UNDERSTANDING DATA MANIPULATION AND HOW TO LEVERAGE IT TO IMPROVE GENERALIZATION

A Dissertation Proposal
Presented to
The Academic Faculty

By

Chi-Heng Lin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

December 2022

UNDERSTANDING DATA MANIPULATION AND HOW TO LEVERAGE IT TO IMPROVE GENERALIZATION

Approved by:

Dr. Eva L. Dyer, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Vidya K. Muthukumar
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mark A. Davenport
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Florian T. Schaefer
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Yao Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: December 6, 2022

To my family, my friends,

and the people who have faith in me.

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my advisor, Dr. Eva L. Dyer, for fully supporting my Ph.D. journey. I appreciate your guidance over these six years. You are a patient advisor and always willing to offer me valuable feedback. Throughout these years, I have been developing the skill set for conducting research under your guidance. This thesis is only possible with your insightful suggestions and faith in my abilities. I am incredibly grateful for your encouragement at some critical moments along the way.

To Dr. Vidya Muthukumar, I sincerely appreciate your guidance on my theoretical research. Your advice, experience, and patience have lit the light during my exploration.

I would like to thank Dr. Mark Davenport for agreeing to be on my thesis committee.I would like to thank Dr. Florian Schaefer for taking time out of his busy schedule to be on my committee. I am thankful to my committee member, Dr. Yao Xie, for providing constructive feedback on my thesis.

It is grateful to meet all kinds of intelligent and responsible colleagues at Georgia Tech. To Mehdi, thank you for being willing to collaborate with me when you first joined the lab, for many following projects, and for the fantastic tips in programming. To Ran, thank you for sharing your experience in life and research and thoughts on all kinds of problems. To Chiraag, thank you for your valuable time discussing and exploring the research topic with me. To all other colleagues, I am grateful to have you with me during my Ph.D. Our communications provide invaluable insights for me to complete my thesis.

Lastly, but most importantly, I am genuinely grateful to my family. To my family, thank you for always having faith in me, and giving me hope and strength to accomplish my degree. To my father and mother, there will never be enough words to describe how blessed I am to be your son.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Augmentations and other transformations of data, either in the input or latent space, are a critical component of modern machine learning systems. While these techniques are widely used in practice and known to provide improved generalization in many cases, it is still unclear how data manipulation impacts learning and generalization. To take a step toward addressing the problem, this thesis focuses on understanding and leveraging data augmentation and alignment for improving machine learning performance and transfer. In the first part of the thesis, we establish a novel theoretical framework to understand how data augmentation (DA) impacts learning in linear regression and classification tasks. The results demonstrate how the augmented transformed data spectrum plays a key role in characterizing the behavior of different augmentation strategies, especially in the overparameterized regime. The tools developed in this aim provide simple guidelines to build new augmentation strategies and a simple framework for comparing the generalization of different types of DA. In the second part of the thesis, we demonstrate how latent data alignment can be used to tackle the domain transfer problem, where training and testing datasets vary in distribution. Our algorithm builds upon joint clustering and data-matching through optimal transport, and outperforms the pure matching algorithm baselines in both synthetic and real datasets. Extension of the generalization analysis and algorithm design for data augmentation and alignment for nonlinear models such as artificial neural networks and random feature models are discussed. This thesis provides tools and analyses for better data manipulation design, which benefit both supervised and unsupervised learning schemes.

CHAPTER 1
INTRODUCTION

Data manipulation is a critical component in machine learning models. Although technologies have advanced model architectures significantly in the recent decade, the understanding of how data manipulation impacts machine learning is still relatively obscure. To take a step toward addressing the problem, this thesis focuses on understanding and leveraging data manipulation for machine learning. In the first part, we will analyze the popular data manipulation scheme, the data augmentation (DA) by establishing a simple framework to study DA's impacts on model generalization. In the second part, we demonstrate how sole data manipulation facilitates machine learning applications by developing a data alignment to tackle the domain transfer problem where training and testing datasets vary in distribution.

The first part of our thesis will be centered around understanding the DA's efficacy. Data augmentation (DA), or the transformation of data samples before or during learning, is quickly becoming a workhorse of both supervised [1, 2, 3] and self-supervised approaches [4, 5, 6, 7] for machine learning (ML). It is critical to the success of modern ML in multiple domains, e.g., computer vision [1], natural language processing [8], time series data [9], and neuroscience [10, 6, 11]. This is especially true in settings where data and/or labels are scarce or in other cases where algorithms are prone to overfitting [12]. While DA is perhaps one of the most widely used tools for regularization, most augmentations are often applied in an ad hoc manner, and it is often unclear exactly how, why, and when a DA strategy will work for a given dataset [13, 14]. Recent theoretical studies have provided insights into the effect of DA on learning and generalization when augmented samples lie close to the original data distribution [15, 16]. However, state-of-the-art augmentations that are used in practice (e.g. data masking [17], cutout [18], mixup [19]) are stochastic and can significantly alter the distribution of the data [20, 17, 21]. In this thesis, we address this challenge by proposing a simple yet flexible theoretical framework for comparing the linear model generalization of a broad class of augmentations. Our framework is simultaneously applicable to:

1. *General stochastic augmentations*, e.g. [16, 22, 18, 17],
2. The classical *underparameterized regime* [23] and the modern *overparameterized regime* [12, 24],
3. *Regression* [25, 26] and *classification tasks* [27, 28], and
4. *Strong* and *weak distributional-shift augmentations* [21].

To do this, we borrow and build on finite-sample analysis techniques of the modern overparameterized regime for linear and kernel models [25, 26, 27, 29]. Our theory reveals that DA induces implicit, training-data-dependent regularization of a twofold type: a) manipulation of the spectrum (i.e. eigenvalues) of the data covariance matrix, and b) the addition of explicit ℓ_2-type regularization to avoid noise overfitting. The first effect of spectral manipulation is often dominant in the overparameterized regime, and we show through several examples how it can either make or break generalization by introducing helpful or harmful biases. In contrast, the explicit ℓ_2 regularization effect always improves generalization by preventing possibly harmful overfitting of noise.

The second part of the study in the thesis is on the domain adaptation problem that leverages the distribution alignment technique called optimal transport (OT) [30, 31]. The idea is to find a map between the data points in both domains so that the model based on one domain can be applied to the other through the map, hence avoiding the model retrain. The method then boils down to finding the map, which can be found by solving a linear programming problem called the optimal transport problem that admits a fast solver [32]. However, ordinary OT has several drawbacks, including robustness issues where the map can be sensitive to outliers and noise and poor sampling rate where the map is inaccurate when the number of data is sparse. Our proposal thus develops a low-rank optimal transport method [33] that solves a robust mapping between two data distributions to remedy these issues. The proposed algorithm also has shown better interpretability and sampling complexity over the ordinary OT.

The rest of the thesis is organized as follows. The background is provided in Chapter 2. The analysis of generalization with DA is presented in Chapter 3. We build up a novel analysis of the DA through our studies and connect our theory to the others in the literature. We develop a robust domain adaptation technique with OT in Chapter 4, where we introduce the latent optimal transport (LOT). Then, we summarize our thesis in Chapter 5. Finally, in Chapter 5.2 we discuss the extensions for future investigations.

CHAPTER 2
BACKGROUNDS

In many machine learning problems, we train a model to predict a target based on the training data in common machine learning problems. The difference between the performance of a machine model on the training and testing data is called its generalizability. When the distribution shift occurs between the training and testing data, the outdated trained model will underperform dramatically, resulting in poor model generalization. For example, concept drift is a common scenario when either the covariates, target variables, or their distributions change through time [34, 35]. Without careful consideration of the distribution change, a machine learning algorithm will fail hard than expected. Therefore, coping with the distribution change in machine learning algorithmic design has become essential to closing the gap between theory and practice.

Conventional techniques to deal with the problem are called domain adaption or transfer learning. There are many great surveys on techniques of domain adaption and transfer learning [36, 37, 38, 39, 40, 41]. The seminal paper [37] categorize the transfer learning setting into inductive, transductive, and unsupervised, depending on the type of the concept shifts and the information of the labels. The most common trend of these methods is to learn common latent features aligning both domains by an artificial neural network (ANN). [42, 43] consider using the shallow network to learn a common space to match the data between the source and target. Other authors leverage deep networks to learn and transfer representations in the inner layers of the deep learning models. For instance, [44] extracts the convolution activation from a CNN to transfer the representations, outperforming traditional baselines. The paper [41] provides a comprehensive survey on the deep domain adaptation. A broader scope of strategies subsumes the deep domain adaptation is the architecture design with the imposition of explicit invariance, e.g., cyclic and shift-invariance. These methods [45, 46, 47, 48, 49, 50, 51, 52, 53] improve the generalizability and lead to many famous models like VGG-16 [45], or ResNet [46].

However, there are two major challenges in these approaches. 1. The domain alignment procedure in most domain adaptation methods can not always be done when the data in the target domain is sparse or lacking. 2. These methods often have high model complexity, e.g., complicated architecture design and multiple classifiers, resulting in increased cost of model training. In our proposal, we study two *data centered* approaches, each copes with one of the aforementioned challenges, respectively. The first method we study is data augmentation (DA), and the second method is domain adaptation with optimal transport (OT). Below we will review the background and the related works of them in a sequence.

2.1 Improving Generalization with Data Augmentation

Data augmentation (DA) [1, 9, 54, 55, 56] has become a standard practice to improve model generalization for both supervised, and unsupervised learning [57, 4] in modern machine learning schemes. It creates synthetic instances based on training data, including color transformation, random crops, and adversarial training with GAN. [1] is a perfect example summarizing the prevalent image data augmentations for deep learning in practice. A research direction strives to find the optimal augmentation among several candidates. [14] uses reinforcement learning to find the optimal composite augmentation. [58] proposes the AutoAugment to search for improved data augmentation policies. By treating the policy as hyperparameter search, [59] uses random search while [60] uses Bayesian optimization to search for the optimal augmentation with lower cost than the reinforcement learning methodology.

2.1.1 Theoretical understanding of data augmentation

Despite DA being simple and flexible, the investigations on when and what data augmentation leads to an effective boost in generalization are mainly conducted on empirical study and remained relatively elusive in theory. Without careful design, the literature has revealed that a non-suitable DA can even lead to negative impact [61, 62]. The issue becomes more problematic when one faces a less understood data domain. Therefore, DA has to be studied under a systematic mathematical framework in theory. In this regard, there are theoretical studies on DA. [63] analyzes the DA of adding Gaussian noise, [15] analyzes DA as feature averaging in a kernel classification task. [64] considers a composite of linear transformation of data and classifies common DA into label-invariant (e.g., dropout, adding of Gaussian noise, random flips) and label-mixing methods (e.g., various mixup augmentations [19, 65, 66]). They show that the label-preserving transformations improve estimation by enlarging the span of the training data while label-mixing transformations improve estimation by inducing a regularization effect. [63] considers the empirical risk minimization of a classification task and answers the question of the robustness improvement in terms of the increase of the margin. [16] uses group theory to show that DA leads to variance reduction by averaging the loss through a group orbit, which implies a generalization improvement. [15] adopts kernel theory to analyze the DA modeled by a Markov process and shows that kernel arises naturally and effect of DA can be approximated by first-order feature averaging and second-order variance regularization components. [67] studies the adversarial training with DA in a linear regression model and characterizes DA's effect on standard error (on unperturbed test inputs) and robust error (over worst-case perturbations). [68] analyzes the covariate shift data augmentation, e.g., DA that only changes the input but not the label) in linear regression, and characterize when the augmented data can lead to a decrease of generalization. Furthermore, they propose X-regularization, which uses unlabeled data to regularize the parameters towards the nonaugmented estimate. Inspired by manifold learning, [69] proposes a new Hessian-based complexity of neural networks and shows that data augmentation can reduce the metric. From an optimization landscape perspective, [70] shows that data augmentation can improve the optimization landscape of neural network training.

2.1.2 Benign overfitting with data augmentation

Although there are many theories on explanation of the benefits of DA in machine learning. The detailed analysis on how different DA and data spectrum affect the model generalization is still unknown. For example, when is a class of DA better than others for a specific data distribution? To unveil the mystery, we focus on the study the generalization of a linear regression model in the overparametrization regime, where the number of model parameters is larger than the number of training sample, in our preliminary work. We study on this model and regime because of the tractability and that modern machine learning algorithms usually have the number of parameters more than the number of samples. Furthermore, to separate the effect of DA from explicit regularization, we will analyze unregularized regression. Unlike the conventional analysis based on complexity generalization bound for underparametrized problem [71], we adopt the modern analysis focusing on the overparametrization regime. These works are summarized in the following. The line of works by [25, 26] strives for explaining the efficacy of the interpolation, i.e., fitting training examples without regularization, in the overparametrization regime that seemly contradicts the common wisdom in statistics that overfitting leads to poor generalization. They term the phenomenon benign overfitting. They study the interpolater and ridge estimators in linear regression and find that some proper spectrum of data covariance can lead to good generalization with overfitting. Following their works, [72] studies the condition of good generalization with constant step stochastic gradient descent in linear regression, which extends the study from a static to a dynamic point of view. These works assume a more general distribution model and consider the data covariance arbitrary, allowing for a full characterization of generalization between the interplay of data covariance and modeling parameters. Another promising analytical tool is the random matrix theory. Seminal works include [73, 74]. Unlike the finite analysis provided by works on benign overfitting, random matrix theory gears toward the asymptotic regime, where the ratio of the number of parameters and the sample approaches a fixed number. Although asymptotic, this approach's plus side is that it can often fully characterize the entire regime, whether it is over or underparametrized. In [73], the authors consider the generalization of a linear regression model and discover several interesting properties, including the occurrence of double-descent. [74] generalizes the model to a random feature model that approximates better to the neural network in practice and characterizes the generalization error asymptotically.

2.2 Domain Adaptation with Optimal Transport

Unlike the DA strategies that modify the training data to simulate possible changes in data distribution, our second focusing approach calibrates the model's prediction when a data distribution change occurs. As distribution shift can occur frequently, we want our calibration to be as cheap as possible. This brings us to the domain adaptation technique with optimal transport (OT) [30, 31]. The approach is to find a map between the data in both domains so that we can apply the trained model on the source domain to the data in the target domain through the map. In this way, the method boils down to finding the map between the data domains, and this is where OT comes into play. Although OT provides an off-the-shelf solution for our need, in our preliminary work [33], we find that it is not robust to various transformations, outliers, and noise. This becomes a severe issue when the testing data set is heavily shifted from the training data set. Hence, we develop a robust distribution matching methodology to improve the ordinary OT in our preliminary work.

In the following, we will introduce our proposed method, the Latent Optimal Transport (LOT), a robust distribution alignment technique. Before the introduction, we will first review the related works and backgrounds of optimal transport.

2.2.1 Related works and background of optimal transport

Optimal transport (OT) [75] is a widely used technique for distribution alignment that learns a *transport plan* which moves mass from one distribution to match another. With recent advances in tools for regularizing and speeding up OT [76], this approach has found applications in many diverse areas of machine learning, including domain adaptation [30, 31], generative modeling [77, 78], document retrieval [79], computer graphics [80, 81, 82], and computational neuroscience [83, 84].

While the ground metric in OT measures detailed variance in data points, it could be fragile to outliers or noise, especially in high dimensions. To overcome this issue, additional cluster/class structure can improve alignment or make transport more robust. Examples of methods that incorporate additional structure into OT include approaches that leverage hierarchical structure or cluster consistency [84, 85, 86], partial class information [31, 30], submodular cost functions [87], and low-rank constraints on the transport plan [88, 89]. Because of the difficulty of incorporating structure into OT, many of these methods need low-dimensional structure in data to be specified in advance (e.g., estimated clusters or labels).

To simultaneously learn the low-dimensional structure and use it to constrain transport, [88, 33, 90] recently introduced low-rank OT that builds a factorization of the transport. The transport rank has a natural interpretation of being the number of clusters/classes of the data. The transports admit fast computations by combining the k-means clustering and iterative Bregman projections. Besides the statistical benefits on the sampling rate studied in [88], [33] has shown that low-rank OT also exhibits robustness against several data transformations, outliers, and noise.

2.2.2 Latent optimal transport

Now we introduce our algorithm Latent Optimal Transport, which belongs to the class of low-rank transport approaches. Most datasets have a low-dimensional latent structure, but OT does not naturally use it during transport. This motivates the idea that distribution alignment methods should both *reveal* the latent structure in the data in addition to aligning these latent structures. Specifically, we assume there exist latent points called anchors representing the source and target data. The number of anchors defines the rank of the transportation, which can also represent the number of clusters in each data domain. LOT then transports the data from the source to the target through these anchors. Because these anchors restrict the way of transportation, the alignment between data is less sensitive to the effects of outliers and various data transformations.

CHAPTER 3

UNDERSTANDING DATA AUGMENTATION IN LINEAR MODELS

3.1 Introduction

Data augmentation (DA), or the transformation of data samples before or during learning, is quickly becoming a workhorse of both supervised [1, 2, 3] and self-supervised approaches [4, 5, 6, 7] for machine learning (ML). It is critical to the success of modern ML in multiple domains, e.g., computer vision [1], natural language processing [8], time series data [9], and neuroscience [10, 6, 11]. This is especially true in settings where data and/or labels are scarce or in other cases where algorithms are prone to overfitting [12]. While DA is perhaps one of the most widely used tools for regularization, most augmentations are often applied in an ad hoc manner, and it is often unclear exactly how, why, and when a DA strategy will work for a given dataset [13, 14].

Recent theoretical studies have provided insights into the effect of DA on learning and generalization when augmented samples lie close to the original data distribution [15, 16]. However, state-of-the-art augmentations that are used in practice (e.g. data masking [17], cutout [18], mixup [19]) are stochastic and can significantly alter the distribution of the data [20, 17, 21]. Despite many efforts to explain the success of DA in the literature [91, 92, 16, 15, 64], there is still a lack of a comprehensive platform to compare different types of augmentations at a quantitative level.

In this paper, we address this challenge by proposing a simple yet flexible theoretical framework for comparing the linear model generalization of a broad class of augmentations. Our framework is simultaneously applicable to: 1. *general stochastic augmentations*, e.g. [16, 22, 18, 17], 2. the classical *underparameterized regime* [23] and the modern *overparameterized regime* [12, 24], 3. *regression* [25, 26] and *classification tasks* [27, 28], and 4. *strong* and *weak distributional-shift augmentations* [21]. To do this, we borrow and build on finite-sample analysis techniques of the modern overparameterized regime for linear and kernel models [25, 26, 27, 29]. Our theory reveals that DA induces implicit, training-data-dependent regularization of a twofold type: a) manipulation of the spectrum (i.e. eigenvalues) of the data covariance matrix, and b) the addition of explicit ℓ_2-type regularization to avoid noise overfitting.

The first effect of spectral manipulation is often dominant in the overparameterized regime, and we show through several examples how it can either make or break generalization by introducing helpful or harmful biases. In contrast, the explicit ℓ_2 regularization effect always improves generalization by preventing possibly harmful overfitting of noise.

3.1.1 Main contributions

Below, we outline and provide a roadmap of the main contributions of the first part of this thesis.

- We propose a new framework for studying generalization with data augmentation for linear models by building on the recent literature on the theory of overparameterized learning [25, 93, 73, 29, 27, 28]. We provide natural definitions of the augmentation mean and covariance operators that capture the impact of change in data distribution on model generalization in Section 3.3.1, and sharply characterize the ensuing performance for both regression and classification tasks in Sections 3.4.3 and 3.4.4, respectively.

- In Section 3.5.1, we apply our theory to provide novel and surprising interpretations of a broad class of randomized DA strategies used in practice; e.g., random-masking [17], cutout [18], noise injection [91], and group-invariant augmentations [16]. An example is as follows: while the classical noise injection augmentation [91] causes only a constant shift in the spectrum, data masking [17, 22], cutout [18] and distribution-preserving augmentations [16] tend to *isotropize* the equivalent data spectrum. This isotropizing effect, as we discuss in Section 3.5.2, can be shown to create an especially high bias and therefore, harm generalization in the overparameterized regime.

- In Section 3.5.3, we directly compare the impact of DA on the downstream tasks of regression and classification and identify strikingly different behaviors. Specifically, we find that, while augmentation bias is mostly harmful in a regression task, its effect can be minimal for classification. This together with the uniform variance improvement can be shown to yield several helpful scenarios for classification. This is consistent with the fact that the empirical benefits of strong augmentation have been observed primarily in classification tasks [21, 94].

- Our framework serves as a testbed for new DA approaches. As a proof-of-concept, in Section 3.5.2, we design a new augmentation method, inspired by isometries in random feature rotation, that can provably achieve smaller bias than the least-squared estimator and variance reduction on the order of the ridge estimator. Moreover, this generalization is *robust* in the sense that it compares favorably with optimally tuned ridge regression for a much wider range of hyperparameters).

- Finally, in Section 3.6 we complement and verify our theoretical insights through a number of empirical studies that examine how multiple factors involving data, model and augmentation type impact generalization. We compare our closed-form expression with augmented SGD [15, 16, 4] and pre-computed augmentations [64, 95] In contrast to augmented SGD, we find that adding more pre-computed augmentations can increase overfitting to noise, thus producing “interpolation peaks” in the sense of [24].

Notation

We use n to denote the number of training examples and p to denote the data dimension. Given a training data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ where each row (representing a training example) is independently and identically distributed (i.i.d.) and has covariance $\mathbf{\Sigma} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, we denote $\mathbf{P}^{\mathbf{\Sigma}}_{1:k-1}$ and $\mathbf{P}^{\mathbf{\Sigma}}_{k:\infty}$ as the projection matrices to the top $k-1$ and the bottom $p-k+1$ eigen-subspaces of $\mathbf{\Sigma}$, respectively. For convenience, we denote the residual Gram matrix by $\mathcal{A}_k(\mathbf{X};\lambda) = \lambda \mathbf{I}_n + \mathbf{X}\boldsymbol{P}^{\mathbf{\Sigma}}_{k:\infty}\mathbf{X}^T$, where λ is some regularization constant. Subscripts denote the subsets of column vectors when applied to a matrix; e.g. for a matrix $\mathbf{V}$ we have $\mathbf{V}_{a:b} := [\mathbf{v}_a, \mathbf{v}_{a+1}, \dots, \mathbf{v}_b]$. A similar definition applies to vectors; e.g. for a vector $\mathbf{x}$ we have $\mathbf{x}_{a:b} = [\mathbf{x}_a, \mathbf{x}_{a+1}, \dots, \mathbf{x}_b]$. The Mahalanobis norm of a vector is defined by $\|\mathbf{x}\|_{\mathbf{H}} = \sqrt{\mathbf{x}^\top \mathbf{H}\mathbf{x}}$. For a matrix $\mathbf{A}$, $\operatorname{diag}(\mathbf{A})$ denotes the diagonal matrix with a diagonal equal to that of $\mathbf{A}$, $\operatorname{Tr}(\mathbf{A})$ denotes its trace and $\mu_i(\mathbf{A})$ its i-th largest eigenvalue. The symbols $\gtrsim$ and $\lesssim$ are used to denote inequality relations that hold up to universal constants that do not depend on n or p. All asymptotic convergence results are stated in probability.

More specific notation corresponding to our signal model is given in Section 3.4.1, and some additional notation that is convenient to define for our analysis is postponed to Section 3.4.2.

3.2 Related work

We organize our discussion of related work into two verticals: a) historical and recent perspectives on the role of data augmentation, and b) recent analyses of minimum-norm and ridge estimators in the over-parameterized regime.

3.2.1 Data augmentation

Classical links between DA and regularization: Early analysis of DA showed that adding random Gaussian noise to data points is equivalent to Tikhonov regularization [91] and *vicinal risk minimization* [19, 92]; in the latter, a local distribution is defined in the neighborhood of each training sample, and new samples are drawn from these local distributions to be used during training. These results established an early link between augmentation and explicit regularization. However, the impact of such approaches on generalization has been mostly studied in the underparameterized regime of ML, where the primary concern is reducing variance and avoiding overfitting of noise. Modern ML practices, by contrast, have achieved great empirical success in overparameterized settings and with a broader range of augmentation strategies [1, 2, 3]. The type of regularization that is induced by these more general augmentation strategies is not well understood. Our work provides a systematic point of view to study this general connection without assuming any additional explicit regularization, or specific operating regime.

In-distribution versus out-of-distribution augmentations: Intuitively, if we could design an augmentation that would produce more virtual but identically distributed samples of our data, we would expect an improvement in generalization. Based on this insight and the inherent structure of many augmentations used in vision (that have symmetries), another set of works explores the common intuition that data augmentation helps insert beneficial group-invariances into the learning process [52, 96, 97, 98, 99]. These studies generally consider cases in which the group structure is explicitly present in the model design via convolutional architectures [52, 98] or feature maps approximating group-invariant kernels [96, 97]. The authors of [16] propose a general group-theoretic framework for DA and explain that an averaging effect helps the model generalize through variance reduction. However, they only consider augmentations that do not alter (or alter by minimal amounts) the original data distribution; consequently, they identify variance reduction as a sole positive effect of DA. Moreover, their analysis applies primarily to underparameterized or explicitly regularized models[1].

Recent empirical studies have highlighted the importance of diverse stochastic augmentations [20]. They argue that in many cases, it is important to introduce samples which are *out-of-distribution* (OOD) [102, 103] (in the sense that they do not resemble the original data). In our framework, we allow for cases in which augmentation leads to significant changes in distribution and provide a path to analysis for such OOD augmentations that encompass empirically popular approaches for DA [17, 18]. We also consider the modern overparameterized regime [24, 104]. We show that the effects of OOD augmentations go far beyond variance reduction, and the spectral manipulation effect introduces interesting biases that can either improve or worsen generalization for overparameterized models.

Analysis of specific types of DA in linear and kernel methods: [15] propose a Markov process-based framework to model compositional DA and demonstrate an asymptotic connection between a Bayes-optimal classifier and a kernel classifier dependent on DA. Furthermore, they study the

[1] More recent studies of invariant kernel methods, trained to interpolation, suggest that invariance could either improve [100] or worsen [101] generalization depending on the precise setting. Our results for the overparameterized linear model (in particular, Corollary 17) also support this message.

augmented empirical risk minimization procedure and show that some types of DA, implemented in this way, induce approximate data-dependent regularization. However, unlike our work, they do not quantitatively study the generalization of these classifiers. [105] also propose a kernel classifier based on a notion of invariance to local translations, which produces competitive empirical performance. In another recent analysis, [64] study the generalization of linear models with DA that constitutes *linear transformations* on the data for regression in the overparameterized regime (but still considering additional explicit regularization). They find that data augmentation can enlarge the span of training data and induce regularization. There are several key differences between their framework and ours. First, they analyze deterministic DA, while we analyze stochastic augmentations used in practice [5, 16]. Second, they assume that the augmentations would not change the labels generated by the ground-truth model, thereby only identifying beneficial scenarios for DA (while we identify scenarios that are both helpful and harmful). Third, they study empirical risk minimization with pre-computed augmentations, in contrast to our study of augmentations applied *on-the-fly* during the optimization process [15, 16], which are arguably more commonly used in practice. Our experiments in Section 3.6.4 identify sizably different impacts of these methods of application of DA even in simple linear models. Finally, the role of DA in linear model optimization, rather than generalization, has also been recently studied; in particular, [106] characterize how DA affects the convergence rate of optimization.

The impact of DA on nonlinear models: Recent works aim to to understand the role of DA in nonlinear models such as neural networks. [107] show that certain local augmentations induce regularization in deep networks via a “rugosity”, or “roughness” complexity measure. While they show empirically that DA reduces rugosity, they leave open the question of whether this alone is an appropriate measure of a model’s generalization capability. Very recently, [95] showed that training a two-layer convolutional neural network with a specific permutation-style augmentation can have a novel *feature manipulation* effect. Assuming the recently posited “multi-view" signal model [108], they show that this permutation-style DA enables the model to better learn the essential feature for a classification task. They also observe that the benefit becomes more pronounced for nonlinear models. Our work provides a similar message, as we also identify the DA-induced data manipulation effect as key to generalization. Because our focus in this work is limited to linear models, the effect of data manipulation manifests itself purely through *spectral regularization* of the data covariance. As a result of this spectral-regularization effect, we are also able to provide a comprehensive general-purpose framework for DA by which we can compare and contrast different augmentations that can either help or hurt generalization (while [95] only analyze a permutation-style augmentation). We believe that combining our general-purpose framework for DA with a more complex nonlinear model analysis is a promising future direction, and we discuss possible analysis paths for this in Section 3.7.

3.2.2 Interpolation and regularization in overparameterized models

Minimum-norm-interpolation analysis: Our technical approach leverages recent results in overparameterized linear regression, where models are allowed to interpolate the training data. Following the definition of [104], we characterize such works by their explicit focus on models that achieve close to zero training loss and which have a high complexity relative to the number of training samples. Specifically, many of these works provide finite sample analysis of the risk of the least squared estimator (LSE) and the ridge estimator [25, 26, 73, 93, 29]. This line of research (most notably, [25, 26]) finds that the mean squared error (MSE), comprising the bias and variance, can be characterized in terms of the effective ranks of the spectrum of the data distribution. The main insight is that, contrary to traditional wisdom, perfect interpolation of the data may not have a harmful effect on the generalization error in highly overparameterized models. In the context of these advances, we identify the principal impact of DA as *spectral manipulation* which directly modifies the effective ranks, thus either improving or worsening generalization. We build in particular on the work of [26], who provide non-asymptotic characterizations of generalization error for general sub-Gaussian design, with some additional technical assumptions that also carry over to our framework[2].

Subsequently, this type of “harmless interpolation” was shown to occur for classification tasks [27, 110, 28, 111, 112, 113, 114]. In particular, [27, 112] showed that classification can be significantly easier than regression due to the relative benignness of the 0-1 test loss. Our analysis also compares classification and regression and shows that the potentially harmful biases generated by DA are frequently nullified with the 0-1 metric. As a result, we identify several beneficial scenarios for DA in classification tasks. At a technical level, we generalize the analysis of [27] to sub-Gaussian design. We also believe that our framework can be combined with the alternative mixture model (where covariates are generated from discrete labels [111, 28, 110]), but we do not formally explore this path in this paper.

Generalized ℓ_2 regularizer analysis: Our framework extends the analyses of least squares and ridge regression to estimators with general Tikhonov regularization, i.e., a penalty of the form $\theta^\top \mathbf{M}\theta$ for arbitrary positive definite matrix $\mathbf{M}$. A closely related work is [115], which analyzes the regression generalization error of general Tikhonov regularization. However, our work differs from theirs in three key respects. First, the analysis of [115] is based on the proportional asymptotic limit (where the sample size n and data dimension p increase proportionally with a fixed ratio) and provides sharp asymptotic formulas for regression error that are exact, but not closed-form and not easily interpretable. On the other hand, our framework is non-asymptotic, and we generally consider $p \gg n$ or $p \ll n$; our expressions are closed-form, match up to universal constants and are easily interpretable. Second, our analysis allows for a more general class of *random* regularizers that themselves depend on the training data; a key technical innovation involves showing that the additional effect of this randomness is, in fact, minimal. Third, we do not explicitly consider the problem of determining an optimal regularizer; instead, we compare and contrast the generalization characteristics of various types of practical augmentations and discuss which characteristics lead to favorable performance.

[2] As remarked on at various points throughout the paper, we believe that the subsequent and very recent work of [109], which weakens these assumptions further, can also be plugged with our analysis framework; we will explore this in the sequel.

In addition to explicitly regularized estimators, [115] also analyze the ridgeless limit for these regularizers, which can be interpreted as the minimum-Mahalanobis-norm interpolator. In Section 3.6.1 we show that such estimators can also be realized in the limit of minimal DA.

The role of explicit regularization and hyperparameter tuning: Research on harmless interpolation and double descent [24] has challenged conventional thinking about regularization and overfitting for overparameterized models; in particular, good performance can be achieved with weak (or even negative) explicit regularization [116, 26], and gradient descent trained to interpolation can sometimes beat ridge regression [117]. These results show that the scale of the ridge regularization significantly affects model generalization; consequently, recent work strives to estimate the optimal scale of ridge regularization using cross-validation techniques [118, 119].

As shown in classical work [91], ridge regularization is equivalent to augmentation with (isotropic) Gaussian noise, and the scale of regularization naturally maps to the variance of Gaussian noise augmentation. Our work links DA to a much more flexible class of regularizers and shows that some types of DA induce an implicit regularization that yields much more robust performance across the hyperparameter(s) dictating the “strength” of the augmentation. In particular, our experiments in Section 3.6.2 show that random mask [17], cutout [18] and our new random rotation augmentation yield comparable generalization error for a wide range of hyperparameters (masking probability, cutout width and rotation angle respectively); the random rotation is a new augmentation proposed in this work and frequently beats ridge regularization as well as interpolation. Thus, our flexible framework enables the discovery of DA with appealing robustness properties not present in the more basic methodology of ridge regularization.

Other types of indirect regularization: We also mention peripherally related but important work on other types of indirect regularization involving *creating fake “knockoff” features* [120, 121] and *dropout in parameter space* [122, 123]. The knockoff methodology creates copies of *features* (rather than augmenting data points) that are uncorrelated with the target to perform variable selection. Dropout also induces implicit regularization by randomly dropping out intermediate neurons (rather than covariates, as does the random mask [17] augmentation) during the learning process, and has been shown to have a close connection with sparsity regularization [123]. Overall, these constitute methods of indirect regularization that are applied to model parameters rather than data. An intriguing question for future work is whether these effects can also be achieved through DA.

3.3 Problem Setup

In this section, we introduce the notation and setup for our analysis of generalization with data augmentation (DA). We review the fundamentals of empirical risk minimization (ERM) without DA and how augmentations affect the ERM procedure. Then, we derive a reduction to ridge regression that paves the way for our analysis in Section 3.4.

3.3.1 Empirical risk minimization with data augmentation

Modern, high-dimensional ML models are commonly trained to minimize a combination of a) prediction error on training data, and b) a measure of model complexity that favors “simpler" or “smaller" models. This is encapsulated in the *regularized empirical risk minimization objective*, expressed for linear models $f_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\theta} \rangle$ as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ \ell(\mathbf{X}\boldsymbol{\theta}, \mathbf{y}) + R(\boldsymbol{\theta}), \tag{3.1}$$

where ℓ is a loss function, $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n \end{bmatrix}^\top \in \mathbb{R}^{n \times p}$ is the training data matrix that stacks the n covariates, $\mathbf{y} \in \mathbb{R}^n$ is the vector of observations/responses, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the linear model parameter that we want to optimize, and $R(\boldsymbol{\theta})$ is an explicit regularizer applied to the model. For example, the popular *ridge regression* procedure uses $R(\boldsymbol{\theta}; \lambda) = \lambda \|\boldsymbol{\theta}\|_2^2$, where λ is a tunable hyperparameter. We will adopt the choice of *squared loss function* $\ell(\mathbf{X}\boldsymbol{\theta}, \mathbf{y}) = \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$ throughout this work, owing to its mathematical tractability and recently observed competitiveness with the cross-entropy loss even in classification tasks [124, 27, 28, 111].

Although the training objective of modern supervised ML models rarely includes explicit regularization of the form (3.1) in practice, it does heavily rely on data augmentation (DA) to achieve state-of-the-art performance [1, 12]. Mathematically speaking, an augmentation $g : \mathbb{R}^p \to \mathbb{R}^p$ is a general mapping from the original data point $\mathbf{x}$ to a transformed data point $g(\mathbf{x})$. In practice, an augmentation function g is often stochastic and drawn at random from an augmentation distribution denoted by $\mathcal{G}$. Each time we augment the data, we randomly draw an instance of $g \sim \mathcal{G}$. For example, the classical *Gaussian noise injection* augmentation [91] is stochastic and takes the form $g(\mathbf{x}) = \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is an isotropic Gaussian random variable.

