'})}. \qquad (7.22)$$

Therefore, by combining (7.20) and (7.22), we obtain

$$\sqrt{\triangle_k(\widetilde{Y'})} \leqslant 2\varepsilon \cdot \sqrt{\triangle_{k-1}(Y')} \leqslant (2+\varepsilon)\cdot\varepsilon\cdot\sqrt{\triangle_{k-1}(\widetilde{Y'})}.$$

Case 2: Suppose $0 \leqslant \frac{1}{4}\sqrt{\triangle_k(Y')} < \varepsilon'$. Combining (7.18), (7.19) and $\|E\|_F \leqslant \varepsilon'$, yields

$$\sqrt{\triangle_k(\widetilde{Y'})} \leqslant \varepsilon' + \sqrt{(1+4\varepsilon')\cdot\triangle_k(Y') + 4(\varepsilon')^2} \leqslant 6\varepsilon'. \qquad (7.23)$$

Further, by Lemma 7.1 and $k/\delta \geqslant 10^9$ we have $\triangle_{k-1}(Y') \geqslant 1/13$, and thus using (7.21) we obtain

$$\sqrt{\triangle_{k-1}(\widetilde{Y'})} \geqslant \sqrt{\triangle_{k-1}(Y')} - \|E\|_F \geqslant 1/4 - \varepsilon' \geqslant 1/5. \qquad (7.24)$$

By combining Equations (7.23) and (7.24), it follows that

$$\sqrt{\triangle_k(\widetilde{Y'})} \leqslant 6\varepsilon' \leqslant 30\varepsilon' \cdot \sqrt{\triangle_{k-1}(\widetilde{Y'})} = \varepsilon \cdot \sqrt{\triangle_{k-1}(\widetilde{Y'})}. \qquad \square$$

7.3 Proof of Approximate Spectral Clustering

We prove now Part (b) of Theorem 5.2. Let $p = \Theta(\frac{\ln n}{\lambda_{k+1}})$. We compute the matrix $M^p S$ in time $O(mkp)$ and its singular value decomposition $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}}$ in time $O(nk^2)$. Based on it, we construct in time $O(mk)$ the approximate normalized SE $\widetilde{Y'}$, see (5.8).

By Theorem 5.6, $\widetilde{Y'}$ is ε-separated for $\varepsilon = 6\cdot 10^{-7}$, i.e. $\triangle_k(\widetilde{Y'}) < 5\varepsilon^2 \cdot \triangle_{k-1}(\widetilde{Y'})$. Let $\alpha = 1 + 10^{-10}$. Then, by Theorem 5.3, there is an algorithm that outputs in time $O(mk^2 + k^4)$ a k-way vector partition with indicator matrix $\widetilde{X'_\alpha}$ such that with probability at least $1 - O(\sqrt{\varepsilon})$, we have

$$\left\| \widetilde{Y'} - \widetilde{X'_\alpha} \left(\widetilde{X'_\alpha} \right)^{\mathrm{T}} \widetilde{Y'} \right\|_F^2 \leqslant \left(1 + \frac{1}{10^{10}}\right) \cdot \left\| \widetilde{Y'} - \widetilde{X'_{\mathrm{opt}}} \left(\widetilde{X'_{\mathrm{opt}}} \right)^{\mathrm{T}} \widetilde{Y'} \right\|_F^2 .$$

Let $\eta \in (n^{-O(1)}, 1)$ be a parameter to be determined soon. Observe that for $\Psi = 20^4 \cdot k^3/\delta$, $\delta \in (0, 1/2]$ and $k/\delta \geqslant 10^9$, by Equations (5.11), (6.3) and Lemma 6.6, it holds

$$\left\| Y' - X'^{(k)}_{\mathrm{opt}} \left(X'^{(k)}_{\mathrm{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 \leqslant \mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leqslant \frac{2k^2}{\Psi} \leqslant \frac{1}{8\cdot 10^{13}}. \qquad (7.25)$$

Combining Theorem 5.6 and Equation (7.25), yields

$$\frac{1}{n^{O(1)}} \leqslant \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F \leqslant \frac{1}{10^6}.$$

Using Lemma 7.11, we apply Theorem 5.5 with $\delta_p = n^{-O(1)}$, $\alpha = 1 + 10^{-10}$, $\delta_\alpha = O(\sqrt{\varepsilon})$ and

$$\varepsilon' = \frac{\sqrt{\eta}}{4} \cdot \frac{1}{n^{O(1)}} \leqslant \frac{\sqrt{\eta}}{4} \cdot \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F ,$$

and obtain with constant probability (close to 1) that

$$\begin{aligned} \left\| \widetilde{Y'} - \widetilde{X'_\alpha} \left(\widetilde{X'_\alpha} \right)^{\mathrm{T}} \widetilde{Y'} \right\|_F^2 &\leqslant (1 + 4\varepsilon') \cdot \alpha \cdot \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 + 4\varepsilon'^2 \\ = &\left[\left(1 + \sqrt{\eta} \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F \right) \alpha + \frac{\eta}{4} \right] \cdot \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 \\ \leqslant &\left[\left(1 + \frac{\sqrt{\eta}}{10^6} \right) \cdot \left(1 + \frac{1}{10^{10}} \right) + \frac{\eta}{4} \right] \cdot \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 , \end{aligned}$$

Then, for $\eta = 1/10^6$ the approximate solution $\widetilde{X'_\alpha}$ yields a multiplicative approximation, satisfying

$$\left\| Y' - \widetilde{X'_\alpha} \left(\widetilde{X'_\alpha} \right)^{\mathrm{T}} Y' \right\|_F^2 \leqslant \left(1 + \frac{1}{10^6} \right) \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 .$$

The statement follows by Part (a) of Theorem 5.2 applied to the k-way partition $(A_1, \ldots, A_k)$ of V that is induced by the indicator matrix $\widetilde{X'_\alpha}$.

Appendix

Let G be an undirected and unweighted graph consisting of k cliques each of size n/k, and let A be its adjacency matrix. Let G^R be the graph G, plus a set E_R of k additional edges connecting the k cliques in the form of a ring such that no two edges in E_R (of the ring) share a vertex. In other words, let A_R be the adjacency matrix of E_R, then $A_R A_R = S$ where S is a diagonal matrix defined as $S_{ii} = 1$ if vertex i is connected to a vertex in a neighboring clique, and $S_{ii} = 0$ otherwise.

Recall that the normalized Laplacian matrix of G is given by $\mathcal{L}_G = I - \mathcal{A}$, where $\mathcal{A} = D^{-1/2} A D^{-1/2}$ and D is a positive diagonal matrix with $D_{ii} = \deg(i)$, for all i. Then, the normalized Laplacian matrix of G^R is defined as

$$\mathcal{L}_{G^R} = I - (D + S)^{-1/2} (A + A_R) (D + S)^{-1/2} .$$

Our analysis relies on the following four lemmas. We note that the first lemma is folklore.

Lemma 7.13. *It holds that $\lambda_1(\mathcal{L}_G) = \cdots = \lambda_k(\mathcal{L}_G) = 0$ and $\lambda_{k+1}(\mathcal{L}_G) = \cdots = \lambda_n(\mathcal{L}_G) = 1$. Moreover, $\lambda_1(\mathcal{A}) = \cdots = \lambda_k(\mathcal{A}) = 1$ and $\lambda_{k+1}(\mathcal{A}) = \cdots = \lambda_{k+1}(\mathcal{A}) = 0$. In particular, $\mathcal{A}$ is a symmetric positive semi-definite matrix.*

Lemma 7.14. *For any symmetric matrix $M \in \mathbb{R}^{n \times n}$ and any positive diagonal matrix $S \in \mathbb{R}^{n \times n}$, it holds that $\lambda_i \left(M S^{-1} \right) = \lambda_i (S^{-1/2} M S^{-1/2})$ for every $i \in \{1, \ldots, n\}$.*

Proof. Let $v_i \in \mathbb{R}^n$ and $\lambda_i \in \mathbb{R}$ be such that $M S^{-1} v_i = \lambda_i v_i$. Let $v_i = S^{1/2} y_i$ for some vector $y_i \in \mathbb{R}^n$ (it exists as S is a positive diagonal matrix), then we have $M S^{-1} \cdot S^{1/2} y_i = \lambda_i \cdot S^{1/2} y_i$ and by multiplying from the left with $S^{-1/2}$, we obtain $S^{-1/2} M S^{-1/2} \cdot y_i = \lambda_i \cdot y_i$. □

Lemma 7.15. *[Kly00, Eig15] Let $Q \in \mathbb{R}^{n \times n}$ be a symmetric positive semi-definite matrix and $S \in \mathbb{R}^{n \times n}$ be a positive diagonal matrix. Let $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ and $\mu_1 \geqslant \ldots \geqslant \mu_n$ be the eigenvalues of Q and $Q S^{-1}$, respectively. Then, it holds that $\lambda_i \cdot \min_j \{S_{jj}^{-1}\} \leqslant \mu_i \leqslant \lambda_i \cdot \max_j \{S_{jj}^{-1}\}$.*

Lemma 7.16. *[Ste90, Corollary 4.10] Let $A, E \in \mathbb{R}^{n \times n}$ be arbitrary symmetric matrices. Then, their eigenvalues satisfy $|\lambda_i(A + E) - \lambda_i(A)| \leqslant \|E\|_2$ for every $i \in \{1, \ldots, n\}$.*

Using the preceding four lemmas, we establish a lower bound on the eigenvalue $\lambda_{k+1}(\mathcal{L}_{G^R})$.

Theorem 7.17. *(Ring of k Cliques) The $(k+1)$-st smallest eigenvalue of the normalized Laplacian matrix of graph G^R satisfies $\lambda_{k+1}(\mathcal{L}_{G^R}) \geqslant 1 - k/n$.*

Proof. Let $Z \overset{\text{def}}{=} D^{-1} \cdot (D+S)$. Note that Z is a positive diagonal matrix. By definition, we have

$$\begin{aligned}\mathcal{L}_{G^R} &= I - (D+S)^{-1/2}(A+A_R)(D+S)^{-1/2} \\ &= I - Z^{-1/2} \cdot D^{-1/2}(A+A_R)D^{-1/2} \cdot Z^{-1/2}.\end{aligned}$$

It suffices to upper bound the $(k+1)$-st largest eigenvalue of matrix $Z^{-1/2} \cdot D^{-1/2}(A+A_R)D^{-1/2} \cdot Z^{-1/2}$. Recall that $\mathcal{A} = D^{-1/2}AD^{-1/2}$ and thus

$$Z^{-1/2} \cdot D^{-1/2}(A+A_R)D^{-1/2} \cdot Z^{-1/2} = Z^{-1/2}\mathcal{A}Z^{-1/2} + Z^{-1/2} \cdot D^{-1/2}A_R D^{-1/2} \cdot Z^{-1/2}.$$

By definition, $\|M\|_2^2 = \max_{\|x\|_2=1} x^T M^T M x$ and since $A_R A_R = S$, we have

$$\|A_R\|_2^2 = \max_{\|x\|_2=1} x^T S x = \max_{\|x\|_2=1} \sum_{j=1}^{2k} x_{i_j}^2 \leqslant 1 \tag{7.26}$$

and

$$\left\|(D+S)^{-1}\right\|_2^2 = \max_{\|x\|_2=1} x^T (D+S)^{-2} x \leqslant \max_i \left\{\frac{1}{(D_{ii}+S_{ii})^2}\right\} \cdot \max_{\|x\|_2=1} x^T x \leqslant \left(\frac{k}{n}\right)^2. \tag{7.27}$$

In order to apply Lemma 7.16, we upper bound first the expression

$$\begin{aligned}&\left\|Z^{-1/2} \cdot D^{-1/2}A_R D^{-1/2} \cdot Z^{-1/2}\right\|_2 \overset{\text{Lem. 7.14}}{=} \left\|A_R \cdot D^{-1}Z^{-1}\right\|_2 \\ &= \left\|A_R \cdot (D+S)^{-1}\right\|_2 \leqslant \|A_R\|_2 \cdot \left\|(D+S)^{-1}\right\|_2 \overset{(7.26)\text{ and }(7.27)}{\leqslant} \frac{k}{n}.\end{aligned} \tag{7.28}$$

Since $\mathcal{A}$ is a symmetric positive semi-definite matrix and Z is a positive diagonal matrix, we have

$$\begin{aligned}&\lambda_{k+1}\left(Z^{-1/2}\mathcal{A}Z^{-1/2}\right) \overset{\text{Lem. 7.14}}{=} \lambda_{k+1}\left(\mathcal{A}Z^{-1}\right) \\ &\overset{\text{Lem. 7.15}}{\leqslant} \lambda_{k+1}(\mathcal{A}) \cdot \max_i \left\{Z_{ii}^{-1}\right\} = \lambda_{k+1}(\mathcal{A}) \overset{\text{Lem. 7.13}}{=} 0.\end{aligned} \tag{7.29}$$

Therefore, the largest $(k+1)$-st eigenvalue

$$\begin{aligned}&\lambda_{k+1}\left(Z^{-1/2} \cdot D^{-1/2}(A+A_R)D^{-1/2} \cdot Z^{-1/2}\right) \\ &\overset{\text{Lem. 7.16}}{\leqslant} \lambda_{k+1}\left(Z^{-1/2}\mathcal{A}Z^{-1/2}\right) + \left\|Z^{-1/2} \cdot D^{-1/2}A_R D^{-1/2} \cdot Z^{-1/2}\right\|_2 \\ &\overset{(7.28)\text{ and }(7.29)}{\leqslant} \frac{k}{n},\end{aligned}$$

and thus the smallest $(k+1)$-st eigenvalue of $\mathcal{L}_{G^R}$ satisfies $\lambda_{k+1}(\mathcal{L}_{G^R}) \geqslant 1 - k/n$. □

Part III

Two Results on Slime Mold Computations

Proceedings

This part is based on an article [BBK+18] that appeared at the journal of Theoretical Computer Science, Series C (Theory of Natural Computing). The full ArXiv version is available online [BBK+17].

Acknowledgements

I would like to thank Kurt Mehlhorn for his guidance, support and fruitful collaboration, Ruben Becker, Vincenzo Bonifaci and Andreas Karrenbauer for fruitful collaboration, and anonymous reviewers of our journal submission for their helpful remarks.

Chapter 8

Introduction

We present two results on slime mold computations, one on the biologically-grounded model and one on the biologically-inspired model. The former model was introduced by biologists to capture the slime's apparent ability to compute shortest paths. We show in Section 8.1 that the dynamics can actually do more. It can solve a wide class of linear programs with nonnegative cost vectors. The latter model was designed as an optimization technique inspired by the former model. We present in Section 8.2 an improved convergence result for its discretization.

8.1 The Biologically-Grounded Model

Physarum polycephalum is a slime mold that apparently is able to solve shortest path problems. Nakagaki, Yamada, and Tóth [NYT00] report about the following experiment; see Figure 8.1. They built a maze, covered it by pieces of Physarum (the slime can be cut into pieces which will reunite if brought into vicinity), and then fed the slime with oatmeal at two locations. After a few hours the slime retracted to a path that follows the shortest path in the maze connecting the food sources. The authors report that they repeated the experiment with different mazes; in all experiments, Physarum retracted to the shortest path.

The paper [TKN07] proposes a mathematical model for the behavior of the slime and argues extensively that the model is adequate. Physarum is modeled as an electrical network with time varying resistors. We have a simple *undirected* graph $G = (N, E)$ with distinguished nodes s_0 and s_1 modeling the food sources. Each edge $e \in E$ has a positive length c_e and a positive capacity $x_e(t)$; c_e is fixed, but $x_e(t)$ is a function of time. The resistance $r_e(t)$ of e is $r_e(t) = c_e/x_e(t)$. In the electrical network defined by these resistances, a current of value 1 is forced from s_0 to s_1. For an (arbitrarily oriented) edge $e = (u, v)$, let $q_e(t)$ be the resulting current over e. Then, the capacity of e evolves according to the differential equation

$$\dot{x}_e(t) = |q_e(t)| - x_e(t), \tag{8.1}$$

where $\dot{x}_e$ is the derivative of x_e with respect to time. In equilibrium ($\dot{x}_e = 0$ for all e), the flow through any edge is equal to its capacity. In non-equilibrium, the capacity grows (shrinks) if the absolute value of the flow is larger (smaller) than the capacity. In the sequel, we will mostly drop the argument t as is customary in the treatment of dynamical systems. We will also write q for the vector with components q_e. It is well-known that the electrical flow q is the feasible flow minimizing energy dissipation $\sum_e r_e q_e^2$ (Thomson's principle).

We refer to the dynamics above as *biologically-grounded*, as it was introduced by biologists to model the behavior of a biological system. Miyaji and Ohnishi were the first to analyze convergence for special graphs (parallel links and planar graphs with source and sink on the same face) in [MO08]. In [BMV12] convergence was proven for *all* graphs. We state the result from [BMV12] for the special case that the shortest path is unique.

Theorem 8.1 ([BMV12]). *Assume $x_e(0) > 0$ and $c_e > 0$ for all e, and that the undirected shortest path P^* from s_0 to s_1 w.r.t. the cost vector c is unique. Then $x(t)$ in (8.1) converges to P^*. Namely, $x_e(t) \to 1$ for $e \in P^*$ and $x_e \to 0$ for $e \notin P^*$ as $t \to \infty$.*

[BMV12] also proves an analogous result for the undirected transportation problem; [Bon13] simplified the argument under additional assumptions. The paper [Bon15] studies a more general dynamics and proves convergence for parallel links.

In this paper, we extend this result to *non-negative undirected linear programs*

$$\min\{c^{\mathrm{T}} x : Af = b,\ |f| \leqslant x\}, \tag{8.2}$$

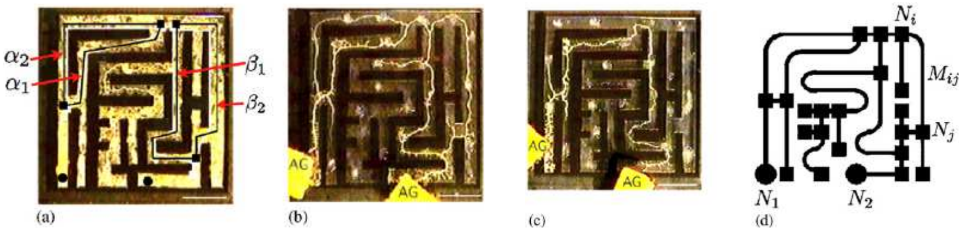

Figure 8.1: The experiment in [NYT00] (reprinted from there): (a) shows the maze uniformly covered by Physarum; yellow color indicates presence of Physarum. Food (oatmeal) is provided at the locations labeled AG. After a while the mold retracts to the shortest path connecting the food sources as shown in (b) and (c). (d) shows the underlying abstract graph. The video [Phy10] shows the experiment.

where $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$, $x \in \mathbb{R}^m$, $c \in \mathbb{R}^m_{\geqslant 0}$, and the absolute values are taken componentwise. Undirected LPs can model a wide range of problems, e.g., optimization problems such as shortest path and min-cost flow in undirected graphs, and the Basis Pursuit problem in signal processing [CDS98].

We use n for the number of rows of A and m for the number of columns, since this notation is appropriate when A is the node-edge-incidence matrix of a graph. A vector $f \in \mathbb{R}^m$ is *feasible* if $Af = b$. We assume that the system $Af = b$ has a feasible solution and every nonzero vector f in the kernel [1] of A has positive cost $\sum_e c_e|f_e| > 0$. The vector $q \in \mathbb{R}^m$ in (8.1) is now the *minimum energy feasible solution*

$$q(t) = \operatorname*{argmin}_{f\in\mathbb{R}^m} \left\{ \sum_{e:x_e\neq 0} \frac{c_e}{x_e(t)} f_e^2 : Af = b \wedge f_e = 0 \text{ whenever } x_e = 0 \right\}. \tag{8.3}$$

We remark that q is unique; see Section 9.2.1. If A is the incidence matrix of a graph (the column corresponding to an edge e has one entry $+1$, one entry -1 and all other entries are equal to zero), (8.2) is a transshipment problem with flow sources and sinks encoded by a demand vector b. The condition that there is no solution in the kernel of A with $c_e f_e = 0$ for all e states that every cycle contains at least one edge of positive cost. In that setting, $q(t)$ as defined by (8.3) coincides with the electrical flow induced by resistors of value $c_e/x_e(t)$. We now state our first main result, which is proved in Chapter 9.

Theorem 8.2. *Let $c \geqslant 0$ satisfy $c^{\mathrm{T}}|f| > 0$ for every non-zero f in the kernel of A. Let x^* be an optimum solution of* (8.2) *and let $X_\star$ be the set of optimum solutions. Assume $x(0) > 0$. The following holds for the dynamics (8.1) with q as in (8.3):*

(i) The solution $x(t)$ exists for all $t \geqslant 0$.
(ii) The cost $c^{\mathrm{T}}x(t)$ converges to $c^{\mathrm{T}}x^*$ as t goes to infinity.
(iii) The vector $x(t)$ converges to $X_\star$, i.e., $\lim_{t\to\infty} \inf\{\, \|x(t) - x'\| : x' \in X_\star \,\} \to 0$.
(iv) For all e with $c_e > 0$, $x_e(t) - |q_e(t)|$ converges to zero as t goes to infinity. [2] *If x^* is unique, $x(t)$ and $q(t)$ converge to x^* as t goes to infinity.*

Item (i) was previously shown in [SV16a] for the case of a strictly positive cost vector. The result in [SV16a] is actually stated only for the all-ones cost vector $c = \mathbf{1}$. The case of a general positive cost vector reduces to this special case by rescaling the solution vector x. Item (i) for the more general cost vector and items (ii) to (iv) are new. We stress that the dynamics (8.1) is biologically-grounded. It was proposed to model a biological system and not as an optimization method. Nevertheless, it can solve a large class of non-negative LPs. Table 8.1 summarizes our first main result and puts it into context.

Sections 9.1 and 9.2 are devoted to the proof of our first main theorem. For ease of exposition, we present the proof in two steps. In Section 9.1, we give a proof under the following simplifying assumptions:

(A) $c > 0$,
(B) The basic feasible solutions of (8.2) have distinct cost,
(C) We start with a positive vector $x(0) \in X_{\text{dom}} := \{\, x \in \mathbb{R}^n : \text{there is a feasible } f \text{ with } |f| \leqslant x \,\}$.

Section 9.1 generalizes [Bon13]. For the undirected shortest path problem, condition (B) states that all simple undirected source-sink paths have distinct cost and condition (C) states that all source-sink cuts have a capacity of at least one at time zero (and hence at all times). The existence of a solution with domain $[0,\infty)$ was already shown in [SV16a]. We will show that X_{dom} is an invariant set, i.e., the solution

[1] The kernel of a matrix A consists of all solutions to the system $Ax = 0$.
[2] We conjecture that this also holds for the indices e with $c_e = 0$.

Reference	**Problem** (Undirected Case)	**Existence of Solution**	**Convergence to OPT**	**Comments**
[MO08]	Shortest Path	Yes	Yes	parallel edges, planar graphs
[BMV12]	Shortest Path	Yes	Yes	all graphs
[SV16a]	Positive LP	Yes	No	$c > 0$
Our Result	Nonnegative LP	Yes	Yes	1) $c \geqslant 0$ 2) $\forall v \in \ker(A) : c^{\mathrm{T}}\lvert v\rvert > 0$

Table 8.1: Convergence results for the continuous undirected Physarum dynamics (8.1).

stays in X_{dom} for all times, and that $E(x) = \sum_e r_e x_e^2 = \sum_e c_e x_e$ is a Lyapunov function [3] [LaS76, Tes12] for the dynamics (8.1), i.e., $\dot{E} \leqslant 0$ and $\dot{E} = 0$ if and only if $\dot{x} = 0$. It follows from general theorems about dynamical systems that the dynamics converges to a fixed point of (8.1). The fixed points are precisely the vectors $|f|$ where f is a feasible solution of (8.2). A final argument establishes that the dynamics converges to a fixed point of minimum cost.

In Section 9.2, we prove the general case of the first main theorem. We assume

(D) $c \geqslant 0$,
(E) $\text{cost}(z) = c^{\mathrm{T}}|z| > 0$ for every non-zero vector z in the kernel of A,
(F) We start with a positive vector $x(0) > 0$.

Section 9.2 generalizes [BMV12] in two directions. First, we treat general undirected LPs and not just the undirected shortest path problem, respectively, the transshipment problem. Second, we replace the condition $c > 0$ by the requirement $c \geqslant 0$ and every non-zero vector in the kernel of A has positive cost. For the undirected shortest path problem, the latter condition states that the underlying undirected graph has no zero-cost cycle. Section 9.2 is technically considerably more difficult than Section 9.1. We first establish the existence of a solution with domain $[0, \infty)$. To this end, we derive a closed formula for the minimum energy feasible solution and prove that the mapping $x \mapsto q$ is locally Lipschitz. Existence of a solution with domain $[0, \infty)$ follows by standard arguments. We then show that X_{dom} is an attractor, i.e., the solution $x(t)$ converges to X_{dom}. We next characterize equilibrium points and exhibit a Lyapunov function. The Lyapunov function is a normalized version of $E(x)$. The normalization factor is equal to the optimal value of the linear program $\max\{\, \alpha : Af = \alpha b,\ |f| \leqslant x \,\}$ in the variables f and α. Convergence to an equilibrium point follows from the existence of a Lyapunov function. A final argument establishes that the dynamics converges to a fixed point of minimum cost.