One approach to implement augmentations is to pre-compute augmented data samples by drawing a fixed number of augmentations before training and then including them along with the original data points when training the model [64, 95]. Nowadays, it is more popular to apply augmentations on the fly during training [16, 5], with different transformations applied stochastically throughout the training procedure. This procedure, typically called *augmented stochastic gradient descent (aSGD)*, is widely used in practice [4, 16]. [16] showed that this algorithm can be viewed as applying SGD to the objective of an *augmented empirical risk minimization (aERM)* problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_G \left[\|G(\mathbf{X})\boldsymbol{\theta} - \mathbf{y}\|_2^2 \right] . \tag{3.2}$$

Above, G denotes a stacked data augmentation function applied to each row of the matrix, i.e., $G(\mathbf{X}) = [g_1(\mathbf{x}_1) \dots, g_n(\mathbf{x}_n)]^T$; we assume that the transformations g_i are stochastic and are drawn i.i.d. from an augmentation function distribution $\mathcal{G}$. We would expect aSGD to converge to the solution of (3.2), and conduct experiments to empirically verify this in Section 3.6.

We begin by defining the first and second-order statistics of an augmentation distribution. We will show that these quantities play a key role in characterizing the solution to the aERM problem.

Definition 1 (**Augmentation Mean and Covariance Operator**). *Consider a stochastic augmentation* $\mathbf{x} \mapsto g(\mathbf{x})$*, where* g *is drawn randomly from an augmentation distribution* $\mathcal{G}$*. We then define* *the augmentation mean and the covariance for a single data point* $\mathbf{x}$ *as*

$$\mu_{\mathcal{G}}(\mathbf{x}) := \mathbb{E}_{g\sim\mathcal{G}}[g(\mathbf{x})], \;\; \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) := \mathbb{E}_{g\sim\mathcal{G}}\left[(g(\mathbf{x}) - \mu_{\mathcal{G}}(\mathbf{x}))\,(g(\mathbf{x}) - \mu_{\mathcal{G}}(\mathbf{x}))^{\top}\right], \tag{3.3}$$

where we use the subscript $\mathcal{G}$ *to emphasize that the expectation is only over the randomness of the augmentation function* g*. Furthermore, for a training data set* $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix}^{\top}$*, we similarly define the augmentation mean and covariance operators with respect to the data set as:*

$$\mu_{\mathcal{G}}(\mathbf{X}) := [\mu_{\mathcal{G}}(\mathbf{x}_1), \mu_{\mathcal{G}}(\mathbf{x}_2), \dots, \mu_{\mathcal{G}}(\mathbf{x}_n)]^{\top}, \;\; \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) := \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}_i). \tag{3.4}$$

Finally, we call an augmentation distribution ***unbiased on average***[3] *if* $\mu_{\mathcal{G}}(\mathbf{x}) = \mathbf{x}$.

With this notation introduced, we now explain why DA gives rise to implicit regularization. For now we consider augmentation distributions that are unbiased on average for conceptual simplicity and leave the extension to distributions that are biased on average to Section 3.4.3. For such unbiased-on-average augmentation distributions, we can simply the objective (3.2) as:

$$\begin{aligned}
\mathbb{E}_{G}[\|G(\mathbf{X})\boldsymbol{\theta} - \mathbf{y})\|_2^2] &= \mathbb{E}_{G}[\|\,(G(\mathbf{X}) - \mu(\mathbf{X}))\,\boldsymbol{\theta} + \mu(\mathbf{X})\boldsymbol{\theta} - \mathbf{y})\|_2^2] \\
&= \|\mu(\mathbf{X})\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \|\boldsymbol{\theta}\|_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})}^2 &(3.5)\\
&= \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \|\boldsymbol{\theta}\|_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})}^2, &(3.6)
\end{aligned}$$

where the last two steps used the assumption that the augmentation distribution is unbiased on average. From this expression, it is clear that DA produces an *implicit*, *data-dependent* regularization $\|\boldsymbol{\theta}\|_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})}^2$, defined by the augmentation covariance we just introduced. The heart of our analysis is a detailed investigation of the implications of this data-dependent regularization on generalization.

[3] Note that this definition of bias is completely different from the bias-variance decomposition that manifests in regression analysis, i.e., (3.12).

3.3.2 Implications of a DA-induced regularizer and connections to ridge regression

In this section, we unpack the effects of the DA-induced regularizer $\|\boldsymbol{\theta}\|^2_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})}$. In general, we note that the objective (3.5) can be viewed as a general Tikhonov regularization problem with a possibly data-dependent regularizer matrix. Using this observation, we will show that this creates the effects of (i) ℓ_2 *regularization* (i.e. Tikhonov regularization with an identity regularizer matrix) and (ii) *data spectrum modification.*

The first step is to explicitly connect the solution to a ridge regression estimator. Since our focus is on stochastic augmentations, we assume that $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) \succ 0$. Then, the objective (3.5) admits a closed-form solution given by

$$\hat{\boldsymbol{\theta}}_{\text{aug}} = (\mathbf{X}^\top\mathbf{X} + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mathbf{X}^\top\mathbf{y}. \tag{3.7}$$

We now use (3.7) to link the estimator $\hat{\boldsymbol{\theta}}_{\text{aug}}$ to a ridge estimator by derivation below. For ease of exposition, we suppress the dependency of $\mathrm{Cov}_{\mathcal{G}}$ on the training data matrix $\mathbf{X}$.

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{aug}} &= (\mathbf{X}^\top\mathbf{X} + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mathbf{X})^\top\mathbf{y} \\
&= \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}(n\mathbf{I}_p + \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\mathbf{X}^\top\mathbf{X}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2})^{-1}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\mathbf{X}^\top\mathbf{y} \\
&= \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}(n\mathbf{I}_p + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y} \quad (\text{where } \widetilde{\mathbf{X}} := \mathbf{X}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}) \\
&= \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\hat{\theta}_{\text{ridge}}, \quad \text{where } \hat{\boldsymbol{\theta}}_{\text{ridge}} := (n\mathbf{I}_p + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y}.
\end{aligned} \tag{3.8}$$

Recall that $\boldsymbol{\Sigma} := \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]$ denotes the original data covariance. Then, it is easy to see that the MSE $\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$ is equivalent to $\|\hat{\boldsymbol{\theta}}_{\text{ridge}} - \mathrm{Cov}_{\mathcal{G}}{}^{1/2}\boldsymbol{\theta}^*\|^2_{\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\boldsymbol{\Sigma}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}}$. Suppose, for a moment, that $\mathrm{Cov}_{\mathcal{G}}$ were fixed (or independent of $\mathbf{X}$). Then, (3.8) demonstrates an equivalence between the solution of aERM and a ridge estimator with data matrix $\widetilde{\mathbf{X}}$, data covariance $\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\boldsymbol{\Sigma}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}$, ridge parameter[4] $\lambda = n$, and true model $\mathrm{Cov}_{\mathcal{G}}{}^{1/2}\boldsymbol{\theta}^*$ (in the sense that both solutions achieve the same MSE). Therefore, in terms of generalization, we can view DA as inducing a two-fold effect: a)

[4] This demonstrates that *negative* regularization, which is studied in some recent work [26, 116] is not possible to achieve through the DA framework.

ℓ_2 regularization at a scale that is proportional to the number of training samples ($\lambda_{\text{reg}} = n$), and b) a modification of the original data covariance from $\mathbf{\Sigma}$ to $\text{Cov}_{\mathcal{G}}{}^{-1/2}\mathbf{\Sigma}\text{Cov}_{\mathcal{G}}{}^{-1/2}$, which can sizably change its *spectrum* (i.e. vector of eigenvalues in decreasing order).

It is important to note that this equivalence between solutions is only approximate since $\text{Cov}_{\mathcal{G}}$ itself depends on $\mathbf{X}$. We will justify and formalize this approximation in Section 3.4.2.

3.3.3 Practically used augmentations

Our framework can accommodate a number of different transformations and augmentations that are used in practice, as long as they are only applied to covariates and not labels. In Table 3.1, we list some common augmentations for which the closed-form expression for the solution to the aERM objective is easily calculatable and interpretable.

Table 3.1: **Examples of common augmentations for which we can compute an interpretable closed-form solution to the aERM objective.**

	Augmentation function: $g(\mathbf{x})$	Covariance operator: $\text{Cov}_{\mathcal{G}}(\mathbf{X})$
Gaussian noise injection	$\mathbf{x} + \mathbf{n}, \mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$	$\sigma^2\mathbf{I}$
Correlated noise injection	$\mathbf{x} + \mathbf{n}, \mathbf{n} \sim \mathcal{N}(0, \mathbf{W})$	$\mathbf{W}$
Unbiased random mask	$\mathbf{b} \odot \mathbf{x}, \mathbf{b}_i \sim \text{Bernoulli}(1-\beta)$	$\frac{\beta}{1-\beta}\frac{1}{n}\text{diag}(\mathbf{X}^\top\mathbf{X})$
Pepper noise injection	$\mathbf{b} \odot \mathbf{x} + (\mathbf{1} - \mathbf{b}) \odot \mathcal{N}(0, \sigma^2)$	$\frac{\beta}{1-\beta}\ \frac{1}{n}\text{diag}\left(\mathbf{X}^\top\mathbf{X}\right) + \frac{\beta\sigma^2}{(1-\beta)^2}\mathbf{I}$
Random Cutout	zero-out k consecutive features	$\frac{p}{p-k}\frac{1}{n}\mathbf{M} \odot \mathbf{X}^\top\mathbf{X}$

Note that, in general, any regularization of the form $\|\boldsymbol{\theta}\|^2_{A(\mathbf{X})}$, where $A(\mathbf{X})$ is some positive semi-definite matrix dependent on $\mathbf{X}$, can be achieved by a simple additive correlated Gaussian noise augmentation where $g(\mathbf{X}) = \mathbf{X} + \mathbf{N}$, $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A}(\mathbf{X}))$. Our focus in this paper is on popular interpretable augmentations used in practice.

3.3.4 Novel augmentation design

Our framework can also serve as a testbed for designing new augmentations. As an example, we introduce a novel augmentation that performs multiple rotations in random planes. Specifically, for an input $\mathbf{x} \in \mathbb{R}^p$, we perform the following steps:

1. Pick an orthonormal basis $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p]$ for the entire p-dimensional space uniformly at random, i.e. from the Haar measure.
2. Divide the basis into sets of $\frac{p}{2}$ orthogonal planes $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{\frac{p}{2}}$, where $\mathbf{U}_i = [\mathbf{u}_{2i-1}, \mathbf{u}_{2i}]$.
3. Rotate $\mathbf{x}$ by an angle α in each of these planes $\mathbf{U}_i$, $i = 1, 2, \ldots, \frac{p}{2}$.

Ultimately, the augmentation mapping is given by

$$\begin{aligned} g(\mathbf{x}) &= \prod_{i=1}^{\frac{p}{2}} \left[\mathbf{I} + \sin\alpha(\mathbf{u}_{2i-1}\mathbf{u}_{2i}^\top - \mathbf{u}_{2i}\mathbf{u}_{2i-1}^\top) + (\cos\alpha - 1)(\mathbf{u}_{2i}\mathbf{u}_{2i}^\top + \mathbf{u}_{2i-1}\mathbf{u}_{2i-1}^\top)\right]\mathbf{x} \\ &= \left[\mathbf{I} + \sum_{i=1}^{\frac{p}{2}} \sin\alpha(\mathbf{u}_{2i-1}\mathbf{u}_{2i}^\top - \mathbf{u}_{2i}\mathbf{u}_{2i-1}^\top) + (\cos\alpha - 1)(\mathbf{u}_{2i}\mathbf{u}_{2i}^\top + \mathbf{u}_{2i-1}\mathbf{u}_{2i-1}^\top)\right]\mathbf{x}. \end{aligned}$$

The induced augmentation covariance is given by $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \frac{4(1-\cos\alpha)}{np}\left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} - \mathbf{X}^\top\mathbf{X}\right)$ (full derivation in Section 3.8.5). Intuitively, this augmentation is composed of several local data transformations that change the data spectrum in a mild way. We quantify its performance in Corollary 19 and demonstrate that it performs favorably compared to optimally-tuned ridge regression while being far more robust to hyperparameter choice, i.e. the value of the angle α.

3.4 Main Results

This section presents our meta theorems for the generalization performance of regression and classification tasks. We consider estimators for augmentations which are unbiased-in-average and biased-in-average separately, as they exhibit significant differences in terms of generalization. The applications of the general theorem will be discussed in detail in Section 3.5. Table 3.2 provides the road map of our main results and their applications in this and the next sections.

3.4.1 Preliminaries

Recall that $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the training data matrix with n i.i.d. rows comprising of the training data. Each data point $\mathbf{x} \in \mathbb{R}^p$ can be written as $\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{z}$, where we assume, without loss of generality, that $\mathbf{\Sigma}$ is a diagonal matrix with non-negative diagonal elements $\lambda_1 \geq \lambda_2, \cdots \geq \lambda_p$, and $\mathbf{z}$ is a latent vector which is zero-mean, isotropic (i.e., $\mathbb{E}[\mathbf{z}] = 0$, $\mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right] = \mathbf{I}$), and sub-Gaussian with sub-Gaussian norm σ_z. (Note that the assumption of diagonal covariance $\mathbf{\Sigma}$ is without loss of generality because sub-Gaussianity is preserved under any unitary transformation; however, the covariance induced by DA will frequently not remain diagonal).

Our analysis applies across the classical underparameterized regime ($n \geq p$) and the modern overparameterized regime ($p > n$); much of our discussion of consequences of DA will be centered on the latter regime. We assume the true data generating model to be $y = \mathbf{x}^T\boldsymbol{\theta}^* + \varepsilon$, where ε denotes the noise, which is also isotropic and sub-Gaussian with sub-Gaussian norm σ_ε and variance σ^2. We believe that our non-asymptotic framework can be extended to more general kernel settings as in the recent work of [109], where features are not assumed to be sub-Gaussian, but we leave this extension to future work.

Error Metrics

In this work, we will focus on the squared loss training objective (3.2) for both regression and classification tasks. While we make this choice for relative mathematical tractability, we note that it is well-justified in practice as recent work [124, 27, 28, 111] has shown that the squared loss can achieve competitive results when compared with the cross-entropy loss in classification tasks[5]. For the regression task, we use the *mean squared error (MSE)*, defined for an estimator $\hat{\boldsymbol{\theta}}$ as:

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}}[(\mathbf{x}^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))^2|\mathbf{X}, \varepsilon], \tag{3.9}$$

[5] We also believe that our analysis of the modified spectrum induced by DA suggests that such equivalences could also be shown for aSGD applied on the cross-entropy v.s. squared loss, but do not pursue this path in this paper.

Table 3.2: **Road map of main results.**

	Regression	**Classification**
Meta-Theorem: Unbiased Estimator	*Theorem* 4	*Theorem* 9
Meta-Theorem: Biased Estimator	*Theorem* 7	*Theorem* 11
Augmentation Case Studies	Cutout: *Cor.* 12, 15, 18 Compositions: *Cor.* 16	Cutout: *Cor.* 13, 14, 15 Group invariant: *Cor.* 17
Interplay with Signal Model	*Corollary* 18	*Corollary* 45
Comparisons between Under- & Over-parameterized regimes	*Corollary* 12, 14, 17	
Comparisons between Regression & Classification	*Proposition* 20, 21	

Recall in the above that $\boldsymbol{\theta}^*$ denotes the true coefficient vector, ε denotes noise in the observed data, and $\mathbf{x}$ denotes a test example that is independent of the training examples $\mathbf{X}$. For classification, we will use the *probability of classification 0-1 error (POE)* as the testing metric:

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}}[\mathbb{I}\{\mathrm{sgn}(\mathbf{x}^\top \hat{\boldsymbol{\theta}}) \neq \mathrm{sgn}(\mathbf{x}^\top \boldsymbol{\theta}^*)\}].$$

Spectral quantities of interest

Recent works studying overparameterized regression and classification tasks [25, 26, 27, 72] have discovered that the *spectrum*, i.e. eigenvalues, of the data covariance play a central role in characterizing the generalization error. In particular, two *effective ranks*, which are functionals of the data spectrum and act as types of effective dimension, dictate the generalization error of both underparameterized and overparameterized models. These are defined below.

Definition 2 (Effective Ranks, [25]). *For any covariance matrix (spectrum)* $\boldsymbol{\Sigma}$*, ridge regularization* *scale given by* c*, and index* $k \in \{0, \ldots, p-1\}$*, two notions of effective ranks are given as below:*

$$\rho_k(\mathbf{\Sigma}; c) := \frac{c + \sum_{i>k} \lambda_i}{n\lambda_{k+1}}, \;\; R_k(\mathbf{\Sigma}; c) := \frac{(c + \sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}.$$

Using this notation, the risk for the minimum-norm least squares estimate from [25, 26] can be sharply characterized as

$$\begin{aligned}
&\text{MSE} \asymp \underbrace{\|\boldsymbol{\theta}^* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}|\mathbf{X}]\|_{\mathbf{\Sigma}}^2}_{\text{Bias}} + \underbrace{\|\hat{\boldsymbol{\theta}} - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}|\mathbf{X}]\|_{\mathbf{\Sigma}}^2}_{\text{Variance}}, \text{ where} \\
&\text{Bias} \lesssim \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\mathbf{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\mathbf{\Sigma}_{0:k}^{-1}} \lambda_{k+1}^2 \rho_k(\mathbf{\Sigma}; 0)^2, \;\; \text{Variance} \asymp \frac{k}{n} + \frac{n}{R_k(\mathbf{\Sigma}; 0)},
\end{aligned}$$

where $k \leq \min(n, p)$ is an index that partitions the spectrum of the data covariance $\mathbf{\Sigma}$ into “spiked" and residual components and can be chosen in the analysis to minimize the above upper bounds. We note that the expression for the bias is matched by a lower bound upto universal constant factors for certain types of signal: either random [26] or sparse [27].

Intuitively, this characterization implies a two-fold requirement on the data spectrum for good generalization (in the sense of statistical consistency: $\text{MSE} \to 0$ as $n \to \infty$): it must a) decay quickly enough to preserve ground-truth signal recovery (i.e. ensure that ρ_k is small, resulting in low bias), but also b) retain a long enough tail to reduce the noise-overfitting effect (i.e. ensure that R_k is large, resulting in low variance).

3.4.2 A deterministic approximation strategy for DA analysis

Our main results show that the DA framework naturally inherits the above principle. In other words, the impact of DA on generalization (in both underparameterized and overparameterized regimes) boils down to understanding the effective ranks of a *modified, augmentation-induced spectrum.* Our starting point is the approximate connection between the aERM estimator and ridge estimator that was established in Section 3.3.2. Out of the box, this *does not* establish a direct equivalence between the MSE of the two estimators. This is because the implicit regularizer Cov_G that is induced by DA intricately depends on the data matrix $\mathbf{X}$, which creates strong dependencies amongst the training examples in the equivalent ridge estimator. A key technical contribution of our work is to show that, in essence, this dependency turns out to be quite weak for a large class of augmentations that are used in practice. Our strategy is to approximate the aERM estimator $\hat{\boldsymbol{\theta}}_{\text{aug}}$ with an idealized estimator $\bar{\boldsymbol{\theta}}_{\text{aug}}$ that uses the *expected* augmentation covariance (over the original data distribution). The two estimators are formally defined below:

$$\hat{\boldsymbol{\theta}}_{\text{aug}} = (\mu_{\mathcal{G}}(\mathbf{X})^\top \mu_{\mathcal{G}}(\mathbf{X}) + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1} \mu_{\mathcal{G}}(\mathbf{X})^\top \mathbf{y}, \tag{3.10}$$

$$\bar{\boldsymbol{\theta}}_{\text{aug}} = (\mu_{\mathcal{G}}(\mathbf{X})^\top \mu_{\mathcal{G}}(\mathbf{X}) + n\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1} \mu_{\mathcal{G}}(\mathbf{X})^\top \mathbf{y}, \tag{3.11}$$

where $\mathbf{x}$ denotes a fresh data point. This admits a decomposition of the MSE into three error terms, given by

$$\mathrm{MSE} \lesssim \underbrace{\|\boldsymbol{\theta}^* - \mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\text{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Bias}} + \underbrace{\|\bar{\boldsymbol{\theta}}_{\text{aug}} - \mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\text{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Variance}} + \underbrace{\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\boldsymbol{\Sigma}}^2}_{\text{Approximation Error}}. \tag{3.12}$$

The bias and variance terms can be analyzed with relative ease through an extension of the techniques of [25, 26] to general positive-semidefinite regularizers that are not dependent on the training data[6] $\mathbf{X}$, as we outlined in Section 3.3.2. We provide a novel analysis of the approximation error term in Section 3.4.3 and show, for an arbitrary data covariance $\boldsymbol{\Sigma}$ and several popular augmentations, that this approximation error is often dominated by either the bias or variance. As described in more detail in Section 3.4.3, this domination implies that we can tightly characterize the MSE with upper and lower bounds that match up to constant factors for these augmentations. Figure 3.1 confirms that the approximation error is indeed negligible. In this plot, we show the decomposition corresponding to the terms in (3.12) for random mask augmentation with different masking probabilities denoted by β. We can see that the approximation error is small compared with other the error components.

[6] For this case, a related contribution lies in the work of [115]. Note that [115] provided precise asymptotics for general regularizers in the proportional regime $p \propto n$ and focused on the question of the optimal Tikhonov regularizer, while our focus is on more interpretable non-asymptotic bounds for the general regularizers that are induced by popular augmentations. We believe that our framework could also yield identical proportional asymptotics for DA under an

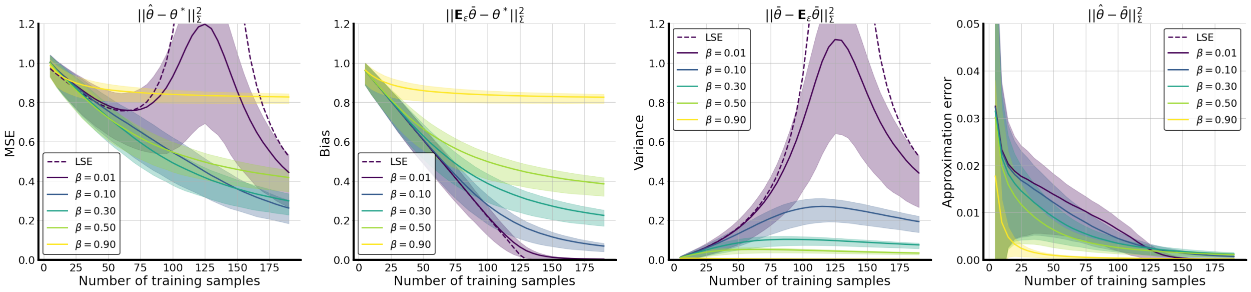


Figure 3.1: **Decomposition of MSE into the bias, variance, and approximation error as in Theorem 4.** Our MSE bound is an extension of the traditional bias-variance decomposition by modifying the bias and variance terms to correspond to an estimator with a deterministic regularizer and adding an additional approximation error term to compensate for the error. In this figure, we show that for random mask augmentation with different dropout probability β, the approximation error is small compared to the bias and variance (being at most $1/10$ of each of these quantities in this case), validating the efficacy of our proposed decomposition.

That the approximation error is negligible is an apriori surprising observation in the high-dimensional regime, as the sample data augmentation covariance $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ and its expectation $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_g(\mathbf{x})]$ are p-dimensional square matrices and $p \gg n$. We critically use the special structure of the augmentations we study to show that despite this high-dimensional structure, it is common for $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ to converge to its expectation at a rate that depends mostly on n and minimally on p.

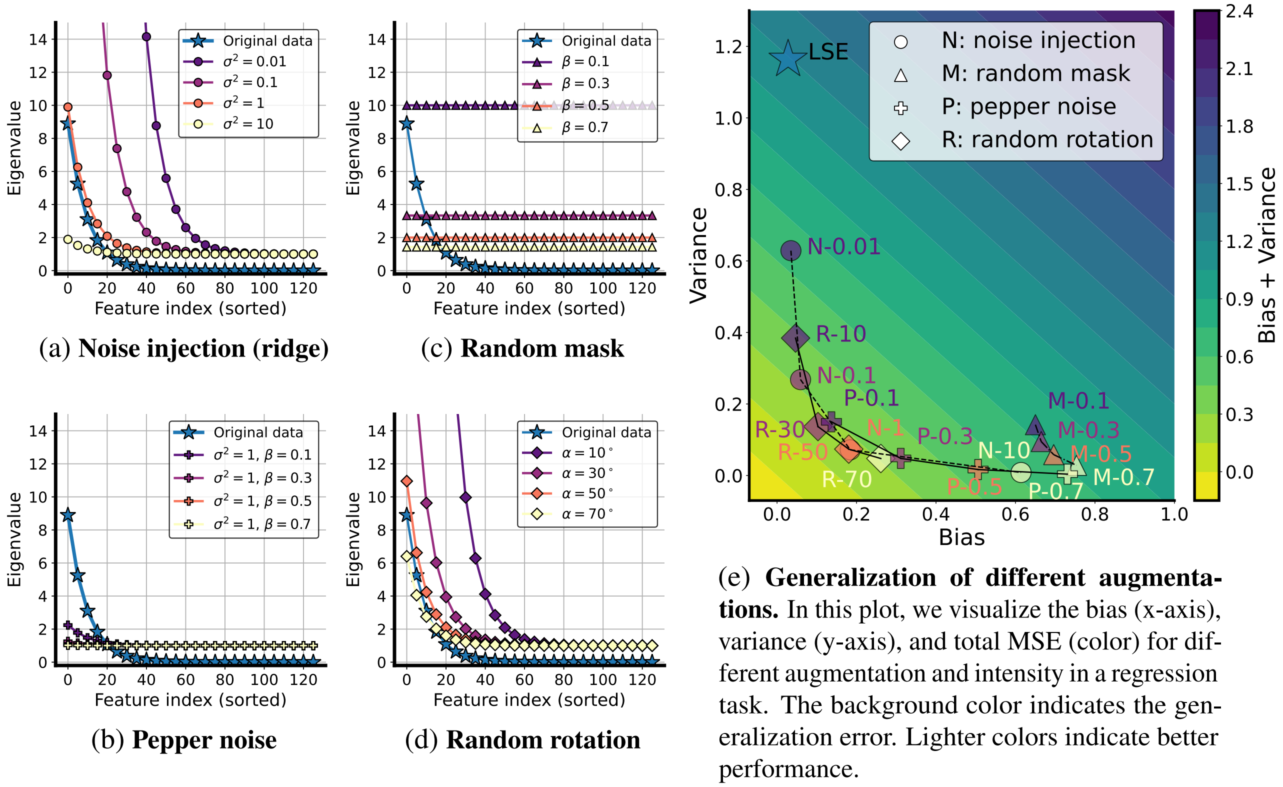
To show that our deterministic approximation is validated, i.e., the approximation error term is negligible, we require the following technical assumption, which shows that a normalized version of the empirical augmentation-induced covariance matrix converges as $n, p \to \infty$.

Assumption 1. *Let the data dimension p grows with n at the polynomial rate $p = n^\alpha$ for some $\alpha > 1$. Then, we assume that for any sequence of data covariance matrices $\{\mathbf{\Sigma}_p\}_{p\geq 1}$, the normalized empirical covariance induced by the augmentation distribution converges to its expectation as $n \to \infty$. More formally, we assume that*

$$\Delta_G := \left\| \frac{1}{n}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} \sum_{i=1}^{n} \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}_i)\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}_p \right\| \to 0 \textit{ as } n \to \infty \textit{ almost surely.}$$

We note here that the above should be interpreted as the limit as both n and p grow together. For our subsequent results to be meaningful, it is further required that this convergence is sufficiently

equivalent version of Assumption 1 for the proportional regime $p \propto n$, but do not pursue this path in this paper.



(a) **Noise injection (ridge)**

(c) **Random mask**

(b) **Pepper noise**

(d) **Random rotation**

(e) **Generalization of different augmentations.** In this plot, we visualize the bias (x-axis), variance (y-axis), and total MSE (color) for different augmentation and intensity in a regression task. The background color indicates the generalization error. Lighter colors indicate better performance.

Figure 3.2: **Equivalent augmented data spectrum and generalization** In plots (a)-(d), we visualize the regularized augmented spectrum (defined in (3.14))) of Gaussian noise injection (N), pepper noise injection (P), random mask (M) and the novel random rotation (R) introduced in Section 3.5.2, and their corresponding generalization in plot (e), where the number followed by the abbreviation in the data point denotes its parameter. The LSE represents the baseline of least-squared estimator without any augmentation.

fast as $n, p \to \infty$. We will show (in Proposition 5, and with several concrete examples) that a wide class of augmentations will satisfy this assumption and converge at the rate $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$. We will see that this rate is sufficient for our results to be tight in non-trivial regimes.

3.4.3 Regression analysis

With the connection of DA to ridge regression established in Section 3.3.2 and the deterministic approximation method established in Section 3.4.2, we are ready to present our meta-theorem for the regression setting. The results for the augmented estimators which are unbiased-in-average are presented in Section 3.4.3, and biased-in-average augmented estimators are studied in Section 3.4.3. The applications of the general theorem in this section will be discussed in detail in Section 3.5.

Regression analysis for general classes of unbiased augmentations

In this section, we present the meta-theorem for estimators induced by unbiased-on-average augmentations (i.e., for which $\mu_{\mathcal{G}}(\mathbf{x}) = \mathbf{x}$) in Theorem 4. All proofs in this section can be found in Section 3.8.2. To state the main result of this section, we introduce new notation for the relevant augmentation-transformed quantities.

Definition 3 (**Augmentation-transformed quantities**). *We define two spectral augmentation transformed quantities, the covariance-of-the-mean-augmentation $\bar{\mathbf{\Sigma}}$, and augmentation-transformed data covariance $\mathbf{\Sigma}_{aug}$, by*

$$\bar{\mathbf{\Sigma}} := \mathbb{E}_{\mathbf{x}}[(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])^\top], \tag{3.13}$$

$$\mathbf{\Sigma}_{aug} := \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\bar{\mathbf{\Sigma}}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}, \tag{3.14}$$

We also denote the eigenvalues of $\mathbf{\Sigma}_{aug}$ by $\lambda_1^{aug} \geq \lambda_2^{aug} \geq \cdots \geq \lambda_p^{aug}$. Similarly, we define the augmentation-transformed data matrix $\mathbf{X}_{aug}$, and augmentation-transformed model parameter $\boldsymbol{\theta}^*_{aug}$

as

$$\mathbf{X}_{aug} := \mu_G(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2},\ \boldsymbol{\theta}^*_{aug} := \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{1/2}\boldsymbol{\theta}^*. \tag{3.15}$$

Note that since the rows of $\mathbf{X}_{aug}$ are still i.i.d., $\mathbf{X}_{aug}$ can be viewed as a modified data matrix with covariance $\mathbf{\Sigma}_{aug}$ and $\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}$ if the augmentation is unbiased in average.

Armed with this notation, we are ready to state our meta-theorem.

Theorem 4 (High probability bound of MSE for unbiased DA). *Consider an unbiased data augmentation g and its corresponding estimator $\hat{\boldsymbol{\theta}}_{aug}$. Recall the definition*

$$\Delta_G := \|\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}_p\|,$$

and let κ be the condition number of $\mathbf{\Sigma}_{aug}$. Assume that the condition numbers for the matrices $\mathcal{A}_{k_1}(\mathbf{X}_{aug}; n)$, $\mathcal{A}_{k_2}(\mathbf{X}_{aug}; n)$ are bounded by L_1 and L_2 respectively with probability $1-\delta'$, and that $\Delta_G \le c'$ for some constant $c' < 1$. Then there exist some constants c, C depending only on σ_x and σ_ε, such that, with probability $1 - \delta' - 4n^{-1}$, the test mean-squared error is bounded by

$$\begin{aligned}
\mathrm{MSE} &\lesssim \mathrm{Bias} + \mathrm{Variance} + \mathrm{ApproximationError}, \\
\frac{\mathrm{Bias}}{C_x L_1^4} &\lesssim \left(\left\| \mathbf{P}^{\mathbf{\Sigma}_{aug}}_{k_1+1:p} \theta^*_{aug} \right\|^2_{\mathbf{\Sigma}_{aug}} + \left\| \mathbf{P}^{\mathbf{\Sigma}_{aug}}_{1:k_1} \theta^*_{aug} \right\|^2_{\mathbf{\Sigma}^{-1}_{aug}} \frac{(\rho^{aug}_{k_1})^2}{(\lambda^{aug}_{k_1+1})^{-2} + (\lambda^{aug}_1)^{-2}(\rho^{aug}_{k_1})^2} \right), \\
\frac{\mathrm{Variance}}{\sigma_\varepsilon^2 L_2^2 \tilde{C}_x} &\lesssim \left(\frac{k_2}{n} + \frac{n}{R^{aug}_k} \right) \log n, \quad \mathrm{Approx.Error} \lesssim \kappa^{\frac{1}{2}} \Delta_G \left(\|\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}} + \sqrt{\mathrm{Bias} + \mathrm{Variance}} \right).
\end{aligned} \tag{3.16}$$

Above, we defined $\rho^{aug}_k := \rho_k(\mathbf{\Sigma}_{aug}; n)$ and $R^{aug}_k := R_k(\mathbf{\Sigma}_{aug}; n)$ as shorthand.

Theorem 4 illustrates the critical role that the spectrum of the augmentation-transformed data covariance $\mathbf{\Sigma}_{\mathrm{aug}}$ plays in generalization. In Fig. 3.2, we visualize this impact for various types of augmentations.

When is our bound in Theorem 4 tight? A natural question is when and whether our bound in Theorem 4 is tight. The tightness of the testing error for an estimator with a fixed regularizer is established (under some additional assumptions on the data distribution, such as sub-Gaussianity and constant condition number) in Theorem 5 of [26]. Hence, as long as the approximation error in our theorem is dominated by either the bias or variance, then our bound will also be tight. Roughly speaking this happens when the convergence of $n^{-1}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ to $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$ is sufficiently fast with respect to n. For interested readers, we have included the full technical condition in Lemma 35 of Section 3.8.1.

An important class of DA in practice involves independently augmenting each of the features. This class subsumes many prevailing augmentations like random mask, salt-and-pepper, and Gaussian noise injection. Because the augmentation covariance $\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})$ is diagonal for such augmentations, we can simplify Theorem 4, as shown in the next proposition. The proposition shows that this class of augmentation has a reordering effect on the magnitude (or importance) of each feature.

Proposition 5 (Independent Feature Augmentations). *Let g be an independent feature augmentation, and $\pi : \{1, 2, \ldots, p\} \to \{1, 2, \ldots, p\}$ be the function that maps the original feature index to the sorted index according to the eigenvalues of $\mathbf{\Sigma}_{aug}$ in a non-increasing order. Then, data augmentation has a spectrum reordering effect which changes the MSE through the bias modification:*

$$\frac{\textit{Bias}}{C_x L_1^4} \lesssim \left\|\theta^*_{\pi(k_1+1:p)}\right\|^2_{\Sigma_{\pi(k_1+1:p)}} + \left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2 \Sigma^{-1}_{\pi(1:k_1)}} \frac{(\rho^{aug}_{k_1})^2}{(\lambda^{aug}_{k_1+1})^{-2} + (\lambda^{aug}_1)^{-2}(\rho^{aug}_{k_1})^2},$$

where $\pi(a : b)$ denotes the indices of $\pi(a), \pi(a+1), \ldots, \pi(b)$. Furthermore, if the variance of each feature augmentation $\mathrm{Var}_{g_i}(g_i(x))$ is a sub-exponential random variable with sub-exponential norm σ_i^2 and mean $\bar{\sigma}_i^2$, $\forall i \in \{1, 2, \ldots, p\}$, and $p = O(n^\alpha)$ for some $\alpha > 0$, then there exists a constant c, *depending only on α, such that with probability* $1-n^{-1}$,

$$\Delta_G \lesssim \max_i \left(\frac{\sigma_i^2}{\bar{\sigma}_i^2}\right)\sqrt{\frac{\log n}{n}}.$$

Proposition 5 gives a bound on the approximation error for independent feature augmentations and finds that $\Delta_G \lesssim \sqrt{\frac{\log n}{n}}$. However, one might wonder whether the approximation error still vanishes for stochastic augmentations that include dependencies between features. While we do not provide such a guarantee for arbitrary augmentations, we present a general technique that we later use to show that the approximation error is indeed vanishing for many popularly used augmentations that include dependencies between features. Specifically, we consider the decomposition $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \mathcal{D} + \mathbf{Q}$, where $\mathcal{D}$ is a diagonal matrix representing the *independent* feature augmentation part. Then, we have

$$\Delta_G \lesssim \frac{\|\mathcal{D}-\mathbb{E}\mathcal{D}\| + \|\mathbf{Q}-\mathbb{E}\mathbf{Q}\|}{\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))}. \tag{3.17}$$

Further discussion on the approximation error for dependent feature augmentation, along with the proof of Eq. (3.17), is provided in Appendix 3.8.6. We use Eq. (3.17) to show that the possibly large error of the non-diagonal part $\|\mathbf{Q}-\mathbb{E}\mathbf{Q}\|$ resulting from a dependent feature augmentation can be mitigated by the denominator $\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))$, for augmentations for which $\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})$ is well-conditioned. We use this in Appendix 3.8.6 to characterize the approximation error for two examples of augmentations that induce dependencies between features: a) the new *random-rotation* augmentation that we introduced in Section 3.3.4, b) the cutout augmentation which is popular in deep learning practice [18].

Regression analysis for general biased on average augmentations

All of our analysis thus far has assumed that the augmentation is *unbiased on average*, i.e. that $\mu_{\mathcal{G}}(\mathbf{x}) = \mathbf{x}$. We now derive and interpret the expression for the estimator that is induced by a general augmentation that can be biased. We introduce the following additional definitions.

Definition 6. *We define the* ***augmentation bias*** *and* ***bias covariance*** *induced by the augmentation g as*

$$\xi(\mathbf{x}) := \mu_g(\mathbf{x}) - \mathbf{x}, \quad \mathrm{Cov}_\xi := \mathbb{E}_{\mathbf{x}}\left[\xi(\mathbf{x})\xi(\mathbf{x})^\top\right]. \tag{3.18}$$

Since $\xi(\mathbf{x})$ is not zero for a biased augmentation, the expression in (3.7) becomes more complicated and we lose the exact equivalence to an ridge regression in (3.8)). This is because biased DA induces a distribution-shift in the training data that does not appear in the test data. Our next result for biased estimators, which is strictly more general than Theorem 4, will show that this distribution-shift affects the test MSE through both *covariate-shift* as well as *label-shift*. To facilitate analysis, we impose the natural assumption that the mean augmentation $\mu(\mathbf{x})$ remains sub-Gaussian.

Assumption 2. *For the input data $\mathbf{x}$, the mean transformation $\mu(\mathbf{x})$ admits the form $\mu(\mathbf{x}) = \bar{\mathbf{\Sigma}}^{1/2}\bar{\mathbf{z}}$, where $\bar{\mathbf{\Sigma}}$ is defined in Definition 3 and $\bar{\mathbf{z}}$ is a centered and isotropic sub-Gaussian vector with sub-Gaussian norm $\sigma_{\bar{z}}$.*

We also recall the definition of the mean augmentation covariance $\bar{\mathbf{\Sigma}} := \mathbb{E}_{\mathbf{x}}[(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])^\top]$. Now we are ready to state our theorem for biased augmentations. The proof is deferred to Appendix 3.8.2.

Theorem 7 (Bounds on the MSE for Biased Augmentations). *Consider the estimator $\hat{\boldsymbol{\theta}}_{aug}$ obtained by solving the aERM in (3.2). Let $\mathrm{MSE}^o(\hat{\boldsymbol{\theta}}_{aug})$ denote the unbiased MSE bound in Eq. (3.16) of Theorem 4, and recall the definition*

$$\Delta_G := \left\| \frac{1}{n}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} \sum_{i=1}^{n} \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}_i) \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}_p \right\|.$$

Suppose the assumptions in Theorem 4 hold for the mean augmentation $\mu(\mathbf{x})$ and that $\Delta_G \leq c < 1$. Then with probability $1 - \delta' - 4n^{-1}$ we have,

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim R_1^2 \cdot \left(\sqrt{\mathrm{MSE}^o(\hat{\boldsymbol{\theta}}_{aug})} + R_2\right)^2,$$

where

$$
\begin{aligned}
R_1 &= 1 + \|\mathbf{\Sigma}^{\frac{1}{2}}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\| \textit{ and} \\
R_2 &= \sqrt{\|\bar{\mathbf{\Sigma}}(\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1}\|}\left(1 + \frac{\Delta_G}{1-c}\right)\left(\sqrt{\Delta_\xi}\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right) \\
&\times \left(\sqrt{\frac{1}{\lambda_k^{aug}}} + \sqrt{\frac{\lambda_{k+1}^{aug}(1+\rho_k^{aug})}{(\lambda_1^{aug}\rho_0^{aug})^2}}\right).
\end{aligned}
$$

Our upper bound for the MSE in the biased augmentation case is a generalization of the bound in [26] to the scenario with distribution-shift. This result shows that two different factors can cause generalization error over and above the unbiased case: 1. *covariate shift*, which is reflected in the multiplicative factor R_1; this term occurs because we are testing the estimator on a distribution with covariance $\mathbf{\Sigma}$ but our training covariates have covariance $\bar{\mathbf{\Sigma}}$ instead, 2. *label shift*, which manifests itself as the additive error given by R_2. This term arises from the training mismatch between the true covariate observation and mean augmented covariate (i.e., $\mathbf{X}$ v.s. $\mu_G(\mathbf{X})$). As a sanity check, we can see that $R_1 = 1$ and $R_2 = 0$ when the augmentation is unbiased in average, i.e., $\mu_{\mathcal{G}}(\mathbf{x}) = \mathbf{x},\ \forall \mathbf{x}$, since $\mathbf{\Sigma} = \bar{\mathbf{\Sigma}}$, $\Delta_\delta = 0$ and $\mathrm{Cov}_\delta = 0$. Thus, we directly recover Theorem 4 in this case. Whether Theorem 7 is tight in general is an interesting open question for future work.

3.4.4 Classification analysis

In this subsection, we state the meta-theorem for generalization of DA in the classification task. We follow a similar path for the analysis as in regression by appealing to the connection between DA and ridge estimators and the deterministic approximation strategy outlined above. While the results in this section operate under stronger assumptions, we provide a similar set of results to the regression case. The primary aim of these results is to compare the generalization behavior of DA between regression and classification settings, which we do in depth in Section 3.5.

Classification analysis setup

We adopt the random signed model from [27], noting that we expect similar analysis to be possible for the Gaussian-mixture-model setting of [111, 28] (we defer such analysis to a companion paper). Given a target vector $\theta^* \in \mathbb{R}^d$ and a label noise parameter $0 \leq \nu^* < 1/2$, we assume the data are generated as binary labels $y_i \in \{-1, 1\}$ according to the signal model

$$y_i = \begin{cases} \operatorname{sgn}(\mathbf{x}_i^\top \boldsymbol{\theta}^*) & \text{with probability } 1 - \nu^* \\ -\operatorname{sgn}(\mathbf{x}_i^\top \boldsymbol{\theta}^*) & \text{with probability } \nu^* \end{cases} \tag{3.19}$$

Just as in [27], we make a *1-sparse* assumption on the true signal $\boldsymbol{\theta}^* = \frac{1}{\sqrt{\lambda_t}}\mathbf{e}_t$. We denote $\mathbf{x}_{\text{sig}} := \mathbf{x}_t$ to emphasize the signal feature. Motivated by recent results which demonstrate the effectiveness of training with the squared loss for classification tasks [124, 27], we study the classification risk of the estimator $\hat{\boldsymbol{\theta}}$ which is computed by solving the aERM objective on the binary labels y_i with respect to the squared loss (Eq. (3.2)).

[27] showed that two quantities, *survival* and *contamination*, play key roles in characterizing the risk, akin to the bias and variance in the regression task (in fact, as shown in the proof of Lemma 39, the contamination term scales identically to the variance from regression analysis). The definitions of these quantities are given below.

Definition 8 (Survival and contamination [27]). *Given an estimator $\hat{\theta}$, its survival (SU) and contamination (CN) are defined as*

$$\operatorname{SU}(\hat{\boldsymbol{\theta}}) = \sqrt{\lambda_t}\hat{\boldsymbol{\theta}}_t, \ \ \operatorname{CN}(\hat{\boldsymbol{\theta}}) = \sqrt{\sum_{j=1, j\neq t}^{p} \lambda_j \hat{\boldsymbol{\theta}}_j^2}. \tag{3.20}$$

For Gaussian data, [27] derived the following closed-form expression for the POE:

$$\text{POE}(\hat{\boldsymbol{\theta}}) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{\operatorname{SU}(\hat{\boldsymbol{\theta}})}{\operatorname{CN}(\hat{\boldsymbol{\theta}})}. \tag{3.21}$$

Thus, the POE depends on the ratio between survival SU and contamination CN, essentially a kind of *signal-to-noise ratio* for the classification task. In this work, we prove that a similar principle arises when we consider training with data augmentation in more general correlated input distributions. Formally, we make the following assumption on the true signal and input distribution for our classification analysis.

Assumption 3. *Assume the target signal is 1-sparse and given by* $\boldsymbol{\theta}^* = \frac{1}{\sqrt{\lambda_t}}\mathbf{e}_t$. *Additionally, assume the input can be factored as* $\mathbf{x} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}$, *where* $\boldsymbol{\Sigma} \succeq 0$ *is diagonal, and* $\mathbf{z}$ *is a sub-Gaussian random vector with norm* σ_z *and uniformly bounded density. We denote* $\mathbf{x}_{sig} = \mathbf{x}_t$ *and* $\mathbf{x}_{noise} = [\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_p]^T$. *We further assume that the signal and noise features are independent*[7], *i.e.,* $\mathbf{x}_{sig} \perp \mathbf{x}_{noise}$.

Similar to the regression case, our classification analysis consists of 1) expressing the excess risk in terms of $\bar{\boldsymbol{\theta}}_{\text{aug}}$, the estimator corresponding to the averaged augmented covariance $\mathbb{E}_{\mathbf{x}}[\text{Cov}_g(\mathbf{x})]$, 2) arguing that the survival and contamination can be viewed as the equivalent quantities for a ridge estimator with a modified data spectrum, and 3) upper and lower bounding the survival and contamination of this ridge estimator. As in the case of regression analysis, step 1) is the most technically involved.

Classification analysis for unbiased augmentations

Now, we present our main theorem for the classification task. The proof of this theorem is deferred to Appendix 3.8.3.

Theorem 9 (Bounds on Probability of Classification Error). *Consider the classification task under the setting in Assumption 3. Recall that* $\hat{\boldsymbol{\theta}}_{aug}$ *is the estimator solving the aERM objective in (3.2) and the definition* $\Delta_G := \|Cov_G(\mathbf{X}) - \mathbb{E}_{\mathbf{x}}[Cov_g(\mathbf{x})]\|$. *Let* $t \leq n$ *be the index (arranged according to the eigenvalues of* $\boldsymbol{\Sigma}_{aug}$*) of the non-zero coordinate of the true signal,* $\widetilde{\boldsymbol{\Sigma}}_{aug}$ *be the*

[7] As mentioned earlier, we expect that our framework can be extended beyond sub-Gaussian features to more general kernel settings. Under the slightly different label model used in [109], we believe that the independence between signal and noise features can also be relaxed.

leave-one-out modified spectrum corresponding to index t, κ be the condition number of $\boldsymbol{\Sigma}_{aug}$, and $\widetilde{\mathbf{X}}_{aug}$ be the leave-one-column-out data matrix corresponding to column t.

Suppose data augmentation is performed independently for $\mathbf{x}_{sig}$ and $\mathbf{x}_{noise}$, and there exists a $t \leq k \leq n$ such that with probability at least $1-\delta$, the condition numbers of $n\mathbf{I}+\widetilde{\mathbf{X}}^{aug}_{k+1:p}(\mathbf{X}^{aug}_{k+1:p})^\top$ and $n\mathbf{I}+\mathbf{X}^{aug}_{k+1:p}(\mathbf{X}^{aug}_{k+1:p})^\top$ are at most L, and that of $\widetilde{\mathbf{X}}_{k+1:p}\boldsymbol{\Sigma}_{k+1:p}\widetilde{\mathbf{X}}^T_{k+1:p}$ is at most L_1. Then as long as $\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\mathrm{SU})$ and $\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\mathrm{CN})$, with probability $1-\delta-\exp(-\sqrt{n})-5n^{-1}$, the probability of classification error (POE) can be bounded in terms of the survival (SU) and contamination (CN), as

$$\mathrm{POE}(\hat{\theta}) \lesssim \frac{\mathrm{CN}}{\mathrm{SU}}\left(1+\sigma_z\sqrt{\log\frac{\mathrm{SU}}{\mathrm{CN}}}\right), \tag{3.22}$$

where

$$\frac{\lambda^{aug}_t(1-2\nu^*)\left(1-\frac{k}{n}\right)}{L\left(\lambda^{aug}_{k+1}\rho_k(\boldsymbol{\Sigma}_{aug};n)+\lambda^{aug}_t L\right)} \lesssim \underbrace{\mathrm{SU}}_{\textit{Survival}} \lesssim \frac{L\lambda^{aug}_t(1-2\nu^*)}{\lambda^{aug}_{k+1}\rho_k(\boldsymbol{\Sigma}_{aug};n)+L^{-1}\lambda^{aug}_t\left(1-\frac{k}{n}\right)}, \tag{3.23}$$

$$\sqrt{\frac{\tilde{\lambda}^{aug}_{k+1}\rho_k(\tilde{\boldsymbol{\Sigma}}^2_{aug};0)}{L'^2(\lambda^{aug}_1)^2(1+\rho_0(\boldsymbol{\Sigma}_{aug};\lambda))^2}} \lesssim \underbrace{\mathrm{CN}}_{\textit{Contamination}} \lesssim \sqrt{(1+\mathrm{SU}^2)L^2\left(\frac{k}{n}+\frac{n}{R_k(\tilde{\boldsymbol{\Sigma}}_{aug};n)}\right)\log n} \tag{3.24}$$

Furthermore, if $\mathbf{x}$ is Gaussian, then we obtain even tighter bounds:

$$\frac{1}{2}-\frac{1}{\pi}\tan^{-1}c\frac{\mathrm{SU}}{\mathrm{CN}} \leq \mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \leq \frac{1}{2}-\frac{1}{\pi}\tan^{-1}\frac{1}{c}\frac{\mathrm{SU}}{\mathrm{CN}} \lesssim \frac{\mathrm{CN}}{\mathrm{SU}}, \tag{3.25}$$

where c is a universal constant.

Remark 10. *Based on the expression for the classification error for Gaussian data, we see that the survival needs to be asymptotically greater than the contamination for the POE to approach 0 in the limit as $n, p \to \infty$. We note that the general upper bound we provide matches the tight upper and lower bounds for the Gaussian case up a log factor. Furthermore, the condition $\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\mathrm{SU})$ and $\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\mathrm{CN})$ is related to our condition for the tightness* *of our regression analysis, but a bit stronger (because our regression analysis only requires one of these relations to be true). We characterize when this stronger condition is met in Lemma 42.*

Based on the upper and lower bounds provided for SU and CN, we see that these quantities depend crucially on the spectral properties of the induced covariance matrix $\mathbf{\Sigma}_{aug}$. For favorable classification performance, Theorem 9 also requires $t \leq n.$ This is a necessary product of our analogy to a ridge estimator and is equivalent to requiring that θ^*_{aug} lies within the eigenspace corresponding to the dominant eigenvalues of the spectrum $\mathbf{\Sigma}_{aug}$. Such requirements have also been used in past analyses of both regression [26] and classification [27].

Classification analysis for general biased augmentations

As a counterpart of our regression analysis for estimators induced by biased-on-average augmentations (i.e. $\mu_g(\mathbf{x}) \neq \mathbf{x}$), we would also like to understand the impact of augmentation-induced bias on classification. Interestingly, the effect of this bias in classification turns out to be much more benign than that in regression. As a simple example, consider a scaling augmentation of the type $g(\mathbf{x}) := 2\mathbf{x}$. The induced bias is $\mu_g(\mathbf{x}) - \mathbf{x} = \mathbf{x}$, and the trained estimator $\hat{\boldsymbol{\theta}}_{\text{aug}}$ is just half the estimator trained with $\mathbf{x}$, which, however, predicts the same labels in a classification task. Therefore, we conclude that even with a large bias, the resultant estimator might be equivalent to the original one for classification tasks. In fact, as we show in the next result, augmentation bias is benign for the classification error metric under relatively mild conditions. The proof of this result is provided in Appendix 3.8.3.

Theorem 11 (POE of biased estimators). *Consider the 1-sparse model $\boldsymbol{\theta}^* = \mathbf{e}_t$. and let $\hat{\boldsymbol{\theta}}_{aug}$ be the estimator that solves the aERM in (3.2) with biased augmentation (i.e., $\mu(\mathbf{x}) \neq \mathbf{x}$). Let Assumption 2 holds, and the assumptions of Theorem 9 be satisfied for data matrix $\mu(\mathbf{X})$. If the mean augmentation $\mu(\mathbf{x})$ modifies the t-th feature independently of other features and the sign of the t-th feature is preserved under the mean augmentation transformation, i.e.,* $\operatorname{sgn}(\mu(\mathbf{x})_t) = \operatorname{sgn}(\mathbf{x}_t)$,

$\forall \mathbf{x}$, *then, the POE($\hat{\boldsymbol{\theta}}_{aug}$) is upper bounded by*

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \leq \mathrm{POE}^o(\hat{\boldsymbol{\theta}}_{aug}), \tag{3.26}$$

where $\mathrm{POE}^o(\hat{\boldsymbol{\theta}}_{aug})$ *is any bound in Theorem 9 with* $\mathbf{X}$ *and* $\boldsymbol{\Sigma}$ *replaced by* $\mu(\mathbf{X})$ *and* $\bar{\boldsymbol{\Sigma}}$*, respectively.*

Note that the sign preservation is only required in expectation and not for every realization of the augmentation, i.e., we only require $\mathbb{E}_g\left[g(\mathbf{x})_t\right]$ has the same sign as $\mathbf{x}_t$, rather than requiring that $g(\mathbf{x})_t$ have the same sign as $\mathbf{x}_t$ for every realization of g. The latter label-preserving property is is much more stringent and has been studied in [64]. At a high level, this result tells us that as long as the signal feature preserves the sign under the mean augmentation, the classification error is purely determined by the modified spectrum induced by DA.

In Fig. 3.3, we simulate the biased and unbiased random mask augmentation [17] and test their performance in regression and classification tasks. We consider the 1-sparse model in $\mathbb{R}^{128}$ (i.e. $p = 128$) with isotropic Gaussian covariates. For the biased variant of random mask, we use the masked estimator without the normalization factor $(1 - \beta)$; therefore, the augmentation mean is equal to $\mu_g(\mathbf{x}) = (1 - \beta)\mathbf{x}$. From the figure, we see the bias can be very harmful in regression, especially in the overparametrized regime ($n \leq 128$), while the performance is identical for classification. This experiment demonstrates the sharp differences in behavior between the settings of Theorems 7 and 11. We discuss this observation further in Section 3.5.3.

3.5 The good, the bad and the ugly sides of data augmentation

In this section, we will use the meta-theorems established in Section 3.4.3 and 3.4.4 to get further insight into the impact of DA on generalization. First, in Section 3.5.1, we derive generalization bounds for many common augmentations. Then, in Section 3.5 we use these bounds to understand when DA can be helpful or harmful. Finally, in Section 3.5.1 we conclude by discussing the complex range of factors (the "ugly") that play an important role in determining the effect of DA.

3.5.1 Case studies: generalization of common DA

In this section, we present and interpret generalization guarantees for commonly used augmentations including *Gaussian noise injection, randomized mask, cutout, and salt-and-pepper noise*. In particular, we discuss whether these augmentations improve or worsen generalization compared to the LSE estimator, beginning with regression tasks.

Gaussian noise injection

As a preliminary example, we note that Proposition 5 generalizes and recovers the existing bounds on the ridge and ridgeless estimators [25, 26]. This is consistent with classical results [91] that show an equivalence between augmented ERM with Gaussian noise injection and ridge regularization. For completeness, we include the generalization bounds for Gaussian noise injection in Appendix 3.8.2.

Randomized masking

Next, we consider the popular randomized masking augmentation (both the biased and unbiased variants), in which each coordinate of each data vector is set to 0 with a given probability, denoted by the masking parameter $\beta \in [0, 1]$. The unbiased variant of randomized masking rescales the features so that the augmented features are unbiased in expectation. This type of augmentation has been widely used in practice [17, 125][8], and is a simplified version of the popular cutout augmentation [18].

The following corollary characterizes the generalization error arising from the randomized masking augmentation in regression tasks.

Corollary 12 (Regression bounds for unbiased randomized masking augmentation). *Consider the unbiased randomized masking augmentation $g(\mathbf{x}) = [b_1\mathbf{x}_1, \ldots, b_p\mathbf{x}_p]/(1-\beta)$, where b_i are i.i.d. Bernoulli$(1-\beta)$. Define $\psi = \frac{\beta}{1-\beta} \in [0, \infty)$. Let L_1, L_2, κ, δ' be universal constants as defined in*

[8] We note that a superficially similar implicit regularization mechanism is at play in *dropout* [126], where the parameters of a neural network are set to 0 at random. In contrast to random masking, dropout zeroes out model parameters rather than data coordinates.

Theorem 4. Assume $p = O(n^\alpha)$ for some $\alpha > 0$. Then, for any set $\mathcal{K} \subset \{1, 2, \ldots, p\}$ consisting of k_1 elements and some choice of $k_2 \in [0, n]$, there exists some constant c', which depends solely on σ_z and σ_ε (the sub-Guassian norms of the covariates and noise), such that the regression MSE is upper-bounded by

$$\mathrm{MSE} \lesssim \underbrace{\|\theta^*_{\mathcal{K}}\|^2_{\Sigma_{\mathcal{K}}} + \|\theta^*_{\mathcal{K}^c}\|^2_{\Sigma_{\mathcal{K}^c}} \frac{(\psi n + p - k_1)^2}{n^2 + (\psi n + p - k_1)^2}}_{\text{Bias}}$$

$$+ \underbrace{\left(\frac{k_2}{n} + \frac{n(p - k_2)}{(\psi n + p - k_2)^2}\right) \log n}_{\text{Variance}} + \underbrace{\sigma_z^2 \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}}_{\text{Approx.Error}}$$

with probability at least $1 - \delta' - n^{-1}$.

Noting that $\psi = \frac{\beta}{1-\beta}$ increases monotonically in the mask probability β, Corollary 12 shows that bias increases with the mask intensity β, while the variance decreases. Figure 3.1 empirically illustrates these phenomena through a bias-variance decomposition. In fact, the regression MSE is proportional to the expression for MSE of the least-squares estimator (LSE) on isotropic data, suggesting that randomized masking essentially has the effect of *isotropizing the data*. As prior work on overparameterized linear models demonstrates [29, 73, 25], the LSE enjoys particularly low variance, but particularly high bias when applied to isotropic, high-dimensional data. For this reason, random masking turns out to be superior to Gaussian noise injection in reducing variance, but much more inferior in mitigating bias. We explore these effects and compare the overall generalization guarantees of the two types of augmentations in depth in Section 3.6.2. Our experiments there show striking differences in the manifested effect of these augmentations on generalization, despite their superficial similarities.

We also not here that the approximation error is relatively minimal, of the order $\sqrt{\frac{\log n}{n}}$. It is easily checked that the approximation error is dominated by the bias and variance as long as $p \ll n^2$ (and hence the lower bounds of [26] imply tightness of our bound in this range). We next present our generalization guarantees for the biased variant of the random masking augmentation. We verify the behavior predicted by this corollary in Figure 3.3.

Corollary 13 (**Regression bounds for biased mask augmentation**). *Consider the biased random mask augmentation $g(\mathbf{x}) = [b_1\mathbf{x}_1, \ldots, b_p\mathbf{x}_p]$, where b_i are i.i.d. Bernoulli $(1-\beta)$), and carry over all the notation from Corollary 12. Then, the regression MSE is upper bounded by*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug}) \leq \left(\sqrt{\mathrm{MSE}^o} + \psi\left(1 + \frac{\log n}{n}\right) \cdot \left(\left(\lambda_1 + \frac{\sum_j \lambda_j}{n}\right)\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}\right)\right)^2,$$

with probability at least $1 - \delta' - n^{-1}$. Above, MSE^o is the RHS of the bound in Corollary 12.

Finally, we characterize the generalization error of the randomized masking augmentation for the classification task.

Corollary 14 (**Classification bounds for random mask augmentation**). *Let $\hat{\boldsymbol{\theta}}_{aug}$ be the estimator computed by solving the aERM objective on binary labels with mask probability β, and denote $\psi := \frac{\beta}{1-\beta}$. Assume $p \ll n^2$. Then, with probability at least $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$*

$$\mathrm{POE} \lesssim Q^{-1}(1 + \sqrt{\log Q}), \tag{3.27}$$

$$\textit{where } Q = (1-2\nu)\sqrt{\frac{n}{p\log n}}\left(1 + \frac{n}{n\psi + p}\right)^{-1}. \tag{3.28}$$

In addition, if we assume the input data has Gaussian features, then we have tight generalization bounds

$$\mathrm{POE} \asymp \frac{1}{2} - \frac{1}{\pi}\tan^{-1} Q \tag{3.29}$$

with the same probability.

Random cutout

Next, we consider the popularly used *cutout* augmentation [18], which picks a set of k (out of p) consecutive data coordinates at random and sets them to zero. Interestingly, our analysis shows that the effect of the cutout augmentation is very similar to the simpler-to-analyze random mask augmentation. The following corollary shows that the generalization error of cutout is equivalent to that of randomized masking with dropout probability $\beta = \frac{k}{p}$. The proof of this corollary can be found in Appendix 3.8.2.

Corollary 15 (Generalization of random cutout). *Let $\hat{\boldsymbol{\theta}}_k^{cutout}$ denote the random cutout estimator that zeroes out k consecutive coordinates (the starting location of which is chosen uniformly at random). Also, let $\hat{\boldsymbol{\theta}}_\beta^{mask}$ be the random mask estimator with the masking probability given by β. We assume that $k = O(\sqrt{\frac{n}{\log p}})$. Then, for the choice $\beta = \frac{k}{p}$ we have*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_k^{cutout}) \asymp \mathrm{MSE}(\hat{\boldsymbol{\theta}}_\beta^{mask}), \;\; \mathrm{POE}(\hat{\boldsymbol{\theta}}_k^{cutout}) \asymp \mathrm{POE}(\hat{\boldsymbol{\theta}}_\beta^{mask}).$$

This result is consistent with our intuition, as the cutout augmentation zeroes out $\frac{k}{p}$ coordinates on average.

Composite augmentation: Salt-and-pepper

Our meta-theorem can also be applied to *compositions* of multiple augmentations. As a concrete example, we consider a “salt-and-pepper” style augmentation in which each coordinate is either replaced by random Gaussian noise with a given probability, or otherwise retained. Specifically, salt-and-pepper augmentation modifies the data as $g(\mathbf{x}) = [\mathbf{x}'_1, \ldots, \mathbf{x}'_p]$, where $\mathbf{x}'_i = \mathbf{x}_i/(1-\beta)$ with probability $1-\beta$ and otherwise $\mathbf{x}'_i = \mathcal{N}(\mu, \sigma^2)/(1-\beta)$. This is clearly a composite augmentation made up of randomized masking and Gaussian noise injection. For simplicity, we only consider the case where $\mu = 0$, since it results in an augmentation which is unbiased on average. The regression error of this composite augmentation is described in the following corollary, which is proved in Appendix 3.8.2.

Corollary 16 (Generalization of Salt-and-Pepper augmentation in regression). *The bias, variance and approximation error of the estimator that are induced by salt-and-pepper augmentation* *(denoted by $\hat{\boldsymbol{\theta}}_{pepper}(\beta, \sigma^2)$) are respectively given by:*

$$\begin{aligned}
\text{Bias}[\hat{\boldsymbol{\theta}}_{pepper}(\beta, \sigma^2)] &\lesssim \left(\frac{\lambda_1(1-\beta)+\sigma^2}{\sigma^2}\right)^2 \text{Bias}\left[\hat{\boldsymbol{\theta}}_{gn}\left(\frac{\beta\sigma^2}{(1-\beta)^2}\right)\right], \\
\text{Variance}[\hat{\boldsymbol{\theta}}_{pepper}(\beta, \sigma^2)] &\lesssim \text{Variance}\left[\hat{\boldsymbol{\theta}}_{gn}\left(\frac{\beta\sigma^2}{(1-\beta)^2}\right)\right], \\
\text{Approx.Error}[\hat{\boldsymbol{\theta}}_{pepper}(\beta, \sigma^2)] &\asymp \text{Approx.Error}[\hat{\boldsymbol{\theta}}_{rm}(\beta)].
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}_{gn}(z^2)$ and $\hat{\boldsymbol{\theta}}_{rm}(\gamma)$ denotes the estimators that are induced by Gaussian noise injection with variance z^2 and random mask with dropout probability γ, respectively. Moreover, the limiting MSE as $\sigma \to 0$ reduces to the MSE of the estimator induced by random masking (denoted by $\hat{\boldsymbol{\theta}}_{rm}(\beta)$):

$$\lim_{\sigma \to 0} \text{MSE}[\hat{\boldsymbol{\theta}}_{pepper}(\beta, \sigma^2)] = \text{MSE}[\hat{\boldsymbol{\theta}}_{rm}(\beta)].$$

Corollary 16 clearly indicates that the generalization performance of the salt-and-pepper augmentation interpolates between that of the random mask and Gaussian noise injections, in the sense that it reduces to random mask in the limit of $\sigma \to 0$, and also has a comparable bias and variance to Gaussian noise injection. More precisely, as we show in the proof of this corollary, this interpolation property is a result of the fact that the eigenvalues of the augmented covariance are the harmonic mean of the eigenvalues induced by random mask and Gaussian noise injection respectively, i.e.

$$\lambda_{pepper}(\beta, \sigma^2)^{-1} = \lambda_{rm}(\beta)^{-1} + \beta^{-1}\lambda_{gn}(\sigma^2)^{-1}. \tag{3.30}$$

3.5.2 The good and bad of DA: are they helpful or harmful?

Armed with generalization guarantees for many common augmentations, we now shift our focus to identifying explicit scenarios in which DA can be helpful or harmful.

The bad: the increase in bias might outweigh the variance reduction

In this section, we consider two different types of common augmentations that suffer from poor generalization in the overparameterized regime. The first is the randomized masking augmentation, whose generalization bounds we provided in Corollaries 12, 13 and 14. At a high level, the random mask drops features uniformly and thus equalizes the importance of each feature. This makes the data spectrum isotropic, i.e., $\mathbf{\Sigma}_{\text{aug}} = \psi^{-1} \cdot \operatorname{diag}(\mathbf{\Sigma})\mathbf{\Sigma} \operatorname{diag}(\mathbf{\Sigma})^{-1/2}$. Our corollaries show that for regression tasks (and either the biased on unbiased variant of randomized masking), the bias and the variance are given by $\mathcal{O}\left(\frac{(\psi n+p)^2}{(n+p)^2}\right)$ and $\mathcal{O}\left(\min(\frac{n}{p}, \frac{p}{n})\right)$ respectively. From this, we can draw the following insights: 1. the variance is always vanishing, and 2. the bias can be controlled in the underparameterized regime $p \ll n$ by adjusting ψ but is otherwise non-vanishing in the overparameterized regime $p \gg n$.

It is worth noting that these conclusions also manifest in the test MSE of the least-squares estimator (LSE) on isotropic data in the overparameterized regime [29, 25, 73]. Specifically, in the language of effective ranks, we observe that either isotropic data or the randomized masking augmentation induces the effective ranks $\rho_k = \Theta\left(\frac{p}{n}\right)$ and $R_k = \Theta\left(\frac{(n+p)^2}{p}\right)$. While the large value of R_k helps in variance reduction, the large value of ρ_k greatly increases the bias. As shown in the experiments in Section 3.6.2, the increase in bias often outweighs the variance reduction and results in the suboptimality of randomized masking relative to the more classical Gaussian noise injection augmentation in many overparameterized settings.

The second class of augmentations that can be harmful is *group-invariant augmentations*, which were extensively studied in [16] in the underparameterized or explicitly regularized regime. An augmentation class $\mathcal{G}$ is said to be group-invariant if $g(\mathbf{x}) \overset{d}{=} \mathbf{x},\ \forall g \in \mathcal{G}$. For such a class, the augmentation modified spectrum $\mathbf{\Sigma}_{\text{aug}}$ in Theorem 9 is given by

$$\mathbf{0} \preceq \mathbf{\Sigma}_{\text{aug}} = \mathbf{\Sigma} - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})\mu_{\mathcal{G}}(\mathbf{x})]^\top \preceq \mathbf{\Sigma}.$$

[16] argued that group invariance is an important reason why DA can help improve gener-

alization and showed that such invariances can greatly reduce the variance of the DA-induced estimator. However, the result below shows that such augmentations could generalize poorly, even for classification tasks, in the overparameterized regime. The proof of this result is contained in Appendix 3.8.3.

Corollary 17. ***[Group invariance augmentation in classification tasks]*** *Consider Gaussian covariates, i.e.* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *and consider the group-invariant augmentation given by* $g(\mathbf{x}) = \frac{1}{\sqrt{2}}\mathbf{x} + \frac{1}{\sqrt{2}}\mathbf{x}'$ *(where* $\mathbf{x}'$ *is an independent copy of* $\mathbf{x}$*). Then, under the assumptions of Theorem 9, the estimator induced by this augmentation has classification error given by*

$$\mathrm{POE} \asymp \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{\mathrm{SU}}{\mathrm{CN}}, \textit{ where} \tag{3.31}$$

$$\mathrm{SU} \asymp (1-2\nu)\frac{n}{2n+p}, \quad \sqrt{\frac{np}{(n+p)^2}} \lesssim \mathrm{CN} \lesssim \sqrt{(1+\mathrm{SU}^2)\frac{np\log n}{(n+p)^2}}. \tag{3.32}$$

with probability at least $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$.