8.2 The Biologically-Inspired Model

Ito et al. [IJNT11] initiated the study of the dynamics

$$\dot{x}(t) = q(t) - x(t). \tag{8.4}$$

We refer to this dynamics as the directed dynamics in contrast to the undirected dynamics (8.1). The directed dynamics is *biologically-inspired* – the similarity to (8.1) is the inspiration. It was never claimed to model the behavior of a biological system. Rather, it was introduced as a biologically-inspired optimization method. The work in [IJNT11] shows convergence of this directed dynamics (8.4) for the directed shortest path problem and [JZ12, SV16c, Bon16] show convergence for general *positive linear programs*, i.e., linear programs with positive cost vector $c > 0$ of the form

$$\min\{c^{\mathrm{T}}x : Ax = b,\ x \geqslant 0\}. \tag{8.5}$$

The *discrete versions* of both dynamics define sequences $x^{(t)}$, $t = 0, 1, 2, \ldots$ through

$$x^{(t+1)} = (1 - h^{(t)})x^{(t)} + h^{(t)}q^{(t)} \qquad \text{discrete directed dynamics;} \tag{8.6}$$

$$x^{(t+1)} = (1 - h^{(t)})x^{(t)} + h^{(t)}|q^{(t)}| \qquad \text{discrete undirected dynamics,} \tag{8.7}$$

where $h^{(t)}$ is the step size and $q^{(t)}$ is the minimum energy feasible solution as in (8.3). For the discrete dynamics, we can ask complexity questions. This is particularly relevant for the discrete directed dynamics as it was designed as an biologically-inspired optimization method.

[3] Lyapunov functions are a main tool for proving convergence of dynamical systems. It is a function $L(t)$ mapping the state $x(t)$ of the system to a non-negative real such that $\dot{L} \leqslant 0$ and $\dot{L} = 0$ iff $\dot{x} = 0$. It is an "art" to find a Lyapunov function for a concrete dynamical system.

For completeness, we review the state-of-the-art results for the discrete undirected dynamics. For the undirected shortest path problem, the convergence of the discrete undirected dynamics (8.7) was shown in [BBD$^+$13]. The convergence proof gives an upper bound on the step size and on the number of steps required until an ε-approximation of the optimum is obtained. [SV16b] extends the result to the transshipment problem and [SV16a] further generalizes the result to the case of positive LPs. The paper [SV16b] is related to our first result. It shows convergence of the discretized undirected dynamics (8.7), we show convergence of the continuous undirected dynamics (8.1) for a more general cost vector.

We come to the discrete directed Physarum-inspired dynamics (8.6). Similarly to the undirected setting, Becchetti et al. [BBD$^+$13] showed the convergence of (8.6) for the shortest path problem. Straszak and Vishnoi extended the analysis to the transshipment problem [SV16b] and positive LPs [SV16c].

Theorem 8.3. *[SV16c, Theorem 1.3] Let $A \in \mathbb{Z}^{n\times m}$ have full row rank ($n \leqslant m$), $b \in \mathbb{Z}^n$, $c \in \mathbb{Z}^m_{>0}$, and let $D_S \overset{\text{def}}{=} \max\{|\det(M)| \,:\, M$ is a square sub-matrix of $A\}$.[4] Suppose the Physarum-inspired dynamics (8.6) is initialized with a feasible point $x^{(0)}$ of (8.5) such that $M^{-1} \leqslant x^{(0)} \leqslant M$ and $c^{\mathrm{T}}x^{(0)} \leqslant M \cdot \mathrm{opt}$ for some $M \geqslant 1$, where opt denotes the optimum cost of (8.5). Then, for any $\varepsilon > 0$ and step size $h \leqslant \varepsilon/(\sqrt{6}\|c\|_1 D_S)^2$, after $k = O((\varepsilon h)^{-2}\ln M)$ steps, $x^{(k)}$ is a feasible solution with $c^{\mathrm{T}}x^{(k)} \leqslant (1+\varepsilon)\mathrm{opt}$.*

Theorem 8.3 gives an algorithm that computes a $(1+\varepsilon)$-approximation to the optimal cost of (8.5). In comparison to [BBD$^+$13, SV16b], it has several shortcomings. First, it requires a feasible starting point. Second, the step size depends linearly on ε. Third, the number of steps required to reach an ε-approximation has a quartic dependence on $\mathrm{opt}/(\varepsilon\Phi)$. In contrast, the analysis in [BBD$^+$13, SV16b] yields a step size independent of ε and a number of steps that depends only logarithmically on $1/\varepsilon$, see Table 8.2.

We overcome these shortcomings in Chapter 10. Before we can state our result, we need some notation. Let $X_\star$ be the set of optimal solutions to (8.5). The distance of a capacity vector x to $X_\star$ is defined as

$$\mathrm{dist}(x, X_\star) \overset{\text{def}}{=} \inf\{\|x - x'\|_\infty \,:\, x' \in X_\star\}.$$

Let $\gamma_A \overset{\text{def}}{=} \gcd(\{A_{ij} : A_{ij} \neq 0\}) \in \mathbb{Z}_{>0}$ and

$$D \overset{\text{def}}{=} \max\left\{|\det(M)| \,:\, M \text{ is a square submatrix of } A/\gamma_A \text{ with dimension } n-1 \text{ or } n\right\}. \tag{8.8}$$

Let $\mathcal{N}$ be the set of non-optimal basic feasible solutions of (8.5) and

$$\Phi \overset{\text{def}}{=} \min_{g\in\mathcal{N}} c^{\mathrm{T}}g - \mathrm{opt} \geqslant 1/(D\gamma_A)^2, \tag{8.9}$$

where the inequality is a well result in combinatorial optimization [PS82, Lemma 8.6]. For completeness, we present a proof in Section 10.5. Informally, our second main result proves the following properties of the Physarum-inspired dynamics (8.6):

(i) For any $\varepsilon > 0$ and any strongly dominating starting point [5] $x^{(0)}$, there is a fixed step size $h(x^{(0)})$ such that the Physarum-inspired dynamics (8.6) initialized with $x^{(0)}$ and $h(x^{(0)})$ converges to $X_\star$, i.e., $\mathrm{dist}(x^{(k)}, X_\star) < \varepsilon/(D\gamma_A)$ for large enough k.

(ii) The step size can be chosen *independently* of ε.

(iii) The number of steps k depends *logarithmically* on $1/\varepsilon$ and *quadratically* on opt/Φ.

(iv) The efficiency bounds depend on the *scale-invariant* determinant [6] D.

In Section 10.8, we establish a corresponding lower bound. We show that for the Physarum-inspired dynamics (8.6) to compute a point $x^{(k)}$ such that $\mathrm{dist}(x^{(k)}, X_\star) < \varepsilon$, the number of steps required for computing an ε-approximation has to grow linearly in $\mathrm{opt}/(h\Phi)$ and $\ln(1/\varepsilon)$, i.e. $k \geqslant \Omega(\mathrm{opt}\cdot(h\Phi)^{-1}\cdot \ln(1/\varepsilon))$. Table 8.2 puts our results into context.

[4] Using Lemma 9.10, the dependence on D_S can be improved to a scale-independent determinant D, defined in (8.8). For further details, we refer the reader to Section 10.2.

[5] We postpone the definition of strongly dominating capacity vector to Section 10.3. Every scaled feasible solution is strongly dominating. In the shortest path problem, a capacity vector x is strongly dominating if every source-sink cut $(S, \overline{S})$ has positive directed capacity, i.e., $\sum_{a\in E(S,\overline{S})} x_a - \sum_{a\in E(\overline{S},S)} x_a > 0$.

[6] Note that $(\gamma_A)^{n-1}D \leqslant D_S \leqslant (\gamma_A)^n D$, and thus D yields an exponential improvement over D_S, whenever $\gamma_A \geqslant 2$.

Reference	**Problem** (Directed Case)	**step size** h	**number of steps** k	**Guarantee**
[BBD+13]	Shortest Path	indep. of ε	$\text{poly}(m, n, \|c\|_1, \|x^{(0)}\|_1)$ $\cdot \ln(1/\varepsilon)$	$\text{dist}(x^{(k)}, X_\star) < \varepsilon$
[SV16b]	Transshipment	indep. of ε	$\text{poly}(m, n, \|c\|_1, \|b\|_1, \|x^{(0)}\|_1)$ $\cdot \ln(1/\varepsilon)$	$\text{dist}(x^{(k)}, X_\star) < \varepsilon$
[SV16c]	Positive LP	depends on ε	$\text{poly}(\|c\|_1, D_S, \ln \|x^{(0)}\|_1)$ $\cdot\ 1/(\Phi\varepsilon)^4$	$c^{\mathrm{T}} x^{(k)} \leqslant (1+\varepsilon)\text{opt}$ $c^{\mathrm{T}} x^{(k)} < \min_{g \in \mathcal{N}} c^{\mathrm{T}} g$
Our Result	Positive LP	indep. of ε	$\text{poly}(\|c\|_1, \|b\|_1, D, \gamma_A, \ln \|x^{(0)}\|_1)$ $\cdot \Phi^{-2} \ln(1/\varepsilon)$	$\text{dist}(x^{(k)}, X_\star) < \frac{\varepsilon}{D\gamma_A}$
Lower Bound	Positive LP	indep. of ε	$\Omega(\text{opt} \cdot (h\Phi)^{-1} \ln(1/\varepsilon))$	$\text{dist}(x^{(k)}, X_\star) < \varepsilon$

Table 8.2: Convergence results for the discrete directed Physarum-inspired dynamics (8.6).

We state now our second main result for the special case of a feasible starting point, and we provide the full version in Theorem 10.2 which applies for arbitrary strongly dominating starting point, see Section 10.1. We use the following constants in the statement of the bounds.

(i) $h_0 \stackrel{\text{def}}{=} c_{\min}/(4D\|c\|_1)$, where $c_{\min} \stackrel{\text{def}}{=} \min_i\{c_i\}$;

(ii) $\Psi^{(0)} \stackrel{\text{def}}{=} \max\{mD^2\|b/\gamma_A\|_1, \|x^{(0)}\|_\infty\}$;

(iii) $C_1 \stackrel{\text{def}}{=} D\|b/\gamma_A\|_1\|c\|_1$, $C_2 \stackrel{\text{def}}{=} 8^2 m^2 n D^5 \gamma_A^2 \|A\|_\infty \|b\|_1$ and $C_3 \stackrel{\text{def}}{=} D^3\gamma_A\|b\|_1\|c\|_1$.

Theorem 8.4. *Suppose* $A \in \mathbb{Z}^{n\times m}$ *has full row rank* ($n \leqslant m$), $b \in \mathbb{Z}^n$, $c \in \mathbb{Z}^m_{>0}$ *and* $\varepsilon \in (0,1)$. *Given a feasible starting point* $x^{(0)} > 0$ *the Physarum-inspired dynamics* (8.6) *with step size* $h \leqslant (\Phi/\text{opt}) \cdot h_0^2/2$ *outputs for any* $k \geqslant 4C_1/(h\Phi) \cdot \ln(C_2\Psi^{(0)}/(\varepsilon \cdot \min\{1, x^{(0)}_{\min}\}))$ *a feasible* $x^{(k)} > 0$ *such that* $\text{dist}(x^{(k)}, X_\star) < \varepsilon/(D\gamma_A)$.

We stated the bounds on h in terms of the unknown quantities Φ and opt. However, $\Phi/\text{opt} \geqslant 1/C_3$ by Lemma 9.10 and hence replacing Φ/opt by $1/C_3$ yields constructive bounds for h. Note that the upper bound on the step size does not depend on ε and that the bound on the number of iterations depends *logarithmically* on $1/\varepsilon$ and *quadratically* on opt/Φ.

What can be done if the initial point is not strongly dominating? For the transshipment problem it suffices to add an edge of high capacity and high cost from every source node to every sink node [BBD+13, SV16b]. This will make the instance strongly dominating and will not affect the optimal solution. We generalize this observation to positive linear programs. We add an additional column equal to b and give it sufficiently high capacity and cost. This guarantees that the resulting instance is strongly dominating and the optimal solution remains unaffected. Moreover, our approach generalizes and improves upon [SV16b, Theorem 1.2], see Section 10.7.

Proof Techniques: The crux of the analysis in [IJNT11, BBD+13, SV16b] is to show that for large enough k, $x^{(k)}$ is close to a *non-negative* flow $f^{(k)}$ and then to argue that $f^{(k)}$ is close to an optimal flow $f^\star$. This line of arguments yields a convergence of $x^{(k)}$ to $X_\star$ with a step size h chosen independently of ε.

In Chapter 10, we extend the preceding approach to positive linear programs, by generalizing the concept of non-negative cycle-free flows to non-negative *feasible kernel-free* vectors (Section 10.4). Although, we use the same high level ideas as in [BBD+13, SV16b], we stress that our analysis generalizes all relevant lemmas in [BBD+13, SV16b] and it uses arguments from linear algebra and linear programming duality, instead of combinatorial arguments. Further, our core efficiency bounds (Section 10.2) extend [SV16c] and yield a *scale-invariant* determinant dependence of the step size and are applicable for any strongly dominating starting point (Section 10.3).

Chapter 9

Biologically-Grounded Physarum Dynamics

9.1 Convergence: Simple Instances

In this section, we prove Theorem 8.2 under the simplifying assumptions (A) to (C), defined in page 88.

9.1.1 Preliminaries

Note that we may assume that A has full row-rank since any equation that is linearly dependent on other equations can be deleted without changing the feasible set. We continue to use n and m for the dimension of A. Thus, A has rank n. We continue by fixing some terms and notation. A *basic feasible solution* of (8.2) is a pair of vectors x and $f = (f_B, f_N)$, where $f_B = A_B^{-1}b$ and A_B is a square $n \times n$ non-singular sub-matrix of A and $f_N = 0$ is the vector indexed by the coordinates not in B, and $x = |f|$. Since f uniquely determines x, we may drop the latter for the sake of brevity and call f a basic feasible solution of (8.2). A feasible solution f is *kernel-free* or *non-circulatory* if it is contained in the convex hull of the basic feasible solutions.[1] We say that a vector f' is *sign-compatible* with a vector f (of the same dimension) or f-sign-compatible if $f'_e \neq 0$ implies $f'_e f_e > 0$. In particular, $\mathrm{supp}(f') \subseteq \mathrm{supp}(f)$. For a given capacity vector x and a vector $f \in \mathbb{R}^m$ with $\mathrm{supp}(f) \subseteq \mathrm{supp}(x)$, we use $E(f) = \sum_e (c_e/x_e) f_e^2$ to denote the *energy* of f. The energy of f is infinite, if $\mathrm{supp}(f) \not\subseteq \mathrm{supp}(x)$. We use $\mathrm{cost}(f) = \sum_e c_e |f_e| = c^{\mathrm{T}}|f|$ to denote the *cost* of f. Note that $E(x) = \sum_e (c_e/x_e) x_e^2 = \sum_e c_e x_e = \mathrm{cost}(x)$. We define the constants $c_{\max} = \|c\|_\infty$ and $c_{\min} = \min_{e: c_e > 0} c_e$.

We use the following corollary of the finite basis theorem for polyhedra.

Lemma 9.1. *Let f be a feasible solution of (8.2). Then f is the sum of a convex combination of at most m basic feasible solutions plus a vector in the kernel of A. Moreover, all elements in this representation are sign-compatible with f.*

Proof. We may assume $f \geqslant 0$. Otherwise, we flip the sign of the appropriate columns of A. Thus, the system $Af = b,\ f \geqslant 0$ is feasible and f is the sum of a convex combination of at most m basic feasible solutions plus a vector in the kernel of A by the finite basis theorem [Sch99, Corollary 7.1b]. By definition, the elements in this representation are non-negative vectors and hence sign-compatible with f. □

Lemma 9.2 (Grönwall's Lemma). *Let $A, B, \alpha, \beta \in \mathbb{R}$, $\alpha \neq 0$, $\beta \neq 0$, and let g be a continuous differentiable function on $[0, \infty)$. If $A + \alpha g(t) \leqslant \dot{g}(t) \leqslant B + \beta g(t)$ for all $t \geqslant 0$, then $-A/\alpha + (g(0) + A/\alpha)e^{\alpha t} \leqslant g(t) \leqslant -B/\beta + (g(0) + B/\beta)e^{\beta t}$ for all $t \geqslant 0$.*

Proof. We show the upper bound. Assume first that $B = 0$. Then

$$\frac{d}{dt}\frac{g}{e^{\beta t}} = \frac{\dot{g}e^{\beta t} - \beta g e^{\beta t}}{e^{2\beta t}} \leqslant 0 \quad \text{implies} \quad \frac{g(t)}{e^{\beta t}} \leqslant \frac{g(0)}{e^{\beta 0}} = g(0).$$

If $B \neq 0$, define $h(t) = g(t) + B/\beta$. Then

$$\dot{h} = \dot{g} \leqslant B + \beta g = B + \beta(h - B/\beta) = \beta h$$

and hence $h(t) \leqslant h(0)e^{\beta t}$. Therefore $g(t) \leqslant -B/\beta + (g(0) + B/\beta)e^{\beta t}$. □

An immediate consequence of Grönwall's Lemma is that the undirected Physarum dynamics (8.1) initialized with any positive starting vector $x(0)$, generates a trajectory $\{x(t)\}_{t \geqslant 0}$ such that each time state $x(t)$ is a positive vector. Indeed, since $\dot{x}_e = |q_e| - x_e \geqslant -x_e$, we have $x_e(t) \geqslant x_e(0) \cdot \exp\{-t\}$ for every index e with $x_e(0) > 0$ and every time t. Further, by (8.1) and (8.3), it holds for indices e with $x_e(0) = 0$ that $x_e(t) = 0$ for every time t. Hence, the trajectory $\{x(t)\}_{t \geqslant 0}$ has a time-invariant support.

[1] For the undirected shortest path problem, we drop the equation corresponding to the sink. Then b becomes the negative indicator vector corresponding to the source node. Note that n is one less than the number of nodes of the graph. The basic feasible solutions are the simple undirected source-sink paths. A circulatory solution contains a cycle on which there is flow.

Lemma 9.3 ([JZ12]). *Let $R = \operatorname{diag}(c_e/x_e)$. Then $q = R^{-1}A^{\mathrm{T}}p$, where $p = (AR^{-1}A^{\mathrm{T}})^{-1}b$.*

Proof. q minimizes $\sum_e r_e q_e^2$ subject to $Aq = b$. The Karush-Kuhn-Tucker (KKT) optimality conditions for constrained optimization [Boy04] imply the existence of a vector p such that $Rq = A^{\mathrm{T}}p$. Substituting into $Aq = b$ yields $p = (AR^{-1}A^{\mathrm{T}})^{-1}b$. □

Lemma 9.4. *X_{dom} is an invariant set, i.e., if $x(0) \in X_{\mathrm{dom}}$ then $x(t) \in X_{\mathrm{dom}}$ for all t.*

Proof. Let $q(t)$ be the minimum energy feasible solution with respect to $R(t) = \operatorname{diag}(c_e/x_e(t))$, and let $f(t)$ be such that $f(0)$ is feasible, $|f(0)| \leqslant x(0)$, and $\dot{f}(t) = q(t) - f(t)$. Then $\frac{d}{dt}(Af - b) = A(q - f) = b - Af$ and hence $Af(t) - b = (Af(0) - b)e^{-t} = 0$. Thus $f(t)$ is feasible for all t. Moreover,

$$\frac{d}{dt}(f - x) = \dot{f} - \dot{x} = q - f - (|q| - x) = q - |q| - (f - x) \leqslant -(f - x).$$

Thus $f(t) - x(t) \leqslant (f(0) - x(0))e^{-t} \leqslant 0$ by Grönwall's Lemma applied with $g(t) = f(t) - x(t)$ and $\beta = -1$, and hence $f(t) \leqslant x(t)$ for all t. Similarly,

$$\frac{d}{dt}(f + x) = \dot{f} + \dot{x} = q - f + (|q| - x) = q + |q| - (f + x) \geqslant -(f + x).$$

Thus $f(t) + x(t) \geqslant (f(0) + x(0))e^{-t} \geqslant 0$ by Grönwall's Lemma applied with $g(t) = f(t) + x(t)$ and $\alpha = -1$ and $A = 0$, and hence $f(t) \geqslant -x(t)$ for all t.

We conclude that $|f(t)| \leqslant x(t)$ for all t. Thus, $x(t) \in X_{\mathrm{dom}}$ for all t. □

9.1.2 The Convergence Proof

We will first characterize the equilibrium points. They are precisely the points $|f|$, where f is a basic feasible solution; the proof uses assumption (B) in page 88. We then show that $E(x)$ is a Lyapunov function for (8.1), in particular, $\dot{E} \leqslant 0$ and $\dot{E} = 0$ if and only if x is an equilibrium point. For this argument, we need that the energy of q is at most the energy of x with equality if and only if x is an equilibrium point. This proof uses assumptions (A) and (C) in page 88. It follows from the general theory of dynamical systems that $x(t)$ approaches an equilibrium point. Finally, we show that convergence to a non-optimal equilibrium is impossible.

Lemma 9.5 (Generalization of Lemma 2.3 in [Bon13]). *Assume (A) to (C). If f is a basic feasible solution of (8.2), then $x = |f|$ is an equilibrium point. Conversely, if x is an equilibrium point, then $x = |f|$ for some basic feasible solution f.*

Proof. Let f be a basic feasible solution, let $x = |f|$, and let q be the minimum energy feasible solution with respect to the resistances c_e/x_e. We have $Aq = b$ and $\operatorname{supp}(q) \subseteq \operatorname{supp}(x)$ by definition of q. Since f is a basic feasible solution there is a subset B of size n of the columns of A such that A_B is non-singular and $f = (A_B^{-1}b, 0)$. Since $\operatorname{supp}(q) \subseteq \operatorname{supp}(x) \subseteq B$, we have $q = (q_B, 0)$ for some vector q_B. Thus, $b = Aq = A_B q_B$ and hence $q_B = f_B$. Therefore $\dot{x} = |q| - x = 0$ and x is an equilibrium point.

Conversely, if x is an equilibrium point, $|q_e| = x_e$ for every e. By changing the signs of some columns of A, we may assume $q \geqslant 0$. Then $q = x$. Since $q_e = x_e/c_e A_e^{\mathrm{T}}p$ where A_e is the e-th column of A by Lemma 9.3, we have $c_e = A_e^{\mathrm{T}}p$, whenever $x_e > 0$. By Lemma 9.1, q is a convex combination of basic feasible solutions and a vector in the kernel of A that are sign-compatible with q. The vector in the kernel must be zero as q is a minimum energy feasible solution. For any basic feasible solution z contributing to q, we have $\operatorname{supp}(z) \subseteq \operatorname{supp}(x)$. Summing over the $e \in \operatorname{supp}(z)$, we obtain $\operatorname{cost}(z) = \sum_{e \in \operatorname{supp}(z)} c_e z_e = \sum_{e \in \operatorname{supp}(z)} z_e A_e^{\mathrm{T}} p = b^{\mathrm{T}}p$. Thus, the convex combination involves only a single basic feasible solution by assumption (B) and hence x is a basic feasible solution. □

The vector $x(t)$ dominates a feasible solution at all times. Since $q(t)$ is the minimum energy feasible solution at time t, this implies $E(q(t)) \leqslant E(x(t))$ at all times. A further argument shows that we have equality if and only if $x = |q|$.

Lemma 9.6 (Generalization of Lemma 3.1 in [Bon13]). *Assume (A) to (C). At all times, $E(q) \leqslant E(x)$. If $E(q) = E(x)$, then $x = |q|$.*

Proof. Recall that $x(t) \in X_{\text{dom}}$ for all t. Thus, at all times, there is a feasible f such that $|f| \leqslant x$. Since q is a minimum energy feasible solution, we have

$$E(q) \leqslant E(f) \leqslant E(x).$$

If $E(q) = E(x)$ then $E(q) = E(f)$ and hence $q = f$ since the minimum energy feasible solution is unique. Also, $|f| = x$ since $|f| \leqslant x$ and $|f_e| < x_e$ for some e implies $E(f) < E(x)$. The last conclusion uses $c > 0$. □

Lyapunov functions are the main tool for proving convergence of dynamical systems. We show that $E(x)$ is a Lyapunov function for (8.1).

Lemma 9.7 (Generalization of Lemma 3.2 in [Bon13])**.** *Assume* (A) *to* (C)*. $E(x)$ is a Lyapunov function for* (8.1)*, i.e., it is continuous as a function of x, $E(x) \geqslant 0$, $\dot{E}(x) \leqslant 0$ and $\dot{E}(x) = 0$ if and only if $\dot{x} = 0$.*

Proof. E is clearly continuous and non-negative. Recall that $E(x) = \text{cost}(x)$. Let R be the diagonal matrix with entries c_e/x_e. Then

$$\begin{aligned}
\frac{d}{dt}\text{cost}(x) &= c^{\mathrm{T}}(|q| - x) && \text{by (8.1)}\\
&= x^{\mathrm{T}}R|q| - x^{\mathrm{T}}Rx && \text{since } c = Rx\\
&= x^{\mathrm{T}}R^{1/2}R^{1/2}|q| - x^{\mathrm{T}}Rx &&\\
&\leqslant (q^{\mathrm{T}}Rq)^{1/2}(x^{\mathrm{T}}Rx)^{1/2} - x^{\mathrm{T}}Rx && \text{by Cauchy-Schwarz}\\
&\leqslant (x^{\mathrm{T}}Rx)^{1/2}(x^{\mathrm{T}}Rx)^{1/2} - x^{\mathrm{T}}Rx && \text{by Lemma 9.6}\\
&= 0.
\end{aligned}$$

Observe that $\frac{d}{dt}\text{cost}(x) = 0$ implies that both inequalities above are equalities. This is only possible if the vectors $|q|$ and x are parallel and $E(q) = E(x)$. Thus, $x = |q|$ by Lemma 9.6. □

It follows now from the general theory of dynamical systems that $x(t)$ converges to an equilibrium point.

Corollary 9.8 (Generalization of Corollary 3.3. in [Bon13].)**.** *Assume* (A) *to* (C)*. As $t \to \infty$, $x(t)$ approaches an equilibrium point and $c^{\mathrm{T}}x(t)$ approaches the cost of the corresponding basic feasible solution.*

Proof. The proof in [Bon13] carries over. We include it for completeness. The existence of a Lyapunov function E implies by [LaS76, Corollary 2.6.5] that $x(t)$ approaches the set $\left\{ x \in \mathbb{R}^m_{\geqslant 0} : \dot{E} = 0 \right\}$, which by Lemma 9.7 is the same as the set $\left\{ x \in \mathbb{R}^m_{\geqslant 0} : \dot{x} = 0 \right\}$. Since this set consists of isolated points (Lemma 9.5), $x(t)$ must approach one of those points, say the point x_0. When $x = x_0$, one has $E(q) = E(x) = \text{cost}(x) = c^{\mathrm{T}}x$. □

It remains to exclude that $x(t)$ converges to a non-optimal equilibrium point.

Theorem 9.9 (Generalization of Theorem 3.4 in [Bon13])**.** *Assume* (A) *to* (C)*. As $t \to \infty$, $c^{\mathrm{T}}x(t)$ converges to the cost of the optimal solution and $x(t)$ converges to the optimal solution.*

Proof. By the corollary, it suffices to prove the second part of the claim. For the second part, assume that $x(t)$ converges to a non-optimal solution z. Let x^* be the optimal solution and let $W = \sum_e x^*_e c_e \ln x_e$. Let $\delta = (\text{cost}(z) - \text{cost}(x^*))/2$. Note that for all sufficiently large t, we have $E(q(t)) \geqslant \text{cost}(z) - \delta \geqslant \text{cost}(x^*) + \delta$. Further, by definition $q_e = (x_e/c_e)A^{\mathrm{T}}_e p$ and thus

$$\dot{W} = \sum_e x^*_e c_e \frac{|q_e| - x_e}{x_e} = \sum_e x^*_e |A^{\mathrm{T}}_e p| - \text{cost}(x^*) \geqslant \sum_e x^*_e A^{\mathrm{T}}_e p - \text{cost}(x^*) \geqslant \delta,$$

where the last inequality follows by $\sum_e x^*_e A^{\mathrm{T}}_e p = b^{\mathrm{T}}p = E(q) \geqslant \text{cost}(x^*) + \delta$. Hence $W \to \infty$, a contradiction to the fact that x is bounded. □

9.2 Convergence: General Instances

In this section, we prove Theorem 8.2 under the more general assumptions (D) to (F), defined in page 89.

9.2.1 Existence of a Solution with Domain $[0, \infty)$

In this subsection we show that a solution $x(t)$ to (8.1) has domain $[0, \infty)$. We first derive an explicit formula for the minimum energy feasible solution q and then show that the mapping $x \mapsto q$ is Lipschitz continuous; this implies existence of a solution with domain $[0, \infty)$ by standard arguments.

The Minimum Energy Solution

Recall that for $\gamma_A = \gcd(\{A_{ij} : A_{ij} \neq 0\}) \in \mathbb{Z}_{>0}$, we defined by

$$D = \max \{|\det(M)| \,:\, M \text{ is a square submatrix of } A/\gamma_A \text{ with dimension } n-1 \text{ or } n\}.$$

We derive now properties of the minimum energy solution. In particular, if every non-zero vector in the kernel of A has positive cost,

(i) the minimum energy feasible solution is kernel-free and unique (Lemma 9.11),
(ii) $|q_e| \leqslant D\|b/\gamma_A\|_1$ for every $e \in [m]$ (Lemma 9.12),
(iii) q is defined by (9.3) (Lemma 9.13), and
(iv) $E(q) = b^{\mathrm{T}}p$, where p is defined by (9.3) (Lemma 9.14).

We note that for positive cost vector $c > 0$, these results are known.

We proceed by establishing some useful properties on basic feasible solutions.

Lemma 9.10. *Suppose $A \in \mathbb{Z}^{n\times m}$ is an integral matrix, and $b \in \mathbb{Z}^n$ is an integral vector. Then, for any basic feasible solutions f with $Af = b$ and $f \geqslant 0$, it holds that $\|f\|_\infty \leqslant D\|b/\gamma_A\|_1$ and $f_j \neq 0$ implies $|f_j| \geqslant 1/(D\gamma_A)$.*

Proof. Since f is a basic feasible solution, it has the form $f = (f_B, 0)$ such that $A_B \cdot f_B = b$ where $A_B \in \mathbb{Z}^{n\times n}$ is an invertible submatrix of A. We write $M_{-i,-j}$ to denote the matrix M with deleted i-th row and j-th column. Let Q_j be the matrix formed by replacing the j-th column of A_B by the column vector b. Then, using the fact that $\det(tA) = t^n \det(A)$ for every $A \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{Z}$, Cramer's rule yields

$$|f_B(j)| = \left|\frac{\det(Q_j)}{\det(A_B)}\right| = \frac{1}{\gamma_A}\left|\sum_{k=1}^{n} \frac{(-1)^{j+k} \cdot b_k \cdot \det\left(\gamma_A^{-1}[A_B]_{-k,-j}\right)}{\det\left(\gamma_A^{-1}A_B\right)}\right|$$

By the choice of γ_A, the values $\det(\gamma_A^{-1}A_B)$ and $\det(\gamma_A^{-1}[A_B]_{-k,-j})$ are integral for all k, it follows that

$$|f_B(j)| \leqslant D\|b/\gamma_A\|_1 \qquad \text{and} \qquad f_B(j) \neq 0 \quad \Longrightarrow \quad \frac{1}{D\gamma_A} \leqslant |f_B(j)|\,. \qquad \square$$

Lemma 9.11. *If every non-zero vector in the kernel of A has positive cost, the minimum energy feasible solution is kernel-free and unique.*

Proof. Let q be a minimum energy feasible solution. Since q is feasible, it can be written as $q_n + q_r$, where q_n is a convex combination of basic feasible solutions and q_r lies in the kernel of A. Moreover, all elements in this representation are sign-compatible with q by Lemma 9.1. If $q_r \neq 0$, the vector $q - q_r$ is feasible and has smaller energy, a contradiction. Thus $q_r = 0$.

We next prove uniqueness. Assume for the sake of a contradiction that there are two distinct minimum energy feasible solutions $q^{(1)}$ and $q^{(2)}$. We show that the solution $(q^{(1)} + q^{(2)})/2$ uses less energy than $q^{(1)}$ and $q^{(2)}$. Since $h \mapsto h^2$ is a strictly convex function from $\mathbb{R}$ to $\mathbb{R}$, the average of the two solutions will be better than either solution if there is an index e with $r_e > 0$ and $q_e^{(1)} \neq q_e^{(2)}$. The difference $z = q^{(1)} - q^{(2)}$ lies in the kernel of A and hence $\mathrm{cost}(z) = \sum_e c_e|z_e| > 0$. Thus there is an e with $c_e > 0$ and $z_e \neq 0$. We have now shown uniqueness. □

Lemma 9.12. *Assume that every non-zero vector in the kernel of A has positive cost. Let q be the minimum energy feasible solution. Then $|q_e| \leqslant D\|b/\gamma_A\|_1$ for every e.*

Proof. Since q is a convex combination of basic feasible solutions, $|q_e| \leqslant \max_z |z_e|$ where z ranges over basic feasible solutions of the form $(z_B, 0)$, where $z_B = A_B^{-1}b$ and $A_B \in \mathbb{R}^{n\times n}$ is a non-singular submatrix of A. Thus, by Lemma 9.10 every component of z is bounded by $D\|b/\gamma_A\|_1$. □

In [SV16c], the bound $|q_e| \leqslant D^2 m \|b\|_1$ was shown. We will now derive explicit formulae for the minimum energy solution q. We will express q in terms of a vector $p \in \mathbb{R}^n$, which we refer to as the *potential*, by analogy with the network setting, in which p can be interpreted as the electric potential of the nodes. The energy of the minimum energy solution is equal to $b^{\mathrm{T}}p$. We show that the mapping $x \mapsto q$ is locally Lipschitz. Note that for $c > 0$ these facts are well-known. Let us split the column indices $[m]$ of A into

$$P \stackrel{\text{def}}{=} \{\, e \in [m] : c_e > 0 \,\} \quad \text{and} \quad Z \stackrel{\text{def}}{=} \{\, e \in [m] : c_e = 0 \,\}. \tag{9.1}$$

Lemma 9.13. *Assume that every non-zero vector in the kernel of A has positive cost. Let $r_e = c_e/x_e$ and let R denote the corresponding diagonal matrix. Let us split A into A_P and A_Z, and q into q_P and q_Z. Since A_Z has linearly independent columns, we may assume that the first $|Z|$ rows of A_Z form a square non-singular matrix. We can thus write $A = \left[\begin{smallmatrix} A'_P & A'_Z \\ A''_P & A''_Z \end{smallmatrix}\right]$ with invertible A'_Z. Then the minimum energy solution satisfies*

$$\begin{bmatrix} A'_P & A'_Z \\ A''_P & A''_Z \end{bmatrix} \begin{bmatrix} q_P \\ q_Z \end{bmatrix} = \begin{bmatrix} b' \\ b'' \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} R_P & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} q_P \\ q_Z \end{bmatrix} = \begin{bmatrix} {A'_P}^{\mathrm{T}} & {A''_P}^{\mathrm{T}} \\ {A'_Z}^{\mathrm{T}} & {A''_Z}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} p' \\ p'' \end{bmatrix} \tag{9.2}$$

for some vector $p = \left[\begin{smallmatrix} p' \\ p'' \end{smallmatrix}\right]$; here p' has dimension $|Z|$. The equation system (9.2) has a unique solution given by

$$\begin{bmatrix} q_Z \\ q_P \end{bmatrix} = \begin{bmatrix} [A'_Z]^{-1}(b' - A'_P q_P) \\ R_P^{-1} A_P^{\mathrm{T}} p \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} p' \\ p'' \end{bmatrix} = \begin{bmatrix} -[[A'_Z]^{\mathrm{T}}]^{-1}[A''_Z]^{\mathrm{T}} p'' \\ M R^{-1} M^{\mathrm{T}} (b'' - A''_Z [A'_Z]^{-1} b') \end{bmatrix}, \tag{9.3}$$

where $M = A''_P - A''_Z [A'_Z]^{-1} A'_P$ is the Schur complement of the block A'_Z of the matrix A.

Proof. q minimizes $E(f) = f^{\mathrm{T}} R f$ among all solutions of $Af = b$. The KKT conditions state that q must satisfy $Rq = A^{\mathrm{T}} p$ for some p. Note that $2Rf$ is the gradient of the energy $E(f)$ with respect to f and that the $-A^{\mathrm{T}} p$ is the gradient of $p^{\mathrm{T}}(b - Af)$ with respect to f. We may absorb the factor -2 in p. Thus q satisfies (9.2).

We show next that the linear system (9.2) has a unique solution. The top $|Z|$ rows of the left system in (9.2) give

$$q_Z = [A'_Z]^{-1}(b' - A'_P q_P). \tag{9.4}$$

Substituting this expression for q_Z into the bottom $n - |Z|$ rows of the left system in (9.2) yields

$$M q_P = b'' - A''_Z [A'_Z]^{-1} b'.$$

From the top $|P|$ rows of the right system in (9.2) we infer $q_P = R_P^{-1} A_P^{\mathrm{T}} \cdot p$. Thus

$$M R_P^{-1} A_P^{\mathrm{T}} \cdot p = b'' - A''_Z [A'_Z]^{-1} \cdot b'. \tag{9.5}$$

The bottom $n - |Z|$ rows of the right system in (9.2) yield $0 = A_Z^{\mathrm{T}} p = [A'_Z]^{\mathrm{T}} p' + [A''_Z]^{\mathrm{T}} p''$ and hence

$$p' = -[[A'_Z]^{\mathrm{T}}]^{-1}[A''_Z]^{\mathrm{T}} p''. \tag{9.6}$$

Substituting (9.6) into (9.5) yields

$$\begin{aligned} b'' - A''_Z [A'_Z]^{-1} b' &= M R_P^{-1} \left([A'_P]^{\mathrm{T}} p' + [A''_P]^{\mathrm{T}} p'' \right) \\ &= M R_P^{-1} \left([A''_P]^{\mathrm{T}} - [A'_P]^{\mathrm{T}} [[A'_Z]^{\mathrm{T}}]^{-1} [A''_Z]^{\mathrm{T}} \right) p'' \\ &= M R_P^{-1} M^{\mathrm{T}} p''. \end{aligned} \tag{9.7}$$

It remains to show that the matrix $M R_P^{-1} M^{\mathrm{T}}$ is non-singular. We first observe that the rows of M are linearly independent. Consider the left system in (9.2). Multiplying the first $|Z|$ rows by $[A'_Z]^{-1}$ and then subtracting A''_Z times the resulting rows from the last $n - |Z|$ rows turns A into the matrix $Q = \left[\begin{smallmatrix} [A'_Z]^{-1} A'_P & I \\ M & 0 \end{smallmatrix}\right]$. By assumption, A has independent rows. Moreover, the preceding operations guarantee that $\operatorname{rank}(A) = \operatorname{rank}(Q)$. Therefore, M has independent rows. Since R_P^{-1} is a positive diagonal matrix, $R_P^{-1/2}$ exists and is a positive diagonal matrix. Let z be an arbitrary non-zero vector of dimension $|P|$. Then $z^{\mathrm{T}} M R_P^{-1} M^{\mathrm{T}} z = (R_P^{-1/2} M^{\mathrm{T}} z)^{\mathrm{T}} (R_P^{-1/2} M^{\mathrm{T}} z) > 0$ and hence $M R_P^{-1} M^{\mathrm{T}}$ is non-singular. It is even positive semi-definite.

There is a shorter proof that the system (9.2) has a unique solution. However, the argument does not give an explicit expression for the solution. In the case of a convex objective function and affine constraints, the KKT conditions are sufficient for being a global minimum. Thus any solution to (9.2) is a global optimum. We have already shown in Lemma 9.11 that the global minimum is unique. □

We next observe that the energy of q can be expressed in terms of the potential.

Lemma 9.14. *Let q be the minimum energy feasible solution and let f be any feasible solution. Then $E(q) = b^{\mathrm{T}}p = f^{\mathrm{T}}A^{\mathrm{T}}p$.*

Proof. As in the proof of Lemma 9.13, we split q into q_P and q_Z, R into R_P and R_Z, and A into A_P and A_Z. Then

$$\begin{aligned} E(q) &= q_P^{\mathrm{T}} R_P q_P && \text{by the definition of } E(q) \text{ and since } R_Z = 0 \\ &= p^{\mathrm{T}} A_P q_p && \text{by the right system in (9.2)} \\ &= p^{\mathrm{T}}(b - A_Z q_Z) && \text{by the left system in (9.2)} \\ &= b^{\mathrm{T}} p && \text{by the right system in (9.2).} \end{aligned}$$

For any feasible solution f, we have $f^{\mathrm{T}}A^{\mathrm{T}}p = b^{\mathrm{T}}p$. □

The Mapping $x \mapsto q$ is Locally Lipschitz

We show that the mapping $x \mapsto q$ is locally Lipschitz continuous; this implies existence of a solution $x(t)$ with domain $[0, \infty)$ by standard arguments. Our analysis builds upon Cramer's rule and the Cauchy-Binet formula. The Cauchy-Binet formula extends Kirchhoff's spanning tree theorem which was used in [BMV12] for the analysis of the undirected shortest path problem.

Lemma 9.15 (Local Lipschitz Condition)**.** *Assume $c \geqslant 0$, no non-zero vector in the kernel of A has cost zero, and that A, b, and c are integral. Let $\alpha, \beta > 0$. For any two vectors x and $\widetilde{x}$ in $\mathbb{R}^m$ with $\alpha \leqslant x_e, \widetilde{x}_e \leqslant \beta$ for all e, define $\gamma \overset{\text{def}}{=} 2m^n(\beta/\alpha)^n c_{\max}^n D^2 \|b/\gamma_A\|_1$. Then $\big||q_e(x)| - |q_e(\widetilde{x})|\big| \leqslant \gamma \|x - \widetilde{x}\|_\infty$ for every $e \in [m]$.*

Proof. First assume that $c > 0$. By Cramer's rule

$$(AR^{-1}A^{\mathrm{T}})^{-1} = \frac{1}{\det(AR^{-1}A^{\mathrm{T}})}((-1)^{i+j}\det(M_{-j,-i}))_{ij},$$

where $M_{-i,-j}$ is obtained from $AR^{-1}A^{\mathrm{T}}$ by deleting the i-th row and the j-th column. For a subset S of $[m]$ and an index $i \in [n]$, let A_S be the $n \times |S|$ matrix consisting of the columns selected by S and let $A_{-i,S}$ be the matrix obtained from A_S by deleting row i. If D is a diagonal matrix of size m, then $(AD)_S = A_S D_S$. The Cauchy-Binet theorem expresses the determinant of a product of two matrices (not necessarily square) as a sum of determinants of square matrices. It yields

$$\begin{aligned} \det(AR^{-1}A^{\mathrm{T}}) &= \sum_{S \subseteq [m];\ |S| = n} (\det((AR^{-1/2})_S))^2 \\ &= \sum_{S \subseteq [m];\ |S| = n} (\prod_{e \in S} x_e/c_e) \cdot (\det A_S)^2. \end{aligned}$$

Similarly,

$$\det(AR^{-1}A^{\mathrm{T}})_{-i,-j} = \sum_{S \subseteq [m];\ |S| = n-1} (\prod_{e \in S} x_e/c_e) \cdot (\det A_{-i,S} \cdot \det A_{-j,S}).$$

Using $p = (AR^{-1}A^{\mathrm{T}})^{-1}b$, we obtain

$$p_i = \frac{\sum_{j \in [n]} (-1)^{i+j} \sum_{S \subseteq [m];\ |S| = n-1} (\prod_{e \in S} x_e/c_e) \cdot (\det A_{-i,S} \cdot \det A_{-j,S}) b_j}{\sum_{S \subseteq [m];\ |S| = n} (\prod_{e \in S} x_e/c_e) \cdot (\det A_S)^2}. \tag{9.8}$$

Substituting into $q = R^{-1}A^{\mathrm{T}}p$ yields

$$\begin{aligned} q_e &= \frac{x_e}{c_e} A_e^{\mathrm{T}} p \\ &= \frac{x_e}{c_e} \sum_i A_{i,e} \cdot \frac{\sum_{j \in [n]} (-1)^{i+j+2n} \sum_{S \subseteq [m];\ |S| = n-1} (\prod_{e' \in S} x_{e'}/c_{e'}) \cdot (\det A_{-i,S} \cdot \det A_{-j,S}) b_j}{\sum_{S \subseteq [m];\ |S| = n} (\prod_{e' \in S} x_{e'}/c_{e'}) \cdot (\det A_S)^2} \\ &= \frac{\sum_{S \subseteq [m];\ |S| = n-1} (\prod_{e' \in S \cup e} x_{e'}/c_{e'}) \cdot \sum_{i \in [n]} (-1)^{i+n} A_{i,e} \det A_{-i,S} \cdot \sum_{j \in [n]} (-1)^{j+n} b_j \det A_{-j,S}}{\sum_{S \subseteq [m];\ |S| = n} (\prod_{e' \in S} x_{e'}/c_{e'}) \cdot (\det A_S)^2} \\ &= \frac{\sum_{S \subseteq [m];\ |S| = n-1} (\prod_{e' \in S \cup e} x_{e'}/c_{e'}) \cdot \det(A_S|A_e) \cdot \det(A_S|b)}{\sum_{S \subseteq [m];\ |S| = n} (\prod_{e' \in S} x_{e'}/c_{e'}) \cdot (\det A_S)^2}, \end{aligned} \tag{9.9}$$

where $(A_S|A_e)$, respectively $(A_S|b)$, denotes the $n \times n$ matrix whose columns are selected from A by S and whose last column is equal to A_e, respectively b.

We are now ready to estimate the derivative $\partial q_e/\partial x_i$. Assume first that $e \neq i$. By the above, $q_e = \frac{x_e}{c_e}\frac{F+Gx_i/c_i}{H+Ix_i/c_i}$, where F, G, H and I are given implicitly by (9.9). Then

$$\left|\frac{\partial q_e}{\partial x_i}\right| = \left|\frac{x_e}{c_e} \cdot \frac{FI/c_i - GH/c_i}{(H+Ix_i/c_i)^2}\right| \leqslant \frac{2\cdot\binom{m}{n-1}\beta^n D^2\|b/\gamma_A\|_1}{(\alpha/c_{\max})^n} \leqslant \gamma.$$

For $e = i$, we have $q_e = \frac{Gx_e/c_e}{H+Ix_e/c_e}$, where G, H, and I are given implicitly by (9.9). Then

$$\left|\frac{\partial q_e}{\partial x_e}\right| = \left|\frac{GH/c_e}{(H+Ix_e/c_e)^2}\right| \leqslant \frac{\binom{m}{n-1}\beta^n D^2\|b/\gamma_A\|_1}{(\alpha/c_{\max})^n} \leqslant \gamma.$$

Finally, consider x and $\widetilde{x}$ with $\alpha \leqslant x_e, \widetilde{x}_e \leqslant \beta$ for all e. Let $\bar{x}_\ell = (\widetilde{x}_1, \ldots, \widetilde{x}_\ell, x_{\ell+1}, \ldots, x_m)$. Then

$$||q_e(x)| - |q_e(\widetilde{x})|| \leqslant |q_e(x) - q_e(\widetilde{x})| \leqslant \sum_{0\leqslant\ell<m} |q_e(\bar{x}_\ell) - q_e(\bar{x}_{\ell+1})| \leqslant \gamma\|x - \widetilde{x}\|_1.$$

In the general case where $c \geqslant 0$, we first derive an expression for p'' similar to (9.8). Then the equations for p' in (9.3) yield p', the equations for q_P in (9.3) yield q_P, and finally the equations for q_Z in (9.3) yield q_Z. □

We are now ready to establish the existence of a solution with domain $[0, \infty)$.

Lemma 9.16. *The solution to the undirected dynamics in* (8.1) *has domain* $[0,\infty)$. *Moreover, for every* $t \geqslant 0$ *and* $e \in [m]$, *we have*

$$x_e(0)\cdot\exp\{-t\} \leqslant x_e(t) \leqslant D\|b/\gamma_A\|_1 + \max(0, x_e(0) - D\|b/\gamma_A\|_1)\cdot\exp\{-t\}.$$

Proof. Consider any $x_0 > 0$ and any $t_0 \geqslant 0$. We first show that there is a positive δ' (depending on x_0) such that a unique solution $x(t)$ with $x(t_0) = x_0$ exists for $t \in (t_0 - \delta', t_0 + \delta')$. By the Picard-Lindelöf Theorem [Tes12, Theorem 2.2], this holds true if the mapping $x \mapsto |q| - x$ is continuous and satisfies a Lipschitz condition in a neighborhood of x_0. Continuity clearly holds. Let $\varepsilon = \min_i(x_0)_i/2$ and let $U = \{\, x : \|x - x_0\|_\infty < \varepsilon \,\}$. Then for every $x, \widetilde{x} \in U$ and every e

$$\big||q_e(x)| - |q_e(\widetilde{x})|\big| \leqslant \gamma\|x - \widetilde{x}\|_1,$$

where γ is as in Lemma 9.15. Local existence implies the existence of a solution which cannot be extended. Since q is bounded (Lemma 9.12), x is bounded at all finite times, and hence the solution exists for all t. The lower bound $x_e(t) \geqslant x_e(0)\cdot\exp\{-t\} > 0$ for all e, holds by Lemma 9.2 with $A = 0$ and $\alpha = -1$. Since $|q_e| \leqslant D\|b/\gamma_A\|_1$, $\dot{x}_e = |q_e| - x_e \leqslant D\|b/\gamma_A\|_1 - x_e$, we have $x_e(t) \leqslant D\|b/\gamma_A\|_1 + \max(0, x_e(0) - D\|b/\gamma_A\|_1)\cdot\exp\{-t\}$ by Lemma 9.2 with $B = D\|b/\gamma_A\|_1$ and $\beta = -1$. □

9.2.2 LP Duality

The energy $E(x)$ is no longer a Lyapunov function, e.g., if $x(0) \approx \mathbf{0}$, $x(t)$ and hence $E(x(t))$ will grow initially. We will show that energy suitably scaled is a Lyapunov function. What is the appropriate scaling factor? In the case of the undirected shortest path problem, [BMV12] used the minimum capacity of any source-sink cut as a scaling factor. The proper generalization to our situation is to consider the linear program $\max\{\alpha : Af = \alpha b, |f| \leqslant x\}$, where x is a fixed positive vector. Linear programming duality yields the corresponding minimization problem which generalizes the minimum cut problem to our situation.