Corollary 17 evaluates a specific type of group-invariant augmentation that is reminiscent of the *knockoff* model augmentation [120]. For this example, it is clear that for the underparameterized regime $p \leq n$, SU=$\Theta(1)$ and CN=$\mathcal{O}(\sqrt{\frac{p\log n}{n}})$ while for the overparameterized regime $p \geq n$, we have SU=$\Theta(\frac{n}{p})$ and CN=$\Omega(\sqrt{\frac{n}{p}})$. Therefore, there is a sharp transition of the survival-to-contamination ratio between the two regimes. As the ratio is asymptotically zero in the overparameterized reigme, we find that group invariant augmentation can be harmful for generalization in this case. Essentially, our result here shows that certain group-invariant augmentations have the same "isotropizing" effect that was also observed in random masking, i.e., $\mathbf{\Sigma}_{\text{aug}} = \mathbf{I}_p$. As already remarked on, this is an undesirable property in overparameterized settings, where it leads to high bias (and low survival for classification [27]).

The good: some types of DA are superior to ridge regularization

In this section, we first use examples to analyze when an augmentation can be effective as a function of the model structure. Then, we demonstrate a usage of our framework as a test bed for DA invention. Concretely, we propose a new augmentation that shows several desirable properties expressed through generalization bounds and numerical simulations.

When is data augmentation helpful? To understand which types of augmentation might yield favorable bounds, we consider, as in Corollary 18, the case of a nonuniform random masking augmentation in which the features that encode signal are masked with a lower probability than the remaining features. Specifically, we consider the k-sparse model where $\boldsymbol{\theta}^* = \sum_{i \in \mathcal{I}_\mathcal{S}} \alpha_i \mathbf{e}_i$ and $|\mathcal{I}_\mathcal{S}| = k$. Define the parameter $\psi := \frac{\beta}{1-\beta}$ where β is the probability of masking a given feature. Suppose that we employ a nonuniform mask across features, i.e. $\psi_i = \psi_1$ if $i \in \mathcal{I}_\mathcal{S}$ and is equal to ψ_0 otherwise. Conceptually, a good mask should retain the semantics of the original data as much as possible while masking the irrelevant parts. We can study this principle analytically through the regression and classification generalization bounds for this type of non-uniform masking. Below we present the regression result, and defer the proofs to Appendix 3.8.2 and the analogous classification result to Corollary 45 in Appendix 3.8.3.

Corollary 18 (Non-uniform random mask in k-sparse model). *Consider the $k-$sparse model and the non-uniform random masking augmentation where $\psi = \psi_1$ if $i \in \mathcal{I}_\mathcal{S}$ and ψ_0 otherwise. Then, if $\psi_1 \leq \psi_0$, we have with probability at least* $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$

$$\text{Bias} \lesssim \frac{\left(\psi_1 n + \frac{\psi_1}{\psi_0}\left(p - |\mathcal{I}_\mathcal{S}|\right)\right)^2}{n^2 + \left(\psi_1 n + \frac{\psi_1}{\psi_0}\left(p - |\mathcal{I}_\mathcal{S}|\right)\right)^2} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2, \ \ \text{Variance} \lesssim \frac{|\mathcal{I}_\mathcal{S}|}{n} + \frac{n\left(p - |\mathcal{I}_\mathcal{S}|\right)}{\left(\psi_0 n + p - |\mathcal{I}_\mathcal{S}|\right)^2},$$
$$\text{Approx.Error} \lesssim \sqrt{\frac{\psi_1}{\psi_0}} \sigma_z^2 \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}.$$

On the other hand, if $\psi_1 > \psi_0$, we have (with the same probability)

$$\text{Bias} \lesssim \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^2}, \ \ \text{Variance} \lesssim \frac{\left(\frac{\psi_1}{\psi_o}\right)^2 + \frac{|\mathcal{I}_\mathcal{S}|}{n}}{\left(\frac{\psi_1}{\psi_o} + \frac{|\mathcal{I}_\mathcal{S}|}{n}\right)^2}, \ \ \text{Approx.Error} \lesssim \sqrt{\frac{\psi_0}{\psi_1}} \sigma_z^2 \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$$

We can see that the bias decreases as the mask ratio ψ_1/ψ_0 between the signal part ($\mathcal{I}_\mathcal{S}$) and the noise part decreases. This corroborates the idea that a successful augmentation should retain semantic information as compared to the noisy parts of the data. Corollary 18 implies that for consistency as $n, p \to \infty$, we require $\frac{1}{n} \ll \frac{\psi_1}{\psi_0} \ll \frac{n}{p}$. This is because we must mask the noise features sufficiently more than the the signal feature for the bias to be small, but the two mask probabilities cannot be too different to allow the approximation error to decay to zero. We note that the bound has a sharp transition—if we mask the signal more than the noise, the bias bound becomes proportional to the null risk (i.e. the bias of an estimator that always predicts 0). Although the previous augmentations that we studied (randomized masking, noise injection, and salt-and-pepper augmentation) generally experience a trade-off between bias and variance as the augmentation intensity increases, we observe that the nonuniform random mask can reduce both bias and variance with appropriate parameter selection. However, while offering useful insight, this scheme relies crucially on knowledge of the target signal's sparsity and may be of limited practical interest. Next, we give a concrete example of how an augmentation, random rotation, can yield favorable performance *without* such oracle knowledge.

Using our framework as a test bed for new DA

We show here that our framework can be used as a testbed to quickly check the generalization of novel augmentation designs. In Section 3.3.4 we introduced a novel augmentation that sequentially rotates high dimensional vectors in $p/2$ independently chosen random planes. We demonstrate here that this “random-rotation" augmentation enjoys good generalization performance *regardless of the signal model*. The derivation of the estimator induced by this random-rotation augmentation is deferred to Appendix 3.8.5.

Corollary 19 (**Generalization of random-rotation augmentation**). *The estimator induced by the random-rotation augmentation (with angle parameter α) can be expressed as*

$$\hat{\boldsymbol{\theta}}_{rot} = \left(\mathbf{X}^\top\mathbf{X} + \frac{4(1-\cos\alpha)}{p}\left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} - \mathbf{X}^\top\mathbf{X}\right)\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

An application of Theorem 4 yields

$$\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{rot}) \asymp \mathrm{Bias}(\hat{\boldsymbol{\theta}}_{lse}),$$

for sufficiently large p *(overparameterized regime), as well as the variance bound*

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{rot}) \lesssim \mathrm{Var}(\hat{\boldsymbol{\theta}}_{ridge,\lambda}),$$

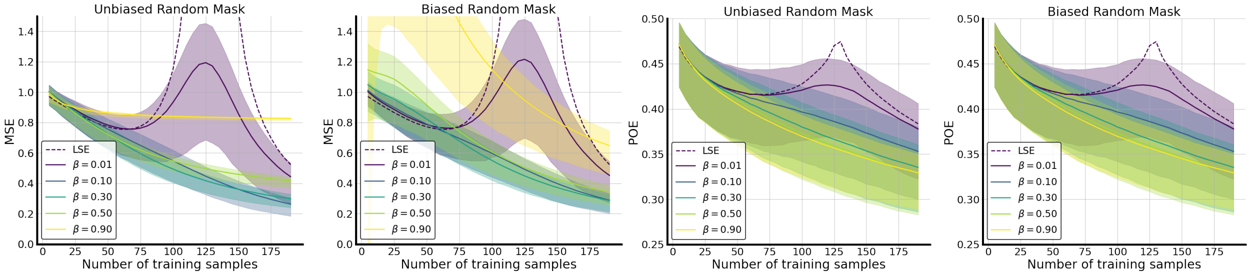
Above, $\hat{\boldsymbol{\theta}}_{lse}$ *and* $\hat{\boldsymbol{\theta}}_{ridge,\lambda}$ *denote the least squared estimator and ridge estimator with ridge intensity* $\lambda = np^{-1}(1-\cos\alpha)\sum_j \lambda_j$*. The approximation error can also be shown to decay as*

$$\mathrm{Approx.Error}(\hat{\boldsymbol{\theta}}_{rot}) \lesssim \max\left(\frac{1}{n}, \frac{\lambda_1}{\sum_{j>1}\lambda_j}\right).$$

The proof of the bias and variance expressions are provided in Appendix 3.8.3, and the proof of the approximation error is provided in Appendix 3.8.6 (this is the most involved step as random-rotation augmentations induce strong dependencies among features). Corollary 19 shows that, surprisingly, this simple augmentation leads to an estimator not only having the best asymptotic bias that matches that of LSE, but also reduces variance on the order of ridge regression. Thus, this estimator inherits the best of both types of estimators. Our experiments in Fig. 3.6 confirm this behavior and also show an appealing robustness property of the estimator across hyperparameter choice (i.e. value of rotation angle α).

3.5.3 The ugly: discrepancies in DA's effect under multiple factors

The results of the previous two sections reveal that several factors influence the effect of DA on generalization. Specifically, Section 3.5.2 shows that generalization performance often depends on whether a problem is in the underparameterized or overparameterized regime. Section 3.5.2 shows that generalization performance also intricately depends on the model structure. In this section, we further show that the impact of DA on generalization also depends on the downstream task



(a) Regression task

(b) Classification task

Figure 3.3: **Comparison of bias impact between regression and classification tasks** In this figure, we simulate the unbiased and biased random mask in regression and classification tasks. In (a), we show that the augmentation bias is mostly harmful to the regression task, especially in the overparameterized regime where the sample number is less than or equal to $p = 128$. In (b), however, the performance is identical with and without bias in the classification task. This verifies the very different conclusions from Theorems 7 and 11.

(i.e. regression or classification).

Augmentation bias is less impactful in classification than regression. A comparison of the generalization errors of biased and unbiased estimators in regression and classification, i.e., Theorems 4, 7, 9 and 11 respectively, reveals that the bias of an estimator has a much more benign effect on classification than regression. We plot the effect of augmentation bias on regression and classification in Fig. 3.3. We observe that the bias is mostly harmful for regression, especially in the overparameterized regime, but has no effect for classification. We also observe that it can be easier to choose augmentation parameters for classification (i.e., a larger range of parameters can lead to favorable performance).

Data augmentation is easier to tune in classification than regression. The results of [27] showed that the choice of test loss function critically impacts generalization. Specifically, they discovered that for the least squared estimator (LSE), there are cases where the model generalizes well for the classification task but not for regression. We complement this study by comparing generalization with DA in the two tasks. Specifically, we find that, for a given DA, the classification loss is always upper-bounded by a lower bound for the regression MSE, implying that regression is easier to train with DA than classification. Furthermore, we provide a concrete example of a simple class of augmentations for which the regression MSE is constant, but the classification POE is asymptotically zero. Our findings are summarized in the following proposition. The proof can be found in Appendix 3.8.4.

Proposition 20 (DA is easier to tune in classification than regression). *Consider the 1-sparse model* $\boldsymbol{\theta}^* = \sqrt{\frac{1}{\lambda_t}}\mathbf{e}_t$ *for Gaussian covariate with independent components and an independent feature augmentation. Suppose that the approximation error is not dominant in the bounds of Theorem 4 (simple sufficient conditions can be found in Lemma 35 in Appendix 3.8.1), and the assumptions in the two theorems hold. Then, we have*

$$\begin{aligned}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) &\lesssim \sqrt{(\lambda_{k+1}^{aug}\rho_k(\boldsymbol{\Sigma}_{aug};n))^2 \cdot \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{aug};n)} + \frac{k}{n}\right)\log n},\\ \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug}) &\gtrsim (\lambda_{k+1}^{aug}\rho_k(\boldsymbol{\Sigma}_{aug};n))^2 + \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{aug};n)} + \frac{k}{n}\right).\end{aligned}$$

As a consequence, the regression risk serves as a surrogate for the classification risk up to a log*-factor:*

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug})\sqrt{\log n}. \tag{3.33}$$

To illustrate the implications of this proposition, let us consider the isotropic Gaussian noise injection augmentation with noise standard deviation σ and random mask with dropout probability β to train the 1-sparse model with a decaying data spectrum $\boldsymbol{\Sigma}_{ii} = \gamma^i$, $\forall i \in \{1, 2, \ldots, p\}$, where γ is some constant satisfying $0 < \gamma < 1$. Let $\hat{\boldsymbol{\theta}}_{\mathrm{gn}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{rm}}$ be the corresponding estimators. Then, a direct consequence of Proposition 20 yields

$$\lim_{n\to\infty}\lim_{\sigma\to\infty}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{\mathrm{gn}}) = 0 \text{ while } \lim_{n\to\infty}\lim_{\sigma\to\infty}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\mathrm{gn}}) = 1. \tag{3.34}$$

Also, when $p\log n \ll n$,

$$\lim_{n\to\infty}\lim_{\beta\to 1}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{\mathrm{rm}}) = 0 \text{ while } \lim_{n\to\infty}\lim_{\beta\to 1}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\mathrm{rm}}) = 1. \tag{3.35}$$

(3.34) and (3.35) show that for both Gaussian noise injection and random mask augmentation, extreme augmentations can achieve perfect generalization in classification but poor generalization in regression.

It is worth noting that (3.34) in particular studies an augmentation that significantly changes the data distribution. In particular, for Gaussian injection augmentations we have

$$\frac{W_2^2(g(\mathbf{x}), \mathbf{x})}{p} \longrightarrow \infty \text{ as } n, \sigma \to \infty, \tag{3.36}$$

where W_2 denotes the 2-Wasserstein distance between the pre- and post-augmented distribution of the data by the Gaussian noise injection. In Figs. 3.6(b) and (d), we compare the Gaussian injection augmentation in the decaying spectrum $\gamma = 0.95$ for regression and classification, respectively. We observe a sharp difference between classification and regression, where, as we increase the augmentation intensity (i.e. variance of injected Gaussian noise), the MSE increases while the POE converges to a stable value (the ratio between SU and CN stays the same), implying that careful tuning is required for regression but not for classification.

Our second example that illustrates special benefits of DA in classification over regression concerns non-uniform random masking. The proof of the following proposition is deferred to Appendix 3.8.4.

Proposition 21 (**Non-uniform random mask is easier to tune in classification than regression**). *Consider the 1-sparse model* $\boldsymbol{\theta}^* = \sqrt{\frac{1}{\lambda_t}}\mathbf{e}_t$. *Suppose the approximation error is not dominant in the bounds of Theorem 4 (simple sufficient conditions can be found in Lemma 35in Appendix 3.8.1) and the assumptions in the two theorems hold. Suppose we apply the non-uniform random mask augmentation and recall the definitions of* ψ *and* ψ_t *as in Corollary 45. Then, if* $\sqrt{\frac{p}{n}} \ll \frac{\psi}{\psi_t} \ll \frac{p}{n}$, *we have*

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{rm}) \xrightarrow{n} 0 \ \textit{while} \ \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{rm}) \xrightarrow{n} 1. \tag{3.37}$$

By allowing the augmentation parameters ψ and ψ_1 to vary with n, the induced spectrum $\boldsymbol{\Sigma}_{\text{aug}}$

recovers the “bi-level” design, which was shown in Theorem 13 of [27] to separate classification and regression performance. It is worth noting that in the case where the true sparsity pattern is known, several augmentations (including nonuniform Gaussian noise injection, which incurs no approximation error in our analysis) can give rise to the same consistency behavior described above.

As a takeaway, the corollaries in this section demonstrate that the choice of augmentation itself can be more benign for classification tasks. Specifically, augmentations that are “biased on average” may perform similarly to their unbiased counterparts; we cannot generally expect this behavior for regression. Finally, we concluded this section by showing that our framework can be applied to strongly *out-of-distribution* (OOD) augmentations, and show that strong distributional shift can sometimes lead to improvements in classification generalization. We observe the same phenomenon empirically in Fig. 3.6, where increasing the intensity of the augmentation improves generalization and also increases the distributional shift.

3.6 Experiments

In this section, we complement our theoretical analysis with empirical investigations. In particular, we explore: 1. Differences between aSGD which is used in practice and the closed-form aERM solution analyzed in this paper, 2. Comparisons between the generalization of different types of augmentations studied in this work, 3. Multiple factors that influence the efficacy of DA, including signal structure and covariate spectrum, and 4. Comparisons between different augmentation strategies, namely precomputed augmentations versus aERM. We provide our Python implementations in https://github.com/nerdslab/augmentation-theory.

3.6.1 Convergence of aSGD solution to the closed-form solution

In this paper, we mathematically study a-ERM (the solution in Equation (3.2)); however, the solution used in practice is obtained by running a-SGD (Algorithm 1). In this set of experiments, we investigate the convergence of Algorithm 1 to the solution of Eq. 3.2 to verify that our theory reflects the solutions obtained in practice. To this end, we use an example in the overparameterized regime with $p = 128 \geq n = 64$ with the random isotropic signal $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and the observation noise $\epsilon \sim \mathcal{N}(0, 0.25)$. We choose a decaying covariate spectrum of the form $\boldsymbol{\Sigma}_{ii} \propto \gamma^i$, where γ is chosen such that $\mu_p(\boldsymbol{\Sigma}) = 0.6\mu_1(\boldsymbol{\Sigma})$. We want to understand the interplay between the convergence rate of aSGD with batch and augmentation size (formally, the augmentation size is the number of augmentations made for each draw of the training examples). We run the aSGD algorithm with different batch sizes and augmentation sizes in the range given by $(64, 1), (32, 2), \ldots, (2, 32), (1, 64)$. Note that the computation cost is proportional to the (batch size) × (augmentation size) per backward pass. Fig. 3.4 illustrates the convergence rate in terms of the number of backward passes. We observe that the convergence rates are fairly robust to different choices of batch and augmentation sizes.

Algorithm 1: Augmented Stochastic Gradient Descent (aSGD)

```
input : Data $\mathbf{x}_i$, $i = 1, \ldots, n$; Learning rates $\eta_t$, $t = 1, \ldots$; transformation
        distribution $\mathcal{G}$; batch size B; aug size H;
init $\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}_0$
while termination condition not satisfied do
    for k=1,...,$\frac{n}{B}$ do
        for i=1,...,B in the batch $\mathcal{B}_k$ do
            Draw H augmentations $g_{ij} \sim \mathcal{G}$, $j = 1, \ldots$,H
        $\hat{\boldsymbol{\theta}}_{t+1} \leftarrow \hat{\boldsymbol{\theta}}_t - \eta_t \sum_{i=1}^{B} \sum_{j=1}^{H} \nabla_{\boldsymbol{\theta}}(\langle \boldsymbol{\theta}, g_{ij}(\mathbf{x}_i) \rangle - y_i)_2^2|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_t}$
```

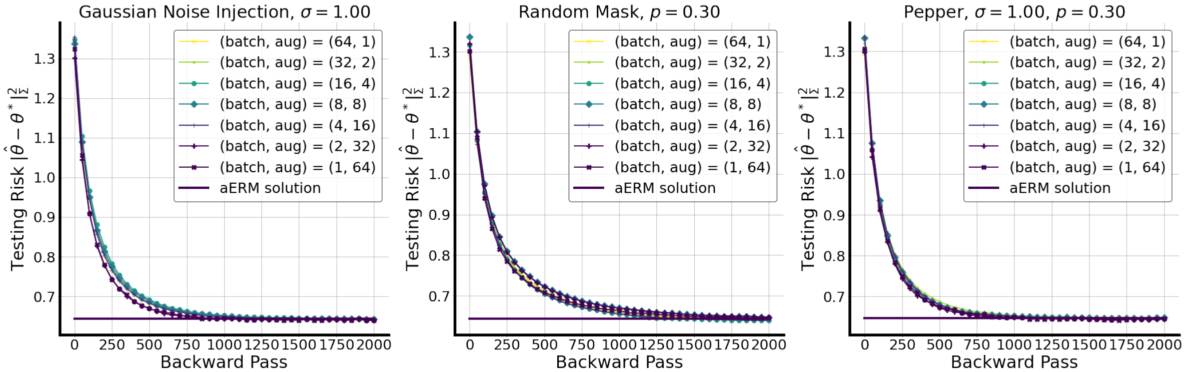


Figure 3.4: **Convergence of augmented stochastic gradient descent (a-SGD, Algorithm 1) as a function of the number of backward passes to the closed-form solution of the a-ERM objective (Equation** (3.2)**).** The result shows fairly stable convergence across different batch sizes and augmentation copies per sample.

A remark on the implicit bias of minimal or “weak" DA: It is well-known that Gaussian noise injection approximates the LSE when the variance of the added noise approaches zero. Surprisingly, however, this does not imply that all kinds of DA approach the LSE in the limit of decreasing augmentation intensity. Suppose that the augmentation g is characterized by some hyperparameter ξ that reflects the intensity of the augmentation (for e.g., mask probability β in the case of randomized mask, or Gaussian noise standard deviation σ in the case of Gaussian noise injection), and that $\mathrm{Cov}_G(\mathbf{X})/\xi \longrightarrow \mathrm{Cov}_\infty$ as $\xi \to 0$ for some positive semidefinite matrix Cov_∞ that does not depend on ξ. Then, the limiting estimator when the augmentation intensity ξ approaches zero is given by

$$\hat{\theta}_{aug} \xrightarrow{\xi \to 0} \mathrm{Cov}_\infty^{-1} \mathbf{X}^\top \left(\mathbf{X} \, \mathrm{Cov}_\infty^{-1} \mathbf{X}^\top\right)^\dagger \mathbf{y}. \tag{3.38}$$

It can be easily checked that this estimator is the minimum-Mahalanobis-norm interpolant of the training data where the positive semi-definite matrix used for the Mahalanobis norm is given by Cov_∞. Formally, the estimator solves the optimization problem

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\mathrm{Cov}_\infty} \text{ s.t. } \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \tag{3.39}$$

Thus, the choice of augmentation impacts the specific interpolator that we obtain in the limit of minimally applied DA. For example, the above formula can be applied to random mask with

$$\mathrm{Cov}_\infty = n^{-1}\mathrm{diag}(\mathbf{X}^T\mathbf{X}) \approx \boldsymbol{\Sigma}.$$

Fig. 3.5 demonstrates that the MSE of the random mask does not converge to that of the LSE. Instead, it converges to the light green curve which we abbreviate as M-LSE (for the *masked least squared estimator*). To test whether these limits appear only in an aERM solution, we plot the convergence path of aSGD with the random mask augmentation with masking probability $\beta = 0.01$. We set the ambient dimension p, noise standard deviation σ_ϵ, number of training examples n, and learning rate η to be 128, 0.5, 64 and 10^{-5} respectively. We choose a decaying covariate spectrum

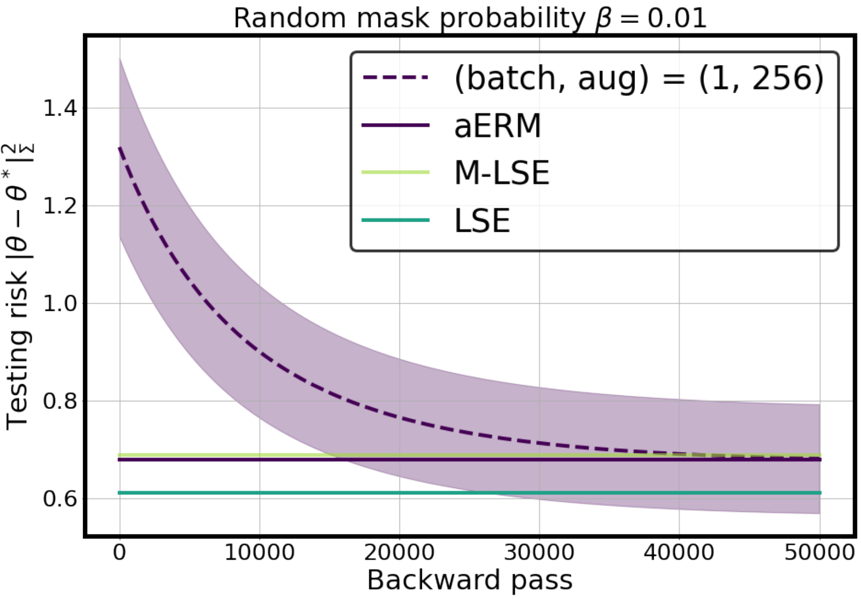


Figure 3.5: **aSGD convergence to aERM for small random mask.** We simulate the convergence of aSGD for random mask with dropout probability 0.01. We compare its converging estimator with the aERM limit (3.38)).

of the form $\mathbf{\Sigma}_{ii} \propto \gamma^i$, where γ is chosen such that $\mu_p(\mathbf{\Sigma}) = 0.2\mu_1(\mathbf{\Sigma})$. It is clear from the plot that both aSGD and aERM converges to the M-LSE solution of (3.38)). The curves and the shaded area denote the averaged result and the 90% confidence interval for 50 experiments. A caveat to this result is that the convergence rate turns out to be relatively slow and highly sensitive to the learning rate. A theoretical investigation of this behavior (and the optimization convergence of aSGD to aERM more generally) is beyond the scope of this work and would be interesting to explore in the future.

3.6.2 Comparisons of different types of augmentations

In this section, we compare the generalization of three canonical augmentations that we analyzed in this work: 1) Gaussian noise injection [91], 2) random mask [17], and 3) random rotation (which we introduced in Section 3.3.4). As in Section 3.6.1, we consider the random isotropic signal $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. We compare regression and classification tasks; in the former, we set the noise standard deviation as $\sigma_\varepsilon = 0.5$ while in the latter, we set the label noise parameter as $\nu^* = 0.1$. We consider diagonal covariance $\mathbf{\Sigma}$ and two choices of spectrum: 1. isotropic (i.e. $\mathbf{\Sigma} = \mathbf{I}_p$) and 2.

decaying spectrum where $\Sigma_{ii} \propto \gamma^i$ with $\gamma = 0.95$.

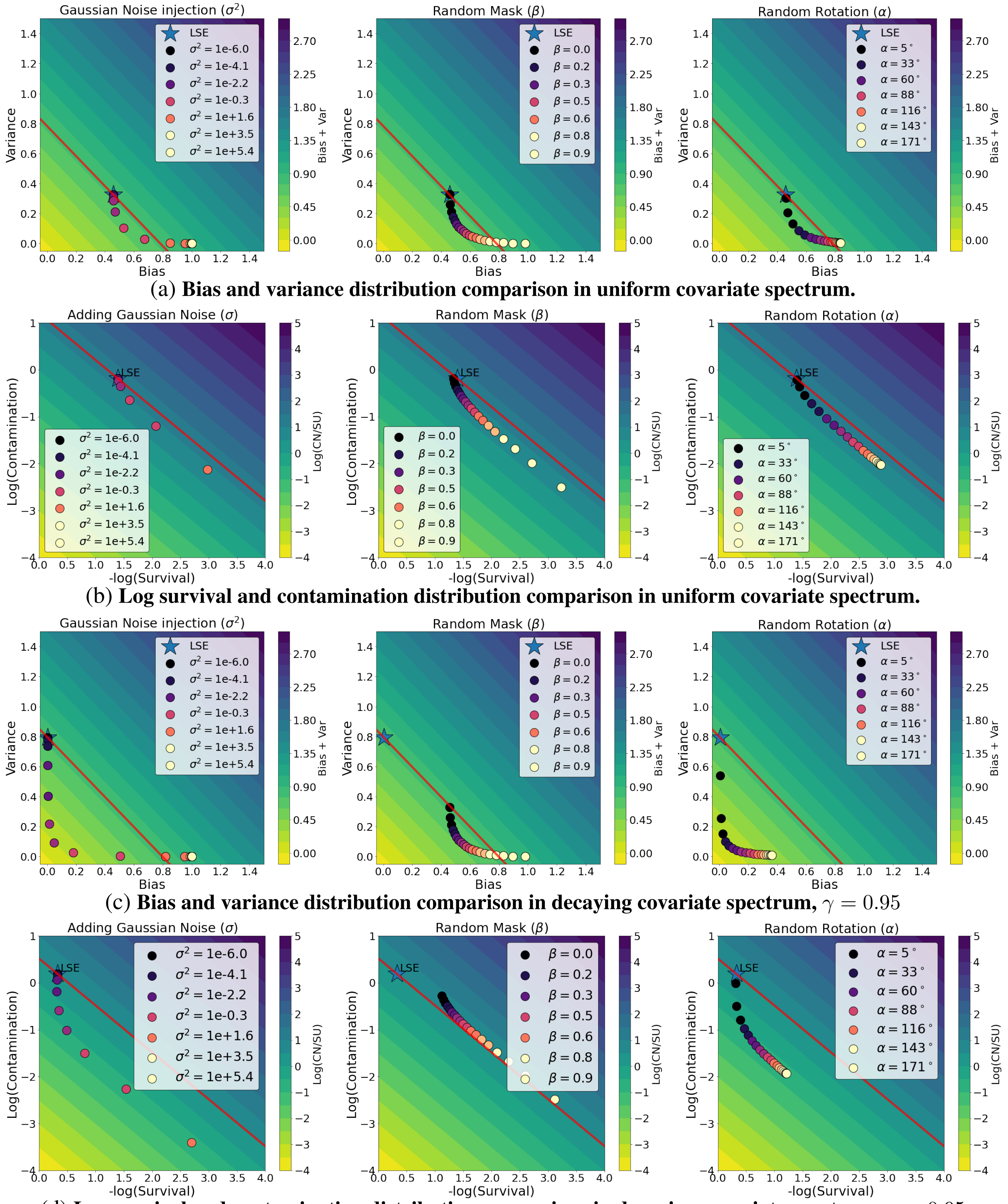
Figure 3.6 illustrates different trade-offs (bias/variance for regression, contamination/survival for classification) for the three canonical augmentations. The hyperparameters for the respective augmentations are: 1) the standard deviation $\sigma \in \mathbb{R}^+$ of the Gaussian noise injection, 2) the masking probability $\beta \in [0, 1]$ of the random mask, and 3) the rotation angle $\alpha \in [0, 90]$. We can make the following observations from Figure 3.6:

1. For isotropic data, all three augmentations achieve similar results in terms of generalization, while for the case of decaying spectrum, Gaussian injection and random rotation outperform the random mask when their respective hyperparameters are all optimally tuned.

2. In the regression task, for both choices of data distribution, Gaussian injection requires careful hyperparameter tuning in the range of $[0, 1.8]$ from $\mathbb{R}^+$ while the random mask and random rotation augmentations are fairly robust in performance in the entire range of the hyperparameter spaces. A possible explanation for this observation is that the random mask and rotation hyperparameters are *scale free* of the data (while the noise injection hyperparameter is not).

3. In the classification task, all the augmentations enjoy robust generalization guarantees with respect to their hyperparameters. This verifies our theoretical observations in Propositions 20 and 21.

4. While noise injection enjoys better generalization when it is optimally tuned, random mask is more robust in terms of generalization across hyperparameter choice. Our novel random rotation augmentation achieves the best of both worlds across different data distributions and tasks. In particular, it not only achieves a comparable generalization guarantee to noise injection when optimally tuned, but also is robust with respect to hyperparameter choice (like the random mask). This observation is consistent with the theoretical prediction of Corollary 19.



(a) **Bias and variance distribution comparison in uniform covariate spectrum.**

(b) **Log survival and contamination distribution comparison in uniform covariate spectrum.**

(c) **Bias and variance distribution comparison in decaying covariate spectrum,** $\gamma = 0.95$

(d) **Log survival and contamination distribution comparison in decaying covariate spectrum,** $\gamma = 0.95$

Figure 3.6: **Adding Gaussian Noise vs. Random Mask v.s Random Rotation in different covariate spectra for the regression and classification tasks.** In this figure we plot the bias/variance (a), (c) and contamination/survival distributions (b), (d) of Gaussian noise injection, random mask, and random rotation. The numbers reflect the respective hyperparameters σ, β, α.

3.6.3 When are augmentations effective?

In this section, we try to understand the impact of the true model $\boldsymbol{\theta}^*$ and the data covariance $\boldsymbol{\Sigma}$ on the efficacy of different augmentations, focusing on the nonuniform random mask introduced in Section 3.5.2. We set the ambient dimension to $p = 128$ and consider the noise standard deviation $\sigma_\epsilon = 0.5$.

Effect of true model We study the impact of nonuniform masking on the 1-sparse model $\boldsymbol{\theta}^* = \mathbf{e}_1$, as depicted in Section 3.4.1 in the regression task and consider isotropic covariance $\boldsymbol{\Sigma} = \mathbf{I}_p$. We vary the probability of the signal feature mask β_{sig} while keeping the probability of the noise feature mask β fixed at 0.2. The results are summarized in Fig. 3.7 and verify our analysis in Corollary 18 that noise features should be masked more compared to signal features so that the semantic component in the data is preserved. Furthermore, we observe that the differences manifest primarily in the bias, and the variance remains roughly the same. This is consistent with our variance bound in Corollary 18, which depends only on the probability of the noise mask β.

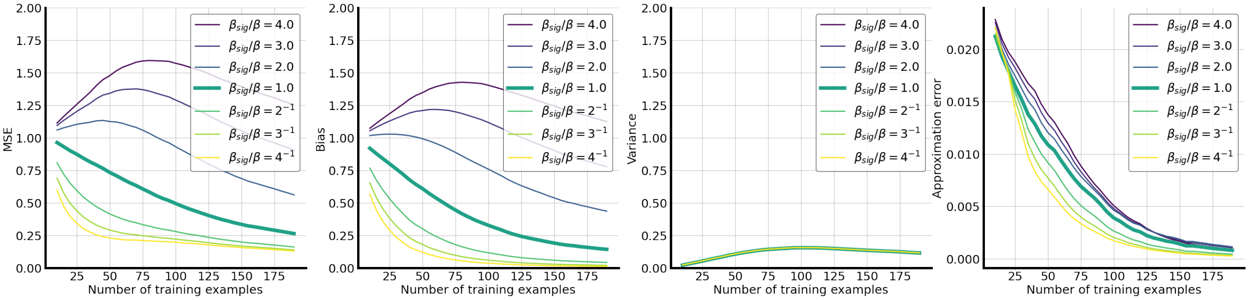


Figure 3.7: **Non-uniform random mask between signal and noise features.** We illustrate different mask strategies by varying the relative mask intensities of the signal and noise features. We see the signal position on the true model has most impact on the bias for data augmentation. Furthermore, the result supports the principle that one should augment the noise features more than the signal feature.

Effect of covariate spectrum Next, to understand the impact of the covariate spectrum, we consider a setting with a decaying data spectrum $\boldsymbol{\Sigma}_{ii} \propto 0.95^i$. We generate the true model using the random isotropic Gaussian $\boldsymbol{\theta}^* \sim \mathcal{N}(0, \mathbf{I}_p)$ and run the experiment 100 times, reporting the average result. We consider a *bilevel masking strategy* where the masking probability for the first half of features is set to β_1, and the second half of features is set to β_p. We vary the ratio between β_p and β_1 to investigate whether a feature with larger eigenvalue should be augmented with stronger intensity or not. The result is presented in Fig. 3.8. We observe from this figure that it is more beneficial to augment more for features with smaller eigenvalues.

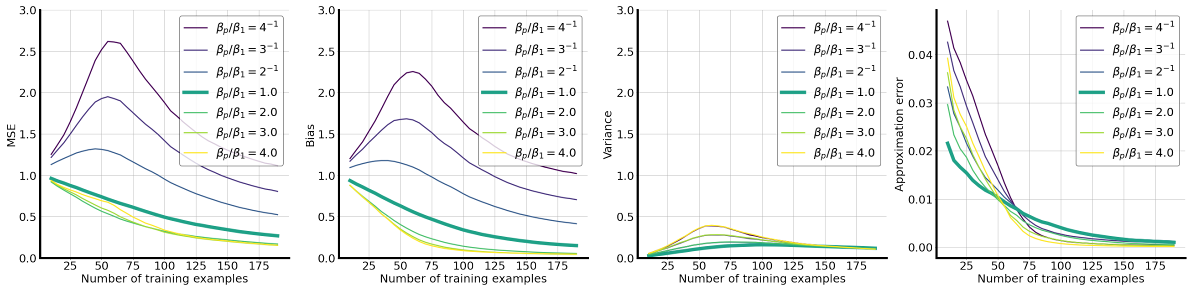


Figure 3.8: **Impact of covariance spectrum on the random mask in $p = 128$ dimensions.** We investigate the bi-level random mask strategies in data with decaying spectrum $\propto 0.95^i$. The first half of features are masked with probability β_1 while the rest are with β_p. We vary the ratio between the intensity β_p/β_1. We observe that augmenting more for features with higher variance benefits generalization.