Lemma 9.17. *Let* $x \in \mathbb{R}^m_{>0}$ *and* $b \neq 0$. *The linear programs*

$$\max\{\alpha : Af = \alpha b, |f| \leqslant x\} \qquad \text{and} \qquad \min\{|y^{\mathrm{T}}A|x : b^{\mathrm{T}}y = -1\} \tag{9.10}$$

are feasible and have the same objective value. Moreover, there is a finite set $Y_A = \{\, d^1, \ldots, d^K \,\}$ *of vectors* $d^i \in \mathbb{R}^m_{\geqslant 0}$ *that are independent of* x *such that the minimum above is equal to* $C_\star = \min_{d\in Y_A} d^{\mathrm{T}}x$. *There is a feasible* f *with* $|f| \leqslant x/C_\star$.[2]

[2] In the undirected shortest path problem, the d's are the incidence vectors of the undirected source-sink cuts. Let S be any set of vertices containing s_0 but not s_1, and let $\mathbf{1}^S$ be its associated indicator vector. The cut corresponding to S contains the edges having exactly one endpoint in S. Its indicator vector is $d^S = |A^{\mathrm{T}}\mathbf{1}^S|$. Then $d^S_e = 1$ iff $|S \cap \{\, u, v \,\}| = 1$, where $e = (u,v)$ or $e = (v,u)$, and $d^S_e = 0$ otherwise. For a vector $x \geqslant 0$, $(d^S)^{\mathrm{T}}x$ is the capacity of the source-sink cut $(S, V\backslash S)$. In this setting, $C_\star$ is the value of a minimum cut.

Proof. The pair $(\alpha, f) = (0, 0)$ is a feasible solution for the maximization problem. Since $b \neq 0$, there exists y with $b^{\mathrm{T}} y = -1$ and thus both problems are feasible. The dual of $\max\{\alpha : Af - \alpha b = 0, f \leqslant x, -f \leqslant x\}$ has unconstrained variables $y \in \mathbb{R}^n$ and non-negative variables $z^+, z^- \in \mathbb{R}^m$ and reads

$$\min\{x^{\mathrm{T}}(z^+ + z^-) : -b^{\mathrm{T}} y = 1, A^{\mathrm{T}} y + z^+ - z^- = 0,\ z^+, z^- \geqslant 0\}. \tag{9.11}$$

From $z^- = A^{\mathrm{T}} y + z^+$, $z^+ \geqslant 0$, $z^- \geqslant 0$ and $x > 0$, we conclude $\min(z^+, z^-) = 0$ in an optimal solution. Thus $z^- = \max(0, A^{\mathrm{T}} y)$ and $z^+ = \max(0, -A^{\mathrm{T}} y)$ and hence $z^+ + z^- = |A^{\mathrm{T}} y|$ in an optimal dual solution. Therefore, (9.11) and the right LP in (9.10) have the same objective value.

We next show that the dual attains its minimum at a vertex of the feasible set. For this it suffices to show that its feasible set contains no line. Assume it does. Then there are vectors $d = (y_1, z_1^+, z_1^-)$, d non-zero, and $p = (y_0, z_0^+, z_0^-)$ such that $(y, z^+, z^-) = p + \lambda d = (y_0 + \lambda y_1, z_0^+ + \lambda z_1^+, z_0^- + \lambda z_1^-)$ is feasible for all $\lambda \in \mathbb{R}$. Thus $z_1^+ = z_1^- = 0$. Note that if either z_1^+ or z_1^- were non-zero then either $z_0^+ + \lambda z_1^+$ or $z_0^- + \lambda z_1^-$ would have a negative component for some λ. Then $A^{\mathrm{T}} y + z^+ + z^- = 0$ implies $A^{\mathrm{T}} y_1 = 0$. Since A has full row rank, $y_1 = 0$. Thus the dual contains no line and the minimum is attained at a vertex of its feasible region. The feasible region of the dual does not depend on x.

Let (y^1, z_1^+, z_1^-) to (y^K, z_K^+, z_K^-) be the vertices of (9.11), and let $Y_A = \{\, |A^{\mathrm{T}} y^1|, \ldots, |A^{\mathrm{T}} y^K| \,\}$. Then

$$\begin{aligned}\min_{d \in Y_A} d^{\mathrm{T}} x &= \min\{x^{\mathrm{T}}(z^+ + z^-) : -b^{\mathrm{T}} y = 1, A^{\mathrm{T}} y + z^+ - z^- = 0,\ z^+, z^- \geqslant 0\} \\ &= \min\{|y^{\mathrm{T}} A| x : b^{\mathrm{T}} y = -1\}.\end{aligned}$$

We finally show that there is a feasible f with $|f| \leqslant x/C_\star$. Let $x' \overset{\text{def}}{=} x/C_\star$. Then $x' > 0$ and $\min_{d \in Y_A} d^{\mathrm{T}} x' = \min_{d \in Y_A} d^{\mathrm{T}} x / C_\star = C_\star / C_\star = 1$ and thus the right LP with $x = x'$ (9.10) has objective value 1. Hence, the left LP has objective value 1 and there is a feasible f with $|f| \leqslant x'$. □

9.2.3 Convergence to Dominance

In the network setting, an important role is played by the set of edge capacity vectors that support a feasible flow. In the LP setting, we generalize this notion to the set of *dominating states*, which is defined as

$$X_{\text{dom}} \overset{\text{def}}{=} \{x \in \mathbb{R}^m : \exists \text{ feasible } f : |f| \leqslant x\}.$$

An alternative characterization, using the set Y_A from Lemma 9.17, is

$$\mathcal{X}_1 \overset{\text{def}}{=} \{x \in \mathbb{R}^m_{\geqslant 0} : d^{\mathrm{T}} x \geqslant 1 \text{ for all } d \in Y_A\}.$$

We now prove that $X_{\text{dom}} = \mathcal{X}_1$ and that the set $\mathcal{X}_1$ is an attractor in the following sense.

Lemma 9.18. *The following statements hold:*

(i) $X_{\text{dom}} = \mathcal{X}_1$. *Moreover,* $\lim_{t\to\infty} \operatorname{dist}(x(t), \mathcal{X}_1) = 0$, *where* $\operatorname{dist}(x, \mathcal{X}_1)$ *is the Euclidean distance between* x *and* $\mathcal{X}_1$.

(ii) If $x(t_0) \in \mathcal{X}_1$, *then* $x(t) \in \mathcal{X}_1$ *for all* $t \geqslant t_0$. *For all sufficiently large* t, $x(t) \in \mathcal{X}_{1/2} \overset{\text{def}}{=} \{x \in \mathbb{R}^n_{\geqslant 0} : d^{\mathrm{T}} x \geqslant 1/2$ *for all* $d \in Y_A\}$, *and if* $x \in \mathcal{X}_{1/2}$ *then there is a feasible* f *with* $|f| \leqslant 2x$.

Proof. (i) If $x \in \mathcal{X}_1$, then $d^{\mathrm{T}} x \geqslant 1$ for all $d \in Y_A$ and hence Lemma 9.17 implies the existence of a feasible solution f with $|f| \leqslant x$. Conversely, if $x \in X_{\text{dom}}$, then there is a feasible f with $|f| \leqslant x$. Thus $d^{\mathrm{T}} x \geqslant 1$ for all $d \in Y_A$ and hence $x \in \mathcal{X}_1$. By the proof of Lemma 9.17, for any $d \in Y_A$, there is a y such that $d = |A^{\mathrm{T}} y|$ and $b^{\mathrm{T}} y = -1$. Let $Y(t) = d^{\mathrm{T}} x$. Then

$$\dot{Y} = |y^{\mathrm{T}} A| \dot{x} = |y^{\mathrm{T}} A|(|q| - x) \geqslant |y^{\mathrm{T}} A q| - Y = |y^{\mathrm{T}} b| - Y = 1 - Y.$$

Thus for any t_0 and $t \geqslant t_0$, $Y(t) \geqslant 1 + (Y(t_0) - 1)e^{-(t - t_0)}$ by Lemma 9.2 applied with $A = 1$ and $\alpha = -1$. In particular, $\liminf_{t\to\infty} Y(t) \geqslant 1$. Thus $\liminf_{t\to\infty} \min_{d \in Y_A} d^{\mathrm{T}} x \geqslant 1$ and hence $\lim_{t\to\infty} \operatorname{dist}(x(t), \mathcal{X}_1) = 0$.

(ii) Moreover, if $Y(t_0) \geqslant 1$, then $Y(t) \geqslant 1$ for all $t \geqslant t_0$. Hence $x(t_0) \in \mathcal{X}_1$ implies $x(t) \in \mathcal{X}_1$ for all $t \geqslant t_0$. Since $x(t)$ converges to $\mathcal{X}_1$, $x(t) \in \mathcal{X}_{1/2}$ for all sufficiently large t. If $x \in \mathcal{X}_{1/2}$ there is f such that $Af = \frac{1}{2} b$ and $|f| \leqslant x$. Thus $2f$ is feasible and $|2f| \leqslant 2x$. □

The next lemma summarizes simple bounds on the values of resistors r, potentials p and states x that hold for sufficiently large t. Recall that $P = \{\, e \in [m] : c_e > 0 \,\}$ and $Z = \{\, e \in [m] : c_e = 0 \,\}$, see (9.1).

Lemma 9.19. *The following statements hold:*

(i) For sufficiently large t, it holds that $r_e \geqslant c_e/(2D\|b/\gamma_A\|_1)$, $b^{\mathrm{T}}p \leqslant 8D\|b/\gamma_A\|_1\|c\|_1$ and $|A_e^{\mathrm{T}}p| \leqslant 8D^2\|b\|_1\|c\|_1$ for all e.

(ii) For all e, it holds that $\dot{x}_e/x_e \geqslant -1$ and for all $e \in P$, it holds that $\dot{x}_e/x_e \leqslant 8D^2\|b\|_1\|c\|_1/c_{\min}$.

(iii) There is a positive constant C such that for all $t \geqslant t_0$, there is a feasible f (depending on t) such that $x_e(t) \geqslant C$ for all indices e in the support of f.

Proof. (i) By Lemma 9.16, $x_e(t) \leqslant 2D\|b/\gamma_A\|_1$ for all sufficiently large t. It follows that $r_e = c_e/x_e \geqslant c_e/(2D\|b/\gamma_A\|_1)$. Due to Lemma 9.18, for large enough t, there is a feasible flow with $|f| \leqslant 2x$. Together with $x_e(t) \leqslant 2D\|b/\gamma_A\|_1$, it follows that

$$b^{\mathrm{T}}p = E(q) \leqslant E(2x) = 4c^{\mathrm{T}}x \leqslant 8D\|b/\gamma_A\|_1\|c\|_1.$$

Now, orient A according to q and consider any index e'. Recall that for all indices e, we have $A_e^{\mathrm{T}}p = 0$ if $e \in Z$, and $q_e = (x_e/c_e) \cdot A_e^{\mathrm{T}}p$ if $e \in P$. Thus $A_e^{\mathrm{T}}p \geqslant 0$ for all e. If $e' \in Z$ or $e' \in P$ and $q_{e'} = 0$, the claim is obvious. So assume $e' \in P$ and $q_{e'} > 0$. Since q is a convex combination of q-sign-compatible basic feasible solutions, there is a basic feasible solution f with $f \geqslant 0$ and $f_{e'} > 0$. By Lemma 9.10, $f_{e'} \geqslant 1/(D\gamma_A)$. Therefore

$$f_{e'}A_{e'}^{\mathrm{T}}p \leqslant \sum_e f_e A_e^{\mathrm{T}}p = b^{\mathrm{T}}p \leqslant 8D\|b/\gamma_A\|_1\|c\|_1$$

for all sufficiently large t. The inequality follows from $f_e \geqslant 0$ and $A_e^{\mathrm{T}}p \geqslant 0$ for all e. Thus $A_{e'}^{\mathrm{T}}p \leqslant 8D^2\|b\|_1\|c\|_1$ for all sufficiently large t.

(ii) We have $\dot{x}_e/x_e = (|q_e| - x_e)/x_e \geqslant -1$ for all e. For e with $c_e > 0$

$$\frac{\dot{x}_e}{x_e} = \frac{|q_e| - x_e}{x_e} \leqslant \frac{|A_e^{\mathrm{T}}p|}{c_e} \leqslant 8D^2\|b\|_1\|c\|_1/c_{\min}.$$

(iii) Let t_0 be such that $d^{\mathrm{T}}x(t) \geqslant 1/2$ for all $d \in Y_A$ and $t \geqslant t_0$. Then for all $t \geqslant t_0$, there is f such that $Af = \frac{1}{2}b$ and $|f| \leqslant x(t)$; f may depend on t. By Lemma 9.1, we can write $2f$ as convex combination of f-sign-compatible basic feasible solutions (at most m of them) and a f-sign-compatible solution in the kernel of A. Dropping the solution in the kernel of A leaves us with a solution which is still dominated by x.

It holds that for every $e \in E$ with $f_e \neq 0$, there is a basic feasible solution g used in the convex decomposition such that $2|f_e| \geqslant |g_e| > 0$. By Lemma 9.10, every non-zero component of g is at least $1/(D\gamma_A)$. We conclude that $x_e \geqslant 1/(2D\gamma_A)$, for every e in the support of g. □

9.2.4 The Equilibrium Points

We next characterize the equilibrium points

$$F = \{\, x \in \mathbb{R}_{\geqslant 0} : |q| = x \,\}. \tag{9.12}$$

Let us first elaborate on the special case of the undirected shortest path problem. Here the equilibria are the flows of value one from source to sink in a network formed by undirected source-sink paths of the same length. This can be seen as follows. Consider any $x \geqslant 0$ and assume $\mathrm{supp}(x)$ is a network of undirected source-sink paths of the same length. Call this network $\mathcal{N}$. Assign to each node u, a potential p_u equal to the length of the shortest undirected path from the sink s_1 to u. These potentials are well-defined as all paths from s_1 to u in $\mathcal{N}$ must have the same length. For an edge $e = (u, v)$ in $\mathcal{N}$, we have $q_e = x_e/c_e(p_u - p_v) = x_e/c_e \cdot c_e = x_e$, i.e., $q = x$ is the electrical flow with respect to the resistances c_e/x_e. Conversely, if x is an equilibrium point and the network is oriented such that $q \geqslant 0$, we have $x_e = q_e = x_e/c_e(p_u - p_v)$ for all edges $e = (u, v) \in \mathrm{supp}(x)$. Thus $c_e = p_u - p_v$ and this is only possible if for every node u, all paths from u to the sink have the same length. Thus $\mathrm{supp}(x)$ must be a network of undirected source-sink paths of the same length. We next generalize this reasoning.

Theorem 9.20. *If $x = |q|$ is an equilibrium point and the columns of A are oriented such that $q \geqslant 0$, then all feasible solutions f with $\mathrm{supp}(f) \subseteq \mathrm{supp}(x)$ satisfy $c^{\mathrm{T}}f = c^{\mathrm{T}}x$. Conversely, if $x = |q|$ for a feasible q, A is oriented such that $q \geqslant 0$, and all feasible solutions f with $\mathrm{supp}(f) \subseteq \mathrm{supp}(x)$ satisfy $c^{\mathrm{T}}f = c^{\mathrm{T}}x$, then x is an equilibrium point.*

Proof. If x is an equilibrium point, $|q_e| = x_e$ for every e. By changing the signs of some columns of A, we may assume $q \geqslant 0$, i.e., $q = x$. Let p be the potential with respect to x. For every index $e \in P$ in the support of x, since $c_e > 0$ we have $q_e = \frac{x_e}{c_e} A_e^{\mathrm{T}} p$ and hence $c_e = A_e^{\mathrm{T}} p$. Further, for the indices $e \in Z$ in the support of x, we have $c_e = 0 = A_e^{\mathrm{T}} p$ due to the second block of equations on the right hand side in (9.2). Let f be any feasible solution whose support is contained in the support of x. Then the first part follows by

$$\sum_{e \in \operatorname{supp}(f)} c_e f_e = \sum_{e \in \operatorname{supp}(f)} f_e A_e^{\mathrm{T}} p = b^{\mathrm{T}} p = E(q) = E(x) = \operatorname{cost}(x).$$

For the second part, we misuse notation and use A to also denote the submatrix of the constraint matrix indexed by the columns in the support of x. We may assume that the rows of A are independent. Otherwise, we simply drop redundant constraints. We may assume $q \geqslant 0$; otherwise we simply change the sign of some columns of A. Then x is feasible. Let A_B be a square non-singular submatrix of A and let A_N consist of the remaining columns of A. The feasible solutions f with $\operatorname{supp}(f) \subseteq \operatorname{supp}(x)$ satisfy $A_B f_B + A_N f_N = b$ and hence $f_B = A_B^{-1}(b - A_N f_N)$. Then

$$c^{\mathrm{T}} f = c_B^{\mathrm{T}} f_B + c_N^{\mathrm{T}} f_N = c_B A_B^{-1} b + (c_N^{\mathrm{T}} - c_B^{\mathrm{T}} A_B^{-1} A_N) f_N.$$

Since, by assumption, $c^{\mathrm{T}} f$ is constant for all feasible solutions whose support is contained in the support of x, we must have $c_N = A_N^{\mathrm{T}} [A_B^{-1}]^{\mathrm{T}} c_B$. Let $p = [A_B^{-1}]^{\mathrm{T}} c_B$. Then it follows that $A^{\mathrm{T}} p = \left[\begin{smallmatrix} A_B^{\mathrm{T}} \\ A_N^{\mathrm{T}} \end{smallmatrix}\right] [A_B^{-1}]^{\mathrm{T}} c_B = \left[\begin{smallmatrix} c_B \\ c_N \end{smallmatrix}\right]$ and hence $Rx = A^{\mathrm{T}} p$. Thus the pair (x, p) satisfies the right hand side of (9.2). Since x is feasible, it also satisfies the left hand side of (9.2). Therefore, x is the minimum energy solution with respect to x. □

Corollary 9.21. *Let g be a basic feasible solution. Then $|g|$ is an equilibrium point.*

Proof. Let g be a basic feasible solution. Orient A such that $g \geqslant 0$. Since g is basic, there is a $B \subseteq [m]$ such that $g = (g_B, g_N) = (A_B^{-1} b, 0)$. Consider any feasible solution f with $\operatorname{supp}(f) \subseteq \operatorname{supp}(g)$. Then $f = (f_B, 0)$ and hence $b = Af = A_B f_B$. Therefore, $f_B = g_B$ and hence $c^{\mathrm{T}} f = c^{\mathrm{T}} g$. Thus $x = |g|$ is an equilibrium point. □

This characterization of equilibria has an interesting consequence.

Lemma 9.22. *The set $L \overset{\text{def}}{=} \{c^{\mathrm{T}} x : x \in F\}$ of costs of equilibria is finite.*

Proof. If x is an equilibrium, $x = |q|$, where q is the minimum energy solution with respect to x. Orient A such that $q \geqslant 0$. Then by Theorem 9.20, $c^{\mathrm{T}} f = c^{\mathrm{T}} x$ for all feasible solutions f with $\operatorname{supp}(f) \subseteq \operatorname{supp}(x)$. In particular, this holds true for all such basic feasible solutions f. Thus L is a subset of the set of costs of all basic feasible solutions, which is a finite set. □

We conclude this part by showing that the optimal solutions of the undirected linear program (8.2) are equilibria.

Theorem 9.23. *Let x be an optimal solution to (8.2). Then x is an equilibrium.*

Proof. By definition, there is a feasible f with $|f| = x$. Let us reorient the columns of A such that $f \geqslant 0$ and let us delete all columns e of A with $f_e = 0$. Consider any feasible g with $\operatorname{supp}(g) \subseteq \operatorname{supp}(x)$. We claim that $c^{\mathrm{T}} x = c^{\mathrm{T}} g$. Assume otherwise and consider the point $y = x + \lambda(g - x)$. If $|\lambda|$ is sufficiently small, $y \geqslant 0$. Furthermore, y is feasible and $c^{\mathrm{T}} y = c^{\mathrm{T}} x + \lambda(c^{\mathrm{T}} g - c^{\mathrm{T}} x)$. If $c^{\mathrm{T}} g \neq c^{\mathrm{T}} x$, x is not an optimal solution to (8.2). The claim now follows from Theorem 9.20. □

9.2.5 Convergence

In order to show convergence, we construct a Lyapunov function. The following functions play a crucial role in our analysis. Let $C_d = d^{\mathrm{T}} x$ for $d \in Y_A$, and recall that $C_\star = \min_{d \in Y_A} d^{\mathrm{T}} x$ denotes the optimum. Moreover, we define

$$h(t) \overset{\text{def}}{=} \sum_e r_e |q_e| \frac{x_e}{C_\star} - E\left(\frac{x}{C_\star}\right) \qquad \text{and} \qquad V_d \overset{\text{def}}{=} \frac{c^{\mathrm{T}} x}{C_d} \text{ for every } d \in Y_A.$$

Theorem 9.24. *(1) For every $d \in Y_A$, $\dot{C}_d \geqslant 1 - C_d$. Thus, if $C_d < 1$ then $\dot{C}_d > 0$.*

(2) If $x(t) \in \mathcal{X}_1$, then $\frac{d}{dt}\mathrm{cost}(x(t)) \leqslant 0$ with equality if and only if $x = |q|$.
(3) Let $d \in Y_A$ be such that $C_\star = d^{\mathrm{T}}x$ at time t. Then it holds that $\dot{V}_d \leqslant h(t)$.
(4) It holds that $h(t) \leqslant 0$ with equality if and only if $|q| = \frac{x}{C_\star}$.

Proof. (i) Recall that for $d \in Y_A$, there is a y such that $b^{\mathrm{T}}y = -1$ and $d = |A^{\mathrm{T}}y|$. Thus $\dot{C}_d = d^{\mathrm{T}}(|q| - x) \geqslant |y^{\mathrm{T}}Aq| - C_d = 1 - C_d$ and hence $\dot{C}_d > 0$, whenever $C_d < 1$.

(ii) Remember that $E(x) = \mathrm{cost}(x)$ and that $x(t) \in \mathcal{X}_1$ implies that there is a feasible f with $|f| = x$. Thus $E(q) \leqslant E(f) \leqslant E(x)$. Let R be the diagonal matrix of entries c_e/x_e. Then

$$\begin{aligned}\frac{d}{dt}\mathrm{cost}(x) &= c^{\mathrm{T}}(|q| - x) && \text{by (8.1)}\\ &= x^{\mathrm{T}}R^{1/2}R^{1/2}|q| - x^{\mathrm{T}}Rx && \text{since } c = Rx\\ &\leqslant (q^{\mathrm{T}}Rq)^{1/2}(x^{\mathrm{T}}Rx)^{1/2} - x^{\mathrm{T}}Rx && \text{by Cauchy-Schwarz}\\ &\leqslant 0 && \text{since } E(q) \leqslant E(x).\end{aligned}$$

If the derivative is zero, both inequalities above have to be equalities. This is only possible if the vectors $|q|$ and x are parallel and $E(q) = E(x)$. Let λ be such that $|q| = \lambda x$. Then $E(q) = \sum_e \frac{c_e}{x_e}q_e^2 = \lambda^2 \sum_e c_e x_e = \lambda^2 E(x)$. Since $E(x) > 0$, this implies $\lambda = 1$.

(iii) By definition of d, $C_\star = C_d$. By the first two items, we have $\dot{C}_\star = d^{\mathrm{T}}|q| - C_\star$ and $\frac{d}{dt}\mathrm{cost}(x) = c^{\mathrm{T}}|q| - \mathrm{cost}(x)$. Thus

$$\begin{aligned}\frac{d}{dt}\frac{\mathrm{cost}(x)}{C_\star} &= \frac{C_\star \frac{d}{dt}\mathrm{cost}(x) - \dot{C}_\star \mathrm{cost}(x)}{C_\star^2} = \frac{C_\star(c^{\mathrm{T}}|q| - \mathrm{cost}(x)) - (d^{\mathrm{T}}|q| - C_\star)\mathrm{cost}(x)}{C_\star^2}\\ &= \frac{C_\star \cdot c^{\mathrm{T}}|q| - d^{\mathrm{T}}|q| \cdot c^{\mathrm{T}}x}{C_\star^2} \leqslant \sum_e r_e|q_e|\frac{x_e}{C_\star} - \sum_e r_e\left(\frac{x_e}{C_\star}\right)^2 = h(t),\end{aligned}$$

where we used $r_e = c_e/x_e$ and hence $c^{\mathrm{T}}|q| = \sum_e r_e x_e|q_e|$, $c^{\mathrm{T}}x = E(x)$, and $d^{\mathrm{T}}|q| \geqslant |y^{\mathrm{T}}Aq| = 1$ since $d = |y^{\mathrm{T}}A|$ for some y with $b^{\mathrm{T}}y = -1$.

(iv) We have

$$\sum_e r_e \tfrac{x_e}{C_\star}|q_e| = \sum_e r_e^{1/2}\tfrac{x_e}{C_\star}r_e^{1/2}|q_e| \leqslant \Big(\sum_e r_e(\tfrac{x_e}{C_\star})^2\Big)^{1/2}\Big(\sum_e r_e q_e^2\Big)^{1/2} = \mathbb{E}(\tfrac{x}{C_\star})^{1/2}\mathbb{E}(q)^{1/2}$$

by Cauchy-Schwarz. Since $h(t) = \sum_e r_e|q_e|\frac{x_e}{C_\star} - E(\frac{x}{C_\star})$ by definition, it follows that

$$h(t) \leqslant \mathbb{E}(\tfrac{x}{C_\star})^{1/2} \cdot \mathbb{E}(q)^{1/2} - \mathbb{E}(\tfrac{x}{C_\star}) = E(\tfrac{x}{C_\star})^{1/2} \cdot \Big(\mathbb{E}(q)^{1/2} - \mathbb{E}(\tfrac{x}{C_\star})^{1/2}\Big) \leqslant 0$$

since $x/C_\star$ dominates a feasible solution and hence $\mathbb{E}(q) \leqslant \mathbb{E}(x/C_\star)$. If $h(t) = 0$, we must have equality in the application of Cauchy-Schwarz, i.e., the vectors $x/C_\star$ and $|q|$ must be parallel, and we must have $E(q) = E(x/C_\star)$ as in the proof of part ii. □

We show now convergence against the set of equilibrium points. We need the following technical Lemma from [BMV12].

Lemma 9.25 (Lemma 9 in [BMV12]). *Let $f(t) = \max_{d \in Y_A} f_d(t)$, where each f_d is continuous and differentiable. If $\dot{f}(t)$ exists, then there is a $d \in Y_A$ such that $f(t) = f_d(t)$ and $\dot{f}(t) = \dot{f}_d(t)$.*

Theorem 9.26. *All trajectories converge to the set F of equilibrium points.*

Proof. We distinguish cases according to whether the trajectory ever enters $\mathcal{X}_1$ or not. If the trajectory enters $\mathcal{X}_1$, say $x(t_0) \in \mathcal{X}_1$, then $\frac{d}{dt}\mathrm{cost}(x) \leqslant 0$ for all $t \geqslant t_0$ with equality only if $x = |q|$. Thus the trajectory converges to the set of fix points. If the trajectory never enters $\mathcal{X}_1$, consider $V = \max_{d \in Y_A}(V_d + 1 - C_d)$. We show that $\dot{V}$ exists for almost all t. Moreover, if $\dot{V}(t)$ exists, then $\dot{V}(t) \leqslant 0$ with equality if and only if $|q_e| = x_e$ for all e. It holds that V is Lipschitz continuous as the maximum of a finite number of continuously differentiable functions. Since V is Lipschitz continuous, the set of t's where $\dot{V}(t)$ does not exist has zero Lebesgue measure (see for example [CLSW98, Ch. 3]). If $\dot{V}(t)$ exists, we have $\dot{V}(t) = \dot{V}_d(t) - \dot{C}_d(t)$ for some $d \in Y_A$ according to Lemma 9.25. Then, it holds that $\dot{V}(t) \leqslant h(t) - (1 - C_d) \leqslant 0$. Thus $x(t)$ converges to the set

$$\Big\{\, x \in \mathbb{R}_{\geqslant 0} : \dot{V} = 0 \,\Big\} = \{\, x \in \mathbb{R}_{\geqslant 0} : |q| = x/C \text{ and } C = 1 \,\} = \{\, x \in \mathbb{R}_{\geqslant 0} : |q| = x \,\}.$$ □

At this point, we know that all trajectories $x(t)$ converge to F. Our next goal is to show that $c^{\mathrm{T}}x(t)$ converges to the cost of an optimum solution of (8.2) and that $|q| - x$ converges to zero. We are only able to show the latter for all indices $e \in P$, i.e. with $c_e > 0$.

9.2.6 Details of the Convergence Process

In the argument to follow, we will encounter the following situation several times. We have a non-negative function $f(t) \geqslant 0$ and we know that $\int_0^\infty f(t)dt$ is finite. We want to conclude that $f(t)$ converges to zero for $t \to \infty$. This holds true if f is Lipschitz continuous. Note that the proof of the following lemma is very similar to the proof in [BMV12, Lemma 11]. However, in our case we apply the local Lipschitz condition that we showed in Lemma 9.15.

Lemma 9.27. *Let $f(t) \geqslant 0$ for all t. If $\int_0^\infty f(t)d(t)$ is finite and $f(t)$ is locally Lipschitz continuous, i.e., for every $\varepsilon > 0$, there is a $\delta > 0$ such that $|f(t') - f(t)| \leqslant \varepsilon$ for all $t' \in [t, t+\delta]$, then $f(t)$ converges to zero as t goes to infinity. The functions $t \mapsto x^{\mathrm{T}}R|q| - x^{\mathrm{T}}Rx = c^{\mathrm{T}}|q| - c^{\mathrm{T}}x$ and $t \mapsto h(t)$ are Lipschitz continuous.*

Proof. If $f(t)$ does not converge to zero, there is $\varepsilon > 0$ and an infinite unbounded sequence $t_1, t_2, \ldots$ such that $f(t_i) \geqslant \varepsilon$ for all i. Since f is Lipschitz continuous there is $\delta > 0$ such that $f(t_i') \geqslant \varepsilon/2$ for $t_i' \in [t_i, t_i+\delta]$ and all i. Hence, the integral $\int_0^\infty f(t)dt$ is unbounded.

Since $\dot{x}_e$ is continuous and bounded (by Lemma 9.16), x_e is Lipschitz continuous. Thus, it is enough to show that q_e is Lipschitz continuous for all e. Since q_Z (recall that $Z = \{\, e : c_e = 0 \,\}$ and $P = [m] \setminus Z$) is an affine function of q_P, it suffices to establish the claim for $e \in P$. So let $e \in P$ be such that $c_e > 0$. First, we claim that $x_e(t+\varepsilon) \leqslant (1 + 2K\varepsilon)x_e$ for all $\varepsilon \leqslant K/4$, where $K = 8D^2\|b\|_1\|c\|_1/c_{\min}$. Assume that this is not the case. Let

$$\varepsilon = \inf\{\delta \leqslant 1/4K : x_e(t+\delta) > (1+2K\delta)x_e(t)\},$$

then $\varepsilon > 0$ (since $\dot{x}_e(t) \leqslant Kx_e(t)$ by Lemma 9.19) and, by continuity, $x_e(t+\varepsilon) \geqslant (1+2K\varepsilon)x_e(t)$. There must be $t' \in [t, t+\varepsilon]$ such that $\dot{x}_e(t') = 2Kx_e(t)$. On the other hand,

$$\dot{x}_e(t') \leqslant Kx_e(t') \leqslant K(1+2K\varepsilon)x_e(t) \leqslant K(1+2K/4K)x_e(t) < 2Kx_e(t),$$

which is a contradiction. Thus, $x_e(t+\varepsilon) \leqslant (1+2K\varepsilon)x_e$ for all $\varepsilon \leqslant 1/4K$. Similarly, $x_e(t+\varepsilon) \geqslant (1-2K\varepsilon)x_e$. Now, let $\alpha = (1-2K\varepsilon)x_e$ and $\beta = (1+2K\varepsilon)x_e$. Then

$$||q_e(t+\delta)| - |q_e(t)|| \leqslant M\|x(t+\delta) - x(t)\|_1 \leqslant Mm(4K\varepsilon)x_e \leqslant 8\varepsilon MmKD\|b/\gamma_A\|_1,$$

since $x_e \leqslant 2D\|b/\gamma_A\|_1$ for sufficiently large t and where M is as in Lemma 9.15. Since $C_\star$ is at least $1/2$ for all sufficiently large t, the division by $C_\star$ and $C_\star^2$ in the definition of $h(t)$ does not affect the claim. □

Lemma 9.28. *For all $e \in [m]$ of positive cost, it holds that $|x_e - |q_e|| \to 0$ as t goes to infinity.*

Proof. For a trajectory ultimately running in $\mathcal{X}_1$, we showed $\frac{d}{dt}\mathrm{cost}(x) \leqslant x^{\mathrm{T}}R|q| - x^{\mathrm{T}}Rx \leqslant 0$ with equality if and only if $x = |q|$. Also, $E(q) \leqslant E(x)$, since x dominates a feasible solution. Furthermore, $x^{\mathrm{T}}R|q| - x^{\mathrm{T}}Rx$ goes to zero using Lemma 9.27. Thus

$$\sum_e r_e(x_e - |q_e|)^2 = \sum_e r_e x_e^2 + \sum_e r_e q_e^2 - 2\sum_e r_e x_e|q_e| \leqslant 2\Big(\sum_e r_e x_e^2 - \sum_e r_e x_e|q_e|\Big)$$

goes to zero. Next observe that there is a constant C such that $x_e(t) \leqslant C$ for all e and t as a result of Lemma 9.16. Also $c_{\min} > 0$ and hence $r_e \geqslant c_{\min}/C$. Thus $\sum_e r_e(x_e - |q_e|)^2 \leqslant \frac{C}{c_{\min}} \cdot \sum_e (x_e - |q_e|)^2$ and hence $|x_e - |q_e|| \to 0$ for every e with positive cost. For trajectories outside $\mathcal{X}_1$, we argue about $||q_e| - \frac{x}{C_\star}|$ and use $C_\star \to 1$, namely

$$\sum_e r_e(\tfrac{x_e}{C_\star} - |q_e|)^2 \leqslant 2\Big(\sum_e r_e(\tfrac{x_e}{C_\star})^2 - \sum_e r_e\tfrac{x_e}{C_\star}|q_e|\Big) \to 0. \qquad □$$

Note that the above does not say anything about the indices $e \in Z$ (with $c_e = 0$). Recall that $A_Pq_P + A_Zq_Z = b$ and that the columns of A_Z are independent. Thus, q_Z is uniquely determined by q_P. For the undirected shortest path problem, the potential difference $p^{\mathrm{T}}b$ between source and sink converges to the length of a shortest source-sink path. If an edge with positive cost is used by some shortest undirected path, then no shortest undirected path uses it with the opposite direction. We prove the natural generalizations.

Let $\mathcal{S}_{\text{OPT}}$ be the set of optimal solutions to (8.2) and let $E_{\text{OPT}} = \cup_{x \in \mathcal{S}_{\text{OPT}}} \operatorname{supp}(x)$ be the set of columns used in some optimal solution. The columns of positive cost in E_{OPT} can be consistently oriented as the following Lemma shows.

Lemma 9.29. *Let x_1^* and x_2^* be optimal solutions to (8.2) and let f and g be feasible solutions with $|f| = x_1^*$ and $|g| = x_2^*$. Then there is no e such that $f_e g_e < 0$ and $c_e > 0$.*

Proof. Assume otherwise. Then $|g_e - f_e| = |g_e| + |f_e| > 0$. Consider $h = (g_e f - f_e g)/(g_e - f_e)$. Then $Ah = (g_e Af - f_e Ag)/(g_e - f_e) = b$ and h is feasible. Also, $h_e = \frac{g_e f_e - f_e g_e}{g_e - f_e} = 0$ and for every index e', it holds that $|h_{e'}| = \frac{|g_e f_{e'} - f_e g_{e'}|}{|g_e - f_e)|} \leqslant \frac{|g_e||f_{e'}| + |f_e||g_{e'}|}{|g_e| + |f_e|}$ and hence

$$\operatorname{cost}(h) < \operatorname{cost}(f) + \operatorname{cost}(g) = \frac{|g_e|}{|g_e| + |f_e|} \operatorname{cost}(x_1^*) + \frac{|f_e|}{|g_e| + |f_e|} \operatorname{cost}(x_2^*) = \operatorname{cost}(x_1^*),$$

a contradiction to the optimality of x_1^* and x_2^*. □

By the preceding lemma, we can orient A such that $f_e \geqslant 0$ whenever $|f|$ is an optimal solution to (8.2) and $c_e > 0$. We then call A *positively oriented.*

Lemma 9.30. *It holds that $p^{\mathrm{T}} b$ converges to the cost of an optimum solution of (8.2). If A is positively oriented, then $\liminf_{t \to \infty} A_e^{\mathrm{T}} p \geqslant 0$ for all e.*

Proof. Let x^* be an optimal solution of (8.2). We first show convergence to a point in L and then convergence to $c^{\mathrm{T}} x^*$. Let $\varepsilon > 0$ be arbitrary. Consider any time $t \geqslant t_0$, where t_0 and C as in Lemma 9.19 and moreover $||q_e| - x_e| \leqslant \frac{C\varepsilon}{c_{\max}}$ for every $e \in P$. Then $x_e \geqslant C$ for all indices e in the support of some basic feasible solution f. For every $e \in P$, we have $q_e = \frac{x_e}{c_e} A_e^{\mathrm{T}} p$. We also assume $q \geqslant 0$ by possibly reorienting columns of A. Hence

$$|c_e - A_e^{\mathrm{T}} p| = \left|1 - \frac{q_e}{x_e}\right| \cdot |c_e = \left|\frac{x_e - q_e}{x_e}\right| \cdot c_e \leqslant \frac{c_{\max}}{C} |q_e - x_e| \leqslant \varepsilon.$$

For indices $e \in Z$, we have $A_e^{\mathrm{T}} p = 0 = c_e$. Since $\|f\|_\infty \leqslant D \|b/\gamma_A\|_1$ (Lemma 9.10), we conclude

$$c^{\mathrm{T}} f - p^{\mathrm{T}} b = \sum_{e \in \operatorname{supp}(f)} (c_e - p^{\mathrm{T}} A_e) f_e \leqslant \varepsilon \sum_e |f_e| \leqslant \varepsilon \cdot m D \|b/\gamma_A\|_1.$$

Since the set L is finite, we can let $\varepsilon > 0$ be smaller than half the minimal distance between elements in L. By the preceding paragraph, there is for all sufficiently large t, a basic feasible solution f such that $|c^{\mathrm{T}} f - b^{\mathrm{T}} p| \leqslant \varepsilon$. Since $b^{\mathrm{T}} p$ is a continuous function of time, $c^{\mathrm{T}} f$ must become constant. We have now shown that $b^{\mathrm{T}} p$ converges to an element in L. We will next show that $b^{\mathrm{T}} p$ converges to the optimum cost. Let x^* be an optimum solution to (8.2) and let $W = \sum_e x_e^* c_e \ln x_e$. Since $x(t)$ is bounded, W is bounded. We assume that A is positively oriented, thus there is a feasible f^* with $|f^*| = x^*$ and $f_e^* \geqslant 0$ whenever $c_e > 0$. By reorienting zero cost columns, we may assume $f_e^* \geqslant 0$ for all e. Then $Ax^* = b$. We have

$$\begin{aligned}
\dot{W} &= \sum_e x_e^* c_e \frac{|q_e| - x_e}{x_e} \\
&= \sum_{e;\ c_e > 0} x_e^* \,|A_e^{\mathrm{T}} p| - \operatorname{cost}(x^*) && \text{since } q_e = \tfrac{x_e}{c_e} A_e^{\mathrm{T}} p \text{ whenever } c_e > 0 \\
&= \sum_e x_e^* \,|A_e^{\mathrm{T}} p| - \operatorname{cost}(x^*) && \text{since } A_e^{\mathrm{T}} p = 0 \text{ whenever } c_e = 0 \\
&= \sum_e x_e^* \Big(|A_e^{\mathrm{T}} p| - A_e^{\mathrm{T}} p \Big) + b^{\mathrm{T}} p - \operatorname{cost}(x^*)
\end{aligned}$$

and hence $b^{\mathrm{T}} p - \operatorname{cost}(x^*)$ must converge to zero; note that $b^{\mathrm{T}} p$ is Lipschitz continuous in t.

Similarly, $|A_e^{\mathrm{T}} p| - A_e^{\mathrm{T}} p$ must converge to zero whenever $x_e^* > 0$. This implies $\liminf A_e^{\mathrm{T}} p \geqslant 0$. Assume otherwise, i.e., for every $\varepsilon > 0$, we have $A_e^{\mathrm{T}} p < -\varepsilon$ for arbitrarily large t. Since p is Lipschitz continuous in t, there is a $\delta > 0$ such that $A_e^{\mathrm{T}} p < -\varepsilon/2$ for infinitely many disjoint intervals of length δ. In these intervals, $|A_e^{\mathrm{T}} p| - A_e^{\mathrm{T}} p \geqslant \varepsilon$ and hence W must grow beyond any bound, a contradiction. □

Corollary 9.31. *$E(x)$ and $\operatorname{cost}(x)$ converge to $c^{\mathrm{T}}x^*$, whereas x and $|q|$ converge to $\mathcal{S}_{\mathrm{OPT}}$. If the optimum solution is unique, x and $|q|$ converge to it. Moreover, if $e \notin E_{\mathrm{OPT}}$, x_e and $|q_e|$ converge to zero.*

Proof. The first part follows from $E(x) = \operatorname{cost}(x) = b^{\mathrm{T}}p$ and the preceding Lemma. Thus x and q converge to the set F of equilibrium points, see (9.12), that are optimum solutions to (8.2). Since every optimum solution is an equilibrium point by Theorem 9.23, x and q converge to $\mathcal{S}_{\mathrm{OPT}}$. For $e \notin E_{\mathrm{OPT}}$, $f_e = 0$ for every $f \in F \cap \mathcal{S}_{\mathrm{OPT}}$. Since x and $|q|$ converge to $F \cap \mathcal{S}_{\mathrm{OPT}}$, x_e and $|q_e|$ converge to zero for every $e \in E_{\mathrm{OPT}}$. □

Chapter 10

Physarum-Inspired Dynamics

In this chapter, we present in its full generality our main technical result on the Physarum-inspired dynamics (8.6).

10.1 Overview

Inspired by the max-flow min-cut theorem, we consider the following primal-dual pair of linear programs: the primal LP is given by $\max\{\, t : Af = t \cdot b;\ 0 \leqslant f \leqslant x \,\}$ in variables $f \in \mathbb{R}^m$ and $t \in \mathbb{R}$, and its dual LP reads $\min\left\{\, x^{\mathrm{T}} z : z \geqslant 0;\ z \geqslant A^{\mathrm{T}} y;\ b^{\mathrm{T}} y = 1 \,\right\}$ in variables $z \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. Since the dual feasible region does not contain a line and the minimum is bounded, the optimum is attained at a vertex, and in an optimum solution we have $z = \max\{0, A^{\mathrm{T}} y\}$. Let V be the set of vertices of the dual feasible region, and let $Y \stackrel{\text{def}}{=} \{\, y : (z, y) \in V \,\}$ be the set of their projections on y-space. Then, the dual optimum is given by $\min\{\, \max\{0, y^{\mathrm{T}} A\} \cdot x : y \in Y \,\}$. The set of *strongly dominating* capacity vectors x is defined as

$$X \stackrel{\text{def}}{=} \left\{\, x \in \mathbb{R}^m_{>0} : y^{\mathrm{T}} A x > 0 \text{ for all } y \in Y \,\right\}.^{1}$$

Note that X contains the set of all scaled feasible solutions $\{x = tf \,:\, Af = b,\ f \geqslant 0,\ t > 0\}$.

We next discuss the choice of step size. For $y \in Y$ and capacity vector x, let $\alpha(y, x) \stackrel{\text{def}}{=} y^{\mathrm{T}} A x$. Further, let $\alpha(x) \stackrel{\text{def}}{=} \min\{\, \alpha(y, x) : y \in Y \,\}$ and $\alpha_\ell \stackrel{\text{def}}{=} \alpha(x^{(\ell)})$. Then, for any $x \in X$ there is a feasible f such that $0 \leqslant f \leqslant x/\alpha(x)$, see Lemma 10.8. In particular, if x is feasible then $\alpha(x) = 1$, since $\alpha(y, x) = 1$ for all $y \in Y$. We partition the Physarum-inspired dynamics (8.6) into the following five regimes and define for each regime a fixed step size, see Section 10.3.

Corollary 10.1. *The Physarum-inspired dynamics (8.6) initialized with $x^{(0)} \in X$ and a step size h satisfies:*

(i) If $\alpha^{(0)} = 1$, we work with $h \leqslant h_0$ and have $\alpha_\ell = 1$ for all ℓ.

(ii) If $1/2 \leqslant \alpha^{(0)} < 1$, we work with $h \leqslant h_0/2$ and have $1 - \delta \leqslant \alpha_\ell < 1$ for $\ell \geqslant h^{-1} \log(1/2\delta)$ and $\delta > 0$.

(iii) If $1 < \alpha^{(0)} \leqslant 1/h_0$, we work with $h \leqslant h_0$ and have $1 < \alpha_\ell \leqslant 1 + \delta$ for $\ell \geqslant h^{-1} \cdot \log(1/\delta h_0)$ and $\delta > 0$.

(iv) If $0 < \alpha^{(0)} < 1/2$, we work with $h \leqslant \alpha^{(0)} h_0$ and have $1/2 \leqslant \alpha_\ell < 1$ for $\ell \geqslant 1/h$.

(v) If $1/h_0 < \alpha^{(0)}$, we work with $h \leqslant 1/4$ and have $1 < \alpha_\ell \leqslant 1/h_0$ for $\ell = \lfloor \log_{1/(1-h)} h_0(\alpha^{(0)} - 1)/(1 - h_0) \rfloor$.

In each regime, we have $1 - \alpha^{(\ell+1)} = (1 - h)(1 - \alpha_\ell)$.

We give now the full version of Theorem 8.4 which applies for any strongly dominating starting point.

Theorem 10.2. *Suppose $A \in \mathbb{Z}^{n \times m}$ has full row rank ($n \leqslant m$), $b \in \mathbb{Z}^n$, $c \in \mathbb{Z}^m_{>0}$ and $\varepsilon \in (0, 1)$. Given $x^{(0)} \in X$ and its corresponding α_0, the Physarum-inspired dynamics (8.6) initialized with $x^{(0)}$ runs in two regimes:*

(i) The first regime is executed when $\alpha^{(0)} \notin [1/2, 1/h_0]$ and it computes a point $x^{(t)} \in X$ such that $\alpha_t \in [1/2, 1/h_0]$. In particular, if $\alpha^{(0)} < 1/2$ then $h \leqslant (\Phi/\mathrm{opt}) \cdot (\alpha_0 h_0)^2$ and $t = 1/h$. Otherwise, if $\alpha^{(0)} > 1/h_0$ then $h \leqslant \Phi/\mathrm{opt}$ and $t = \lfloor \log_{1/(1-h)} [h_0(\alpha^{(0)} - 1)/(1 - h_0)] \rfloor$.

(ii) The second regime starts from a point $x^{(t)} \in X$ with $\alpha_t \in [1/2, 1/h_0]$, it has a step size $h \leqslant (\Phi/\mathrm{opt}) \cdot h_0^2/2$ and outputs for any $k \geqslant 4C_1/(h\Phi) \cdot \ln(C_2 \Psi^{(0)}/(\varepsilon \cdot \min\{1, x^{(0)}_{\min}\}))$ a vector $x^{(t+k)} \in X$ such that $\mathrm{dist}(x^{(t+k)}, X_\star) < \varepsilon/(D\gamma_A)$.

We stated the bounds on h in terms of the unknown quantities Φ and opt. However, $\Phi/\mathrm{opt} \geqslant 1/C_3$ by Lemma 9.10 and hence replacing Φ/opt by $1/C_3$ yields constructive bounds for h.

[1] In the shortest path problem (recall that $b = e_1 - e_n$) the set Y consists of all $y \in \{\, -1, +1 \,\}^n$ such that $y_1 = 1 = -y_n$, i.e., y encodes a cut with $S = \{\, i : y_i = -1 \,\}$ and $\overline{S} = \{\, i : y_i = +1 \,\}$. The condition $y^{\mathrm{T}} A x > 0$ translates into $\sum_{a \in E(S, \overline{S})} x_a - \sum_{a \in E(\overline{S}, S)} x_a > 0$, i.e., every source-sink cut must have positive directed capacity.

Organization: This chapter is devoted to proving Theorem 10.2. It is organized as follows: Section 10.2 establishes core efficiency bounds that extend [SV16c] and yield a *scale-invariant* determinant dependence of the step size and are applicable to strongly dominating points. Section 10.3 gives the definition of strongly dominating points and shows that the Physarum-inspired dynamics (8.6) initialized with such a point is well defined. Section 10.4 extends the analysis in [BBD$^+$13, SV16b, SV16c] to positive linear programs, by generalizing the concept of non-negative flows to non-negative *feasible kernel-free* vectors. Section 10.5 shows that $x^{(\ell)}$ converges to $X_\star$ for large enough ℓ. Section 10.6 concludes the proof of Theorem 10.2.

10.2 Useful Lemmas

Recall that $R^{(\ell)} = \mathrm{diag}(c)\cdot(X^{(\ell)})^{-1}$ is a positive diagonal matrix and $L^{(\ell)} \overset{\text{def}}{=} A(R^{(\ell)})^{-1}A^{\mathrm{T}}$ is invertible. Let $p^{(\ell)}$ be the unique solution of $L^{(\ell)}p^{(\ell)} = b$. We improve the dependence on D_S in [SV16c, Lemma 5.2] to D.

Lemma 10.3. *[SV16c, extension of Lemma 5.2] Suppose $x^{(\ell)} > 0$, $R^{(\ell)}$ is a positive diagonal matrix and $L = A(R^{(\ell)})^{-1}A^{\mathrm{T}}$. Then for every $e \in [m]$, it holds that $\|A^{\mathrm{T}}(L^{(\ell)})^{-1}A_e\|_\infty \leqslant D\cdot c_e/x_e^{(\ell)}$.*

Proof. The statement follows by combining the proof in [SV16c, Lemma 5.2] with Lemma 9.10. □

We show next that [SV16b, Corollary 5.3] holds for x-capacitated vectors, which extends the class of feasible starting points, and further yields a bound in terms of D.