3.6.4 Comparisons of pre-computing samples vs. augmented ERM

In our final set of experiments, we dig into the differences between pre-computing augmented samples and creating augmentations on the fly (our analysis concerns the latter). Because modern deep learning training usually relies on GPU computation, the overhead of doing augmentation with CPU on the fly with optimization could become a bottleneck during training time. Hence, pre-processing of the data becomes a plausible alternative. This is essentially equivalent to optimizing an *empirical* notion of aERM:

$$\frac{1}{\text{aug size}} \sum_{i=1}^{\text{aug size}} [\|G_i(\mathbf{X}) - \mathbf{y}\|_2^2] = \hat{\mathbb{E}}_G[\|G(\mathbf{X}) - \mathbf{y}\|_2^2]. \tag{3.40}$$

For this experiment, we generate isotropic random signal $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{128})$ and observation noise with standard deviation $\sigma = 0.5$. For simplicity, we choose the isotropic covariate spectrum $\mathbf{\Sigma} = \mathbf{I}_{128}$. Fig. 3.9 shows regression performance in terms of MSE, bias, and variance as we vary the augmentation size, which is the number of augmented copies of each original sample. In Figs. 3.9 (a)-(b), we observe the well-known double descent peaks [93, 127] when the training number approaches the ambient dimension $n = p = 128$ for LSE, and observe that adding pre-computed augmentation shifts these peaks to the left. The peak for a pre-computing method with an augmentation size k is observed to be approximately at $n = 128/k$. Intuitively, this mode of augmentation virtually increases the size of the training data: in particular, if we had $128/k$ original data points the induced total training size (including original data points and augmentations) becomes equal to $(128/k) \times k = 128$.

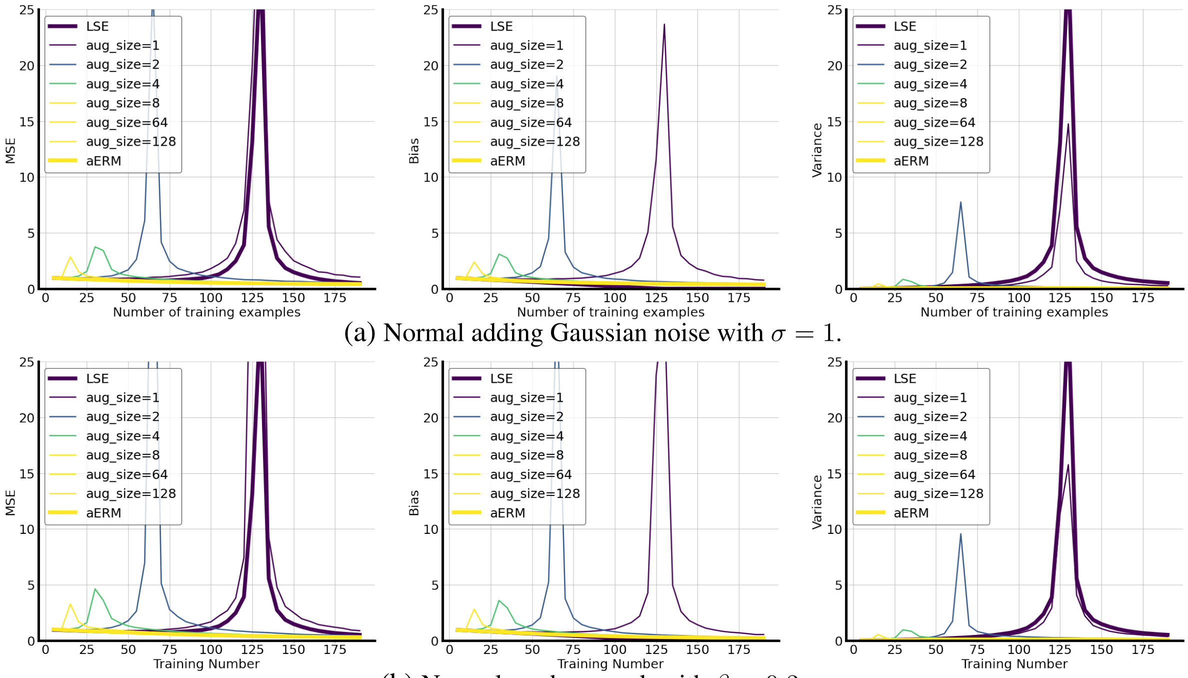
Interestingly, both the magnitude of the peak and the width decrease as we increase the augmentation size, and the peak almost disappears when $k > 8$. The general behavior of pre-computing is observed to approach aERM as k increases. Another interesting observation is that, unlike LSE which only has a double descent peak in the variance, pre-computing augmentations induces peaks in both the bias and the variance. A possible explanation for peaks appearing even in the bias term is that the *variance induced by a finite number of augmentations* is itself embedded in the bias term. In more detail: let $\hat{\boldsymbol{\theta}}_{\text{aug}} = \hat{\boldsymbol{\theta}}_{\text{aug}}(\varepsilon, g, \mathbf{X})$ be the augmentation estimator that depends on the observation noise ε, finite augmentation g, and training data $\mathbf{X}$. Then the bias term can be decomposed as

$$\underbrace{\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \mathbb{E}_{\varepsilon}[\hat{\boldsymbol{\theta}}_{\text{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Bias}} \lesssim \underbrace{\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \mathbb{E}_{g,\varepsilon}[\hat{\boldsymbol{\theta}}_{\text{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Average augmentation bias}} + \underbrace{\|\mathbb{E}_{\varepsilon}[\hat{\boldsymbol{\theta}}_{\text{aug}}|\mathbf{X}] - \mathbb{E}_{g,\varepsilon}[\hat{\boldsymbol{\theta}}_{\text{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Finite augmentation variance}} .$$

We defer a detailed mathematical study of these intriguing observations to future work.

3.7 Discussion

In this paper, we establish a new framework to analyze the generalization error of the linear model with data augmentation in underparameterized and overparameterized regimes. We characterize generalization error for both regression and classification tasks in terms of the interplay between the characteristics of the data augmentation and spectrum of the data covariance. As a side product, our results also generalize the recent line of research on *harmless interpolation* from ridge/ridgeless regression to settings where the learning objectives are penalized by data dependent regularizers. Through our analysis, we characterize when a DA can help or hurt generalization based on the



(a) Normal adding Gaussian noise with $\sigma = 1$.

(b) Normal random mask with $\beta = 0.3$.

Figure 3.9: **Pre-augmentation versus aERM in Gaussian Noise Injection and Random Mask.** The estimators based on aERM have monotonicity in generalization error with respect to the number of training samples, while the pre-computing methods exhibit the double-descent phenomenon like least-squared estimators. We note that the pre-computing methods shifts the error peak left compared with LSE. Also, the peak appears approximately at the sample number equals to $\frac{p}{k}$, where k is the augmentation size.

effective ranks of the augmented data spectrum. As concrete case studies, we show that in the overparametrized regime, random mask and group invariant augmentations can be harmful due to their *isotropizing* effect; on the other hand, our proposed random rotation augmentation is provably beneficial for generalization and highly robust. Our framework also uncovers the nuanced impact of DA on generalization as multiple factors, including the operating regimes, downstream tasks, and signal positions, come into play. We find that generalization can exhibit huge discrepancies even when the same type of DA has been employed.

There are several promising future directions that arise from our work. A first natural question is to what extent our insights extend to the nonlinear realm. A near-term future direction consists of extending our framework to kernel methods, which is often regarded as the first step toward understanding complex nonlinear models. As remarked at through various points in this paper, we believe the generalization analysis for a fixed estimator can be naturally extended; the more interesting question lies in the understanding the regularizer and estimator induced by DA for kernel models (that remain linear in feature-space but can be highly nonlinear in the data). Second, while our work focuses on the aERM objective, pre-computing augmentations is a natural alternative for which our preliminary experiments in Section 3.6.4 show intriguing differences in behavior. Understanding the fundamental differences between the two paradigms is an essential next step in comprehensively characterizing the effects of DA.

3.8 Technical Proofs and Derivations

We provide the detailed proofs in the following subsections.

3.8.1 General Auxiliary Lemmas

Notation For a data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with i.i.d. rows with covariance $\mathbf{\Sigma}$, recall we denote $\mathbf{P}^{\mathbf{\Sigma}}_{1:k-1}$ and $\mathbf{P}^{\mathbf{\Sigma}}_{k:\infty}$ as the projection matrices to the first $k-1$ and the remaining eigen-subspaces of $\mathbf{\Sigma}$, respectively. In addition, we have defined two effective ranks $\rho_k(\mathbf{\Sigma}; c) = \frac{c+\sum_{i>k}\lambda_i}{n\lambda_{k+1}}$, $R_k(\mathbf{\Sigma}; c) = \frac{(c+\sum_{i>k}\lambda_i)^2}{\sum_{i>k}\lambda_i^2}$. For convenience, we denote the residual Gram matrix by $\mathcal{A}_k(\mathbf{X}; \lambda) = \lambda\mathbf{I}_n + \mathbf{X}\boldsymbol{P}^{\mathbf{\Sigma}}_{k:\infty}\mathbf{X}^T$.

Lemma 22 (An useful identity for the ridge estimator [26]). *For any matrix $\mathbf{V} \in \mathbb{R}^{p \times k}$ composed of k independent orthonormal columns (therefore, $\mathbf{V}$ represents a k-dimensional subspace), the ridge estimator $\hat{\boldsymbol{\theta}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^\top\mathbf{X}^\top\mathbf{y}$ has the property:*

$$(\mathbf{I}_k + \mathbf{V}^\top\mathbf{X}^\top\mathbf{P}_k^{-1}\mathbf{X}\mathbf{V})\mathbf{V}^\top\hat{\theta} = \mathbf{V}^\top\mathbf{X}^\top\mathbf{P}_k^{-1}\mathbf{y}, \tag{3.41}$$

where $\mathbf{P}_k := \lambda\mathbf{I}_n + \mathbf{X}\mathbf{V}^\perp(\mathbf{V}^\perp)^\top\mathbf{X}^\top$ and $\mathbf{V}^\perp$ is a p by $p-k$ matrix satisfying $(\mathbf{V}^\perp)^\top\mathbf{V} = \mathbf{0}$ and $(\mathbf{V}^\perp)^\top\mathbf{V}^\perp = \mathbf{I}_{p-k}$.

Lemma 23 (Bernstein-type inequality for sum of sub-exponential variables). *Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be independent zero-mean sub-exponential random variables with sub-exponential norm at most σ_x^2. Then for every $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i\mathbf{x}_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\sigma_x^4\|a\|_2^2}, \frac{t}{\sigma_x^2\|a\|_\infty}\right)\right]$$

where $c > 0$ is an absolute constant.

Lemma 24 (Concentration of regularized truncated empirical covariance, Lemma 21 in [26]). *Suppose $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_p] \in \mathbb{R}^{n \times p}$ is a matrix with independent isotropic sub-gaussian rows with norm σ. Consider $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ for some positive non-increasing sequence $\{\lambda_i\}_{i=1}^p$.*

Denote $\mathbf{A}_k = \lambda\mathbf{I}_n + \sum_{i>k}\lambda_i\mathbf{z}_i\mathbf{z}_i^\top$ for some $\lambda \geq 0$. Suppose that it is known that for some $\delta, L > 0$ independent of n, p and some $k < n$ with probability at least $1 - \delta$, the condition number of the matrix $\mathbf{A}_k$ is at most L. Then, for some absolute constant c with probability at least $1 - \delta - 2\exp(-ct)$

$$\frac{(n - t\sigma^2)}{L}\lambda_{k+1}\rho_k(\boldsymbol{\Sigma}; \lambda) \leq \mu_n(\mathbf{A}_k) \leq \mu_1(\mathbf{A}_k) \leq \left(n + t\sigma^2\right)L\lambda_{k+1}\rho_k(\boldsymbol{\Sigma}; \lambda)$$

Lemma 25 (Concentration of leave-one-out empirical covariance). *Under the same notations and assumptions in Lemma 24, denote $\mathbf{A}_{-t} := \lambda\mathbf{I}_n + \sum_{i \neq t}\lambda_i\mathbf{z}_i\mathbf{z}_i^\top$ for some $\lambda \geq 0$. Then for any*

$t \leq k \leq n$ such that the condition number of $\mathbf{A}_k$ is bounded by L, we have

$$\frac{(n-t\sigma^2)}{L}\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda) \leq \mu_n(\mathbf{A}_{-t}) \leq \mu_1(\mathbf{A}_{-t}) \leq \left(n+t\sigma^2\right) L\lambda_1\rho_0(\mathbf{\Sigma};\lambda)$$

Proof The lemma follows by Lemma 24 and the observations of $\mu_1(\mathbf{A}_{-t}) \leq \mu_1(\mathbf{A}_0))$ and $\mathbf{A}_{-t} \succeq \mathbf{A}_k$. ■

Lemma 26 (Concentration of matrix with independent sub-gaussian rows, Theorem 5.39 in [128]). *Let $\mathbf{X}$ be an $n \times k$ matrix (with $n > k$) whose rows $\mathbf{x}_i$ are independent sub-gaussian isotropic random vectors in $\mathbb{R}^k$. Then for every $t \geq 0$ such that $\sqrt{n} - C\sqrt{k} - t > 0$ for some constant $C > 0$, we have with probability at least $1 - 2\exp\left(-ct^2\right)$ that*

$$\sqrt{n} - C\sqrt{k} - t \leq s_{\min}(\mathbf{X}) \leq s_{\max}(\mathbf{X}) \leq \sqrt{n} + C\sqrt{k} + t$$

Here $s_{\min}$ and $s_{\max}$ denotes the minimum and maximum singular values and $C, c > 0$ are some constants depend only on the sub-gaussian norm of the rows.

Lemma 27 (Concentration of the sum of squared norms, Lemma 17 in [26]). *Suppose $\mathbf{Z} \in \mathbb{R}^{n\times p}$ is a matrix with independent isotropic sub-gaussian rows with norm σ. Consider $\mathbf{\Sigma} = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_p\right)$ for some positive non-decreasing sequence $\{\lambda_i\}_{i=1}^p$. Then for some absolute constant c and any $t \in (0, n)$ with probability at least $1 - 2\exp(-ct)$*

$$\left(n - t\sigma^2\right)\sum_{i>k}\lambda_i \leq \sum_{i=1}^{n}\left\|\mathbf{\Sigma}_{k:\infty}^{1/2}\mathbf{Z}_{i,k:\infty}\right\|^2 \leq \left(n + t\sigma^2\right)\sum_{i>k}\lambda_i$$

Lemma 28 (Applications of Hanson-Wright inequality as done in [27]). *Let ε be a random vector composed of n i.i.d. zero-mean sub-gaussian variables with norm 1. Then,*

1. *there exists universal constant $c > 0$ such that for any fixed positive semi-definite matrix $\mathbf{A}$,*

with probability $1-2\exp(-\sqrt{n})$*, we have*

$$\left|\varepsilon^\top \mathbf{A}\varepsilon - \mathbb{E}\left[\varepsilon^\top \mathbf{A}\varepsilon\right]\right| \le c\|\mathbf{A}\|n^{\frac{3}{4}}.$$

2. there exists some universal constant $C > 0$ *such that with probability at least* $1-\frac{1}{n}$

$$\varepsilon^\top \mathbf{A}\varepsilon \le C\operatorname{tr}(\mathbf{A})\log n.$$

Lemma 29 (Operator norm bound of matrix with sub-gaussian rows [26]). *Suppose* $\{\mathbf{z}_i\}_{i=1}^n$ *is a sequence of independent sub-gaussian vectors in* $\mathbb{R}^p$ *with* $\|\mathbf{z}_i\| \le \sigma$*. Consider* $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ *for some positive non-decreasing sequence* $\{\lambda_i\}_{i=1}^p$ *. Denote* $\mathbf{X}$ *to be the matrix with rows* $\boldsymbol{\Sigma}^{1/2}\mathbf{z}_i$*. Then for some absolute constant c, for any* $t > 0$ *with probability at least* $1-4e^{-t/c}$

$$\|\mathbf{X}\| \le c\sigma\sqrt{\lambda_1(t+n) + \sum_{j=1}^p \lambda_j}.$$

3.8.2 Proofs of Regression Results

In this section, we will include essential lemmas in 3.8.2 to prove the main theorems for regression analysis in the sections 3.8.2 and 3.8.2. Then, we will use these theorems to prove the propositions and corollaries in sections 3.8.2 and 3.8.2, respectively.

Regression Lemmas

Lemma 30 (Sharpened bias of ridge regression, extension of [26]).

$$\frac{\text{Bias}}{C_x L_1^4} \lesssim \left\|\mathbf{P}_{k_1+1:p}^{\boldsymbol{\Sigma}}\theta^*\right\|_{\boldsymbol{\Sigma}}^2 + \left\|\mathbf{P}_{1:k_1}^{\boldsymbol{\Sigma}}\theta^*\right\|_{\boldsymbol{\Sigma}^{-1}}^2 \frac{\rho_{k_1}^2(\boldsymbol{\Sigma}; n)}{(\lambda_{k_1+1})^{-2} + (\lambda_1)^{-2}\rho_{k_1}^2(\boldsymbol{\Sigma}; n)} \tag{3.42}$$

Remark 31. *The reason we modify the bound from [26] is twofold: 1. we consider non-diagonal covariance matrix* $\boldsymbol{\Sigma}$*. This is because even if the original data covariance is diagonal, the equivalent* *spectrum might become non-diagonal after the data augmentation. Therefore, we modify the bound so that the eigenspaces of the data covariance matrix do not have to be aligned with the standard basis. 2. As we show in our work, some augmentations, e.g. random mask, have the effect of making the equivalent data spectrum isotropic. However, in this case, the bias bound in [26], as shown below, can be vacuous as being almost the same as the null estimator so we modify the bound to remedy the case.*

$$\begin{aligned}\text{Bias bound} &\asymp \left\|\mathbf{P}^{\mathbf{\Sigma}}_{k_1+1:p}\theta^*\right\|^2_{\mathbf{\Sigma}} + \left\|\mathbf{P}^{\mathbf{\Sigma}}_{1:k_1}\theta^*\right\|^2_{\mathbf{\Sigma}^{-1}} \lambda^2_{k_1+1}\rho^2_{k_1}(\mathbf{\Sigma}; n) \\ &= \left\|\mathbf{P}^{\mathbf{\Sigma}}_{k_1+1:p}\theta^*\right\|^2 + \left\|\mathbf{P}^{\mathbf{\Sigma}}_{1:k_1}\theta^*\right\|^2 \frac{p-k_1}{n} \gtrsim \|\boldsymbol{\theta}^*\|_2^2,\end{aligned}$$

Proof This lemma is a modification to Theorem 1 in [26], where we only change slightly in the estimation of the lower tail of the bias. For self-containment, we illustrate where we make the change. Consider the diagonalization $\mathbf{\Sigma} = \mathbf{V}\mathcal{D}\mathbf{V}^\top$. Let $\mathbf{V}_1,\ \mathbf{V}_2$ be the matrices with columns consisting of the top k eigenvectors of $\mathbf{\Sigma}$ and the remaining eigenvectors, respectively. Note that we have $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$, $\boldsymbol{P}^{\mathbf{\Sigma}}_{1:k-1} = \mathbf{V}_1\mathbf{V}_1^\top$, and $\boldsymbol{P}^{\mathbf{\Sigma}}_{k:\infty} = \mathbf{V}_2\mathbf{V}_2^\top$. Moreover, we have $\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$. Now, for the ridge estimator $\hat{\theta} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$, apply Lemma 22 with $\mathbf{V} = \mathbf{V}_1$ to obtain

$$(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathcal{A}_k(\mathbf{\Sigma};\lambda)^{-1}\mathbf{X}\mathbf{V}_1)\mathbf{V}_1^\top\hat{\theta} = \mathbf{V}_1^\top\mathbf{X}^\top\mathcal{A}_k(\mathbf{\Sigma};\lambda)^{-1}\mathbf{y}, \tag{3.43}$$

where $\mathcal{A}_k(\mathbf{\Sigma};\lambda) := \lambda\mathbf{I}_p + \mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top$. As there will be no ambiguity of which covariance matrix the residual spectrum corresponds to, we will just write $\mathbf{A}_k$ from now on.

To bound the bias, we split it into

$$\text{Bias} \le 2\|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*)\|^2_{\mathbf{\Sigma}} + 2\|\mathbf{V}_2\mathbf{V}_2^\top(\mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*)\|^2_{\mathbf{\Sigma}}, \tag{3.44}$$

where the expectations are over the noise ε. Observe that the averaged estimator is $\mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}] = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$, so we can apply Lemma 22 with $\hat{\boldsymbol{\theta}}$ and $\mathbf{y}$ replaced by $\mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}]$ and $\mathbf{X}\boldsymbol{\theta}^*$, respectively. As a result, we can write

$$\begin{aligned}(\mathbf{I}_k + \mathbf{V}_1^\top \mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)\mathbf{V}_1^\top \mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}] &= \mathbf{V}_1^\top \mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}\boldsymbol{\theta}^* \\ &= \mathbf{V}_1^\top \mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}(\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top)\boldsymbol{\theta}^*.\end{aligned}$$

Now, subtracting $\mathbf{V}_1^\top\boldsymbol{\theta}^* + \mathbf{V}_1^\top \mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*$ from both sides of the above equation followed by a left multiplication of $\mathbf{V}_1$ gives

$$\begin{aligned}&\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \mathbf{V}_1\mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= \mathbf{V}_1\mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^* - \mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*,\end{aligned}$$

where we use the identity $\mathbf{I}_p = \mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top$.

Now multiply both sides with $(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top$, the R.H.S. is

$$\begin{aligned}&= (\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^* - (\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}^* \\ &\le \|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}\mu_n(\mathbf{A}_k)^{-1}\sqrt{\mu_1\left(\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\mathbf{V}_1^\top\right)}\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\| \\ &+ \|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}\|\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^{-1}}.\end{aligned} \tag{3.45}$$

Note that in the last term of the inequality, we have use the fact that

$$\begin{aligned}(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}^* &= (\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}(\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top)\boldsymbol{\theta}^* \\ &= (\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*.\end{aligned}$$

On the other hand, the L.H.S. is

$$\ge \lambda_1^{-1}\|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 + (\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\mathbf{V}_1\mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{3.46}$$

in which the second term is

$$
\begin{aligned}
&= (\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{V}_1\mathbf{V}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{V}_1\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\mathbf{V}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{V}_1\mathbf{V}_1^\top (\mathbf{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\
&\geq \|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \|\mathbf{V}_1\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\mathbf{V}_1^\top\| \\
&\geq \|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \mu_k(\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top \mathbf{A}_k^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1) \\
&\geq \|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \mu_1(\mathbf{A}_k)^{-1} \mu_k(\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top \mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1). \qquad (3.47)
\end{aligned}
$$

Therefore, combining e.q. (3.45), (3.46) and (3.47), we have

$$
\begin{aligned}
&\|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}} \\
&\leq \frac{\mu_n^{-1}(\mathbf{A}_k)\sqrt{\mu_1\left(\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top \mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\right)}\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\| + \|\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^{-1}}}{\lambda_1^{-1} + \mu_1^{-1}(\mathbf{A}_k)\mu_k(\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top \mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1)}.
\end{aligned}
$$

Now, we turn to bound $\|\mathbf{V}_2\mathbf{V}_2^\top(\mathbf{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2$. The proof follows the same step as [26] except we use projection matrices to accommodate for the non-diagonal covariance:

$$
\begin{aligned}
\|\mathbf{V}_2\mathbf{V}_2^\top(\mathbf{E}_\varepsilon\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \lesssim \underbrace{\|\boldsymbol{P}_{k:\infty}^{\boldsymbol{\Sigma}}\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2}_{T_1} + \underbrace{\|\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2}_{T_2} \\
+ \underbrace{\|\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2}_{T_3}
\end{aligned}
$$

T_2 is bounded by

$$
\mu_n^{-2}(\mathbf{A}_k)\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2. \qquad (3.48)
$$

For T_3 on the other hand, recall $\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_p = \mathbf{X}\mathbf{V}_1\mathbf{V}_1^\top\mathbf{X}^\top + \mathbf{A}_k$. Then by the

Sherman–Morrison–Woodbury formula, we have

$$\begin{aligned}
&(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_p)^{-1}\mathbf{X}\mathbf{V}_1 \\
&= \left(\mathbf{A}_k^{-1} - \mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}\mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\right)\mathbf{X}\mathbf{V}_1 \\
&= \mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_p)^{-1}\mathbf{X}\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 \\
\leq\ & \mu_n^{-2}(\mathbf{A}_k)\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|\|\mathbf{X}\mathbf{V}_1(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_2^2,
\end{aligned}$$

where

$$\begin{aligned}
&\mathbf{X}\mathbf{V}_1(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}\mathbf{V}_1^\top\boldsymbol{\theta}^* \\
&\overset{(a)}{=} \mathbf{X}\mathbf{V}_1(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2})(\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1)(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2})(\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1)\mathbf{V}_1^\top\boldsymbol{\theta}^* \\
&\overset{(b)}{=} \mathbf{X}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1)(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2})(\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1)\mathbf{V}_1^\top\boldsymbol{\theta}^* \\
&\overset{(c)}{=} \mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1}\mathbf{V}_1 + \mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1)^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*,
\end{aligned}$$

where (a) follows from $\mathbf{V}_1^\top\mathbf{V}_1 = \mathbf{I}_k$, (b) from

$$\mathbf{X}\mathbf{V}_1(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2})(\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1) = \mathbf{X}(\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1 = \mathbf{X}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1$$

as $\mathbf{V}_1^\top\mathbf{V}_2 = 0$ and $\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top = \mathbf{I}_p$, and (c) follows from the facts

$$\begin{aligned}
&\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1 = \mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\left(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1\right) \\
&(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{1/2}\mathbf{V}_1)^{-1} = \mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1.
\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
&\|\mathbf{X}\mathbf{V}_1(\mathbf{I}+\mathbf{V}_1^\top\mathbf{X}^\top\mathbf{A}_k^{-1}\mathbf{X}\mathbf{V}_1)^{-1}\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_2^2\\
&\le \frac{\mu_1\left(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\right)}{\lambda_1^{-2}+\mu_1^{-2}(\mathbf{A}_k)\mu_k^2(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1)}\|\boldsymbol{P}^{\boldsymbol{\Sigma}}_{1:k-1}\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^{-1}}.
\end{aligned}
$$

Now, adding all the terms above together, the bias is

$$
\begin{aligned}
\text{Bias} \lesssim\ & \frac{\mu_n^{-2}(\mathbf{A}_k)\mu_1\left(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\right)\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\|_2^2+\|\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^{-1}}^2}{\lambda_1^{-2}+\mu_1^{-2}(\mathbf{A}_k)\mu_k^2(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1)}\\
&+\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|\frac{\mu_n^{-2}(\mathbf{A}_k)\mu_1\left(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\right)\|\mathbf{V}_1\mathbf{V}_1^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^{-1}}^2}{\lambda_1^{-2}+\mu_1^{-2}(\mathbf{A}_k)\mu_k\left(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\right)^2}\\
&+\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|\mu_n^{-2}(\mathbf{A}_k)\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2+\|\boldsymbol{P}^{\boldsymbol{\Sigma}}_{k:\infty}\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2,
\end{aligned}
$$

where for the diagonal covariance $\boldsymbol{\Sigma}$, the first two terms are sharpened with additional λ_1^{-2} in the denominators as compared to [26]. As in [26], these terms can be bounded by concentration bounds: $\mu_i\left(\mathbf{V}_1^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}^{-1/2}\mathbf{V}_1\right)$ by Lemma 26, $\mu_j(\mathbf{A}_k)$ by Lemma 24, $\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2\|_2^2$ and $\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|$ by Lemma 27. The details can be found in the proof of MSE bound of [26]. ■

Lemma 32 (Variance bound of ridge regression for non-diagonal covariance data [26]). *Consider the regression task with the model setting in Section 3 where the input variable* $\mathbf{x}$ *possibly has non-diagonal covariance* $\boldsymbol{\Sigma}$ *with eigenvalues* $\lambda_1 \ge \lambda_2 \dots \lambda_p$*. Given a ridge estimator* $\hat{\theta} = (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ *and* $\lambda \ge 0$*, if we know that for some* k_2*, the condition number of* $\mathcal{A}_{k_2}(\mathbf{X};\lambda)$ *is bounded by* L_2 *with probability* $1-\delta$*, where* $\delta < 1-\exp(-n/c_x^2)$*, then there exists some constant* $\tilde{C}_x$ *depending only on* σ_x *such that with probability at least* $1-\delta-n^{-1}$*,*

$$
\frac{\text{Variance}}{\sigma_\varepsilon^2 L_2^2\tilde{C}_x} \lesssim \left(\frac{k_2}{n}+\frac{n}{R_{k_2}(\boldsymbol{\Sigma};n)}\right)\log n. \tag{3.49}
$$

Lemma 33 (Generalization bound of ridge regression for non-diagonal covariance data, extension of [26]). *Consider the regression task with the model setting in Section 3 where the input variable $\mathbf{x}$ has possibly non-diagonal covariance $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \dots$. Then, given a ridge regression estimator $\hat{\theta} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ and $\lambda \geq 0$, suppose we know that for some k_1 and k_2, the condition numbers of $\mathcal{A}_{k_1}(\mathbf{X}; \lambda)$ and $\mathcal{A}_{k_2}(\mathbf{X}; \lambda)$ are bounded by L_1 and L_2 with probability $1 - \delta$, where $\delta < 1 - \exp(-n/c_x^2)$, then there exists some constants $C_x, \tilde{C}_x$ depending only on σ_x such that with probability at least $1 - n^{-1}$,*

$$\begin{aligned}\mathrm{MSE} \lesssim & \underbrace{C_x L_1^4 \left(\|\mathbf{P}^{\mathbf{\Sigma}}_{k_1+1:p} \theta^*\|^2_{\mathbf{\Sigma}} + \|\mathbf{P}^{\mathbf{\Sigma}}_{1:k_1} \theta^*\|^2_{\mathbf{\Sigma}^{-1}} \frac{\rho^2_{k_1}(\mathbf{\Sigma}; n)}{(\lambda_{k_1+1})^{-2} + (\lambda_1)^{-2} \rho^2_{k_1}(\mathbf{\Sigma}; n)} \right)}_{\text{Bias}} \\ & + \underbrace{\sigma_\varepsilon^2 L_2^2 \tilde{C}_x \left(\frac{k_2}{n} + \frac{n}{R_{k_2}(\mathbf{\Sigma}; n)} \right) \log n}_{\text{Variance}} \end{aligned} \tag{3.50}$$

Proof The statement is a direct combination of Lemma 30, 32 and the bias-variance decomposition of MSE from [26]. ∎

Lemma 34 (Bounds on the approximation error for regression). *Denote*

$$\hat{\boldsymbol{\theta}}_{aug} := (\mathbf{X}^\top \mathbf{X} + n \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1} \mathbf{X}^\top \mathbf{y}, \;\; \bar{\boldsymbol{\theta}}_{aug} := (\mathbf{X}^\top \mathbf{X} + n \mathbb{E}_{\mathbf{x}} \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))^{-1} \mathbf{X}^\top \mathbf{y},$$

and κ the condition number of $\mathbf{\Sigma}_{aug}$. Assume for some constant $c < 1$ that

$$\Delta_G := \|\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}\| \leq c.$$

Then the approximation error is bounded by,

$$\|\hat{\boldsymbol{\theta}}_{aug} - \bar{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}} \lesssim \kappa^{\frac{1}{2}} \Delta_G \left(\|\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}} + \sqrt{\mathrm{Bias}(\bar{\boldsymbol{\theta}}_{aug})} + \sqrt{\mathrm{Variance}(\bar{\boldsymbol{\theta}}_{aug})} \right).$$

Proof *For ease of notation, we denote* $\mathcal{D} = \mathrm{Cov}_{\mathcal{G}}$*,* $\bar{\mathcal{D}} = \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$*, and* $\boldsymbol{\Delta} = \bar{\mathcal{D}}^{-\frac{1}{2}}\mathcal{D}\bar{\mathcal{D}}^{-\frac{1}{2}} - \mathbf{I}$*. Then*

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{aug} - \bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} &= \|(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\mathbf{X}^\top\mathbf{y} - (\mathbf{X}^\top\mathbf{X} + n\bar{\mathcal{D}})^{-1}\mathbf{X}^\top\mathbf{y}\|_{\boldsymbol{\Sigma}} \\
&= \|(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}(\mathbf{X}^\top\mathbf{X} + n\bar{\mathcal{D}} - \mathbf{X}^\top\mathbf{X} - n\mathcal{D})(\mathbf{X}^\top\mathbf{X} + n\bar{\mathcal{D}})^{-1}\mathbf{X}^\top\mathbf{y}\|_{\boldsymbol{\Sigma}} \\
&= n\|\boldsymbol{\Sigma}^{\frac{1}{2}}\bar{\mathcal{D}}^{-\frac{1}{2}}\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\boldsymbol{\Delta}\bar{\mathcal{D}}^{\frac{1}{2}}\bar{\boldsymbol{\theta}}_{aug}\|_2, \\
&\lesssim n\|\boldsymbol{\Sigma}^{\frac{1}{2}}\bar{\mathcal{D}}^{-\frac{1}{2}}\|\|\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\|\|\boldsymbol{\Delta}\|\|\bar{\mathcal{D}}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|\|\bar{\boldsymbol{\theta}}_{aug}\|_2 \\
&\lesssim n\kappa^{\frac{1}{2}}\Delta_G\|\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}}\|\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\| \qquad (3.51)
\end{aligned}$$

By (3.57), $\|\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}}$ *can be bounded as,*

$$\|\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} \leq \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} + \|\bar{\boldsymbol{\theta}}_{aug} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} \lesssim \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} + \sqrt{\mathrm{Bias}(\bar{\boldsymbol{\theta}}_{aug})} + \sqrt{\mathrm{Variance}(\bar{\boldsymbol{\theta}}_{aug})}.$$

It remains to bound $\|\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\|$*.*

Now, observe

$$\begin{aligned}
&\|\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\| = \left(\mu_p\left(\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\right)^{-1}\right)^{-1} \\
= &\left(\mu_p\left(\bar{\mathcal{D}}^{-\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})\bar{\mathcal{D}}^{-\frac{1}{2}}\right)\right)^{-1} \\
\leq &\left(\mu_p\left(\bar{\mathcal{D}}^{-\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\bar{\mathcal{D}})\bar{\mathcal{D}}^{-\frac{1}{2}}\right) - \|\bar{\mathcal{D}}^{-\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\bar{\mathcal{D}} - \mathbf{X}^\top\mathbf{X} - n\mathcal{D})\bar{\mathcal{D}}^{-\frac{1}{2}}\|\right)^{-1}.
\end{aligned}$$

However,

$$\left(\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\bar{\mathcal{D}})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\right)^{-1} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + n\mathbf{I}),$$

where $\tilde{\mathbf{X}}$ *has sub-gaussian rows with covariance* $\boldsymbol{\Sigma}_{aug}$*. Hence, the first term is at least* n*, while the* *second term is just $n\Delta_G$ by definition. So by the assumption that $\Delta_G < c$ for some $c < 1$, we have,*

$$\|\bar{\mathcal{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathcal{D})^{-1}\bar{\mathcal{D}}^{\frac{1}{2}}\| \lesssim \frac{1}{n},$$

and finally we have,

$$\|\hat{\boldsymbol{\theta}}_{aug} - \bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} \lesssim \kappa^{\frac{1}{2}}\Delta_G \left(\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} + \sqrt{\mathrm{Bias}(\bar{\boldsymbol{\theta}}_{aug})} + \sqrt{\mathrm{Variance}(\bar{\boldsymbol{\theta}}_{aug})}\right).$$

■

Lemma 35 (Condition on bias/variance dominating error approximation). *Suppose the conditions of Theorem 4 hold. If*

$$\kappa^{\frac{1}{2}}\Delta_G \overset{n}{\ll} \min\left(\mathrm{Bias} + \mathrm{Variance}, \sqrt{\mathrm{Bias} + \mathrm{Variance}}\right). \tag{3.52}$$

Then there exists $c'' > 0$ such that,

$$\frac{1}{c''} \leq \frac{\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{aug}) + \mathrm{Variance}(\hat{\boldsymbol{\theta}}_{aug})}{\mathrm{Bias}(\bar{\boldsymbol{\theta}}_{aug}) + \mathrm{Variance}(\bar{\boldsymbol{\theta}}_{aug})} \leq c''. \tag{3.53}$$

Proof The lemma follows from Theorem 4 with the observations:

$$\kappa^{\frac{1}{2}}\Delta_G \left(\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} + \sqrt{\mathrm{Bias}(\bar{\boldsymbol{\theta}}_{\mathrm{aug}})} + \sqrt{\mathrm{Variance}(\bar{\boldsymbol{\theta}}_{\mathrm{aug}})}\right) \overset{n}{\ll} \mathrm{Bias}(\bar{\boldsymbol{\theta}}_{\mathrm{aug}}) + \mathrm{Variance}(\bar{\boldsymbol{\theta}}_{\mathrm{aug}})$$

■

Proof of Theorem 4

Theorem 4 (Bounds of Mean-Squared Error for Regression). *Consider an unbiased data augmentation g and its corresponding estimator $\hat{\boldsymbol{\theta}}_{aug}$. Recall the definition*

$$\Delta_G := \|\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}_p\|,$$

and let κ be the condition number of $\boldsymbol{\Sigma}_{aug}$. Assume with probability $1-\delta'$, we have that the condition numbers for the matrices $\mathcal{A}_{k_1}(\mathbf{X}_{aug}; n)$, $\mathcal{A}_{k_2}(\mathbf{X}_{aug}; n)$ are bounded by L_1 and L_2 respectively, and that $\Delta_G \le c'$ for some constant $c' < 1$. Then there exist some constants c, C depending only on σ_x and σ_ε, such that, with probability $1 - \delta' - 4n^{-1}$, the testing mean-squared error is bounded by

$$\begin{aligned}
\mathrm{MSE} &\lesssim \mathrm{Bias} + \mathrm{Variance} + \mathrm{ApproximationError},\\
\frac{\mathrm{Bias}}{C_x L_1^4} &\lesssim \left(\left\|\mathbf{P}^{\boldsymbol{\Sigma}_{aug}}_{k_1+1:p}\theta^*_{aug}\right\|^2_{\boldsymbol{\Sigma}_{aug}} + \left\|\mathbf{P}^{\boldsymbol{\Sigma}_{aug}}_{1:k_1}\theta^*_{aug}\right\|^2_{\boldsymbol{\Sigma}^{-1}_{aug}} \frac{(\rho^{aug}_{k_1})^2}{(\lambda^{aug}_{k_1+1})^{-2} + (\lambda^{aug}_1)^{-2}(\rho^{aug}_{k_1})^2}\right),\\
\frac{\mathrm{Variance}}{\sigma_\varepsilon^2 L_2^2 \tilde{C}_x} &\lesssim \left(\frac{k_2}{n} + \frac{n}{R^{aug}_k}\right)\log n, \quad \mathrm{Approx.Error} \lesssim \kappa^{\frac{1}{2}}\Delta_G\left(\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} + \sqrt{\mathrm{Bias} + \mathrm{Variance}}\right),
\end{aligned}$$

where $\rho^{aug}_k := \rho_k(\boldsymbol{\Sigma}_{aug}; n)$ and $R^{aug}_k := R_k(\boldsymbol{\Sigma}_{aug}; n)$.