Lemma 10.4. *[SV16b, extension of Corollary 5.3] Let $p^{(\ell)}$ be the unique solution of $L^{(\ell)}p^{(\ell)} = b$ and assume $x^{(\ell)}$ is a positive vector with corresponding positive scalar α_ℓ such that there is a vector f satisfying $Af = \alpha_\ell\cdot b$ and $0 \leqslant f \leqslant x^{(\ell)}$. Then $\|A^{\mathrm{T}}p^{(\ell)}\|_\infty \leqslant D\|c\|_1/\alpha_\ell$.*

Proof. By assumption, f satisfies $\alpha_\ell b = Af = \sum_e f_e A_e$ and $0 \leqslant f \leqslant x^{(\ell)}$. This yields

$$\begin{aligned}\alpha_\ell\|A^{\mathrm{T}}p^{(\ell)}\|_\infty &= \|A^{\mathrm{T}}(L^{(\ell)})^{-1}\cdot\alpha_\ell b\|_\infty = \|\sum_e f_e A^{\mathrm{T}}(L^{(\ell)})^{-1}A_e\|_\infty \\ &\leqslant \sum_e f_e\|A^{\mathrm{T}}(L^{(\ell)})^{-1}A_e\|_\infty \overset{\text{(Lem. 10.3)}}{\leqslant} D\sum_e f_e\frac{c_e}{x_e^{(\ell)}} \leqslant D\|c\|_1.\end{aligned}$$ □

We note that applying Lemma 10.3 and Lemma 10.4 into the analysis of [SV16c, Theorem 1.3] yields an improved result that depends on the scale-invariant determinant D. Moreover, we show in the next Section 10.3 that the Physarum-inspired dynamics (8.6) can be initialized with any strongly dominating point.

We establish now an upper bound on q that does not depend on x. We then use this upper bound on q to establish a uniform upper bound on x.

Lemma 10.5. *For any $x^{(\ell)} > 0$, $\|q^{(\ell)}\|_\infty \leqslant mD^2\|b/\gamma_A\|_1$.*

Proof. Let f be a basic feasible solution of $Af = b$. By definition, $q_e^{(\ell)} = (x_e^{(\ell)}/c_e)A_e^{\mathrm{T}}(L^{(\ell)})^{-1}b$ and thus

$$\left|q_e^{(\ell)}\right| = \left|\frac{x_e^{(\ell)}}{c_e}\sum_u A_e^{\mathrm{T}}(L^{(\ell)})^{-1}A_u f_u\right| \leqslant \frac{x_e^{(\ell)}}{c_e}\sum_u |f_u|\cdot\left|A_e^{\mathrm{T}}(L^{(\ell)})^{-1}A_u\right| \leqslant D\|f\|_1,$$

where the last inequality follows by

$$\left|A_e^{\mathrm{T}}(L^{(\ell)})^{-1}A_u\right| = \left|A_u^{\mathrm{T}}(L^{(\ell)})^{-1}A_e\right| \leqslant \|A^{\mathrm{T}}(L^{(\ell)})^{-1}A_e\|_\infty \overset{\text{(Lem. 10.3)}}{\leqslant} D\cdot c_e/x_e^{(\ell)}.$$

By Cramer's rule and Lemma 9.10, we have $|q_e^{(\ell)}| \leqslant D\|f\|_1 \leqslant mD^2\|b/\gamma_A\|_1$. □

Let $k, t \in \mathbb{N}$. We denote by

$$\overline{q}^{(t,k)} = \sum_{i=t}^{t+k-1}\frac{h\,(1-h)^{t+k-1-i}}{1-(1-h)^k}q^{(i)} \quad\text{and}\quad \overline{p}^{(t,k)} = \sum_{i=t}^{t+k-1}p^{(i)}. \tag{10.1}$$

Straightforward checking shows that $A\overline{q}^{(t,k)} = b$. Further, for $C \overset{\text{def}}{=} \operatorname{diag}(c)$, $t \geqslant 0$ and $k \geqslant 1$, we have

$$x^{(t)} \prod_{i=t}^{t+k-1} [1 + h(C^{-1}A^{\mathrm{T}}p^{(i)} - 1)] \;=\; x^{(t+k)} \;=\; (1-h)^k x^{(t)} + [1-(1-h)^k]\overline{q}^{(t,k)}.$$

We give next an upper bound on $x^{(k)}$ that is independent of k.

Lemma 10.6. *Let $\Psi^{(0)} = \max\{mD^2\|b/\gamma_A\|_1, \|x^{(0)}\|_\infty\}$. Then $\|x^{(k)}\|_\infty \leqslant \Psi^{(0)}$, $\forall k \in \mathbb{N}$.*

Proof. We prove the statement by induction. The base case $\|x^{(0)}\|_\infty \leqslant \Psi^{(0)}$ is clear. Suppose the statement holds for some $k > 0$. Then, triangle inequality and Lemma 10.5 yield

$$\|x^{(k+1)}\|_\infty \leqslant (1-h)\|x^{(k)}\|_\infty + h\|q^{(k)}\|_\infty \leqslant (1-h)\Psi^{(0)} + h\Psi^{(0)} \leqslant \Psi^{(0)}. \qquad \square$$

We show now convergence to feasibility.

Lemma 10.7. *Let $r^{(k)} = b - Ax^{(k)}$. Then $r^{(k+1)} = (1-h)r^{(k)}$ and hence $r^{(k)} = (1-h)^k(b - Ax^{(0)})$.*

Proof. By definition $x^{(k+1)} = (1-h)x^{(k)} + hq^{(k)}$, and thus the statement follows by

$$r^{(k+1)} = b - Ax^{(k+1)} = b - (1-h)Ax^{(k)} - hb = (1-h)r^{(k)}. \qquad \square$$

10.3 Strongly Dominating Capacity Vectors

For the shortest path problem, it is known that one can start from any capacity vector x for which the directed capacity of every source-sink cut is positive, where the directed capacity of a cut is the total capacity of the edges crossing the cut in source-sink direction minus the total capacity of the edges crossing the cut in the sink-source direction. We generalize this result. We start with the max-flow like LP

$$\max\{\, t : Af = t \cdot b;\; 0 \leqslant f \leqslant x \,\} \tag{10.2}$$

in variables $f \in \mathbb{R}^m$ and $t \in \mathbb{R}$ and its dual

$$\min\left\{\, x^{\mathrm{T}}z : z \geqslant 0;\; z \geqslant A^{\mathrm{T}}y;\; b^{\mathrm{T}}y = 1 \,\right\} \tag{10.3}$$

in variables $z \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. The feasible region of the dual contains no line. Assume otherwise; say it contains $(z, y) = (z^{(0)}, y^{(0)}) + \lambda(z^{(1)}, y^{(1)})$ for all $\lambda \in \mathbb{R}$. Then, $z \geqslant 0$ implies $z^{(1)} = 0$ and further $z \geqslant A^{\mathrm{T}}y$ implies $z^{(0)} \geqslant A^{\mathrm{T}}y^{(0)} + \lambda A^{\mathrm{T}}y^{(1)}$ and hence $A^{\mathrm{T}}y^{(1)} = 0$. Since A has full row rank, we have $y^{(1)} = 0$. The optimum of the dual is therefore attained at a vertex. In an optimum solution, we have $z = \max\{0, A^{\mathrm{T}}y\}$. Let V be the set of vertices of the feasible region of the dual (10.3), and let

$$Y \overset{\text{def}}{=} \{\, y : (z, y) \in V \,\}$$

be the set of their projections on y-space. Then, the optimum of the dual (10.3) is given by

$$\min_{y \in Y} \left\{\max\{0, y^{\mathrm{T}}A\} \cdot x\right\}. \tag{10.4}$$

The set of strongly dominating capacity vectors x is defined by

$$X \overset{\text{def}}{=} \left\{\, x \in \mathbb{R}^m_{>0} : y^{\mathrm{T}}Ax > 0 \text{ for all } y \in Y \,\right\}. \tag{10.5}$$

We next show that for all $x^{(0)} \in X$ and sufficiently small step size, the sequence $\{x^{(k)}\}_{k \in \mathbb{N}}$ stays in X. Moreover, $y^{\mathrm{T}}Ax^{(k)}$ converges to 1 for every $y \in Y$. We define by

$$\alpha(y, x) \overset{\text{def}}{=} y^{\mathrm{T}}Ax \quad \text{ and } \quad \alpha(x) \overset{\text{def}}{=} \min\{\, \alpha(y, x) : y \in Y \,\}.$$

Let $\alpha_\ell \overset{\text{def}}{=} \alpha(x^{(\ell)})$. Then, $x^{(\ell)} \in X$ iff $\alpha_\ell > 0$. We summarize the discussion in the following Lemma.

Lemma 10.8. *Suppose $x^{(\ell)} \in X$. Then, there is a vector f such that $Af = \alpha_\ell \cdot b$ and $0 \leqslant f \leqslant x^{(\ell)}$.*

Proof. By the strong duality theorem applied on (10.2) and (10.3), it holds by (10.4) that

$$t = \min_{y\in Y}\left\{\max\{0, y^{\mathrm{T}}A\}\cdot x^{(\ell)}\right\} \geqslant \min_{y\in Y} y^{\mathrm{T}}Ax^{(\ell)} = \alpha_\ell.$$

The statement follows by the definition of (10.2). □

We demonstrate now that α_ℓ converges to 1.

Lemma 10.9. *Assume $x^{(\ell)} \in X$. Then, for any $h^{(\ell)} \leqslant \min\{1/4, \alpha_\ell h_0\}$ we have $x^{(\ell+1)} \in X$ and*

$$1-\alpha_{\ell+1} = (1-h^{(\ell)})\cdot(1-\alpha_\ell).$$

Proof. By applying Lemma 10.4 and Lemma 10.8 with $x^{(\ell)} \in X$, we have $\|A^{\mathrm{T}}p^{(\ell)}\|_\infty \leqslant D\|c\|_1/\alpha_\ell$ and hence for every index e it holds $-h^{(\ell)}\cdot c_e^{-1}x_e^{(\ell)}A_e^{\mathrm{T}}p^{(\ell)} \geqslant -(h^{(\ell)}x_e^{(\ell)})/(2\alpha_\ell h_0) \geqslant -x_e^{(\ell)}/2$. Thus,

$$x_e^{(\ell+1)} = (1-h^{(\ell)})x_e^{(\ell)} + h^{(\ell)}\cdot[R_e^{(\ell)}]^{-1}A_e^{\mathrm{T}}p^{(\ell)} \geqslant \frac{3}{4}x_e^{(\ell)} - \frac{1}{2}x_e^{(\ell)} = \frac{1}{4}x_e^{(\ell)} > 0.$$

Let $y \in Y$ be arbitrary. Then $y^{\mathrm{T}}b = 1$ and hence $y^{\mathrm{T}}r^{(\ell)} = y^{\mathrm{T}}(b - Ax^{(\ell)}) = 1 - y^{\mathrm{T}}Ax^{(\ell)} = 1 - \alpha(y, x^{(\ell)})$. The second claim now follows from Lemma 10.7. □

We note that the convergence speed crucially depends on the initial point $x^{(0)} \in X$, and in particular to its corresponding value α_0. Further, this dependence naturally partitions the Physarum-inspired dynamics (8.6) into the five regimes given in Corollary 10.1.

10.4 $x^{(k)}$ is Close to a Non-Negative Kernel-Free Vector

In this section, we generalize [SV16b, Lemma 5.4] to positive linear programs. We achieve this in two steps. First, we generalize a result by Ito et al. [IJNT11, Lemma 2] to positive linear programs and then we substitute the notion of a non-negative cycle-free flow with a non-negative feasible kernel-free vector.

Throughout this and the consecutive section, we denote by $\rho_A \overset{\mathrm{def}}{=} \max\left\{D\gamma_A,\, nD^2\|A\|_\infty\right\}$.

Lemma 10.10. *Suppose a matrix $A \in \mathbb{Z}^{n\times m}$ has full row rank and vector $b \in \mathbb{Z}^n$. Let g be a feasible solution to $Ag = b$ and $S \subseteq [n]$ be a subset of row indices of A such that $\sum_{i\in S}|g_i| < 1/\rho_A$. Then, there is a feasible solution f such that $g_i \cdot f_i \geqslant 0$ for all $i \in [n]$, $f_i = 0$ for all $i \in S$ and $\|f - g\|_\infty < 1/(D\gamma_A)$.*

Proof. W.l.o.g. we can assume that $g \geqslant 0$ as we could change the signs of the columns of A accordingly. Let $\mathbf{1}_S$ be the indicator vector of S. We consider the linear program

$$\min\{\mathbf{1}_S^{\mathrm{T}}x \,:\, Ax = b,\ x \geqslant 0\}$$

and let *opt* be its optimum value. Notice that $0 \leqslant opt \leqslant \mathbf{1}_S^{\mathrm{T}}g < 1/\rho_A$. Since the feasible region does not contain a line and the minimum is bounded, the optimum is attained at a basic feasible solution, say f. Suppose that there is an index $i \in S$ with $f_i > 0$. By Lemma 9.10, we have $f_i \geqslant 1/(D\gamma_A)$. This is a contradiction to the optimality of f and hence $f_i = 0$ for all $i \in S$.

Among the feasible solutions f such that $f_ig_i \geqslant 0$ for all i and $f_i = 0$ for all $i \in S$, we choose the one that minimizes $\|f - g\|_\infty$. For simplicity, we also denote it by f. Note that f satisfies $\mathrm{supp}(f) \subseteq \overline{S}$, where $\overline{S} = [m]\backslash S$. Further, since $f_S = 0$ and

$$A_Sg_S + A_{\overline{S}}g_{\overline{S}} = Ag = b = Af = A_Sf_S + A_{\overline{S}}f_{\overline{S}} = A_{\overline{S}}f_{\overline{S}}$$

we have $A_{\overline{S}}(f_{\overline{S}} - g_{\overline{S}}) = A_Sg_S$. Let A_B be a linearly independent column subset of $A_{\overline{S}}$ with maximal cardinality, i.e. the column subset A_N, where $N = \overline{S}\setminus B$, is linearly dependent on A_B. Hence, there is an invertible square submatrix $A'_B \in \mathbb{Z}^{|B|\times|B|}$ of A_B and a vector $v = (v_B, 0_N)$ such that

$$\begin{pmatrix} A'_B \\ A''_B \end{pmatrix} v_B = A_Bv_B = A_Sg_S.$$

Let $r = (A_Sg_S)_B$. Since A'_B is invertible, there is a unique vector v_B such that $A'_Bv_B = r$. Observe that

$$|r_i| = \left|\sum_{j\in S}A_{i,j}g_j\right| \leqslant \|A\|_\infty\sum_{j\in S}|g_j| < \frac{\|A\|_\infty}{nD^2\|A\|_\infty} = \frac{1}{nD^2}.$$

By Cramer's rule $v_B(e)$ is quotient of two determinants. The denominator is $\det(A'_B)$ and hence at least one in absolute value. For the numerator, the e-th column is replaced by r. Expansion according to this column shows that the absolute value of the numerator is bounded by

$$\frac{D}{\gamma_A}\sum_{i\in B}|r_i| < \frac{D}{\gamma_A}\cdot\frac{|B|}{nD^2} \leqslant \frac{1}{D\gamma_A}.$$

Therefore, $\|f-g\|_\infty \leqslant 1/(D\gamma_A)$ and the statement follows. □

Lemma 10.11. *Let $q\in\mathbb{R}^m$, $p\in\mathbb{R}^n$ and $N=\{e\in[m]: q_e\leqslant 0 \text{ or } p^{\mathrm{T}}A_e\leqslant 0\}$, where $Aq=b$ and $p=L^{-1}b$. Suppose $\sum_{e\in N}|q_e|<1/\rho_A$. Then there is a non-negative feasible kernel-free vector f such that $\operatorname{supp}(f)\subseteq E\backslash N$ and $\|f-q\|_\infty<1/(D\gamma_A)$.*

Proof. We apply Lemma 10.10 to q with $S=N$. Then, there is a non-negative feasible vector f such that $\operatorname{supp}(f)\subseteq E\backslash N$ and $\|f-q\|_\infty<1/(D\gamma_A)$. By Lemma 9.1, f can be expressed as a sum of a convex combination of basic feasible solutions plus a vector w in the kernel of A. Moreover, all vectors in this representation are sign compatible with f, and in particular w is non-negative too.

Suppose for contradiction that $w\neq 0$. By definition, $0=p^{\mathrm{T}}Aw=\sum_{e\in[m]}p^{\mathrm{T}}A_ew_e$ and since $w\geqslant 0$ and $w\neq 0$, it follows that there is an index $e\in[m]$ satisfying $w_e>0$ and $p^{\mathrm{T}}A_e\leqslant 0$. Since f and w are sign compatible, $w_e>0$ implies $f_e>0$. On the other hand, as $p^{\mathrm{T}}A_e\leqslant 0$ we have $e\in N$ and thus $f_e=0$. This is a contradiction, hence $w=0$. □

Using Corollary 10.1, for any point $x^{(0)}\in X$ there is a point $x^{(t)}\in X$ such that $\alpha_t\in[1/2,1/h_0]$. Thus, we can assume that $\alpha_0\in[1/2,1/h_0]$ and work with $h\leqslant h_0/2$, where $h_0=c_{\min}/(2D\|c\|_1)$. We generalize next [SV16b, Lemma 5.4].

Lemma 10.12. *Suppose $x^{(t)}\in X$ such that $\alpha_t\in[1/2,1/h_0]$, $h\leqslant h_0/2$ and $\varepsilon\in(0,1)$. Then, for any $k\geqslant h^{-1}\ln(8m\rho_A\Psi^{(0)}/\varepsilon)$ there is a non-negative feasible kernel-free vector f such that $\|x^{(t+k)}-f\|_\infty<\varepsilon/(D\gamma_A)$.*

Proof. Let $\beta^{(k)}\overset{\text{def}}{=}1-(1-h)^k$. By (10.1), vector $\overline{q}^{(t,k)}$ satisfies $A\overline{q}^{(t,k)}=b$ and thus Lemma 10.6 yields

$$\|x^{(t+k)}-\beta^{(k)}\overline{q}^{(t,k)}\|_\infty=(1-h)^k\cdot\|x^{(t)}\|_\infty\leqslant\exp\{-hk\}\cdot\Psi^{(0)}\leqslant\varepsilon/(8m\rho_A). \tag{10.6}$$

Using Corollary 10.1, we have $x^{(t+k)}\in X$ such that $\alpha^{(t+k)}\in(1/2,1/h_0)$ for every $k\in\mathbb{N}_+$. Let $F_k=Q_k\cup P_k$, where $Q_k=\{e\in[m]:\overline{q}_e^{(t,k)}\leqslant 0\}$ and $P_k=\{e\in[m]:A_e^{\mathrm{T}}\overline{p}^{(t,k)}\leqslant 0\}$. Then, for every $e\in Q_k$ it holds

$$|\overline{q}_e^{(t,k)}|\leqslant[\beta^{(k)}]^{-1}\cdot|x_e^{(t+k)}-\beta^{(k)}\overline{q}_e^{(t,k)}|\leqslant\varepsilon/(7m\rho_A). \tag{10.7}$$

By Lemma 10.6, $\|x^{(\cdot)}\|\leqslant\Psi^{(0)}$. Moreover, by (10.1) for every $e\in P_k$ we have

$$\begin{aligned}
x_e^{(t+k)} &= x_e^{(t)}\prod_{i=t}^{k+t-1}\left[1+h\left(c_e^{-1}A_e^{\mathrm{T}}p^{(i)}-1\right)\right]\\
&\leqslant x_e^{(t)}\cdot\exp\left\{-hk+(h/c_e)\cdot A_e^{\mathrm{T}}\overline{p}^{(t,k)}\right\}\\
&\leqslant \exp\{-hk\}\cdot\Psi^{(0)}\\
&\leqslant \varepsilon/(8m\rho_A),
\end{aligned}$$

and by combining the triangle inequality with (10.6), it follows for every $e\in P_k$ that

$$\begin{aligned}
|\overline{q}_e^{(t,k)}| &\leqslant [\beta^{(k)}]^{-1}\cdot\left[|x_e^{(t+k)}-\beta^{(k)}\overline{q}_e^{(t,k)}|+|x_e^{(t+k)}|\right]\\
&\leqslant [\beta^{(k)}]^{-1}\cdot\varepsilon/(4m\rho_A)\\
&\leqslant \varepsilon/(3m\rho_A).
\end{aligned}\tag{10.8}$$

Therefore, (10.7) and (10.8) yields that

$$\sum_{e\in F_k}|\overline{q}_e^{(t,k)}|\leqslant m\cdot\varepsilon/(3m\rho_A)\leqslant\varepsilon/(3\rho_A). \tag{10.9}$$

By Lemma 10.11 applied with $\overline{q}_e^{(t,k)}$ and $N = F_k$, it follows by (10.9) that there is a non-negative feasible kernel-free vector f such that $\mathrm{supp}(f) \subseteq E \backslash N$ and

$$\|f - \overline{q}^{(t,k)}\|_\infty < \varepsilon/(3D\gamma_A).$$

By Lemma 10.5, we have $\|\overline{q}^{(t,k)}\|_\infty \leqslant mD^2\|b/\gamma_A\|_1$ and since $\Psi^{(0)} \geqslant mD^2\|b/\gamma_A\|_1$, it follows that

$$\begin{aligned}
\|x^{(t+k)} - f\|_\infty &= \|x^{(t+k)} - \beta^{(k)}\overline{q}^{(t+k)} + \beta^{(k)}\overline{q}^{(t+k)} - f\|_\infty \\
&\leqslant \|x^{(t+k)} - \beta^{(k)}\overline{q}^{(t+k)}\|_\infty + \|f - \overline{q}^{(t+k)}\|_\infty + (1-h)^k\,\|\overline{q}^{(t+k)}\|_\infty \\
&\leqslant \frac{\varepsilon}{8m\rho_A} + \frac{\varepsilon}{3D\gamma_A} + \frac{\varepsilon \cdot mD^2\|b/\gamma_A\|_1}{8m\rho_A \cdot \Psi^{(0)}} \leqslant \frac{\varepsilon}{D\gamma_A}.
\end{aligned}$$

□

10.5 $x^{(k)}$ is ε-Close to an Optimal Solution

Recall that $\mathcal{N}$ denotes the set of non-optimal basic feasible solutions of (8.5) and $\Phi = \min_{g\in\mathcal{N}} c^{\mathrm{T}}g - \mathrm{opt}$. For completeness, we prove next a well known inequality [PS82, Lemma 8.6] that lower bounds the value of Φ.

Lemma 10.13. *Suppose $A \in \mathbb{R}^{n\times m}$ has full row rank, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}^m$ are integral. Then, $\Phi \geqslant 1/(D\gamma_A)^2$.*

Proof. Let $g = (g_B, 0)$ be an arbitrary basic feasible solution with basis matrix A_B, where $g_B(e) \neq 0$ and $|\mathrm{supp}(g_B)| = n$. We write $M_{-i,-j}$ to denote the matrix M with deleted i-th row and j-th column. Let Q_e be the matrix formed by replacing the e-th column of A_B by the column vector b. Then, by Cramer's rule, we have

$$|g_B(e)| = \left|\frac{\det(Q_e)}{\det(A_B)}\right| = \frac{1}{\gamma_A}\left|\sum_{k=1}^{n}\frac{(-1)^{j+k}\cdot b_k\cdot\det\left(\gamma_A^{-1}[A_B]_{-k,-j}\right)}{\det\left(\gamma_A^{-1}A_B\right)}\right| \geqslant \frac{1}{D\gamma_A}.$$

Note that all components of vector g_B have denominator with equal value, i.e. $\det(A_B)$. Consider an arbitrary non-optimal basic feasible solution g and an optimal basic feasible solution $f^\star$. Then, $g_e = G_e/G$ and $f_e^\star = F_e/F$ are rationals such that $G_e, G, F_e, F \leqslant D\gamma_A$ for every e. Further, let $r_e = c_e\,(G_eF - F_eF) \in \mathbb{Z}$ for every $e \in [m]$, and observe that

$$c^{\mathrm{T}}\,(g - f^\star) = \sum_e c_e\,(g_e - f_e^\star) = \frac{1}{GF}\sum_e r_e \geqslant 1/(D\gamma_A)^2,$$

where the last inequality follows by $c^{\mathrm{T}}(g - f^\star) > 0$ implies $\sum_e r_e \geqslant 1$. □

Lemma 10.14. *Let f be a non-negative feasible kernel-free vector and $\varepsilon \in (0,1)$ a parameter. Suppose for every non-optimal basic feasible solution g, there exists an index $e \in [m]$ such that $g_e > 0$ and $f_e < \varepsilon/(2mD^3\gamma_A\|b\|_1)$. Then, $\|f - f^\star\|_\infty < \varepsilon/(D\gamma_A)$ for some optimal $f^\star$.*

Proof. Let $C = 2D^2\|b\|_1$. Since f is kernel-free, by Lemma 9.1 it can be expressed as a convex combination of sign-compatible basic feasible solutions $f = \sum_{i=1}^{\ell}\alpha_i f^{(i)} + \sum_{i=\ell+1}^{m}\alpha_i f^{(i)}$, where $f^{(1)}, \ldots, f^{(\ell)}$ denote the optimal solutions. By Lemma 9.10, $f_e^{(i)} > 0$ implies $f_e^{(i)} \geqslant 1/(D\gamma_A)$. By the hypothesis, for every non-optimal $f^{(i)}$, i.e. $i \geqslant \ell + 1$, there exists an index $e(i) \in [m]$ such that

$$1/(D\gamma_A) \leqslant f_{e(i)}^{(i)} \quad\text{and}\quad f_{e(i)} < \varepsilon/(mD\gamma_A\cdot C).$$

Therefore, we have

$$\alpha_i/(D\gamma_A) \leqslant \alpha_i f_{e(i)}^{(i)} \leqslant \sum_{j=1}^{m}\alpha_j f_{e(i)}^{(j)} = f_{e(i)} < \varepsilon/(mD\gamma_A\cdot C),$$

and hence $\sum_{i=\ell+1}^{m}\alpha_i \leqslant \varepsilon/C$. Further, by Lemma 9.10, for every j we have

$$\|f^{(j)}\|_\infty \leqslant D\|b/\gamma_A\|_1 = C/(2D\gamma_A).$$

Let $\beta \geqslant 0$ be an arbitrary vector satisfying $\sum_{i=1}^{\ell} \beta_i = \sum_{i=\ell+1}^{m} \alpha_i$. Let $\nu_i = \alpha_i + \beta_i$ for every $i \in [\ell]$ and let $f^\star = \sum_{i=1}^{\ell} \nu_i f^{(i)}$. Then, $f^\star$ is an optimal solution and we have

$$\begin{aligned} \|f^\star - f\|_\infty &= \left\| \sum_{i=1}^{\ell} \beta_i f^{(i)} - \sum_{i=\ell+1}^{m} \alpha_i \cdot f^{(i)} \right\|_\infty \\ &\leqslant \max_{i \in [1:m]} \left\| f^{(i)} \right\|_\infty \cdot \left(\sum_{i=1}^{\ell} \beta_i + \sum_{i=\ell+1}^{m} \alpha_i \right) \\ &\leqslant \frac{2\varepsilon}{C} \cdot \frac{C}{2D\gamma_A} = \frac{\varepsilon}{D\gamma_A}. \end{aligned}$$

□

In the following lemma, we extend the analysis in [SV16b, Lemma 5.6] from the transshipment problem to positive linear programs. Our result crucially relies on an argument that uses the parameter $\Phi = \min_{g \in \mathcal{N}} c^{\mathrm{T}} g - \mathrm{opt}$. It is here, where our analysis incurs the linear step size dependence on Φ/opt and the quadratic dependence on opt/Φ for the number of steps.

An important technical detail is that the first regime incurs an extra (Φ/opt)-factor dependence. At first glance, this might seem unnecessary due to Corollary 10.1, however it is crucial for an inductive argument in our analysis to hold, see (10.11) and (10.12). Further, we note that the undirected Physarum dynamics (8.7) satisfies $x_{\min}^{(t)} \geqslant (1-h)^t \cdot x_{\min}^{(0)}$, whereas the directed Physarum-inspired dynamics (8.6) might yield a value $x_{\min}^{(t)}$ which decreases with faster than exponential rate. As our analysis incurs a logarithmic dependence on $1/x_{\min}^{(0)}$, it is prohibitive to decouple the two regimes and give bounds in terms of $\log(1/x_{\min}^{(t)})$, which would be necessary as $x^{(t)}$ is the initial point of the second regime.

Lemma 10.15. *Let g be an arbitrary non-optimal basic feasible solution. Given $x^{(0)} \in X$ and its corresponding α_0, the Physarum-inspired dynamics (8.6) initialized with $x^{(0)}$ runs in two regimes:*

1. *The first regime is executed when $\alpha^{(0)} \notin [1/2, 1/h_0]$ and computes a point $x^{(t)} \in X$ such that $\alpha_t \in [1/2, 1/h_0]$. In particular, if $\alpha^{(0)} < 1/2$ then $h \leqslant (\Phi/\mathrm{opt}) \cdot (\alpha_0 h_0)^2$ and $t = 1/h$. Otherwise, if $\alpha^{(0)} > 1/h_0$ then $h \leqslant \Phi/\mathrm{opt}$ and $t = \lfloor \log_{1/(1-h)} [h_0(\alpha^{(0)} - 1)/(1-h_0)] \rfloor$.*
2. *The second regime starts from a point $x^{(t)} \in X$ such that $\alpha_t \in [1/2, 1/h_0]$, it has step size $h \leqslant (\mathrm{opt}/\Phi) \cdot h_0^2/2$ and for any $k \geqslant 4 \cdot c^{\mathrm{T}} g/(h\Phi) \cdot \ln(\Psi^{(0)}/\varepsilon x_{\min}^{(0)})$, guarantees the existence of an index $e \in [m]$ such that $g_e > 0$ and $x_e^{(t+k)} < \varepsilon$.*

Proof. Similar to the work of [BBD+13, SV16b], we use a potential function that takes as input a basic feasible solution g and a step number ℓ, and is defined by

$$\mathcal{B}_g^{(\ell)} \overset{\mathrm{def}}{=} \sum_{e \in [m]} g_e c_e \ln x_e^{(\ell)}.$$

Since $x_e^{(\ell+1)} = x_e^{(\ell)}(1 + h^{(\ell)}[c_e^{-1} \cdot A_e^{\mathrm{T}} p^{(\ell)} - 1])$, we have

$$\begin{aligned} \mathcal{B}_g^{(\ell+1)} - \mathcal{B}_g^{(\ell)} &= \sum_e g_e c_e \ln \frac{x_e^{(\ell+1)}}{x_e^{(\ell)}} = \sum_e g_e c_e \ln \left(1 + h^{(\ell)} \left[\frac{A_e^{\mathrm{T}} p^{(\ell)}}{c_e} - 1 \right] \right) \\ &\leqslant h^{(\ell)} \sum_e g_e c_e \left[\frac{A_e^{\mathrm{T}} p^{(\ell)}}{c_e} - 1 \right] = h^{(\ell)} \left[-c^{\mathrm{T}} g + [p^{(\ell)}]^{\mathrm{T}} A g \right] = h^{(\ell)} \left[-c^{\mathrm{T}} g + b^{\mathrm{T}} p^{(\ell)} \right]. \end{aligned} \tag{10.10}$$

Let $f^\star$ be an optimal solution to (8.5). In order to lower bound $\mathcal{B}_{f^\star}^{(\ell+1)} - \mathcal{B}_{f^\star}^{(\ell)}$, we use the inequality $\ln(1+x) \geqslant x - x^2$, for all $x \in [-\frac{1}{2}, \frac{1}{2}]$. Then, we have

$$\begin{aligned} \mathcal{B}_{f^\star}^{(\ell+1)} - \mathcal{B}_{f^\star}^{(\ell)} &= \sum_e f_e^\star c_e \ln \frac{x_e^{(\ell+1)}}{x_e^{(\ell)}} = \sum_e f_e^\star c_e \ln \left(1 + h^{(\ell)} \left[\frac{A_e^{\mathrm{T}} p^{(\ell)}}{c_e} - 1 \right] \right) \\ &\geqslant \sum_e f_e^\star c_e \left(h^{(\ell)} \left[\frac{A_e^{\mathrm{T}} p^{(\ell)}}{c_e} - 1 \right] - [h^{(\ell)}]^2 \left[\frac{A_e^{\mathrm{T}} p^{(\ell)}}{c_e} - 1 \right]^2 \right) \\ &\geqslant h^{(\ell)} \left(b^{\mathrm{T}} p^{(\ell)} - \mathrm{opt} - h^{(\ell)} \cdot (1/2\alpha_\ell h_0)^2 \cdot \mathrm{opt} \right), \end{aligned} \tag{10.11}$$

where the last inequality follows by combining

$$\sum_e f_e^\star c_e[(c_e^{-1}A_e^{\mathrm{T}}p^{(\ell)}) - 1] = [p^{(\ell)}]^{\mathrm{T}}Af^\star - \mathrm{opt} = b^{\mathrm{T}}p^{(\ell)} - \mathrm{opt},$$

$\|A^{\mathrm{T}}p^{(\ell)}\|_\infty \leqslant D\|c\|_1/\alpha_\ell$ (by Lemma 10.4 and Lemma 10.8 applied with $x^{(\ell)} \in X$), $h_0 = c_{\min}/(4D\|c\|_1)$ and

$$h^{(\ell)}\sum_e f_e^\star c_e \cdot (c_e^{-1}A_e^{\mathrm{T}}p^{(\ell)} - 1)^2 \leqslant h^{(\ell)}(2D\|c\|_1/\alpha_\ell c_{\min})^2\mathrm{opt} = h^{(\ell)}(1/2\alpha_\ell h_0)^2\mathrm{opt}.$$

Further, by combining (10.10), (10.11), $c^{\mathrm{T}}g - \mathrm{opt} \geqslant \Phi$ for every non-optimal basic feasible solution g and provided that the inequality $h^{(\ell)}(1/2\alpha_\ell h_0)^2\mathrm{opt} \leqslant \Phi/2$ holds, we obtain

$$\mathcal{B}_{f^\star}^{(\ell+1)} - \mathcal{B}_{f^\star}^{(\ell)} \geqslant h^{(\ell)}\left(b^{\mathrm{T}}p^{(\ell)} - c^{\mathrm{T}}g\right) + h^{(\ell)}\left(c^{\mathrm{T}}g - \mathrm{opt} - \frac{\Phi}{2}\right) \geqslant \mathcal{B}_g^{(\ell+1)} - \mathcal{B}_g^{(\ell)} + \frac{h^{(\ell)}\Phi}{2}. \tag{10.12}$$

Using Corollary 10.1, we partition the Physarum-inspired dynamics (8.6) execution into three regimes, based on α_0. For every $i \in \{1,2,3\}$, we show next that the i-th regime has a fixed step size $h^{(\ell)} = h_i$ such that $h^{(\ell)}(1/2\alpha_\ell h_0)^2\mathrm{opt} \leqslant \Phi/2$, for every step ℓ in this regime.

By Lemma 10.9, for every $i \in \{1,2,3\}$ it holds for every step ℓ in the i-th regime that

$$\alpha_\ell = 1 - (1-h_i)^\ell \cdot (1-\alpha_0). \tag{10.13}$$

Case 1: Suppose $\alpha_0 > 1/h_0$. Notice that $h^{(\ell)} = \Phi/\mathrm{opt}$ suffices, since $1/(2\alpha_\ell h_0) < 1/2$ for every $\alpha_\ell > 1/h_0$. Further, by applying (10.13) with $\alpha_t \overset{\text{def}}{=} 1/h_0$, we have $t = \lfloor \log_{1/(1-h^{(\ell)})}[h_0(\alpha^{(0)} - 1)/(1 - h_0)]\rfloor \leqslant (\mathrm{opt}/\Phi)\cdot\log(\alpha_0 h_0)$. Note that by (10.13) the sequence $\{\alpha_\ell\}_{\ell\leqslant t}$ is decreasing, and by Corollary 10.1 we have $1 < \alpha_t \leqslant 1/h_0$.

Case 2: Suppose $\alpha_0 \in (0,1/2)$. By (10.13) the sequence $\{\alpha_\ell\}_{\ell\in\mathbb{N}}$ is increasing and by Corollary 10.1 the regime is terminated once $\alpha_\ell \in [1/2,1)$. Observe that $h^{(\ell)} = (\Phi/\mathrm{opt})\cdot(\alpha_0 h_0)^2$ suffices, since $\alpha_0 \leqslant \alpha_\ell$. Then, by (10.13) applied with $\alpha_t \overset{\text{def}}{=} 1/2$, this regime has at most $t = (\mathrm{opt}/\Phi)\cdot(1/\alpha_0 h_0)^2$ steps.

Case 3: Suppose $\alpha_0 \in [1/2, 1/h_0]$. By (10.13) the sequence $\{\alpha_\ell\}_{\ell\in\mathbb{N}}$ converges to 1 (decreases if $\alpha_0 \in (1,1/h_0]$ and increases when $\alpha_0 \in [1/2,1)$. Notice that $h^{(\ell)} = (\Phi/\mathrm{opt})\cdot h_0^2/2$ suffices, since $1/2 \leqslant \alpha_\ell \leqslant 1/h_0$ for every $\ell \in \mathbb{N}$. We note that the number of steps in this regime is to be determined soon.

Hence, we conclude that inequality (10.12) holds. Further, using Case 1 and Case 2 there is an integer $t \in \mathbb{N}$ such that $\alpha_t \in [1/2, 1/h_0]$. Let $k \in \mathbb{N}$ be the number of steps in Case 3, and let $h \overset{\text{def}}{=} (\Phi/\mathrm{opt})\cdot h_0^2/2$. Then, for every $\ell \in \{t,\ldots,t+k-1\}$ it holds that $h^{(\ell)} = h$ and thus

$$\mathcal{B}_{f^\star}^{(t+k)} - \mathcal{B}_{f^\star}^{(0)} \geqslant \mathcal{B}_g^{(t+k)} - \mathcal{B}_g^{(0)} + \sum_{\ell=0}^{t+k-1}\frac{h^{(\ell)}\Phi}{2} \geqslant \mathcal{B}_g^{(t+k)} - \mathcal{B}_g^{(0)} + k\cdot\frac{h\Phi}{2}. \tag{10.14}$$

By Lemma 10.6, $\mathcal{B}_g^{(\ell)} \leqslant c^{\mathrm{T}}g\cdot\ln\Psi^{(0)}$ for every basic feasible solution g and every $\ell \in \mathbb{N}$, and thus

$$\begin{aligned}
\mathcal{B}_g^{(t+k)} &\leqslant -k\cdot\frac{h\Phi}{2} + \mathcal{B}_g^{(0)} + \mathcal{B}_{f^\star}^{(t+k)} - \mathcal{B}_{f^\star}^{(0)}\\
&\leqslant -k\cdot\frac{h\Phi}{2} + c^{\mathrm{T}}g\cdot\ln\Psi^{(0)} + \mathrm{opt}\cdot\ln\Psi^{(0)} - \mathrm{opt}\cdot\ln x_{\min}^{(0)}\\
&\leqslant -k\cdot\frac{h\Phi}{2} + 2c^{\mathrm{T}}g\cdot\ln\frac{\Psi^{(0)}}{x_{\min}^{(0)}}.
\end{aligned}$$

Suppose for the sake of a contradiction that for every $e \in [m]$ with $g_e > 0$ it holds $x_e^{(t+k)} > \varepsilon$. Then, $\mathcal{B}_g^{(t+k)} > c^{\mathrm{T}}g\cdot\ln\varepsilon$ yields $k < 4\cdot c^{\mathrm{T}}g/(h\Phi)\cdot\ln(\Psi^{(0)}/(\varepsilon x_{\min}^{(0)}))$, a contradiction to the choice of k. □

10.6 Proof of Theorem 10.2

By Corollary 10.1 and Lemma 10.15, if $x^{(0)} \in X$ such that $\alpha^{(0)} > 1/h_0$, we work with $h \leqslant \Phi/\mathrm{opt}$ and after $t = \lfloor\log_{1/(1-h)}[h_0(\alpha^{(0)} - 1)/(1-h_0)]\rfloor \leqslant (\mathrm{opt}/\Phi)\cdot\log(\alpha_0 h_0)$ steps, we obtain $x^{(t)} \in X$ such that $\alpha_t \in (1, 1/h_0]$. Otherwise, if $\alpha^{(0)} \in (0,1/2)$ we work with $h \leqslant (\Phi/\mathrm{opt})\cdot(\alpha_0 h_0)^2$ and after $t = 1/h$

steps, we obtain $x^{(t)} \in X$ such that $\alpha_t \in [1/2, 1)$. Hence, we can assume that $\alpha_t \in [1/2, 1/h_0]$ and set $h \leqslant (\Phi/\mathrm{opt}) \cdot h_0^2$. Then, the Lemmas in Section 10.4 and 10.5 are applicable.

Let $E_1 \overset{\mathrm{def}}{=} D\|b/\gamma_A\|_1\|c\|_1$, $E_2 \overset{\mathrm{def}}{=} 8m\rho_A\Psi^{(0)}$, $E_3 \overset{\mathrm{def}}{=} 2mD^3\gamma_A\|b\|_1$ and $E_4 \overset{\mathrm{def}}{=} 8mD^2\|b\|_1$. Consider an arbitrary non-optimal basic feasible solution g.

By Lemma 9.10, we have $c^{\mathrm{T}}g \leqslant E_1$ and thus both Lemma 10.12 and Lemma 10.15 are applicable with h, $\varepsilon^\star \overset{\mathrm{def}}{=} \varepsilon/E_4$ and any $k \geqslant k_0 \overset{\mathrm{def}}{=} 4E_1/(h\Phi) \cdot \ln[(E_2/\min\{1, x_{\min}^{(0)}\}) \cdot (D\gamma_A/\varepsilon^\star)]$. Hence, by Lemma 10.15, the Physarum-inspired dynamics (8.6) guarantees the existence of an index $e \in [m]$ such that $g_e > 0$ and $x_e^{(t+k)} < \varepsilon^\star/(D\gamma_A)$. Moreover, by Lemma 10.12 there is a non-negative feasible kernel-free vector f such that $\|x^{(t+k)} - f\|_\infty < \varepsilon^\star/(D\gamma_A)$. Thus, for the index e it follows that $g_e > 0$ and $f_e < 2\varepsilon^\star/D\gamma_A = (\varepsilon/2) \cdot (4/E_4 D\gamma_A) = \varepsilon/(2E_3)$. Then Lemma 10.14, yields $\|f - f^\star\|_\infty < \varepsilon/(2D\gamma_A)$ and by triangle inequality we have $\|x^{(k)} - f^\star\|_\infty < \varepsilon/(D\gamma_A)$.

By construction, $\rho_A = \max\{D\gamma_A, nD^2\|A\|_\infty\} \leqslant nD^2\gamma_A\|A\|_\infty$. Let $E_2' = 8mnD^2\gamma_A\|A\|_\infty\Psi^{(0)}$ and $E_5 = E_2'E_4 \cdot D\gamma_A = 8^2m^2nD^5\gamma_A^2\|A\|_\infty\|b\|_1$. Further, let $C_1 = E_1$ and $C_2 = E_5$. Then, the statement follows for any $k \geqslant k_1 \overset{\mathrm{def}}{=} 4C_1/(h\Phi) \cdot \ln(C_2\Psi^{(0)}/(\varepsilon \cdot \min\{1, x_{\min}^{(0)}\}))$. □

10.7 Preconditioning

In this section, we generalize the preconditioning technique developed in [BBD⁺13, SV16b] for flow problems, to the setting of positive linear programs.

Theorem 10.16. *Given an integral LP $(A, b, c > 0)$, a positive $x^{(0)} \in \mathbb{R}^m$ and a parameter $\varepsilon \in (0, 1)$. Let $([A\,|\,b], b, (c, c'))$ be an extended LP with $c' = 2C_1$ and $z^{(0)} \overset{\mathrm{def}}{=} 1 + D_S\|x\|_\infty\|A\|_1\|b\|_1$.*[2] *Then, $(x^{(0)}; z^{(0)})$ is a strongly dominating starting point of the extended problem such that $y^{\mathrm{T}}[A\,|\,b](x^{(0)}, z^{(0)}) \geqslant 1$, for all $y \in Y$. In particular, the Physarum-inspired dynamics (8.6) initialized with $(x^{(0)}, z^{(0)})$ and a step size $h \leqslant h_0^2/C_3$, outputs for any $k \geqslant 4C_1 \cdot (D\gamma_A)^2/h \cdot \ln(C_2\Upsilon^{(0)}/(\varepsilon \cdot \min\{1, x_{\min}^{(0)}\}))$ a vector $(x^{(k)}, z^{(k)}) > 0$ such that $\mathrm{dist}(x^{(k)}, X_\star) < \varepsilon/(D\gamma_A)$ and $z^{(k)} < \varepsilon/(D\gamma_A)$, where $\Upsilon^{(0)} \overset{\mathrm{def}}{=} \max\{\Psi^{(0)}, z^{(0)}\}$.*

Theorem 10.16 subsumes [SV16b, Theorem 1.2] for flow problems by giving a tighter asymptotic convergence rate, since for the transshipment problem A is a totally unimodular matrix and satisfies $D = D_S = 1$, $\gamma_A = 1$, $\|A\|_\infty = 1$ and $\Phi = 1$. We note that the scalar $z^{(0)}$ depends on the scaled determinant D_S, see Theorem 8.3.

10.7.1 Proof of Theorem 10.16

In the extended problem, we concatenate to matrix A a column equal to b such that the resulting constraint matrix becomes $[A\,|\,b]$. Let c' be the cost and let x' be the initial capacity of the newly inserted constraint column. We will determine c' and x' in the course of the discussion. Consider the dual of the max-flow like LP for the extended problem. It has an additional variable z' and reads

$$\min\left\{\, x^{\mathrm{T}}z + x'z' : z \geqslant 0;\; z' \geqslant 0;\; z \geqslant A^{\mathrm{T}}y;\; z' \geqslant b^{\mathrm{T}}y;\; b^{\mathrm{T}}y = 1 \,\right\}.$$

In any optimal solution, $z' = b^{\mathrm{T}}y = 1$ and hence the dual is equivalent to

$$\min\left\{\, x^{\mathrm{T}}z + x' : z \geqslant 0;\; z \geqslant A^{\mathrm{T}}y;\; b^{\mathrm{T}}y = 1 \,\right\}. \tag{10.15}$$

The strongly dominating set of the extended problem is therefore equal to

$$X = \left\{ \begin{pmatrix} x \\ x' \end{pmatrix} \in \mathbb{R}^{m+1}_{>0} : y^{\mathrm{T}}[A\,|\,b]\begin{pmatrix} x \\ x' \end{pmatrix} > 0 \text{ for all } y \in Y \right\}. \tag{10.16}$$

The defining condition translates into $x' > -y^{\mathrm{T}}Ax$ for all $y \in Y$. We summarize the discussion in the following Lemma.

Lemma 10.17. *Given a positive $x \in \mathbb{R}^m$, let $\rho \overset{\mathrm{def}}{=} \|b\|_1 D_S$ and $x' \overset{\mathrm{def}}{=} 1 + \rho\|A\|_1\|x\|_\infty$, where $\|A\|_1 \overset{\mathrm{def}}{=} \sum_{i,j}|A_{i,j}|$ and $D_S \overset{\mathrm{def}}{=} \max\{|\det(A')| : A'$ is a square sub-matrix of $A\}$. Then, $(x;\ x')$ is a strongly dominating starting point of the extended problem such that $y^{\mathrm{T}}[A\,|\,b](x;\ x') = y^{\mathrm{T}}Ax + x' \geqslant 1$, for all $y \in Y$.*

[2] We denote by $\|A\|_1 \overset{\mathrm{def}}{=} \sum_{i,j}|A_{i,j}|$, i.e. we interpret matrix A as a vector and apply to it the standard ℓ_1 norm.

Proof. We show first that $\max_{y\in Y}\|y\|_\infty \leqslant \rho$ implies the statement. Let $y \in Y$ be arbitrary. Since $|y^{\mathrm{T}}Ax| \leqslant \|A\|_1\|x\|_\infty\|y\|_\infty$, we have $\max_{y\in Y}|y^{\mathrm{T}}Ax| \leqslant \rho\|A\|_1\|x\|_\infty = x' - 1$ and hence $y^{\mathrm{T}}[A\,|\,b](x;\ x') \geqslant 1$.

It remains to show that $\max_{y\in Y}\|y\|_\infty \leqslant \rho$. The constraint polyhedron of the dual (10.15) is given in matrix notation as

$$P^{(\mathrm{ext})} \stackrel{\mathrm{def}}{=} \left\{ \begin{pmatrix} z \\ y \end{pmatrix} \in \mathbb{R}^{m+n} : \begin{bmatrix} I_{m\times m} & -A^{\mathrm{T}} \\ 0_m^{\mathrm{T}} & b^{\mathrm{T}} \\ I_{m\times m} & 0_{m\times n} \end{bmatrix} \begin{pmatrix} z \\ y \end{pmatrix} \begin{matrix} \geqslant \\ = \\ \geqslant \end{matrix} \begin{pmatrix} 0_m \\ 1 \\ 0_m \end{pmatrix} \right\}.$$

Let us denote the resulting constraint matrix and vector by $M \in \mathbb{R}^{2m+1\times m+n}$ and $d \in \mathbb{R}^{2m+1}$, respectively.

Note that if $b = 0$ then the primal LP (10.2) is either unbounded or infeasible. Hence, we consider the non-trivial case when $b \neq 0$. Observe that the polyhedron $P^{(\mathrm{ext})}$ is not empty, since for any y such that $b^{\mathrm{T}}y = 1$ there is $z = \max\{0, A^{\mathrm{T}}y\}$ satisfying $(z;\ y) \in P^{(\mathrm{ext})}$. Further, $P^{(\mathrm{ext})}$ does not contain a line (see Section 10.3) and thus $P^{(\mathrm{ext})}$ has at least one extreme point $p' \in P^{(\mathrm{ext})}$. As the dual LP (10.3) has a bounded value (the target function is lower bounded by 0) and an extreme point exists ($p' \in P^{(\mathrm{ext})}$), the optimum is attained at an extreme point $p \in P^{(\mathrm{ext})}$. Moreover, as every extreme point is a basic feasible solution and matrix M has linearly independent columns (A has full row rank), it follows that p has $m+n$ tight linearly independent constraints.