Proof

$$\mathrm{MSE} = \mathbb{E}_{\mathbf{x}}[(\mathbf{x}^\top(\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*))^2 | \mathbf{X}, \varepsilon] = \|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}. \tag{3.54}$$

Because the possible dependency of $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ on $\mathbf{X}$, we approximate the $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ with the estimator $\bar{\boldsymbol{\theta}}_{\mathrm{aug}} := (\mathbf{X}^\top\mathbf{X} + n\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1}\mathbf{X}^\top\mathbf{y}$. Now, by the triangle inequality, the MSE can be bounded as

$$\mathrm{MSE} \le 2\|\bar{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}} + 2\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|^2_{\boldsymbol{\Sigma}} \tag{3.55}$$

We can bound the first term by using its connection to ridge regression:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{aug}} &= (\mathbf{X}^\top\mathbf{X} + n\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1}\mathbf{X}^\top\mathbf{y} \\
&= \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}(n\mathbf{I}_p + \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\mathbf{X}^\top\mathbf{X}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2})^{-1}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\mathbf{X}^\top\mathbf{y} \\
&= \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}(n\mathbf{I}_p + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{y} \quad (\tilde{\mathbf{X}} := \mathbf{X}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}) \\
&= \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\hat{\theta}_{\text{ridge}}, \quad (\hat{\boldsymbol{\theta}}_{\text{ridge}} := (n\mathbf{I}_p + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{y}).
\end{aligned} \tag{3.56}
$$

So the MSE becomes $\|\hat{\boldsymbol{\theta}}_{\text{ridge}} - \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{1/2}\boldsymbol{\theta}^*\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}}$. These observations have shown an approximate equivalence to a ridge estimator with data matrix $\tilde{\mathbf{X}}$, which has data covariance = $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}$, ridge intensity $\lambda = n$, and true model parameter $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{1/2}\boldsymbol{\theta}^*$. Hence, we can apply Lemma 33 to bound $\|\bar{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$, where $\|\mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\text{aug}}] - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$ and $\|\mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\text{aug}}] - \bar{\boldsymbol{\theta}}_{\text{aug}}\|^2_{\boldsymbol{\Sigma}}$ are exactly the bias and variance, in Theorem 4, respectively. Specifically, we have,

$$
\begin{aligned}
&\|\mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\text{aug}}] - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}} \lesssim \\
&C_x L_1^4 \left(\left\| \mathbf{P}^{\boldsymbol{\Sigma}_{\text{aug}}}_{k_1+1:p} \theta^*_{\text{aug}} \right\|^2_{\boldsymbol{\Sigma}_{\text{aug}}} + \left\| \mathbf{P}^{\boldsymbol{\Sigma}_{\text{aug}}}_{1:k_1} \theta^*_{\text{aug}} \right\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{aug}}} \frac{\rho^2_{k_1}(\boldsymbol{\Sigma}_{\text{aug}}; n)}{(\lambda^{\text{aug}}_{k_1+1})^{-2} + (\lambda^{\text{aug}}_1)^{-2}\rho^2_{k_1}(\boldsymbol{\Sigma}_{\text{aug}}; n)} \right), && (3.57) \\
&\|\mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\text{aug}}] - \bar{\boldsymbol{\theta}}_{\text{aug}}\|^2_{\boldsymbol{\Sigma}} \lesssim \ \sigma^2_\varepsilon t L_2^2 \tilde{C}_x \left(\frac{k_2}{n} + \frac{n}{R_{k_2}(\boldsymbol{\Sigma}_{\text{aug}}; n)} \right). && (3.58)
\end{aligned}
$$

For the second error term $\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|^2_{\boldsymbol{\Sigma}}$, we apply Lemma 34.

■

Proof of Theorem 7

Theorem 7 (Bounds of MSE for Biased Estimator)**.** *Consider the estimator $\hat{\boldsymbol{\theta}}_{aug}$ obtained by solving the aERM in (3.2). Let $\mathrm{MSE}^o(\hat{\boldsymbol{\theta}}_{aug})$ denote the unbiased MSE bound in Eq. (3.16) of Theorem 4, $\bar{\mathbf{C}} := \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$, and*

$$\Delta_G = \|n^{-1}\bar{\mathbf{C}}^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\bar{\mathbf{C}}^{-\frac{1}{2}} - \mathbf{I}_p\|.$$

Suppose the assumptions in Theorem 4 hold for the mean augmentation $\mu(\mathbf{x})$ and that $\Delta_G \leq c < 1$. Recall the definition of the mean augmentation covariance $\bar{\mathbf{\Sigma}} := \mathbb{E}_{\mathbf{x}}[(\mu_g(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mathbf{x}])(\mu_g(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mathbf{x}])^\top]$.Then with probability $1 - \delta' - 4n^{-1}$ we have,

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim R_1^2 \cdot \left(\sqrt{\mathrm{MSE}^o(\hat{\boldsymbol{\theta}}_{aug})} + R_2\right)^2,$$

where

$$R_1 = 1 + \|\mathbf{\Sigma}^{\frac{1}{2}}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\|,$$
$$R_2 = \sqrt{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}\left(1 + \frac{\Delta_G}{1-c}\right)\left(\sqrt{\Delta_\xi}\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right)\left(\sqrt{\frac{1}{\lambda_k^{aug}}} + \sqrt{\frac{\lambda_{k+1}^{aug}(1+\rho_k(\mathbf{\Sigma}_{aug};n))}{(\lambda_1^{aug}\rho_0(\mathbf{\Sigma}_{aug};n))^2}}\right).$$

Proof

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\mathrm{aug}}) = \|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\mathbf{\Sigma}}^2 \leq \left(\underbrace{\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\bar{\mathbf{\Sigma}}}}_{L_1} + \underbrace{\left|\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\mathbf{\Sigma}} - \|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\bar{\mathbf{\Sigma}}}\right|}_{L_2}\right)^2.$$

Now we will bound L_2 and L_1 in a sequence. For the L_2, denote $\Delta = \hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*$, then

$$\begin{aligned}
&\left|\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\mathbf{\Sigma}} - \|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\bar{\mathbf{\Sigma}}}\right| = \left|\sqrt{\Delta^\top\mathbf{\Sigma}\Delta} - \sqrt{\Delta^\top\bar{\mathbf{\Sigma}}\Delta}\right| \\
= \ & \frac{\left|\Delta^\top(\mathbf{\Sigma} - \bar{\mathbf{\Sigma}})\Delta\right|}{\|\Delta\|_{\mathbf{\Sigma}} + \|\Delta\|_{\bar{\mathbf{\Sigma}}}} \leq \frac{\|\Delta^\top(\mathbf{\Sigma}^{\frac{1}{2}} - \bar{\mathbf{\Sigma}}^{\frac{1}{2}})\|\|(\mathbf{\Sigma}^{\frac{1}{2}} + \bar{\mathbf{\Sigma}}^{\frac{1}{2}})\Delta\|}{\|\Delta\|_{\mathbf{\Sigma}} + \|\Delta\|_{\bar{\mathbf{\Sigma}}}} \\
\leq \ & \|\Delta^\top(\mathbf{\Sigma}^{\frac{1}{2}} - \bar{\mathbf{\Sigma}}^{\frac{1}{2}})\| \leq \|\Delta\|_{\bar{\mathbf{\Sigma}}}\|\mathbf{\Sigma}^{\frac{1}{2}}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\| = \|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\bar{\mathbf{\Sigma}}}\|\mathbf{\Sigma}^{\frac{1}{2}}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\|.
\end{aligned}$$

Hence, it remains to bound $\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\bar{\mathbf{\Sigma}}}$ because

$$L_1 + L_2 \leq (1 + \|\mathbf{\Sigma}^{\frac{1}{2}}\bar{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\|)\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|_{\bar{\mathbf{\Sigma}}}. \tag{3.59}$$

Now observe that $\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}}$ is just like the test error of an estimator where the covariate has the distribution of $\mu_{\mathcal{G}}(\mathbf{x})$. However, recall the caveat that when g is biased, there will be both a covariate shift and a misalignment of the observations in the estimator. Therefore, we have to take the latter into account. Specifically, recall that our observations $\mathbf{y}$ are, in fact, $\mathbf{X}\boldsymbol{\theta}^* + \mathbf{n}$. To match the covariate distribution $\mu_{\mathcal{G}}(\mathbf{x})$, we define $\tilde{\mathbf{y}} = \mu(\mathbf{X})\boldsymbol{\theta}^* + \mathbf{n}$. Although we do not actually observe $\tilde{\mathbf{y}}$, we can bound the error between observing $\mathbf{y}$ and $\tilde{\mathbf{y}}$. Therefore, we denote $\tilde{\boldsymbol{\theta}}_{\text{aug}} := (\mu(\mathbf{X})^\top \mu(\mathbf{X}) + \text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top \tilde{\mathbf{y}}$. This is the biased estimator that uses the biased augmentation g and also has an observation distribution that matches the covariate distribution. Then,

$$\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}} \lesssim \underbrace{\|\tilde{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}}}_{L_3} + \underbrace{\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \tilde{\boldsymbol{\theta}}_{\text{aug}}\|_{\bar{\boldsymbol{\Sigma}}}}_{L_4}. \tag{3.60}$$

Now, since $\tilde{\boldsymbol{\theta}}_{\text{aug}}$ has observations matching its covariate distribution $\mu_{\mathcal{G}}(\mathbf{x})$, we can apply Theorem 4 to bound L_3:

$$\|\tilde{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}} \leq \sqrt{\text{MSE}^o}, \tag{3.61}$$

where MSE^o is the bound in E.q. (3.16). It remains to bound L_4. Note that this error arises from the additive error between $\mathbf{y}$ and $\tilde{\mathbf{y}}$. Recall $\bar{\mathbf{C}} := \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]$, then,

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \tilde{\boldsymbol{\theta}}_{\text{aug}}\|_{\bar{\boldsymbol{\Sigma}}} &= \|(\mu(\mathbf{X})^\top \mu(\mathbf{X}) + \text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top(\mathbf{y} - \tilde{\mathbf{y}})\|_{\bar{\boldsymbol{\Sigma}}} \\
&= \|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top \mu(\mathbf{X}) + \text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top(\mathbf{y} - \tilde{\mathbf{y}})\| \\
&\leq \underbrace{\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top \mu(\mathbf{X}) + \text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top\|}_{L_5} \underbrace{\|(\mathbf{y} - \tilde{\mathbf{y}})\|}_{L_6}.
\end{aligned}$$

We first bound L_5,

$$\begin{aligned}
&\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top\| \\
\leq\ & \underbrace{\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\mu(\mathbf{X})^\top\|}_{L_7} \\
+\ & \underbrace{\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top-\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\mu(\mathbf{X})^\top\|}_{L_8}.
\end{aligned}$$

Observe that

$$\begin{aligned}
L_7 = \|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\mu(\mathbf{X})^\top\| &= \|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top+n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}\| \\
\leq\ \underbrace{\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top+n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{1:k}\|}_{L_9} &+ \underbrace{\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top+n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{k+1:p}\|}_{L_{10}},
\end{aligned}$$

where $\tilde{\mathbf{X}}$ has sub-gaussian rows with covariance $\boldsymbol{\Sigma}_{\text{aug}}$ as defined in E.q. (3.14).

Now, we bound L_9 and L_{10}. For convenience, denote $\mathbf{A}=\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top+n\mathbf{I}_n$ and $\mathbf{A}_k=\tilde{\mathbf{X}}_{k+1:p}\tilde{\mathbf{X}}_{k+1:p}^\top+n\mathbf{I}_n$. By Woodbury matrix identity, we have

$$\mathbf{A}^{-1}\tilde{\mathbf{X}}_{1:k}=\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k}(\mathbf{I}_p+\tilde{\mathbf{X}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k})^{-1}.$$

Hence, L_9 is bounded by

$$\begin{aligned}
&\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top+n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{1:k}\| = \|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k}(\mathbf{I}_p+\tilde{\mathbf{X}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k})^{-1}\| \\
\leq\ & \mu_n(\mathbf{A}_k)^{-1}\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\|\|\tilde{\mathbf{X}}_{1:k}(\mathbf{I}_p+\tilde{\mathbf{X}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k})^{-1}\| \\
=\ & \mu_n(\mathbf{A}_k)^{-1}\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\|\|\tilde{\mathbf{Z}}_{1:k}(\boldsymbol{\Sigma}_{\text{aug,1:k}}^{-1}+\tilde{\mathbf{Z}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{Z}}_{1:k})^{-1}\boldsymbol{\Sigma}_{\text{aug, 1:k}}^{-\frac{1}{2}}\| \\
\leq\ & \mu_n(\mathbf{A}_k)^{-1}\|\bar{\boldsymbol{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\|\|\boldsymbol{\Sigma}_{\text{aug,1:k}}^{-\frac{1}{2}}\|\|\tilde{\mathbf{Z}}_{1:k}(\boldsymbol{\Sigma}_{\text{aug,1:k}}^{-1}+\tilde{\mathbf{Z}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{Z}}_{1:k})^{-1}\|, \qquad (3.62)
\end{aligned}$$

where $\tilde{\mathbf{Z}}$ has sub-gaussian rows with isotropic covariance $\mathbf{I}_p$. Now applying Lemma 26, we have, with probability $1-5n^{-3}$,

$$\begin{aligned}\|\tilde{\mathbf{Z}}_{1:k}(\mathbf{\Sigma}_{\text{aug,1:k}}^{-1}+\tilde{\mathbf{Z}}_{1:k}^{\top}\mathbf{A}_k^{-1}\tilde{\mathbf{Z}}_{1:k})^{-1}\| &\lesssim \|\tilde{\mathbf{Z}}_{1:k}\|\mu_k^{-1}(\tilde{\mathbf{Z}}_{1:k}^{\top}\mathbf{A}_k^{-1}\tilde{\mathbf{Z}}_{1:k})\\ &\lesssim \mu_1(\mathbf{A}_k)\frac{\sqrt{n}}{\mu_k^{-1}(\tilde{\mathbf{Z}}_{1:k}^{\top}\tilde{\mathbf{Z}}_{1:k})} \lesssim \frac{\mu_1(\mathbf{A}_k)}{\sqrt{n}}.\end{aligned}$$

Combining the above and E.q. (3.62) with Lemma 24, we have with probability $1-\delta-2n^{-3}$ that

$$L_9=\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\top}+n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{1:k}\| \lesssim \sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{\lambda_k^{\text{aug}}n}}, \tag{3.63}$$

where λ_k^{aug} is the k-th eigenvalue of $\mathbf{\Sigma}_{\text{aug}}$. On the other hand, by Lemma 24 and 29,

$$\begin{aligned}L_{10}=\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\top}+n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{k+1:p}\| &\lesssim \frac{1}{\lambda_1^{\text{aug}}\rho_0(\mathbf{\Sigma}_{\text{aug}};n)}\sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|(\lambda_{k+1}^{\text{aug}}n+\sum_{j>k}\lambda_j^{\text{aug}})}{n^2}}\\ &= \sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|\lambda_{k+1}^{\text{aug}}(1+\rho_k(\mathbf{\Sigma}_{\text{aug}};n))}{n(\lambda_1^{\text{aug}}\rho_0(\mathbf{\Sigma}_{\text{aug}};n))^2}},\end{aligned}$$

with probability $1-\delta'-\exp(-ct)$ (where we set $t:=\log n$ for the final theorem statement). Hence,

$$L_7 \le L_9+L_{10} \lesssim \sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{n}}\left(\sqrt{\frac{1}{\lambda_k^{\text{aug}}}}+\sqrt{\frac{\lambda_{k+1}^{\text{aug}}(1+\rho_k(\mathbf{\Sigma}_{\text{aug}};n))}{(\lambda_1^{\text{aug}}\rho_0(\mathbf{\Sigma}_{\text{aug}};n))^2}}\right). \tag{3.64}$$

Next, we bound L_8:

$$\begin{aligned}&\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^{\top}\mu(\mathbf{X})+\text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^{\top}-\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^{\top}\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\mu(\mathbf{X})^{\top}\|\\ =\;& n\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^{\top}\mu(\mathbf{X})+\text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\left(n^{-1}\text{Cov}_{\mathcal{G}}(\mathbf{X})-\bar{\mathbf{C}}\right)(\mu(\mathbf{X})^{\top}\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\mu(\mathbf{X})^{\top}\|\\ \lesssim\;& n\underbrace{\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^{\top}\mu(\mathbf{X})+\text{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\|}_{L_{11}}\|n^{-1}\bar{\mathbf{C}}^{-\frac{1}{2}}\text{Cov}_{\mathcal{G}}(\mathbf{X})\bar{\mathbf{C}}^{-\frac{1}{2}}-\mathbf{I}_p\|\\ &\cdot\underbrace{\|\bar{\mathbf{C}}^{\frac{1}{2}}(\mu(\mathbf{X})^{\top}\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\mu(\mathbf{X})^{\top}\|}_{L_{12}}.\end{aligned}$$

The term L_{11} is identical to (3.64) and can be bounded with that inequality. In the meantime, the term $L_{12} = \|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\|$ can be bounded by noting that,

$$\begin{aligned}
&\mu_p\left(\left(\bar{\mathbf{C}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\right)^{-1}\right) \\
\gtrsim\ &\mu_p\left(\left(\bar{\mathbf{C}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + n\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\right)^{-1}\right) \\
&- \|\bar{\mathbf{C}}^{-\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + n\bar{\mathbf{C}})\bar{\mathbf{C}}^{-\frac{1}{2}} - \bar{\mathbf{C}}^{-\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))\bar{\mathbf{C}}^{-\frac{1}{2}}\|.
\end{aligned}$$

Here, by Lemma 24

$$\mu_p\left(\left(\bar{\mathbf{C}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + n\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\right)^{-1}\right) = \mu_p\left(\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + n\mathbf{I}_p\right)\right) \geq n \tag{3.65}$$

Also,

$$\begin{aligned}
&\|\bar{\mathbf{C}}^{-\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + n\bar{\mathbf{C}})\bar{\mathbf{C}}^{-\frac{1}{2}} - \bar{\mathbf{C}}^{-\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))\bar{\mathbf{C}}^{-\frac{1}{2}}\| \\
= &\|\bar{\mathbf{C}}^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\bar{\mathbf{C}}^{-\frac{1}{2}} - n\mathbf{I}_p\| = n\Delta_G
\end{aligned}$$

Adding the above inequalities together, L_8 is bounded by

$$\begin{aligned}
&\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\mu(\mathbf{X})(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\bar{\mathbf{C}}^{\frac{1}{2}} - \bar{\mathbf{\Sigma}}^{\frac{1}{2}}\mu(\mathbf{X})(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + n\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\| \\
\lesssim\ &\frac{\Delta_G}{1-\Delta_G}\sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{n}}\left(\sqrt{\frac{1}{\lambda_k^{\text{aug}}}} + \sqrt{\frac{\lambda_{k+1}^{\text{aug}}(1+\rho_k(\mathbf{\Sigma}_{\text{aug}};n))}{(\lambda_1^{\text{aug}}\rho_0(\mathbf{\Sigma}_{\text{aug}};n))^2}}\right),
\end{aligned} \tag{3.66}$$

by our assumption that $\Delta_G \leq c$ for some $c < 1$. E.q. (3.64) and (3.66) now imply

$$\begin{aligned}
L_5 &= \|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X}) + \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top\| \leq L_7 + L_8 \\
&\lesssim \sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{n}}\left(\sqrt{\frac{1}{\lambda_k^{\text{aug}}}} + \sqrt{\frac{\lambda_{k+1}^{\text{aug}}(1+\rho_k(\mathbf{\Sigma}_{\text{aug}};n))}{(\lambda_1^{\text{aug}}\rho_0(\mathbf{\Sigma}_{\text{aug}};n))^2}}\right)\cdot\left(1+\frac{\Delta_G}{1-c}\right).
\end{aligned} \tag{3.67}$$

On the other hand,

$$\begin{aligned}
L_6 = \|\mathbf{y} - \tilde{\mathbf{y}}\| &= \|(\mu(\mathbf{X}) - \mathbf{X})\boldsymbol{\theta}^*\| = \sqrt{n}\|\boldsymbol{\theta}^*\|_{n^{-1}(\mu(\mathbf{X})-\mathbf{X})(\mu(\mathbf{X})-\mathbf{X})^\top} \\
&\leq \sqrt{n}\left(\|\boldsymbol{\theta}^*\|\sqrt{\|n^{-1}(\mu(\mathbf{X}) - \mathbf{X})(\mu(\mathbf{X}) - \mathbf{X})^\top - \mathrm{Cov}_\xi\|} + \|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right) \\
&\leq \sqrt{n}\left(\sqrt{\Delta_\delta}\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right), \qquad (3.68)
\end{aligned}$$

where Cov_δ is defined in Definition 6.

Combining E.q. (3.67) and (3.68), we obtain the following:

$$\begin{aligned}
L_4 = \|\hat{\boldsymbol{\theta}}_{\text{aug}} - \tilde{\boldsymbol{\theta}}_{\text{aug}}\|_{\bar{\boldsymbol{\Sigma}}} = L_5 \cdot L_6 \;&\lesssim\; \sqrt{\|\bar{\boldsymbol{\Sigma}}\bar{\mathbf{C}}^{-1}\|}\left(1 + \frac{\Delta_G}{1-c}\right)\left(\sqrt{\Delta_\xi}\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right) \\
&\cdot \left(\sqrt{\frac{1}{\lambda_k^{\text{aug}}}} + \sqrt{\frac{\lambda_{k+1}^{\text{aug}}(1 + \rho_k(\boldsymbol{\Sigma}_{\text{aug}}; n))}{(\lambda_1^{\text{aug}}\rho_0(\boldsymbol{\Sigma}_{\text{aug}}; n))^2}}\right) \qquad (3.69)
\end{aligned}$$

Finally, putting together the results of Eq. (3.59), (3.60), (3.61) and (3.69) completes the proof. ■

Proof of Proposition 5

Proposition 5 (Independent Feature Augmentations). *Let g be an independent feature augmentation, and $\pi : \{1, 2, \ldots, p\} \to \{1, 2, \ldots, p\}$ be the function that maps the original feature index to the sorted index according to the eigenvalues of $\boldsymbol{\Sigma}_{aug}$ in a non-increasing order. Then, data augmentation has a spectrum reordering effect which changes the MSE through the bias modification:*

$$\frac{\mathit{Bias}}{C_x L_1^4} \lesssim \left\|\theta^*_{\pi(k_1+1:p)}\right\|^2_{\Sigma_{\pi(k_1+1:p)}} + \left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2\Sigma^{-1}_{\pi(1:k_1)}} \frac{(\rho_{k_1}^{aug})^2}{(\lambda_{k_1+1}^{aug})^{-2} + (\lambda_1^{aug})^{-2}(\rho_{k_1}^{aug})^2},$$

where $\pi(a : b)$ denotes the indices of $\pi(a), \pi(a+1), \ldots, \pi(b)$. Furthermore, if the variance of each feature augmentation $\mathrm{Var}_{g_i}(g_i(x))$ is a sub-exponential random variable with sub-exponential norm σ_i^2 and mean $\bar{\sigma}_i^2$, $\forall i \in \{1, 2, \ldots, p\}$, and $p = O(n^\alpha)$ for some $\alpha > 0$, then there exists a constant c, *depending only on* α*, such that with probability* $1 - n^{-3}$,

$$\Delta_G \lesssim \max_i \left(\frac{\sigma_i^2}{\bar{\sigma}_i^2}\right) \sqrt{\frac{\log n}{n}}.$$

Proof For independent feature augmentation, $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$ is a diagonal matrix. Since the original covariance $\mathbf{\Sigma}$ is also diagonal by our model assumption, the augmentation modified spectrum $\mathbf{\Sigma}_{\text{aug}}$ is diagonal. Furthermore, the diagonal implies the projections to $\mathbf{\Sigma}_{\text{aug}}$'s first $k-1$ and the rest eigenspaces are to the features $\pi(1:k-1)$ and $\pi(k,p)$. Lastly, because $P^{\mathbf{\Sigma}_{\text{aug}}}$ commutes with $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$, we have

$$\begin{aligned}
\left\|\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\theta^*_{\text{aug}}\right\|^2_{\mathbf{\Sigma}_{\text{aug}}} &= (\theta^*_{\text{aug}})^\top \mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\theta^*_{\text{aug}} \\
&= (\boldsymbol{\theta}^*)^\top \bar{\mathcal{D}}^{1/2}\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\bar{\mathcal{D}}^{-1/2}\mathbf{\Sigma}\bar{\mathcal{D}}^{-1/2}\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\bar{\mathcal{D}}^{1/2}\boldsymbol{\theta}^* \\
&= (\boldsymbol{\theta}^*)^\top \mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\bar{\mathcal{D}}^{1/2}\bar{\mathcal{D}}^{-1/2}\mathbf{\Sigma}\bar{\mathcal{D}}^{-1/2}\bar{\mathcal{D}}^{1/2}\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\boldsymbol{\theta}^* \\
&= \|\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{k_1+1:p}\boldsymbol{\theta}^*\|^2_{\mathbf{\Sigma}} = \left\|\theta^*_{\pi(k_1+1:p)}\right\|^2_{\Sigma_{\pi(k_1+1:p)}}, \\
\left\|\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{1:k_1}\theta^*_{\text{aug}}\right\|^2_{\mathbf{\Sigma}^{-1}_{\text{aug}}} &= (\boldsymbol{\theta}^*)^\top \mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{1:k_1}\bar{\mathcal{D}}^{1/2}\bar{\mathcal{D}}^{1/2}\mathbf{\Sigma}^{-1}\bar{\mathcal{D}}^{1/2}\bar{\mathcal{D}}^{1/2}\mathbf{P}^{\mathbf{\Sigma}_{\text{aug}}}_{1:k_1}\boldsymbol{\theta}^* \\
&= \left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2\Sigma^{-1}_{\pi(1:k_1)}},
\end{aligned}$$

where $\bar{\mathcal{D}} = \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$.

To prove the approximation error bound, we proceed as follows. By independence assumption on feature augmentation, $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ is diagonal. Hence, to bound Δ_G, we only need to control the diagonals of $\mathbf{Q} := n^{-1}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}$. Now, denoting $\mathcal{D} = n^{-1}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}$, we have $\mathbf{Q} = \mathcal{D} - \mathbf{I}$. For any $i \in \{1, 2, \ldots, p\}$, $\mathcal{D}_{ii} = n^{-1}\sum_{j=1}^n \frac{\mathrm{Var}_{g_i}(\mathbf{x}_{ji})}{\mathbb{E}_{\mathbf{x}}[\mathrm{Var}_{g_i}(\mathbf{x})]}$, where $\mathbf{x}_{ji}$ is the i-th element of the j-th row of $\mathbf{X}$. By our assumptions of $\mathrm{Var}_{g_i}(\mathbf{x}_{ji})$, $j = 1, 2, \ldots, n$, being identical and independent sub-exponential random variables with sub-exponential norm σ_i^2 and mean $\bar{\sigma}_i^2$. we can apply concentration bounds to $\mathbf{Q}_{ii} = \frac{1}{\bar{\sigma_i}^2}\left(n^{-1}\sum_{j=1}^n \mathrm{Var}_{g_i}(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}}[\mathrm{Var}_{g_i}(\mathbf{x})]\right)$ as it is a sum of i.i.d. sub-exponential random variables with sub-exponential norm ${\sigma_i}^2/\bar{\sigma}_i^2$. Specifically, we apply the Bernstein inequality in

Lemma 23 with $t \propto \sigma_i^2\sqrt{\frac{\log n}{n}}$ to conclude that there exists a constant c' such that, with probability $1 - n^{-1}$, we have,

$$\mathbf{Q}_{ii} = \frac{1}{\bar{\sigma_i}^2}\left(n^{-1}\sum_{j=1}^{n}\mathrm{Var}_{g_i}(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}}[\mathrm{Var}_{g_i}(\mathbf{x})]\right) \leq c'\frac{\sigma_i^2}{\bar{\sigma}_i^2}\sqrt{\frac{\log n}{n}}. \tag{3.70}$$

Then, we apply a union bound over i and obtain

$$\|\mathbf{\Delta}_G\| \leq \max_i \|\mathbf{Q}_{ii}\| \lesssim \max_i\left(\frac{\sigma_i^2}{\bar{\sigma}_i^2}\right)\sqrt{\frac{\log n}{n}},$$

with probability $1 - n^{-1}$. Note that we can get the same error rate after the union bound as long as p grows polynomially with n.

■

Proofs of Corollaries

Corollary 36 (Generalization of Gaussian Noise Injection). *Consider the data augmentation which adds samples with independent additive Gaussian noise: $g(\mathbf{x}) = \mathbf{x} + \mathbf{n}$, where $\mathbf{n} \sim \mathcal{N}(0, \sigma^2)$. The estimator is given by $\hat{\boldsymbol{\theta}} = (\mathbf{X}^\top\mathbf{X} + \sigma^2 n\mathbf{I}_p)^{-1}\mathbf{X}^\top y$. Let L denote the condition number of $n\sigma^2\mathbf{I} + \mathbf{X}_{1:k}\mathbf{X}_{1:k}^\top$. Then, we can bound the error as MSE $\leq$ Bias + Variance, where with high probability*

$$\mathrm{MSE} \lesssim \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\mathbf{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\mathbf{\Sigma}^{-1}_{0:k}}\lambda^2_{k+1}\rho^2_k(\mathbf{\Sigma}; n\sigma^2) + R_k^{-1}(\mathbf{\Sigma}; n\sigma^2) + kn^{-1}.$$

Proof Since this belongs to the independent feature augmentation class, we can apply Corollary 5. Below are the quantities in the corollary.