Let $M_{B(p)} \in \mathbb{R}^{m+n\times m+n}$ be the basis submatrix of M satisfying $M_{B(p)}p = d_{B(p)}$. Since A, b are integral and $M_{B(p)}$ is invertible, using Laplace expansion we have $1 \leqslant |\det(M_{B(p)})| \leqslant \|b\|_1 D_S = \rho$. Let Q_i denotes the matrix formed by replacing the i-th column of $M_{B(p)}$ by the column vector $d_{B(p)}$. Then, by Cramer's rule, it follows that $|y_i| = |\det(Q_i)/\det(M_{B(p)})| \leqslant |\det(M_{B(p)})| \leqslant \rho$, for all $i \in [n]$. □

It remains to fix the cost of the new column. Using Lemma 9.10, $\mathrm{opt} \leqslant c^{\mathrm{T}}x^{(k)} \leqslant C_1$ for every $k \in \mathbb{N}$, and thus we set $c' \stackrel{\mathrm{def}}{=} 2C_1$.

10.8 A Simple Lower Bound

Building on [SV16b, Lemma B.1], we give a lower bound on the number of steps required for computing an ε-approximation to the optimum shortest path. In particular, we show that for the Physarum-inspired dynamics (8.6) to compute a point $x^{(k)}$ such that $\mathrm{dist}(x^{(k)}, X_\star) < \varepsilon$, the required number of steps k has to grow linearly in $\mathrm{opt}/(h\Phi)$ and $\ln(1/\varepsilon)$.

Theorem 10.18. *Let (A, b, c) be a positive LP instance such that $A = [1\ 1]$, $b = 1$ and $c = (\mathrm{opt},\ \mathrm{opt}{+}\Phi)^{\mathrm{T}}$, where $\mathrm{opt} > 0$ and $\Phi > 0$. Then, for any $\varepsilon \in (0,1)$ the discrete directed Physarum-inspired dynamics (8.6) initialized with $x^{(0)} = (1/2, 1/2)$ and any step size $h \in (0, 1/2]$, requires at least $k = (1/2h)\cdot\max\{\mathrm{opt}/\Phi, 1\}\cdot \ln(2/\varepsilon)$ steps to guarantee $x_1^{(k)} \geqslant 1-\varepsilon$, $x_2^{(k)} \leqslant \varepsilon$. Moreover, if $\varepsilon \leqslant \Phi/(2\mathrm{opt})$ then $c^{\mathrm{T}}x^{(k)} \geqslant (1+\varepsilon)\mathrm{opt}$ as long as $k \leqslant (1/2h)\cdot\max\{\mathrm{opt}/\Phi, 1\}\cdot\ln(2\Phi/(\varepsilon\cdot\mathrm{opt}))$.*

Proof. Let $c_1 = \mathrm{opt}$ and $c_2 = \gamma\mathrm{opt}$, where $\gamma = 1 + \Phi/\mathrm{opt}$. We first derive closed-form expressions for $x_1^{(k)}$, $x_2^{(k)}$, and $x_1^{(k)} + x_2^{(k)}$. Let $s^{(k)} = \gamma x_1^{(k)} + x_2^{(k)}$. For any $k \in \mathbb{N}$, we have $q_1^{(k)} + q_2^{(k)} = 1$ and $q_1^{(k)}/q_2^{(k)} = (x_1^{(k)}/c_1)/(x_2^{(k)}/c_2) = \gamma x_1^{(k)}/x_2^{(k)}$. Therefore, $q_1^{(k)} = \gamma x_1^{(k)}/s^{(k)}$ and $q_2^{(k)} = x_2^{(k)}/s^{(k)}$, and hence

$$x_1^{(k)} = (1 + h(-1 + \gamma/s^{(k-1)}))x_1^{(k-1)} \quad\text{and}\quad x_2^{(k)} = (1 + h(-1 + 1/s^{(k-1)}))x_2^{(k-1)}. \tag{10.17}$$

Further, $x_1^{(k)} + x_2^{(k)} = (1-h)(x_1^{(k-1)} + x_2^{(k-1)}) + h$, and thus by induction $x_1^{(k)} + x_2^{(k)} = 1$ for all $k \in \mathbb{N}$.

Therefore, $s^{(k)} \leqslant \gamma$ for all $k \in \mathbb{N}$ and hence $x_1^{(k)} \geqslant x_1^{(k-1)}$, i.e. the sequence $\{x_1^{(k)}\}_{k\in\mathbb{N}}$ is increasing and the sequence $\{x_2^{(k)}\}_{k\in\mathbb{N}}$ is decreasing. Moreover, since $h(-1 + 1/s^{(k-1)}) \geqslant h(1-\gamma)/\gamma = -h\Phi/(\mathrm{opt}+\Phi)$ and using the inequality $1 - z \geqslant e^{-2z}$ for every $z \in [0, 1/2]$, it follows by (10.17) and induction on k that

$$x_2^{(k)} \geqslant \left(1 - \frac{h\Phi}{\mathrm{opt}+\Phi}\right)^k x_2^{(0)} \geqslant \frac{1}{2}\exp\left\{-k\cdot\frac{2h\Phi}{\mathrm{opt}+\Phi}\right\}.$$

Thus, $x_2^{(k)} \geqslant \varepsilon$ whenever $k \leqslant (1/2h)\cdot(\mathrm{opt}/\Phi + 1)\cdot\ln(2/\varepsilon)$. This proves the first claim.

For the second claim, observe that $c^{\mathrm{T}}x^{(k)} = \mathrm{opt}\cdot x_1^{(k)} + \gamma\mathrm{opt}\cdot x_2^{(k)} = \mathrm{opt}\cdot(1 + (\gamma-1)x_2^{(k)})$. This is greater than $(1+\varepsilon)\mathrm{opt}$ iff $x_2^{(k)} \geqslant \varepsilon\cdot\mathrm{opt}/\Phi$. Thus, $c^{\mathrm{T}}x^{(k)} \geqslant (1+\varepsilon)\mathrm{opt}$ as long as $k \leqslant (1/2h)\cdot(\mathrm{opt}/\Phi + 1)\cdot\ln(2/(\varepsilon\cdot\mathrm{opt}/\Phi))$. □

Chapter 11

Bibliography

[ABH$^+$05] Sanjeev Arora, Eli Berger, Elad Hazan, Guy Kindler, and Muli Safra. On non-approximability for quadratic programs. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005), 23-25 October 2005, Pittsburgh, PA, USA, Proceedings*, pages 206–215, 2005.

[ABS10] Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6(4):59:1–59:26, 2010.

[Ale11] Michael Alekhnovich. More on average case vs approximation complexity. *Computational Complexity*, 20(4):755–786, 2011.

[APY09] Noga Alon, Rina Panigrahy, and Sergey Yekhanin. Deterministic approximation algorithms for the nearest codeword problem. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings*, pages 339–351, 2009.

[AS99] Noga Alon and Benny Sudakov. On two segmentation problems. *J. Algorithms*, 33(1):173–184, 1999.

[AW17] Josh Alman and R. Ryan Williams. Probabilistic rank and matrix rigidity. In *STOC*, pages 641–652, 2017.

[AY95] Charles J. Alpert and So-Zen Yao. Spectral partitioning: The more eigenvectors, the better. In *Proceedings of the 32st Conference on Design Automation, San Francisco, California, USA, Moscone Center, June 12-16, 1995.*, pages 195–200, 1995.

[BBB$^+$18] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P. Woodruff. A PTAS for $\mathrm{l_p}$-low rank approximation. *CoRR*, abs/1807.06101, 2018.

[BBB$^+$19] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P. Woodruff. A PTAS for $\mathrm{l_p}$-low rank approximation. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, Westin San Diego, San Diego, California, USA, January 6-9*, 2019. To appear.

[BBD$^+$13] Luca Becchetti, Vincenzo Bonifaci, Michael Dirnberger, Andreas Karrenbauer, and Kurt Mehlhorn. Physarum can compute shortest paths: Convergence proofs and complexity bounds. In *ICALP*, volume 7966 of *LNCS*, pages 472–483, 2013.

[BBK$^+$17] Ruben Becker, Vincenzo Bonifaci, Andreas Karrenbauer, Pavel Kolev, and Kurt Mehlhorn. Two results on slime mold computations. *CoRR*, abs/1707.06631, 2017.

[BBK$^+$18] Ruben Becker, Vincenzo Bonifaci, Andreas Karrenbauer, Pavel Kolev, and Kurt Mehlhorn. Two results on slime mold computations. *Theoretical Computer Science*, 2018.

[BHH$^+$06] James C. Bezdek, Richard J. Hathaway, Jacalyn M. Huband, Christopher Leckie, and Kotagiri Ramamohanarao. Approximate clustering in very large relational data. *Int. J. Intell. Syst.*, 21(8):817–841, 2006.

[BHI02] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257, 2002.

[BK02] Piotr Berman and Marek Karpinski. Approximating minimum unsatisfiability of linear equations. In *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 6-8, 2002, San Francisco, CA, USA.*, pages 514–516, 2002.

[BKG15] Christos Boutsidis, Prabhanjan Kambadur, and Alex Gittens. Spectral clustering via the power method - provably. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 40–48, 2015.

[BKW17a] Karl Bringmann, Pavel Kolev, and David P. Woodruff. Approximation algorithms for l_0-low rank approximation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NIPS 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6651–6662, 2017.

[BKW17b] Karl Bringmann, Pavel Kolev, and David P. Woodruff. Approximation algorithms for l_0-low rank approximation. *CoRR*, abs/1710.11253, 2017. Full version of [BKW17a].

[BM14] Christos Boutsidis and Malik Magdon-Ismail. Faster svd-truncated regularized least-squares. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014*, pages 1321–1325, 2014.

[BMV12] Vincenzo Bonifaci, Kurt Mehlhorn, and Girish Varma. Physarum can compute shortest paths. *Journal of Theoretical Biology*, 309:121 – 133, 2012.

[BN01] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 585–591, 2001.

[Bon13] Vincenzo Bonifaci. Physarum can compute shortest paths: A short proof. *Inf. Process. Lett.*, 113(1-2):4–7, 2013.

[Bon15] Vincenzo Bonifaci. A revised model of network transport optimization in Physarum Polycephalum. November 2015.

[Bon16] Vincenzo Bonifaci. On the convergence time of a natural dynamics for linear programming. *CoRR*, abs/1611.06729, 2016.

[Boy04] Stephen Boyd. *Convex Optimization.* Cambridge University Press, 2004.

[BV10] Radim Belohlávek and Vilém Vychodil. Discovery of optimal factors in binary data via a novel method of matrix decomposition. *J. Comput. Syst. Sci.*, 76(1):3–20, 2010.

[BY02] Ziv Bar-Yossef. *The complexity of massive data set computations.* PhD thesis, University of California, Berkeley, 2002.

[CCDL14] Jiang-Zhong Cao, Pei Chen, Qingyun Dai, and Bingo Wing-Kuen Ling. Local information-based fast approximate spectral clustering. *Pattern Recognition Letters*, 38:63–69, 2014.

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[CGK$^+$17] Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P. Woodruff. Algorithms for ℓ_p low-rank approximation. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 806–814, 2017.

[Chu97] Fan R. K. Chung. *Spectral Graph Theory.* American Mathematical Society, 1997.

[CIK16] L. Sunil Chandran, Davis Issac, and Andreas Karrenbauer. On the parameterized complexity of biclique cover and partition. In *IPEC*, volume 63 of *LIPIcs*, pages 11:1–11:13, 2016.

[CKC$^+$16] Mihai Cucuringu, Ioannis Koutis, Sanjay Chawla, Gary L. Miller, and Richard Peng. Simple and scalable constrained clustering: a generalized spectral method. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 445–454, 2016.

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[CLSW98] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1998.

[DAJ+15] C. Dan, K. Arnsfelt Hansen, H. Jiang, L. Wang, and Y. Zhou. Low Rank Approximation of Binary Matrices: Column Subset Selection and Generalizations. *ArXiv e-prints*, 2015.

[DKLR00] Paul Dagum, Richard M. Karp, Michael Luby, and Sheldon M. Ross. An optimal algorithm for monte carlo estimation. *SIAM J. Comput.*, 29(5):1484–1496, 2000.

[DS01] Kenneth R. Davidson and Stanislaw J. Szarek. Chapter 8 local operator theory, random matrices and banach spaces. volume 1 of *Handbook of the Geometry of Banach Spaces*, pages 317 – 366. Elsevier Science B.V., 2001.

[Eig15] A multiplicative version of horn's inequality, 2015. https://math.stackexchange.com/questions/20302/eigenvalues-of-product-of-a-matrix-and-a-diagonal-matrix.

[FBCM04] Charless C. Fowlkes, Serge J. Belongie, Fan R. K. Chung, and Jitendra Malik. Spectral grouping using the nyström method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):214–225, 2004.

[Fei14] Uriel Feige. Np-hardness of hypercube 2-segmentation. *CoRR*, http://arxiv.org/abs/1411.0821, 2014.

[FGL+18] Fedor V. Fomin, Petr A. Golovach, Daniel Lokshtanov, Fahad Panolan, and Saket Saurabh. Approximation schemes for low-rank binary matrix approximation problems. volume abs/1807.07156, 2018.

[FGP18] Fedor V. Fomin, Petr A. Golovach, and Fahad Panolan. Parameterized low-rank binary matrix approximation. *CoRR*, abs/1803.06102, 2018.

[FLM+17] Fedor V. Fomin, Daniel Lokshtanov, Syed Mohammad Meesum, Saket Saurabh, and Meirav Zehavi. Matrix rigidity from the viewpoint of parameterized complexity. In *STACS*, volume 66 of *LIPIcs*, pages 32:1–32:14, 2017.

[FMS07] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8, 2007*, pages 11–18, 2007.

[For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010.

[GGYT12] Harold W. Gutch, Peter Gruber, Arie Yeredor, and Fabian J. Theis. ICA over finite fields - separability and algorithms. *Signal Processing*, 92(8):1796–1808, 2012.

[Gri80] D. Grigoriev. Using the notions of separability and independence for proving the lower bounds on the circuit complexity. *Journal of Soviet Math.*, 14(5):1450–1456, 1980.

[GT14] Shayan Oveis Gharan and Luca Trevisan. Partitioning into expanders. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1256–1266, 2014.

[GV15] Nicolas Gillis and Stephen A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *CoRR*, abs/1509.09236, 2015.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.).* Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[GVL12] Gene Howard Golub and Charles F. Van Loan. *Matrix computations.* Johns Hopkins studies in the mathematical sciences. The Johns Hopkins University Press, Baltimore, London, 2012.

[HK05] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. In *Proceedings of the 21st ACM Symposium on Computational Geometry, Pisa, Italy, June 6-8, 2005*, pages 126–134, 2005.

[IJNT11] Kentaro Ito, Anders Johansson, Toshiyuki Nakagaki, and Atsushi Tero. Convergence properties for the physarum solver. arXiv:1101.5249v1, January 2011.

[JK15] Gorav Jindal and Pavel Kolev. An efficient parallel algorithm for spectral sparsification of laplacian and sddm matrix polynomials. *CoRR*, abs/1507.07497, 2015.

[JKPS17a] Gorav Jindal, Pavel Kolev, Richard Peng, and Saurabh Sawlani. Density Independent Algorithms for Sparsifying k-Step Random Walks. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*, volume 81 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[JKPS17b] Gorav Jindal, Pavel Kolev, Richard Peng, and Saurabh Sawlani. Density independent algorithms for sparsifying k-step random walks. *CoRR*, abs/1702.06110, 2017.

[JPHY14] Peng Jiang, Jiming Peng, Michael Heath, and Rui Yang. A clustering approach to constrained binary matrix factorization. In *Data Mining and Knowledge Discovery for Big Data*, pages 281–303. Springer, 2014.

[JZ12] A. Johannson and J. Zou. A slime mold solver for linear programming problems. In *CiE*, pages 344–354, 2012.

[KG03] Mehmet Koyutürk and Ananth Grama. Proximus: a framework for analyzing very high dimensional discrete-attributed datasets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 147–156. ACM, 2003.

[KGR05] Mehmet Koyutürk, Ananth Grama, and Naren Ramakrishnan. Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets. *IEEE Trans. Knowl. Data Eng.*, 17(4):447–461, 2005.

[KGR06] Mehmet Koyutürk, Ananth Grama, and Naren Ramakrishnan. Nonorthogonal decomposition of binary matrices for bounded-error data compression and analysis. *ACM Transactions on Mathematical Software (TOMS)*, 32(1):33–69, 2006.

[KLL17] Tsz Chiu Kwok, Lap Chi Lau, and Yin Tat Lee. Improved cheeger's inequality and analysis of local graph partitioning using vertex expansion and expansion profile. *SIAM J. Comput.*, 46(3):890–910, 2017.

[Kly00] Alexander A. Klyachko. Random walks on symmetric spaces and inequalities for matrix spectra. *Linear Algebra and its Applications*, 319(1):37 – 59, 2000.

[KM16] Pavel Kolev and Kurt Mehlhorn. A note on spectral clustering. In *24th Annual European Symposium on Algorithms, ESA 2016, August 22-24, 2016, Aarhus, Denmark*, pages 57:1–57:14, 2016.

[KM18] Pavel Kolev and Kurt Mehlhorn. Approximate spectral clustering: Efficiency and guarantees. *CoRR*, abs/1509.09188, 2018. Submitted. A preliminary version of this paper was presented at the 24th Annual European Symposium on Algorithms (ESA 2016).

[KPR04] Jon M. Kleinberg, Christos H. Papadimitriou, and Prabhakar Raghavan. Segmentation problems. *J. ACM*, 51(2):263–280, 2004.

[KSS04] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Symposium on Foundations of Computer Science (FOCS 2004), 17-19 October 2004, Rome, Italy, Proceedings*, pages 454–462, 2004.

[KSS05] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear time algorithms for clustering problems in any dimensions. In *ICALP*, pages 1374–1385, 2005.

[KV09] Ravi Kannan and Santosh Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3-4):157–288, 2009.

[KVV04] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[LaS76] J. B. LaSalle. *The Stability of Dynamical Systems.* SIAM, 1976.

[LC10] Frank Lin and William W. Cohen. Power iteration clustering. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 655–662, 2010.

[LGT12] James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 1117–1130, 2012.

[Li05] Tao Li. A general model for clustering binary data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 188–197. ACM, 2005.

[LRTV12] Anand Louis, Prasad Raghavendra, Prasad Tetali, and Santosh Vempala. Many sparse cuts via higher eigenvalues. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 1131–1140, 2012.

[LZ04] Rong Liu and Hao Zhang. Segmentation of 3d meshes through spectral clustering. In *12th Pacific Conference on Computer Graphics and Applications (PG 2004), 6-8 October 2004, Seoul, Korea*, pages 298–305, 2004.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[MBLS01] Jitendra Malik, Serge J. Belongie, Thomas K. Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[MGNR06] Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 977–984, 2006.

[MGT15] Seyed Hamid Mirisaee, Éric Gaussier, and Alexandre Termier. Improved local search for binary matrix factorization. In *AAAI*, pages 1198–1204, 2015.

[Mie] Pauli Miettinen. Matrix decomposition methods for data mining: Computational complexity and algorithms. PhD Thesis, University of Helsinki, Finland, 2009.

[MMG$^+$08] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. The discrete basis problem. *IEEE Trans. Knowl. Data Eng.*, 20(10):1348–1362, 2008.

[MNV12] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi R. Varadarajan. The planar k-means problem is np-hard. *Theor. Comput. Sci.*, 442:13–21, 2012.

[MO08] T. Miyaji and Isamu Ohnishi. Physarum can solve the shortest path problem on Riemannian surface mathematically rigourously. *International Journal of Pure and Applied Mathematics*, 47:353–369, 2008.

[MS90] David W. Matula and Farhad Shahrokhi. Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, 27(1-2):113–123, 1990.

[MV14] Pauli Miettinen and Jilles Vreeken. MDL4BMF: minimum description length for boolean matrix factorization. *TKDD*, 8(4):18:1–18:31, 2014.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 849–856, 2001.

[Nys30] EJ Nystrm. On the practical resolution of integral equations with applications on boundary value tasks. *Acta Math.*, 54:185–204, 1930.

[NYT00] T. Nakagaki, H. Yamada, and Á. Tóth. Maze-solving by an amoeboid organism. *Nature*, 407:470, 2000.

[OR00] Rafail Ostrovsky and Yuval Rabani. Polynomial time approximation schemes for geometric k-clustering. In *FOCS*, pages 349–358, 2000.

[ORSS13] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28:1–28:22, January 2013.

[Phy10] http://people.mpi-inf.mpg.de/~mehlhorn/ftp/SlimeAusschnitt.webm, 2010.

[PP04] Massimiliano Pavan and Marcello Pelillo. Efficient out-of-sample extension of dominant-set clusters. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1057–1064, 2004.

[Pre81] P. M. Prenter. The numerical treatment of integral equations (c. t. h. baker). *SIAM Review*, 23(2):266–267, 1981.

[PRF16] Amichai Painsky, Saharon Rosset, and Meir Feder. Generalized independent component analysis over finite alphabets. *IEEE Trans. Information Theory*, 62(2):1038–1053, 2016.

[PS82] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1982.

[PSZ17] Richard Peng, He Sun, and Luca Zanetti. Partitioning well-clustered graphs: Spectral clustering works! *SIAM J. Comput.*, 46(2):710–743, 2017.

[RPG16] Siamak Ravanbakhsh, Barnabás Póczos, and Russell Greiner. Boolean matrix factorization and noisy completion via message passing. In *ICML*, volume 48, pages 945–954, 2016.

[SBM03] Jouni K. Seppänen, Ella Bingham, and Heikki Mannila. A simple algorithm for topic identification in 0-1 data. In *Knowledge Discovery in Databases: PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings*, pages 423–434, 2003.

[Sch99] Alexander Schrijver. *Theory of linear and integer programming*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1999.

[SH06] Tomás Singliar and Milos Hauskrecht. Noisy-or component analysis and its application to link analysis. *Journal of Machine Learning Research*, 7:2189–2213, 2006.

[SJY09] Bao-Hong Shen, Shuiwang Ji, and Jieping Ye. Mining discrete patterns via binary matrix factorization. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 757–766, 2009.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[SST06] Arvind Sankar, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Analysis Applications*, 28(2):446–476, 2006. Available at: http://www.cs.yale.edu/homes/spielman/Research/nopivotdas.pdf.

[ST14] Daniel A. Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Analysis Applications*, 35(3):835–885, 2014.

[Ste90] G. W. Stewart. *Matrix Perturbation Theory*. Academic Press, Inc., New York, USA, 1990.

[SV16a] Damian Straszak and Nisheeth K. Vishnoi. IRLS and slime mold: Equivalence and convergence. *CoRR*, abs/1601.02712, 2016.

[SV16b] Damian Straszak and Nisheeth K. Vishnoi. Natural algorithms for flow problems. In *SODA*, pages 1868–1883, 2016.

[SV16c] Damian Straszak and Nisheeth K. Vishnoi. On a natural dynamics for linear programming. In *ITCS*, pages 291–291, New York, NY, USA, 2016. ACM.

[SWZ18] Zhao Song, David P. Woodruff, and Peilin Zhong. Entrywise low rank approximation of general functions, 2018. Manuscript.

[Tas12] Kadim Tasdemir. Vector quantization based approximate spectral clustering of large datasets. *Pattern Recognition*, 45(8):3034–3044, 2012.

[Tes12] Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*. Graduate studies in mathematics. American Mathematical Society, 2012.

[TKN07] A. Tero, R. Kobayashi, and T. Nakagaki. A mathematical model for adaptive transport network in path finding by true slime mold. *Journal of Theoretical Biology*, pages 553–564, 2007.

[VAG07] Jaideep Vaidya, Vijayalakshmi Atluri, and Qi Guo. The role mining problem: finding a minimal descriptive set of roles. In *12th ACM Symposium on Access Control Models and Technologies, SACMAT 2007, Sophia Antipolis, France, June 20-22, 2007, Proceedings*, pages 175–184, 2007.

[Val77] Leslie G. Valiant. Graph-theoretic arguments in low-level complexity. In *Mathematical Foundations of Computer Science 1977, 6th Symposium, Tatranska Lomnica, Czechoslovakia, September 5-9, 1977, Proceedings*, pages 162–176, 1977.

[vL07] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[Wal74] A. J. Walker. New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronics Letters*, 10(8):127–128, April 1974.

[WD12] Lijun Wang and Ming Dong. Multi-level low-rank approximation-based spectral clustering for image segmentation. *Pattern Recognition Letters*, 33(16):2206–2215, 2012.

[WLRB09] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James C. Bezdek. Approximate spectral clustering. In *Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27-30, 2009, Proceedings*, pages 134–146, 2009.

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

[WS05] Scott White and Padhraic Smyth. A spectral clustering approach to finding communities in graph. In *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005, Newport Beach, CA, USA, April 21-23, 2005*, pages 274–285, 2005.

[Yer11] Arie Yeredor. Independent component analysis over galois fields of prime order. *IEEE Trans. Information Theory*, 57(8):5342–5359, 2011.

[YHJ09] Donghui Yan, Ling Huang, and Michael I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 907–916, 2009.

[ZLD$^+$10] Zhong-Yuan Zhang, Tao Li, Chris Ding, Xian-Wen Ren, and Xiang-Sun Zhang. Binary matrix factorization for analyzing gene expression data. *Data Mining and Knowledge Discovery*, 20(1):28–52, 2010.

[ZLDZ07] Zhongyuan Zhang, Tao Li, Chris Ding, and Xiangsun Zhang. Binary matrix factorization with applications. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 391–400. IEEE, 2007.

[ZP04] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1601–1608, 2004.