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \sigma^2\mathbf{I},\ \boldsymbol{\theta}^*_{\text{aug}} = \sigma\boldsymbol{\theta}^*,\ \mathbf{\Sigma}_{\text{aug}} = \sigma^{-2}\mathbf{\Sigma},\ \lambda^{\text{aug}} = \sigma^{-2}\lambda,$$

hence,

$$\begin{aligned}
\rho_k^{\text{aug}} &= \rho_k(\mathbf{\Sigma}_{\text{aug}}; n) = \frac{n + \sum\limits_{i=k+1}^{p} \lambda_i^{\text{aug}}}{n\lambda_{k+1}^{\text{aug}}} = \frac{n\sigma^2 + \sum\limits_{i=k+1}^{p} \lambda_i}{n\lambda_{k+1}} = \rho_k(\mathbf{\Sigma}; n\sigma^2), \\
R_k^{\text{aug}} &= R_k(\mathbf{\Sigma}_{\text{aug}}; n) = \frac{\left(n + \sum\limits_{i=k+1}^{p} \lambda_i^{\text{aug}}\right)^2}{\sum\limits_{i=k+1}^{p} (\lambda_{k+1}^{\text{aug}})^2} = \frac{\left(n\sigma^2 + \sum\limits_{i=k+1}^{p} \lambda_i\right)^2}{n\sum\limits_{i=k+1}^{p} \lambda_{k+1}^2} = R_k(\mathbf{\Sigma}; n\sigma^2).
\end{aligned}$$

Note that $R_k(\mathbf{\Sigma}; n\sigma^2)$ and $\rho_k(\mathbf{\Sigma}_{\text{aug}}; n\sigma^2)$ are the effective dimensions of the original spectrum for ridge regression with regularization parameter $n\sigma^2$, as defined in [26]. Finally, the approximation error term is zero because $\Delta_G = 0$. ■

Corollary 12 (Generalization of random mask augmentation). *Consider the unbiased randomized masking augmentation $g(\mathbf{x}) = [b_1\mathbf{x}_1, \dots, b_p\mathbf{x}_p]/(1-\beta)$, where b_i are i.i.d. Bernoulli$(1-\beta)$. Define $\psi = \frac{\beta}{1-\beta} \in [0, \infty)$. Let L_1, L_2, κ, δ' be universal constants as defined in Theorem 4. Assume $p = O(n^\alpha)$ for some $\alpha > 0$. Then, for any set $\mathcal{K} \subset \{1, 2, \dots, p\}$ consisting of k_1 elements and $k_2 \in [0, n]$, there exists some constant c', which depends solely on σ_z and σ_ε (the sub-Guassian norms of the covariates and noise), such that the regression MSE is upper-bounded by*

$$\begin{aligned}
\text{MSE} \lesssim & \underbrace{\|\theta^*_{\mathcal{K}}\|^2_{\Sigma_{\mathcal{K}}} + \|\theta^*_{\mathcal{K}^c}\|^2_{\Sigma_{\mathcal{K}^c}} \frac{(\psi n + p - k_1)^2}{n^2 + (\psi n + p - k_1)^2}}_{\text{Bias}} \\
& + \underbrace{\left(\frac{k_2}{n} + \frac{n(p-k_2)}{(\psi n + p - k_2)^2}\right) \log n}_{\text{Variance}} + \underbrace{\sigma_z^2 \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}}}_{\text{Approx.Error}}
\end{aligned}$$

with probability at least $1 - \delta' - n^{-1}$.

Proof Random mask belongs to independent feature augmentation class, so we can apply Proposition 5. We calculate the quantities used in the corollary.

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \psi \mathrm{diag}(\boldsymbol{\Sigma}) = \psi \boldsymbol{\Sigma},\ \boldsymbol{\theta}^*_{\text{aug}} = \psi^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*,\ \boldsymbol{\Sigma}_{\text{aug}} = \psi^{-1}\mathbf{I},\ \lambda^{\text{aug}} = \psi^{-1}.$$

The effective ranks of the augmentation modified spectrum are

$$\rho_k^{\text{aug}} = \frac{\psi n + p - k}{n}, \tag{3.71}$$

$$R_k^{\text{aug}} = \frac{(\psi n + p - k)^2}{p - k}. \tag{3.72}$$

Now, we apply Proposition 5. Because random mask has effectively isotropized the spectrum, the mapping π in the proposition can be chosen arbitrarily. Hence, we can chose $\pi(1 : k_1)$ to be any set with k elements. For the approximation error term, we first note that $\kappa = 1$. Furthermore, $\mathrm{Var}_{g_i}(\mathbf{x}_j) = \psi \mathbf{x}_j^2$. So, its subexponential norm is bounded by $\psi\lambda_j\sigma_z^2$, and its expectation is given by $\psi\lambda_j$. Putting all the pieces together, we derive the MSE bound as

$$\begin{aligned} \text{Bias} &\lesssim \|\theta^*_{\mathcal{K}}\|^2_{\Sigma_{\mathcal{K}}} + \|\theta^*_{\mathcal{K}^c}\|^2_{\Sigma_{\mathcal{K}^c}} \frac{(\psi n + p - k_1)^2}{n^2 + (\psi n + p - k_1)^2}, \\ \text{Variance} &\lesssim \frac{k_2}{n} + \frac{n(p - k_2)}{(\psi n + p - k_2)^2}, \\ \text{Approx. Error} &\lesssim \sigma_z^2 \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}. \end{aligned}$$

■

Corollary 15 (Bounds of random cutout). *Let $\hat{\boldsymbol{\theta}}_k^{cutout}$ denote the random cutout estimator that zeroes out k consecutive coordinates (the starting location of which is chosen uniformly at random). Also, let $\hat{\boldsymbol{\theta}}_\beta^{mask}$ be the random mask estimator with the masking probability given by β. We assume that $k = O(\sqrt{\frac{n}{\log p}})$. Then, for the choice $\beta = \frac{k}{p}$ we have*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_k^{cutout}) \asymp \mathrm{MSE}(\hat{\boldsymbol{\theta}}_\beta^{mask}),\ \ \mathrm{POE}(\hat{\boldsymbol{\theta}}_k^{cutout}) \asymp \mathrm{POE}(\hat{\boldsymbol{\theta}}_\beta^{mask}).$$

Proof This can be verified directly by noticing that for random cutout

$$\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) = \frac{k}{p-k}\mathrm{diag}(\boldsymbol{\Sigma}),$$

while for random mask

$$\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) = \psi\mathrm{diag}(\boldsymbol{\Sigma}).$$

Furthermore, the approximation is negligible when $k \ll \min(\sqrt{\frac{n}{\log p}}, \frac{p}{\sqrt{n}})$ as shown in Appendix 3.8.6. Now, setting $\psi = \frac{k}{d-k}$ gives $\beta = \frac{k}{p}$. ■

Corollary 18 (Non-uniform random mask). *Consider a general $k-$sparse model where $\boldsymbol{\theta}^* = \sum_{i\in\mathcal{I}_{\mathcal{S}}} \alpha_i \mathbf{e}_i$, where $|\mathcal{I}_{\mathcal{S}}| = k$. Suppose we employ non-uniform random mask where $\psi = \psi_1$ if $i \in \mathcal{I}_{\mathcal{S}}$ and $= \psi_0$ otherwise. Then, if $\psi_1 \leq \psi_0$, we have*

$$\begin{aligned}
\mathrm{Bias} &\lesssim \frac{\left(\psi_1 n + \frac{\psi_1}{\psi_0}\left(p - |\mathcal{I}_{\mathcal{S}}|\right)\right)^2}{n^2 + \left(\psi_1 n + \frac{\psi_1}{\psi_0}\left(p - |\mathcal{I}_{\mathcal{S}}|\right)\right)^2}\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2, \\
\mathrm{Variance} &\lesssim \frac{|\mathcal{I}_{\mathcal{S}}|}{n} + \frac{n\left(p - |\mathcal{I}_{\mathcal{S}}|\right)}{\left(\psi_0 n + p - |\mathcal{I}_{\mathcal{S}}|\right)^2}, \\
\mathrm{Approx.Error} &\lesssim \sqrt{\frac{\psi_0}{\psi_1}\sigma_z^2}\sqrt{\frac{\log n}{n}}\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}
\end{aligned}$$

while if $\psi_1 > \psi_0$, we have

$$\mathrm{Bias} \lesssim \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2, \;\; \mathrm{Variance} \lesssim \frac{\left(\frac{\psi_1}{\psi_o}\right)^2 + \frac{|\mathcal{I}_{\mathcal{S}}|}{n}}{\left(\frac{\psi_1}{\psi_o} + \frac{|\mathcal{I}_{\mathcal{S}}|}{n}\right)^2}, \;\; \mathrm{Approx.Error} \lesssim \sqrt{\frac{\psi_1}{\psi_0}\sigma_z^2}\sqrt{\frac{\log n}{n}}\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$$

Proof Let Ψ denote the diagonal matrix with $\Psi_{i,i} = \psi_1$ if $i \in \mathcal{I}_{\mathcal{S}}$ and ψ_0 otherwise. Then, we apply

Corollary 5 with:

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \Psi \mathrm{diag}(\mathbf{\Sigma}) = \Psi\mathbf{\Sigma},\ \boldsymbol{\theta}^*_{\text{aug}} = \Psi^{1/2}\mathbf{\Sigma}^{1/2}\boldsymbol{\theta}^*,\ \mathbf{\Sigma}_{\text{aug}} = \Psi^{-1}.$$

Now as in the proof of Proposition 12, we calculate the effective ranks. For the k^* partitioning the spectrum, we choose $k^* = |\mathcal{I}_{\mathcal{S}}|$ when $\psi_1 \leq \psi_0$, while $k^* \asymp n$ for $\psi_1 > \psi_0$. The proof for the approximation error term is identical to in the uniform random mask case. ■

Corollary 16 (Generalization of Pepper/Salt augmentation). *The MSE components of the estimator that are induced by salt-and-pepper augmentation (denoted by $\hat{\boldsymbol{\theta}}_{pepper}(\beta, \sigma^2)$) have the properties,*

$$\begin{aligned}
\mathrm{Bias}[\hat{\boldsymbol{\theta}}_{pepper}(\beta,\sigma^2)] &\lesssim \left(\frac{\lambda_1(1-\beta)+\sigma^2}{\sigma^2}\right)^2 \mathrm{Bias}\left[\hat{\boldsymbol{\theta}}_{gn}\left(\frac{\beta\sigma^2}{(1-\beta)^2}\right)\right],\\
\mathrm{Variance}[\hat{\boldsymbol{\theta}}_{pepper}(\beta,\sigma^2)] &\lesssim \mathrm{Variance}\left[\hat{\boldsymbol{\theta}}_{gn}\left(\frac{\beta\sigma^2}{(1-\beta)^2}\right)\right],\\
\mathrm{Approx.Error}[\hat{\boldsymbol{\theta}}_{pepper}(\beta,\sigma^2)] &\asymp \mathrm{Approx.Error}[\hat{\boldsymbol{\theta}}_{rm}(\beta)].
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}_{gn}(z^2)$ and $\hat{\boldsymbol{\theta}}_{rm}(\gamma)$ denotes the estimators that are induced by Gaussian noise injection with variance z^2 and random mask with dropout probability γ, respectively. Moreover, the limiting MSE as $\sigma \to 0$ reduces to the MSE of the estimator induced by random masking (denoted by $\hat{\boldsymbol{\theta}}_{rm}(\beta)$):

$$\lim_{\sigma\to 0} \mathrm{MSE}[\hat{\boldsymbol{\theta}}_{pepper}(\beta,\sigma^2)] = \mathrm{MSE}[\hat{\boldsymbol{\theta}}_{rm}(\beta)].$$

Proof Proposition 5 is applicable to salt/pepper augmentation. The related quantities in the proposition are:

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \psi\mathbf{\Sigma} + \frac{\psi\sigma^2}{1-\beta}\mathbf{I},\ \boldsymbol{\theta}^*_{\text{aug}} = \sqrt{\psi\mathbf{\Sigma} + \frac{\psi\sigma^2}{1-\beta}\mathbf{I}}\boldsymbol{\theta}^*,\ \lambda_i^{\text{aug}} = \frac{\lambda_i}{\psi(\lambda_i + \frac{\sigma^2}{1-\beta})}.$$

Observe that the expression of λ_i^{aug} implies that the augmented eigenvalues of salt/pepper augmentation is a harmonic sum of that of random mask and Gaussian noise injection,

$$\lambda_{pepper}(\beta, \sigma^2)^{-1} = \lambda_{rm}(\beta)^{-1} + \beta^{-1}\lambda_{gn}(\sigma^2)^{-1}. \tag{3.73}$$

Hence, the statement of MSE limit is clear as we take $\sigma \to 0$ in (3.73) along with the fact that $\lambda_{\text{gn}} \to \infty$. Now we prove the bias statement. By Proposition 5,

$$\hat{\boldsymbol{\theta}}_{\text{pepper}}(\beta, \sigma) \lesssim \|\boldsymbol{\theta}^*_{k+1:p}\|^2_{\boldsymbol{\Sigma}_{k+1:p}} + \left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2 \Sigma^{-1}_{\pi(1:k_1)}} (\lambda^{\text{aug}}_{k+1}\rho^{\text{aug}}_k)^2. \tag{3.74}$$

In particular,

$$\left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2 \Sigma^{-1}_{\pi(1:k_1)}} = \sum_{i \le k} \frac{\left(\psi\lambda_i + \frac{\psi\sigma^2}{1-\beta}\right)^2}{\lambda_i} (\boldsymbol{\theta}^*_i)^2, \tag{3.75}$$

$$\lambda^{\text{aug}}_{k+1}\rho^{\text{aug}}_k = \frac{n + \sum_{i>k} \frac{\lambda_i}{\psi(\lambda_i + \frac{\sigma^2}{1-\beta})}}{n} \le \frac{n + \sum_{i>k} \frac{\lambda_i}{\psi\frac{\sigma^2}{1-\beta}}}{n}. \tag{3.76}$$

Now the result follows by combining Eq. (3.74), (3.75) and (3.76).

The variance statement can be proved using similar calculations. From Corollary 5, we only need to compare R_k of salt/pepper with that of Gaussian noise injection. Without lose of generality, we assume k is chosen in the corollary such that $\lambda_i \le c' \frac{\sigma^2}{1-\beta}$ for all $i \ge k$ for some constant c'. Then,

$$R_k \ge \frac{\left(n + \sum_{i \ge k} \frac{\lambda_i}{\psi(\lambda_i + \frac{\sigma^2}{1-\beta})}\right)^2}{\sum_{i \ge k}\left(\frac{\lambda_i}{\psi(\lambda_i + \frac{\sigma^2}{1-\beta})}\right)^2} \ge \frac{\left(n + \sum_{i \ge k} \frac{\lambda_i}{\psi((c'+1)\frac{\sigma^2}{1-\beta})}\right)^2}{\sum_{i \ge k}\left(\frac{\lambda_i}{\psi(\frac{\sigma^2}{1-\beta})}\right)^2} \ge \frac{1}{(c'+1)^2} \frac{\left(n + \sum_{i \ge k} \frac{\lambda_i}{\frac{\beta\sigma^2}{(1-\beta)^2}}\right)^2}{\sum_{i \ge k}\left(\frac{\lambda_i}{\frac{\beta\sigma^2}{(1-\beta)^2}}\right)^2},$$

The statement now follows by noting that the last quantity is the R_k of Gaussian noise injection with noise variance $\frac{\beta\sigma^2}{(1-\beta)^2}$ up to a constant scaling factor.

Finally, the approximation error statement holds because the augmented covariance is that of random mask summed with a constant matrix. ■

Corollary 13 (Generalization of biased mask augmentation). *Consider the biased random mask augmentation* $g(\mathbf{x}) = [b_1\mathbf{x}_1, \dots, b_p\mathbf{x}_p]$*, where* b_i *are i.i.d. Bernoulli(1-*β*). Define* $\psi = \frac{\beta}{1-\beta} \in [0, \infty)$*. Assume the assumptions in Corollary 12 hold. Then with probability* $1 - \delta' - 3pn^{-5}$*, the generalization error is upper bounded by*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug}) \leq \left(\sqrt{\mathrm{MSE}^o} + \psi\left(1 + \frac{\log n}{n}\right) \cdot \left(\left(\lambda_1 + \frac{\sum_j \lambda_j}{n}\right)\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}\right)\right)^2,$$

where MSE^o *is the bound given in Corollary 12.*

Proof This proof is a direct application of Theorem 7 by the two steps: First, plugging in

$$\boldsymbol{\Sigma}_{\text{aug}} = \frac{1-\beta}{\beta}\mathbf{I},\ \bar{\boldsymbol{\Sigma}} = (1-\beta)^2\boldsymbol{\Sigma},\ \mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) = \beta(1-\beta)\boldsymbol{\Sigma}.$$

Secondly, observing $\delta(\mathbf{x}) = -\beta\mathbf{x}$, $\mathrm{Cov}_\delta = \beta^2\boldsymbol{\Sigma}$, so concentration bound in Lemma 29 gives that

$$\Delta_\delta \lesssim \beta^2\left(\frac{\lambda_1 n + \sum_j \lambda_j}{n}\right).$$

■

3.8.3 Proofs of Classification Results

Classification Lemmas

Lemma 37 (Upper bound on probability of classification error for correlated sub-Gaussian input). *Consider the 1-sparse model* $\boldsymbol{\theta}^* = \frac{1}{\sqrt{\lambda_t}}\mathbf{e}_t$ *described in Section 3.4.4 and input distribution satisfying Assumption 3, where* $\mathbf{x}_{sig} = \mathbf{x}_t$ *is the feature corresponding to the non-zero coordinate* *of θ^*. Given any estimator $\hat{\theta}$ having $\hat{\boldsymbol{\theta}}_t \geq 0$, the probability of classification error (POE) is upper bounded by*

$$\mathrm{POE}(\hat{\theta}) \lesssim \frac{\mathrm{CN}}{\mathrm{SU}}\left(1+\sigma_z\sqrt{\log\frac{\mathrm{SU}}{\mathrm{CN}}}\right). \tag{3.77}$$

Furthermore, if we assume $\mathbf{x}$ is Gaussian, then

$$\mathrm{POE}(\hat{\theta}) = \frac{1}{2}-\frac{1}{\pi}\tan^{-1}\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})}{\mathrm{CN}(\hat{\boldsymbol{\theta}})} \leq \frac{\mathrm{CN}(\hat{\boldsymbol{\theta}})}{\mathrm{SU}(\hat{\boldsymbol{\theta}})}. \tag{3.78}$$

Proof

We first note that the assumption that $\hat{\theta}_t \geq 0$ is satisfied in the situations we consider, based on the lower bounds on survival which we provide in Lemma 38. Assume without loss of generality that $\mathbf{x}_{sig} = \mathbf{x}_t = \mathbf{x}_1$.

$$\begin{aligned}
\mathrm{POE}(\hat{\boldsymbol{\theta}}) &= \mathbb{P}\left(\mathrm{sgn}(\mathbf{x}_{sig}) \neq \mathrm{sgn}(\langle \mathbf{x}, \hat{\boldsymbol{\theta}}\rangle)\right) \\
&= \mathbb{P}\left(\mathrm{sgn}(\mathbf{x}_{sig}) \neq \mathrm{sgn}(\mathbf{x}_{sig}(\hat{\boldsymbol{\theta}}_1 + \frac{\mathbf{x}_2}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_2 + \cdots + \frac{\mathbf{x}_p}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_p))\right) \\
&= \mathbb{P}\left(\hat{\boldsymbol{\theta}}_1 + \frac{\mathbf{x}_2}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_2 + \cdots + \frac{\mathbf{x}_p}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_p < 0\right) \\
&= \mathbb{E}_{\mathbf{x}_{sig}}\mathbb{P}\left(\frac{\mathbf{x}_2}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_2 + \cdots + \frac{\mathbf{x}_p}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_p < -|\hat{\boldsymbol{\theta}}_1|\right).
\end{aligned}$$

Now, because $\mathbf{z}' := [\frac{\mathbf{x}_2}{\sqrt{\lambda_2}}, \frac{\mathbf{x}_3}{\sqrt{\lambda_3}}, \ldots, \frac{\mathbf{x}_p}{\sqrt{\lambda_p}}]$ is a sub-Gaussian vector with norm σ_z, $\langle \mathbf{z}', \mathbf{u}\rangle$ is a sub-Gaussian variable with norm $\|\mathbf{u}\|$ for any fixed $\mathbf{u}$. Let $\mathbf{u} = \frac{1}{\mathbf{x}_{sig}}[\sqrt{\lambda_2}\hat{\boldsymbol{\theta}}_2, \sqrt{\lambda_3}\hat{\boldsymbol{\theta}}_3, \ldots, \sqrt{\lambda_p}\hat{\boldsymbol{\theta}}_p]$, which, *by assumption, is independent of* $\mathbf{z}'$*. Then,*

$$\mathbb{E}_{\mathbf{x}_{sig}}\mathbb{P}\left(\frac{\mathbf{x}_2}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_2+\cdots+\frac{\mathbf{x}_p}{\mathbf{x}_{sig}}\hat{\boldsymbol{\theta}}_p<-|\hat{\boldsymbol{\theta}}_1|\right)=\mathbb{E}_{\mathbf{x}_{sig}}\mathbb{P}\left(\langle\mathbf{z}',\mathbf{u}\rangle\leq-|\hat{\boldsymbol{\theta}}_1|\right)$$
$$\leq\mathbb{E}_{\mathbf{x}_{sig}}\exp\left(-\frac{\hat{\boldsymbol{\theta}}_1^2}{\sum_{j\geq 2}\lambda_j(\frac{\hat{\boldsymbol{\theta}}_j}{\mathbf{x}_{sig}})^2\sigma_z^2}\right)$$
$$=\mathbb{E}_{\mathbf{x}_{sig}}\exp\left(-\frac{\mathbf{x}_{sig}^2}{\lambda_1\sigma_z^2}\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})^2}{\mathrm{CN}(\hat{\boldsymbol{\theta}})^2}\right)$$
$$\leq\mathbb{P}(\mathbf{x}_{sig}^2<\delta)+3\exp\left(-\frac{\delta}{\lambda_1\sigma_z^2}\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})^2}{\mathrm{CN}(\hat{\boldsymbol{\theta}})^2}\right)$$
$$\lesssim\sqrt{\frac{\delta}{\lambda_1}}+3\exp\left(-\frac{\delta}{\lambda_1\sigma_z^2}\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})^2}{\mathrm{CN}(\hat{\boldsymbol{\theta}})^2}\right),$$

where the last inequality follows from the assumption that $\mathbf{z}_{sig}$ *has bounded density and a small ball probability bound from [128, Exercise 2.2.10]. Choosing* $\delta=\lambda_1\sigma_z^2\log\frac{SU}{CN}/\left(\frac{SU}{CN}\right)^2$ *yields the result.*

The second statement follows from Proposition 17 in [27] and the bound $\tan^{-1}(x)\geq\frac{\pi}{2}-\frac{1}{x}$*, for all* $x>0$*.* ■

Lemma 38 (Survival of ridge estimator for dependent features). *Consider the classification task under the model and assumption described in Section 3.4.4 where* $\boldsymbol{\Sigma}=diag(\lambda_1,\ldots,\lambda_p)$ *and the true signal* $\theta^*=\frac{1}{\sqrt{\lambda_t}}\mathbf{e}_t$ *is 1-sparse in coordinate* t*. Let* $\hat{\boldsymbol{\theta}}=\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-1}\mathbf{y}$ *be a ridge estimator. Suppose for some* $t\leq k\leq n$ *that* $\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda)\geq c$ *for some constant* $c>0$*, and with probability at least* $1-\delta$ *that the condition number of* $\lambda\mathbf{I}+\mathbf{X}_{k+1:p}\mathbf{X}_{k+1:p}^T$ *is at most* L*, then with probability* $1-\delta-\exp(-\sqrt{n})$*, we have:*

$$\frac{\lambda_t(1-2\nu^*)\left(1-\frac{k}{n}\right)}{L\left(\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda)+\lambda_t L\right)}\lesssim\mathrm{SU}(\hat{\boldsymbol{\theta}})\lesssim\frac{L\lambda_t(1-2\nu^*)}{\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda)+L^{-1}\lambda_t\left(1-\frac{k}{n}\right)}. \tag{3.79}$$

Proof *Our bound is a generalization to Theorem 22 in [27] for correlated features and ridge estimator. We only require the signal and noise features to be independent.*

Denote $\tilde{\mathbf{X}}$ to be the matrix consisting of the columns of $\mathbf{X}$ except for the t-th column, and $\mathbf{A}_{-t} := \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda\mathbf{I}$. As the proof in [27], our proof begins with writing the SU in terms of a quadratic form of signal vector and applying Hanson-Wright inequality, Lemma 28, by invoking the independence between the signal and noise. The result is that, with probability $1 - \exp(-\sqrt{n})$,

$$\mathrm{SU} \gtrsim \frac{\lambda_t \cdot \left((1-2\nu^*)\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) - 2c_1 \left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)}{1 + \lambda_t\left(\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) + c_1\left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)} \text{ and} \tag{3.80}$$

$$\mathrm{SU} \lesssim \frac{\lambda_t \cdot \left((1-2\nu^*)\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) + 2c_1 \left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)}{1 + \lambda_t\left(\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) - c_1\left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)}, \tag{3.81}$$

Now observe $\|\mathbf{A}_{-t}^{-1}\| = \mu_n(\mathbf{A}_{-t})^{-1}$, so by Lemma 25, we have

$$\|\mathbf{A}_{-t}^{-1}\| \lesssim \frac{L}{n\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)}. \tag{3.82}$$

By our assumption $\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda) \geq c$, we have

$$\lambda_1\rho_0(\mathbf{\Sigma};\lambda) = n^{-1}\sum_{i=1}^{k}\lambda_i + \lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda) \leq \lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)(1 + \frac{k\lambda_1}{nc}). \tag{3.83}$$

Also, using the same Lemma and (3.83),

$$\begin{aligned}\frac{(1-k/n)(1+\frac{k\lambda_1}{nc})^{-1}}{L\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)} \lesssim \frac{1-k/n}{L\lambda_1\rho_0(\mathbf{\Sigma};\lambda)} \lesssim \frac{n-k}{\mu_{k+1}(\mathbf{A}_{-t})} \lesssim \operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) =& \sum_{i=1}^{n}\frac{1}{\mu_i(\mathbf{A}_{-t})} \lesssim \frac{n}{\mu_n(\mathbf{A}_{-t})} \\ \lesssim& \frac{L}{\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)}.\end{aligned} \tag{3.84}$$

Finally, plugging in the bounds in (3.84) and (3.82) into (3.80) completes the proof. ■

Lemma 39 (Contamination of ridge estimator for dependent features). *Consider the classification task under the model and assumption described in Section 3.4.4 where $\mathbf{\Sigma} = diag(\lambda_1, \ldots, \lambda_p)$ and the true signal $\theta^* = \frac{1}{\sqrt{\lambda_t}}\mathbf{e}_t$ is 1-sparse in coordinate t. Denote the leave-signal-out covariance and data matrix as $\tilde{\mathbf{\Sigma}} = diag(\lambda_1, \ldots, \lambda_{t-1}, \lambda_{t+1}, \ldots, \lambda_p) = diag(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{p-1})$ and*

$\tilde{\mathbf{X}} = [\mathbf{X}_{:1}, \dots, \mathbf{X}_{:t-1}, \mathbf{X}_{:t+1}, \dots, \mathbf{X}_{:p}]$, *respectively. Let* $\hat{\boldsymbol{\theta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}$ *be a ridge regression estimator. Suppose for some* $k \leq n$*, with probability at least* $1-\delta$*, the condition numbers of* $\tilde{\mathbf{X}}_{k+1:p}\mathbf{\Sigma}_{k+1:p}\tilde{\mathbf{X}}_{k+1:p}^T$ *and* $\lambda\mathbf{I} + \tilde{\mathbf{X}}_{k+1:p}\tilde{\mathbf{X}}_{k+1:p}^T$ *are at most* L' *and* L*, respectively. Then with probability* $1-\delta-5n^{-1}$*, we have:*

$$\sqrt{\frac{\tilde{\lambda}_{k+1}\rho_k(\tilde{\mathbf{\Sigma}}^2;0)}{L'^2\lambda_1^2(1+\rho_0(\mathbf{\Sigma};\lambda))^2}} \lesssim \mathrm{CN}(\hat{\boldsymbol{\theta}}) \lesssim \sqrt{(1+\mathrm{SU}(\hat{\boldsymbol{\theta}})^2)L^2\left(\frac{k}{n}+\frac{n}{R_k(\tilde{\mathbf{\Sigma}};\lambda)}\right)\log n}. \tag{3.85}$$

Proof We begin with the same argument as in Lemma 28 in [27] to write the CN as a quadratic form of signal vector. For notation convenience, we denote the columns of $\mathbf{X}$ to be $\mathbf{X}_{:i},\ i \in \{1, 2, \dots, p\}$, and define the leave-one-out quantities $\tilde{\mathbf{X}} := [\mathbf{X}_{:1}, \dots, \mathbf{X}_{:t-1}, \mathbf{X}_{:t+1}, \dots, \mathbf{X}_{:p}]$, $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\lambda_1, \dots, \lambda_{t-1}, \lambda_{t+1}, \dots, \lambda_p)$, and $\tilde{\mathbf{A}} := \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top + \lambda\mathbf{I}$,. Then,

$$\mathrm{CN}(\hat{\boldsymbol{\theta}})^2 \leq 2\mathbf{y}^\top\widetilde{\mathbf{C}}\mathbf{y} + 2\mathrm{SU}^2\mathbf{z}^\top\widetilde{\mathbf{C}}\mathbf{z},$$

where $\mathbf{z} = \lambda_t^{-1/2}\mathbf{X}_{:t}$ and $\widetilde{\mathbf{C}} := \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{X}}\tilde{\mathbf{A}}^{-1}$. Because of the sparsity assumption and the independence between signal and noise features in Assumption 2, $\mathbf{y}$ and $\mathbf{z}$ are independent of $\widetilde{\mathbf{C}}$. Furthermore, $\mathbf{y}$ and $\mathbf{z}$ are both sub-Gaussian random vector with norm 1 and independent features.

Now consider an ridge estimator with the observation vector ε without looking at the t-feature:

$$\hat{\boldsymbol{\theta}}_{-t}(\varepsilon) = (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top + \lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}^\top\varepsilon.$$

The first key observation here is that

$$\mathbf{y}^\top\widetilde{\mathbf{C}}\mathbf{y} = \|\hat{\boldsymbol{\theta}}_{-t}(\mathbf{y})\|_{\tilde{\mathbf{\Sigma}}}^2, \ \ \mathbf{z}^\top\widetilde{\mathbf{C}}\mathbf{z} = \|\hat{\boldsymbol{\theta}}_{-t}(\mathbf{z})\|_{\tilde{\mathbf{\Sigma}}}^2, \tag{3.86}$$

so we can bound CN as long as we bound the $\|\hat{\boldsymbol{\theta}}_{-t}(\varepsilon)\|_{\tilde{\mathbf{\Sigma}}}^2$ for any sub-Gaussian vector ε independent of $\tilde{\mathbf{X}}$ and has unit norm. The second key observation is that $\|\hat{\boldsymbol{\theta}}_{-t}(\varepsilon)\|_{\tilde{\mathbf{\Sigma}}}^2$ is in fact the variance in the regression analysis.

As shown in Lemma 12 of [26],

$$\|\hat{\boldsymbol{\theta}}_{-t}(\varepsilon)\|_{\tilde{\boldsymbol{\Sigma}}}^2 \leq \frac{\varepsilon^\top \tilde{\mathbf{A}}_k^{-1} \tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1} \tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{A}}_k^{-1} \varepsilon}{\mu_n \left(\tilde{\mathbf{A}}_k^{-1}\right)^2 \mu_k \left(\tilde{\boldsymbol{\Sigma}}_{0:k}^{-1/2} \tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1/2}\right)^2} + \varepsilon^\top \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{X}}_{k:\infty} \tilde{\boldsymbol{\Sigma}}_{k:\infty} \tilde{\mathbf{X}}_{k:\infty}^\top \tilde{\mathbf{A}}^{-1} \varepsilon, \quad (3.87)$$

where $\tilde{\mathbf{A}}_k = \tilde{\mathbf{X}}_{k+1:p} \tilde{\mathbf{X}}_{k+1:p}^\top + \lambda \mathbf{I}$. For self-containment, we sketch the proof on the variance bound. For the first term, by Lemma 28, for some constant c_1, with probability $1 - 2n^{-1}$,

$$\begin{aligned} & \varepsilon^\top \tilde{\mathbf{A}}_k^{-1} \tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1} \tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{A}}_k^{-1} \varepsilon \lesssim \operatorname{tr}\left(\tilde{\mathbf{A}}_k^{-1} \tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1} \tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{A}}_k^{-1}\right) \log n \\ \lesssim\ & \mu_n(\tilde{\mathbf{A}}_k)^{-2} \operatorname{tr}\left(\tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1} \tilde{\mathbf{X}}_{0:k}^\top\right) \log n \lesssim \mu_n(\tilde{\mathbf{A}}_k)^{-2} \cdot nk \log n, \end{aligned}$$

where the last follows from the concentration of sum of sub-Gaussian variables. On the other hand, by Lemma 26, for some constant $c_2 > 0$,

$$\begin{aligned} \mu_n \left(\tilde{\mathbf{A}}_k^{-1}\right)^2 \mu_k \left(\tilde{\boldsymbol{\Sigma}}_{0:k}^{-1/2} \tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1/2}\right)^2 &= \mu_1 \left(\tilde{\mathbf{A}}_k\right)^{-2} \mu_k \left(\tilde{\boldsymbol{\Sigma}}_{0:k}^{-1/2} \tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{X}}_{0:k} \tilde{\boldsymbol{\Sigma}}_{0:k}^{-1/2}\right)^2 \\ &\gtrsim \mu_1 \left(\tilde{\mathbf{A}}_k\right)^{-2} \cdot (n)^2, \end{aligned}$$

with probability $1 - 8\exp(-c_2 t)$.

So the first term is, for some constant $c_3 > 0$, bounded by $L^2 \frac{k}{n}$ with probability $1 - 16\exp(-c_3 t)$. Similarly for the second term, again by Lemma 28, Lemma 24, and Lemma 27, we have for some constant $c_4 > 0$,

$$\begin{aligned} & \varepsilon^\top \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{X}}_{k:\infty} \tilde{\boldsymbol{\Sigma}}_{k:\infty} \tilde{\mathbf{X}}_{k:\infty}^\top \tilde{\mathbf{A}}^{-1} \varepsilon \lesssim \operatorname{tr}\left(\tilde{\mathbf{A}}^{-1} \tilde{\mathbf{X}}_{k:\infty} \tilde{\boldsymbol{\Sigma}}_{k:\infty} \tilde{\mathbf{X}}_{k:\infty}^\top \tilde{\mathbf{A}}^{-1}\right) \log n \\ \lesssim\ & \frac{L^2}{n^2} \frac{1}{\tilde{\lambda}_{k+1}^2 \rho_k^2(\tilde{\boldsymbol{\Sigma}}; \lambda)} \cdot n \sum_{i>k} \tilde{\lambda}_i^2 \log n \lesssim \frac{L^2 n}{R_k(\tilde{\boldsymbol{\Sigma}}; \lambda)} \log n, \end{aligned}$$

with probability $1 - 16\exp(-c_4 t)$.

Combining all above, we deduce that

$$\mathrm{CN}(\hat{\boldsymbol{\theta}})^2 \lesssim (1+\mathrm{SU}(\hat{\boldsymbol{\theta}})^2)L^2\left(\frac{k}{n}+\frac{n}{R_k(\tilde{\boldsymbol{\Sigma}};\lambda)}\right)\log n. \tag{3.88}$$

For the lower bound of $\mathrm{CN}(\hat{\boldsymbol{\theta}})^2$, as shown in [27],

$$\mathrm{CN}(\hat{\boldsymbol{\theta}})^2 = \mathbf{y}^\top\mathbf{C}\mathbf{y} \geq \mu_n(\mathbf{C})\|\mathbf{y}\|_2^2 = n\mu_n(\mathbf{C}), \tag{3.89}$$

where $\mathbf{C} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{X}}^T(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$. Now, by Lemma 29, we have

$$\mu_1(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-2} \lesssim \frac{1}{(\lambda_1 n + \sum_{j=1}^p \lambda_j + \lambda)^2} \lesssim \frac{1}{\lambda_1^2 n^2(1+\rho_0(\boldsymbol{\Sigma};\lambda))^2}. \tag{3.90}$$

Also, by the boundness assumption on the condition number of $\tilde{\mathbf{X}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{X}}^\top$ and Lemma 24 we have

$$\mu_n(\tilde{\mathbf{X}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{X}}^\top) \gtrsim \frac{n}{L'}\tilde{\lambda}_{k+1}\rho_k(\tilde{\boldsymbol{\Sigma}}^2;\lambda), \tag{3.91}$$

with probability $1-\delta-n^{-1}$. Finally, the lower bound in the theorem is established by combining eq. (3.90) and (3.91):

$$\mu_n(\mathbf{C}) \geq \mu_1(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-2}\mu_n(\tilde{\mathbf{X}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{X}}^\top) \gtrsim \frac{\tilde{\lambda}_{k+1}\rho_k(\tilde{\boldsymbol{\Sigma}}^2;0)}{L'^2 n\lambda_1^2(1+\rho_0(\boldsymbol{\Sigma};\lambda))^2}.$$

■

Lemma 40 (Probability of classification error of ridge estimator for dependent features). *Consider the classification task under the model and assumption described in Section 3.4.4 where* $\boldsymbol{\Sigma} = \mathit{diag}(\lambda_1,\ldots,\lambda_p)$ *and the true signal* $\theta^* = \frac{1}{\sqrt{\lambda_t}}\mathbf{e}_t$ *is 1-sparse in coordinate t. Denote the leave-one-out covariance and data matrix as* $\tilde{\boldsymbol{\Sigma}} = \mathit{diag}(\lambda_1,\ldots,\lambda_{t-1},\lambda_{t+1},\ldots,\lambda_p) = \mathit{diag}(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{p-1})$ *and* $\tilde{\mathbf{X}} = [\mathbf{X}_{:1},\ldots,\mathbf{X}_{:t-1},\mathbf{X}_{:t+1},\ldots,\mathbf{X}_{:p}]$*, respectively. Let* $\hat{\boldsymbol{\theta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}$ *be a ridge* *estimator. Suppose for some $t \leq k \leq n$, with probability at least $1-\delta$, the condition numbers of $\tilde{\mathbf{X}}_{k+1:p}\mathbf{\Sigma}_{k+1:p}\tilde{\mathbf{X}}_{k+1:p}^T$ and $\lambda\mathbf{I}+\tilde{\mathbf{X}}_{k+1:p}\tilde{\mathbf{X}}_{k+1:p}^T$ are at most L' and L, respectively and $\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda) \geq c$ for some constant $c > 0$. Then with probability $1-\delta-5n^{-1}$, we have:*

$$\mathrm{POE}(\hat{\theta}) \lesssim \frac{\mathrm{CN}(\hat{\boldsymbol{\theta}})}{\mathrm{SU}(\hat{\boldsymbol{\theta}})}\left(1+\sigma_z\sqrt{\log\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})}{\mathrm{CN}(\hat{\boldsymbol{\theta}})}}\right), \tag{3.92}$$

$$\frac{\lambda_t(1-2\nu^*)\left(1-\frac{k}{n}\right)}{L\left(\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)+\lambda_t L\right)} \lesssim \underbrace{\mathrm{SU}(\hat{\boldsymbol{\theta}})}_{\textit{Survival}} \lesssim \frac{L\lambda_t(1-2\nu^*)}{\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)+L^{-1}\lambda_t\left(1-\frac{k}{n}\right)}, \tag{3.93}$$

$$\sqrt{\frac{\tilde{\lambda}_{k+1}\rho_k(\tilde{\mathbf{\Sigma}}^2;0)}{L'^2\lambda_1^2(1+\rho_0(\mathbf{\Sigma};\lambda))^2}} \lesssim \underbrace{\mathrm{CN}(\hat{\boldsymbol{\theta}})}_{\textit{Contamination}} \lesssim \sqrt{(1+\mathrm{SU}(\hat{\boldsymbol{\theta}})^2)L^2\left(\frac{k}{n}+\frac{n}{R_k(\tilde{\mathbf{\Sigma}};\lambda)}\right)\log n}. \tag{3.94}$$

Furthermore, if the distribution of the covariate $\mathbf{x}$ is Gaussian with independent features, then

$$\mathrm{POE}(\hat{\theta}) = \frac{1}{2}-\frac{1}{\pi}\tan^{-1}\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})}{\mathrm{CN}(\hat{\boldsymbol{\theta}})} \leq \frac{\mathrm{CN}(\hat{\boldsymbol{\theta}})}{\mathrm{SU}(\hat{\boldsymbol{\theta}})}.$$

Proof This is a direct combination of Lemma 37, 38, and 39. ■

Lemma 41 (Bounds on the survival-to-contamination ratio between $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ and $\bar{\boldsymbol{\theta}}_{\mathrm{aug}}$). *Consider an estimator $\hat{\boldsymbol{\theta}}_{aug}$ that solves the objective (3.2). Denote its averaged approximation $\bar{\boldsymbol{\theta}}_{aug}$ as in (3.10). Suppose $\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}} = O(\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug}))$ and $\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}} = O(\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug}))$. Then, the probability of classification error of $\hat{\boldsymbol{\theta}}_{aug}$ can be bounded by:*

$$\frac{1}{\mathrm{EM}}\frac{\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})}{\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})} \leq \frac{\mathrm{SU}(\hat{\boldsymbol{\theta}}_{aug})}{\mathrm{CN}(\hat{\boldsymbol{\theta}}_{aug})} \leq \mathrm{EM}\frac{\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})}{\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})}, \tag{3.95}$$

where EM:= $\exp\left(\left(1+\frac{\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})}\right)\left(1+\frac{\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})}\right)-1\right) \in [1,\infty]$ denotes the error multiplier.

Proof Without ambiguity, we will denote $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ and $\bar{\boldsymbol{\theta}}_{\mathrm{aug}}$ as $\hat{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$, respectively. Define $f(\theta) = \log\frac{\|\mathbf{V}^T\theta\|}{\|\mathbf{U}^T\theta\|}$, where $\mathbf{V} = \mathbf{e}_1$ and $\mathbf{U} = [\mathbf{e}_2, \mathbf{e}_3, \cdots, \mathbf{e}_p]$. Then, for any estimator θ, $\frac{\mathrm{SU}(\theta)}{\mathrm{CN}(\theta)} = \exp(f(\mathbf{\Sigma}^{1/2}\hat{\theta}))$

By the mean value theorem we have

$$f(\mathbf{\Sigma}^{1/2}\hat{\theta}) = f(\mathbf{\Sigma}^{1/2}\bar{\theta}) + \nabla f(\mathbf{\Sigma}^{1/2}\eta)\mathbf{\Sigma}^{1/2}(\hat{\theta} - \bar{\theta}), \tag{3.96}$$

where η is on the line segment between $\hat{\theta}$ and $\bar{\theta}$. Our goal is to show that $\|\nabla f(\Sigma^{1/2}\eta)\|\|\hat{\theta} - \bar{\theta}\|_{\Sigma}$ is small. To this end, firstly, observe that the norm of f's gradient has a clean expression,

$$\begin{aligned}
\|\nabla f(\theta)\| &= \frac{1}{\|\mathbf{U}^\top\theta\|\|\mathbf{V}^\top\theta\|}\left\|\frac{\|\mathbf{U}^\top\theta\|\mathbf{V}\mathbf{V}^\top\theta}{\|\mathbf{V}^\top\theta\|} - \frac{\|\mathbf{V}^\top\theta\|\mathbf{U}\mathbf{U}^\top\theta}{\|\mathbf{U}^\top\theta\|}\right\| && (3.97)\\
&= \frac{1}{\|\mathbf{U}^\top\theta\|\|\mathbf{V}^\top\theta\|}\sqrt{\frac{\|\mathbf{U}^\top\theta\|^2}{\|\mathbf{V}^\top\theta\|^2}\|\mathbf{V}\mathbf{V}^\top\theta\|^2 + \frac{\|\mathbf{V}^\top\theta\|^2}{\|\mathbf{U}^\top\theta\|^2}\|\mathbf{U}\mathbf{U}^\top\theta\|^2} && (3.98)\\
&= \frac{\|\theta\|}{\|\mathbf{U}^T\theta\|\|\mathbf{V}^T\boldsymbol{\theta}\|}. && (3.99)
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\nabla f(\mathbf{\Sigma}^{1/2}\eta)\|\|\hat{\theta} - \bar{\theta}\|_{\Sigma} &\leq \frac{(\|\mathbf{\Sigma}^{1/2}\bar{\theta}\| + t\|\mathbf{\Sigma}^{1/2}(\hat{\theta} - \bar{\theta})\|)\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{(\|\mathbf{U}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\| - t\|\mathbf{U}^T\mathbf{\Sigma}^{1/2}(\hat{\theta} - \bar{\theta})\|)(\|\mathbf{V}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\| - t\|\mathbf{V}^T\mathbf{\Sigma}^{1/2}(\hat{\theta} - \bar{\theta})\|)}\\
&\leq \frac{(\|\bar{\theta}\|_{\mathbf{\Sigma}} + t\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}})\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{(\|\mathbf{U}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\| - t\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}})(\|\mathbf{V}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\| - t\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}})}, && (3.100)
\end{aligned}$$

for some $t \in [0, 1]$. Secondly, we use the assumption that $\mathrm{CN}(\bar{\theta}) = \|\mathbf{U}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\| \gg \|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}$ and $\mathrm{SU}(\bar{\theta}) = \|\mathbf{V}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\| \gg \|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}$ for large enough n. Then, using the fact that $\|\bar{\theta}\|_{\mathbf{\Sigma}} \asymp SU(\bar{\theta}) + CN(\bar{\theta})$, eq. (3.100) is bounded by

$$\begin{aligned}
&\lesssim \left(\frac{1}{\mathrm{SU}(\bar{\theta})} + \frac{1}{\mathrm{CN}(\bar{\theta})} + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})\mathrm{SU}(\bar{\theta})}\right)\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}\\
&= \left(1 + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})}\right)\left(1 + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\theta})}\right) - 1. && (3.101)
\end{aligned}$$

Hence,

$$f(\mathbf{\Sigma}^{1/2}\hat{\theta}) \geq f(\mathbf{\Sigma}^{1/2}\bar{\theta}) - \left(1 + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})}\right)\left(1 + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\theta})}\right) + 1, \tag{3.102}$$

and

$$\frac{\mathrm{SU}(\hat{\theta})}{\mathrm{CN}(\hat{\theta})} \geq \frac{\mathrm{SU}(\bar{\theta})}{\mathrm{CN}(\bar{\theta})} \exp\left(1 - \left(1 + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})}\right)\left(1 + \frac{\|\hat{\theta} - \bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\theta})}\right)\right) := \frac{\mathrm{SU}(\bar{\theta})}{\mathrm{CN}(\bar{\theta})}\frac{1}{\mathrm{EM}}. \tag{3.103}$$

The upper bound follows by an identical argument. ■

Lemma 42. *Let $\hat{\boldsymbol{\theta}}_{aug}$ and $\bar{\boldsymbol{\theta}}_{aug}$ be defined as in (3.10) for a classification task. Recall*

$$\Delta_G := \|\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}\|,$$

and let κ be the condition number of $\mathbf{\Sigma}_{aug}$. Assume $\Delta_G < c$ for some constant $c < 1$. Then,

$$\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}}^2 \leq \kappa\Delta_G^2\left(\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})^2 + \mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})^2\right). \tag{3.104}$$

Proof For ease of notation, we denote $\bar{\mathcal{D}} := \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]$ and $\mathcal{D} = \mathrm{Cov}_G[\mathbf{X}]$. Then,

$$\begin{aligned}
\|\bar{\boldsymbol{\theta}}_{\text{aug}} - \hat{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}}^2 &= \Delta_G^2\|\mathbf{\Sigma}^{1/2}(\mathbf{X}^\top\mathbf{X} + \mathcal{D})^{-1}\bar{\mathcal{D}}^{1/2}\boldsymbol{n}\bar{\mathcal{D}}^{1/2}(\mathbf{X}^\top\mathbf{X} + \bar{\mathcal{D}})^{-1}\mathbf{X}^\top\mathbf{y}\|_2^2 \\
&= n^2\Delta_G^2\|\mathbf{\Sigma}^{1/2}(\mathbf{X}^\top\mathbf{X} + \mathcal{D})^{-1}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + \bar{\mathcal{D}})^{-1}\mathbf{X}^\top\mathbf{y}\|_2^2 \\
&= n^2\Delta_G^2\|\mathbf{\Sigma}^{1/2}(\mathbf{X}^\top\mathbf{X} + \mathcal{D})^{-1}\bar{\mathcal{D}}^{1/2}\bar{\mathcal{D}}^{1/2}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\bar{\boldsymbol{\theta}}_{\text{aug}}\|_2^2 \\
&\leq \frac{\kappa\Delta_G^2 n^2}{\mu_p((\mathbf{X}^\top\mathbf{X} + \mathcal{D})\,\bar{\mathcal{D}}^{-1})^2}\|\bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}}^2 \leq \kappa\Delta_G^2\|\bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}}^2,
\end{aligned}$$

where, by the assumption $\Delta_G < c$, one can prove $\mu_p(\left(\mathbf{X}^\top\mathbf{X} + \mathcal{D}\right)\bar{\mathcal{D}}^{-1})^2 \gtrsim n^2$ similarly as in Lemma 34. Finally, recalling Definition 8, we observe that

$$\|\bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}}^2 = \sum_{i=1}^{p}\lambda_i(\bar{\boldsymbol{\theta}}_{\text{aug}})_i^2 = \mathrm{SU}(\bar{\boldsymbol{\theta}}_{\text{aug}})^2 + \mathrm{CN}(\bar{\boldsymbol{\theta}}_{\text{aug}})^2.$$

■

Remark 43. *Comparing with Lemma 34, we see that the error between $\hat{\boldsymbol{\theta}}_{aug}$ and $\bar{\boldsymbol{\theta}}_{aug}$ for classification and regression are exactly the same with SU^2 and CN^2 replaced by Bias and Var.*

Proof of Theorem 9

Theorem 9 (Bounds on Probability of Classification Error). *Consider the classification task under the model and assumption described in Section 3.4.4 where the true signal θ^* is 1-sparse. Let $\hat{\boldsymbol{\theta}}_{aug}$ be the estimator solving the aERM objective in (3.2). Denote $\Delta_G := \|Cov_G(\mathbf{X}) - \mathbb{E}_{\mathbf{x}}[Cov_g(\mathbf{x})]\|$, let $t \leq n$ be the index (arranged according to the eigenvalues of $\boldsymbol{\Sigma}_{aug}$) of the non-zero coordinate of the true signal, $\tilde{\boldsymbol{\Sigma}}_{aug}$ be the leave-one-out modified spectrum corresponding to index t, κ be the condition number of $\boldsymbol{\Sigma}_{aug}$, and $\tilde{\mathbf{X}}_{aug}$ be the leave-one-column-out data matrix corresponding to column t.*

Suppose data augmentation is performed independently for $\mathbf{x}_{sig}$ and $\mathbf{x}_{noise}$, and there exists a $t \leq k \leq n$ such that with probability at least $1-\delta$, the condition numbers of $n\mathbf{I} + \tilde{\mathbf{X}}^{aug}_{k+1:p}(\mathbf{X}^{aug}_{k+1:p})^\top$ and $n\mathbf{I} + \mathbf{X}^{aug}_{k+1:p}(\mathbf{X}^{aug}_{k+1:p})^\top$ are at most L, and that of $\tilde{\mathbf{X}}_{k+1:p}\boldsymbol{\Sigma}_{k+1:p}\tilde{\mathbf{X}}^T_{k+1:p}$ is at most L_1. Then as long as $\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(SU)$ and $\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(CN)$, with probability $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$, the probability of classification error (POE) can be bounded in terms of the survival (SU) and contamination (CN), as

$$\mathrm{POE}(\hat{\theta}) \lesssim \frac{\mathrm{CN}}{\mathrm{SU}}\left(1 + \sigma_z\sqrt{\log\frac{\mathrm{SU}}{\mathrm{CN}}}\right), \tag{3.105}$$

where

$$\frac{\lambda_t^{aug}(1-2\nu^*)\left(1-\frac{k}{n}\right)}{L\left(\lambda_{k+1}^{aug}\rho_k(\boldsymbol{\Sigma}_{aug};n) + \lambda_t^{aug}L\right)} \lesssim \underbrace{\mathrm{SU}}_{\mathit{Survival}} \lesssim \frac{L\lambda_t^{aug}(1-2\nu^*)}{\lambda_{k+1}^{aug}\rho_k(\boldsymbol{\Sigma}_{aug};n) + L^{-1}\lambda_t^{aug}\left(1-\frac{k}{n}\right)}, \tag{3.106}$$

$$\sqrt{\frac{\tilde{\lambda}_{k+1}^{aug}\rho_k(\tilde{\boldsymbol{\Sigma}}^2_{aug};0)}{L'^2(\lambda_1^{aug})^2(1+\rho_0(\boldsymbol{\Sigma}_{aug};\lambda))^2}} \lesssim \underbrace{\mathrm{CN}}_{\mathit{Contamination}} \lesssim \sqrt{(1+\mathrm{SU}^2)L^2\left(\frac{k}{n} + \frac{n}{R_k(\tilde{\boldsymbol{\Sigma}}_{aug};n)}\right)\log n} \tag{3.107}$$

Furthermore, if $\mathbf{x}$ is Gaussian and the augmentation modified spectrum $\boldsymbol{\Sigma}_{aug}$ is diagonal then we *have tighter bounds of*

$$\frac{1}{2} - \frac{1}{\pi}\tan^{-1} c\frac{\mathrm{SU}}{\mathrm{CN}} \lesssim \mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{1}{c}\frac{\mathrm{SU}}{\mathrm{CN}} \lesssim \frac{\mathrm{CN}}{\mathrm{SU}}, \tag{3.108}$$

where c is a universal constant.

Proof We can prove the theorem by carefully walking through the proofs of Lemma 37, 38, 39, and 41 and noting that the error multiplier defined in Lemma 41 is on the order of a constant under the assumptions made in this theorem. ■

Proof of Theorem 11

Theorem 11 (POE of biased estimators). *Consider the 1-sparse model $\boldsymbol{\theta}^* = \mathbf{e}_t$. and let $\hat{\boldsymbol{\theta}}_{aug}$ be the estimator that solves the aERM in (3.2) with biased augmentation (i.e., $\mu(\mathbf{x}) \neq \mathbf{x}$). Assume that Assumption 2 holds, and the assumptions of Theorem 9 are satisfied for data matrix $\mu(\mathbf{X})$. If the mean augmentation $\mu(\mathbf{x})$ modifies the t-th feature independently of other features and the sign of the t-th feature is preserved under the mean augmentation transformation, i.e.,* $\mathrm{sgn}\left(\mu(\mathbf{x})_t\right) = \mathrm{sgn}\left(\mathbf{x}_t\right)$, *$\forall\mathbf{x}$, then, the POE($\hat{\boldsymbol{\theta}}_{aug}$) is upper bounded by*

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim \mathrm{POE}^o(\hat{\boldsymbol{\theta}}_{aug}), \tag{3.109}$$

where $\mathrm{POE}^o(\hat{\boldsymbol{\theta}}_{aug})$ is any bound in Theorem 9 with $\mathbf{X}$ and $\boldsymbol{\Sigma}$ replaced by $\mu(\mathbf{X})$ and $\bar{\boldsymbol{\Sigma}}$, respectively.

Proof First, from Lemma 37, we know that the POE can be written as a function of the SU and CN of $\hat{\boldsymbol{\theta}}_{\text{aug}}$. Next, recall that from E. q. (3.7), the biased estimator is given by

$$\hat{\boldsymbol{\theta}}_{\text{aug}} = (\mu_{\mathcal{G}}(\mathbf{X})^T\mu_{\mathcal{G}}(\mathbf{X}) + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu_{\mathcal{G}}(\mathbf{X})^T\mathbf{y}.$$

Now, observe that this estimator is almost equivalent to the one with training covariates

$$\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), \ldots, \mu(\mathbf{x}_n),$$

except that the observation vector $\mathbf{y}$ consists of the signs of $\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \ldots, \mathbf{x}_{n,t}$ instead of $\tilde{\mathbf{y}}$, the signs of $\mu(\mathbf{x}_{1,t}), \mu(\mathbf{x}_{2,t}), \ldots, \mu(\mathbf{x}_{n,t})$. However, $\mathbf{y}$ equals $\tilde{\mathbf{y}}$ by our assumption that the sign of the t-th feature is preserved under the mean augmentation transform. So we can bound the SU and CN of $\hat{\boldsymbol{\theta}}_{\text{aug}}$ by just utilizing the bounds in Theorem 9 with $\mathbf{X}$ and $\boldsymbol{\Sigma}$ replaced by $\mu(\mathbf{X})$ and $\bar{\boldsymbol{\Sigma}}$, respectively. ■

Proofs of Corollaries

Corollary 14 (Classification bounds for uniform random mask augmentation). *Let $\hat{\boldsymbol{\theta}}_{aug}$ be the estimator computed by solving the aERM objective on binary labels with mask probability β, and denote $\psi := \frac{\beta}{1-\beta}$. Assume $p \ll n^2$. Then, with probability at least $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$*

$$\text{POE} \lesssim Q^{-1}(1 + \sqrt{\log Q}) \textit{ where} \tag{3.110}$$

$$Q = (1 - 2\nu)\sqrt{\frac{n}{p \log n}} \left(1 + \frac{n}{n\psi + p}\right)^{-1}. \tag{3.111}$$

In addition, if we assume the input data has independent Gaussian features, then we have tight generalization bounds

$$\text{POE} \asymp \frac{1}{2} - \frac{1}{\pi} \tan^{-1} Q \tag{3.112}$$

with the same probability.

Proof We first note the following key quantities:

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \psi \mathrm{diag}(\mathbf{\Sigma}) = \psi \mathbf{\Sigma},\ \boldsymbol{\theta}^*_{\text{aug}} = \psi^{1/2}\mathbf{\Sigma}^{1/2}\boldsymbol{\theta}^*,\ \mathbf{\Sigma}_{\text{aug}} = \psi^{-1}\mathbf{I},\ \lambda^{\text{aug}} = \psi^{-1},$$

and the effective ranks of the augmentation modified spectrum are

$$\rho_k^{\text{aug}} = \frac{\psi n + p - k}{n}, \tag{3.113}$$

$$R_k^{\text{aug}} = \frac{(\psi n + p - k)^2}{p - k}. \tag{3.114}$$

Substituting into Theorem 9 yields the formulas for the components of POE

$$\mathrm{SU} \asymp (1 - 2\nu)\frac{n}{n\psi + n + p}, \tag{3.115}$$

$$\sqrt{\frac{np}{(n\psi + p)^2}} \lesssim \mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\frac{np\log n}{(n\psi + p)^2}} \tag{3.116}$$

$$\tag{3.117}$$

It remains to check when the conditions $\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}} = O(SU)$ and $\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}} = O(CN)$ are met. When p grows faster than n, we will have $SU \asymp \frac{n}{p}$ and $CU \lesssim \sqrt{\frac{n}{p}}$. Then, using Lemma 42, we have

$$\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|_{\mathbf{\Sigma}} \lesssim \kappa^{1/2}\Delta_G(SU + CN) \tag{3.118}$$

$$\lesssim \sigma_{\mathbf{z}}^2\sqrt{\frac{\log n}{n}}\sqrt{\frac{n}{p}} \tag{3.119}$$

So, the condition is met for $p \ll n^2$. ■

Corollary 17 (Group invariant augmentation). *An augmentation class $\mathcal{G}$ is said to be group-invariant if $g(\mathbf{x}) \overset{d}{=} \mathbf{x},\ \forall g \in \mathcal{G}$. For such a class, the augmentation modified spectrum $\mathbf{\Sigma}_{aug}$ in*

Theorem 9 is given by

$$\mathbf{0} \preceq \mathbf{\Sigma}_{aug} = \mathbf{\Sigma} - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})\mu_{\mathcal{G}}(\mathbf{x})]^{\top} \preceq \mathbf{\Sigma}.$$

Consider the case where the input covariates satisfy $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. *Let* $\mathbf{x}'$ *be i.i.d. with* $\mathbf{x}$ *and consider the group-invariant augmentation given by* $g(\mathbf{x}) = \frac{1}{\sqrt{2}}\mathbf{x} + \frac{1}{\sqrt{2}}\mathbf{x}'$*. Then, under the assumptions of 9 and with probability at least* $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$*, this augmented estimator has generalization error*

$$\mathrm{POE} \asymp \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{\mathrm{SU}}{\mathrm{CN}}, \text{ where} \tag{3.120}$$

$$\mathrm{SU} \asymp (1-2\nu)\frac{n}{2n+p}, \quad \sqrt{\frac{np}{(n+p)^2}} \lesssim \mathrm{CN} \lesssim \sqrt{(1+\mathrm{SU}^2)\frac{np\log n}{(n+p)^2}}. \tag{3.121}$$

Proof By definition and the assumption of group invariance,

$$\begin{aligned}\mathbf{\Sigma}_{aug} &= \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \mathbb{E}_{\mathbf{x}}\mathbb{E}_g[g(\mathbf{x})g(\mathbf{x})^{\top} - \mathbb{E}_g[g(\mathbf{x})]\mathbb{E}_g[g(\mathbf{x})]^{\top}]\\ &= \mathbb{E}_g\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})g(\mathbf{x})^{\top} - \mu_{\mathcal{G}}(\mathbf{x})\mu_{\mathcal{G}}(\mathbf{x})^{\top}] = \mathbb{E}_g\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^{\top} - \mu_{\mathcal{G}}(\mathbf{x})\mu_{\mathcal{G}}(\mathbf{x})^{\top}]\\ &= \mathbf{\Sigma} - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})\mu_{\mathcal{G}}(\mathbf{x})]^{\top}.\end{aligned}$$

The change of the expectation order follows from the Tonelli's theorem, while the last inequality is by the group invariance assumption. Now applying Theorem 9 completes the proof for $\mathbf{\Sigma}_{\mathrm{aug}}$.

Now, for the example in this corollary, first note that this is a group-invariant augmentation as $g(\mathbf{x})$ is Gaussian with the same mean and covariance as $\mathbf{x}$. Direct calculations show that $\mu_{\mathcal{G}}(\mathbf{x}) = \frac{1}{\sqrt{2}}\mathbf{x}$ and $\mathbf{\Sigma}_{\mathrm{aug}} = \frac{1}{2}\mathbf{\Sigma}$. Furthermore, $\mathrm{Cov}_G(\mathbf{X}) = \frac{1}{2}\mathbf{\Sigma}$ is a constant matrix so $\Delta_G = 0$ and the approximation error is zero. Now applying Theorem 9 and 11 yields the result. ■

Corollary 19 (Generalization of random rotation). *The estimator induced by the random-rotation* *augmentation (with angle parameter α) can be expressed as*

$$\hat{\boldsymbol{\theta}}_{\textit{rot}} = \left(\mathbf{X}^\top\mathbf{X} + \frac{4(1-\cos\alpha)}{p}\left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} - \mathbf{X}^\top\mathbf{X}\right)\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

An application of Theorem 4 yields

$$\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\textit{rot}}) \asymp \mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\textit{lse}}),$$

for sufficiently large p (overparameterized regime), as well as the variance bound

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\textit{rot}}) \lesssim \mathrm{Var}(\hat{\boldsymbol{\theta}}_{\textit{ridge},\lambda}),$$

Above, $\hat{\boldsymbol{\theta}}_{\textit{lse}}$ and $\hat{\boldsymbol{\theta}}_{\textit{ridge},\lambda}$ denote the least squared estimator and ridge estimator with ridge intensity $\lambda = np^{-1}(1-\cos\alpha)\sum_j \lambda_j$. The approximation error can also be shown to decay as

$$\mathrm{Approx.Error}(\hat{\boldsymbol{\theta}}_{\textit{rot}}) \lesssim \max\left(\frac{1}{n}, \frac{1}{\mathrm{tr}(\boldsymbol{\Sigma})}\right).$$

Proof The proof is based on the application of Theorem 4, where

$$\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) = \frac{4(1-\cos\alpha)}{p}(\mathrm{Tr}(\boldsymbol{\Sigma})\mathbf{I} - \boldsymbol{\Sigma}),\ \boldsymbol{\Sigma}_{\mathrm{aug}} = \frac{p}{4(1-\cos\alpha)}\boldsymbol{\Sigma}(\mathrm{Tr}(\boldsymbol{\Sigma})\mathbf{I} - \boldsymbol{\Sigma})^{-1}.$$

Hence, $\lambda_i^{\mathrm{aug}} \asymp \frac{p}{4(1-\cos\alpha)}\frac{\lambda_i}{\sum_j \lambda_j}$, and

$$\begin{aligned}
&\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\mathrm{rot}}) \\
&\lesssim \|\boldsymbol{\theta}^*_{k+1:\infty}\|^2_{\boldsymbol{\Sigma}_{k+1:\infty}} + \sum_{i=1}^{k}\frac{(\boldsymbol{\theta}^*_i\sum_{j\neq i}\lambda_j)^2}{\lambda_i}\left(1 + \frac{p}{4(1-\cos\alpha)n}\frac{\sum_{j>k}\lambda_j}{\sum_j\lambda_j}\right)^2\left(\frac{4(1-\cos\alpha)}{p}\right)^2 \\
&\lesssim \|\boldsymbol{\theta}^*_{k+1:\infty}\|^2_{\boldsymbol{\Sigma}_{k+1:\infty}} + \sum_{i=1}^{k}\frac{(\boldsymbol{\theta}^*_i\sum_{j\neq i}\lambda_j)^2}{\lambda_i}\left(\frac{\sum_{j>k}\lambda_j}{n\sum_j\lambda_j}\right)^2 \text{, for sufficiently large } p \\
&\asymp \|\boldsymbol{\theta}^*_{k+1:\infty}\|^2_{\boldsymbol{\Sigma}_{k+1:\infty}} + \|\boldsymbol{\theta}^*_{1:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{1:k}}\lambda^2_{k+1}\rho_k(\boldsymbol{\Sigma};0)^2 = \mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\mathrm{lse}}),
\end{aligned}$$

where the last equality is by Corollary 36 with $\lambda = 0$. The variance part can be proved similarly. The approximation error bound is proved in Appendix 3.8.6. ■

Corollary 44 (Classification bounds for Gaussian noise injection). *Consider the independent, additive Gaussian noise augmentation:* $g(\mathbf{x}) = \mathbf{x} + \mathbf{n}$*, where* $\mathbf{n} \sim \mathcal{N}(0, \sigma^2)$*. Let* $\tilde{\boldsymbol{\Sigma}}$ *be the leave-one-out spectrum corresponding to index* t*. Then, with probability at least* $1 - \exp(\sqrt{n}) - 5n^{-1}$*,*

$$\mathrm{SU} \asymp (1 - 2\nu^*) \frac{\lambda_t}{\lambda_{k+1}\rho_k(\boldsymbol{\Sigma}; n\sigma^2) + \lambda_t}, \tag{3.122}$$

$$\mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\left(\frac{k}{n} + \frac{n}{R_k(\tilde{\boldsymbol{\Sigma}}; n\sigma^2)}\right) \log n}, \tag{3.123}$$

$$\tag{3.124}$$

and $\mathrm{EM} = 1$.

Proof *As in the regression analysis, we note that in this case, the key quantities are given by*

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \sigma^2 \mathbf{I},\ \boldsymbol{\theta}^*_{aug} = \sigma \boldsymbol{\theta}^*,\ \boldsymbol{\Sigma}_{aug} = \sigma^{-2}\boldsymbol{\Sigma},\ \lambda^{aug} = \sigma^{-2}\lambda,$$

and the effective ranks are given by

$$\rho_k(\boldsymbol{\Sigma}_{aug}; n) = \rho_k(\boldsymbol{\Sigma}; n\sigma^2),$$
$$R_k(\boldsymbol{\Sigma}_{aug}; n) = R_k(\boldsymbol{\Sigma}; n\sigma^2).$$

Finally, $\log(EM)$ *is zero because* $\Delta_G = 0$*. Substituting the above quantities into the Theorem 9 yields the result.* ■

Corollary 45 (Classification bounds for non-uniform random mask). *Consider the case where* *the dropout parameter* $\psi_j = \frac{\beta_j}{1-\beta_j}$ *is applied to the* j*-th feature, and assume the conditions of Theorem 9 are met. For simplicity, we consider the bi-level case where* $\psi_j = \psi$ *for* $j \neq t$*. Then, with probability at least* $1 - \delta - \exp(\sqrt{n}) - 5n^{-1}$*,*

$$\mathrm{SU} \asymp \frac{1}{\psi_1 + \frac{p\psi_t}{n\psi} + 1} \tag{3.125}$$

$$\mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\frac{np\log n}{(n\psi + p)^2}} \tag{3.126}$$

Proof

Let Ψ denote the diagonal matrix with $\Psi_{i,i} = \psi$ if $i \neq t$ and $\Psi_{t,t} = \psi_t$.

We can then compute the following key quantities:

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \Psi\boldsymbol{\Sigma},\ \boldsymbol{\theta}^*_{\text{aug}} = \Psi^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*,\ \boldsymbol{\Sigma}_{\text{aug}} = \Psi^{-1},$$

and the effective ranks of the augmentation modified spectrum are

$$\rho_k^{\text{aug}} = \frac{\psi n + p - k}{n}, \tag{3.127}$$

$$R_k^{\text{aug}} = \frac{(\psi n + p - k)^2}{p - k}. \tag{3.128}$$

The approximation error bound proceeds as in the uniform random mask case. Substituting the above quantities into Theorem 9 completes the proof. ■

3.8.4 Comparisons between Regression and Classification

Proof of Proposition 20

Proposition 20 (DA is easier to tune in classification than regression). *Consider the 1-sparse model* $\boldsymbol{\theta}^* = \sqrt{\frac{1}{\lambda_t}}\mathbf{e}_t$ *for Gaussian covariate with independent components and an independent feature augmentation. Suppose that the approximation error is not dominant in the bounds of*

Theorem 4 (simple sufficient conditions can be found in Lemma 35 in Appendix 3.8.1) and the assumptions in the two theorems hold, then,

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim \sqrt{(\lambda^{aug}_{k+1}\rho_k(\boldsymbol{\Sigma}_{aug};n))^2 \cdot \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{aug};n)} + \frac{k}{n}\right)\log n},$$
$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug}) \gtrsim (\lambda^{aug}_{k+1}\rho_k(\boldsymbol{\Sigma}_{aug};n))^2 + \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{aug};n)} + \frac{k}{n}\right).$$

As a consequence, the regression risk serves as a surrogate for the classification risk up to a log-factor:

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug})\sqrt{\log n}. \tag{3.129}$$

As concrete examples of the regression risk being a surrogate of classification risk, consider Gaussian noise injection augmentation with noise standard deviation σ and random mask with dropout probability β to train the 1-sparse model in the decaying data spectrum $\boldsymbol{\Sigma}_{ii} = \gamma^i$, $\forall i \in \{1, 2, \dots, p\}$, where γ is some constant satisfying $0 < \gamma < 1$. Let $\hat{\boldsymbol{\theta}}_{gn}$ and $\hat{\boldsymbol{\theta}}_{rm}$ be the corresponding estimators, then

$$\lim_{n\to\infty}\lim_{\sigma\to\infty}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{gn}) = 0 \;\; \textit{while} \;\; \lim_{n\to\infty}\lim_{\sigma\to\infty}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{gn}) = 1. \tag{3.130}$$

Also, when $p\log n \ll n$,

$$\lim_{n\to\infty}\lim_{\beta\to 1}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{rm}) = 0 \;\; \textit{while} \;\; \lim_{n\to\infty}\lim_{\beta\to 1}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{rm}) = 1. \tag{3.131}$$

Furthermore, the augmentation of Gaussian injection has gone through significant distributional shift where

$$\frac{W_2^2(g(\mathbf{x}), \mathbf{x})}{p} \xrightarrow{n,\sigma} \infty, \tag{3.132}$$

in which W_2 denotes the 2-Wasserstein distance between the pre- and post-augmented distribution of the data by the Gaussian noise injection.

Proof We begin with proving the first statement. By our assumption that the approximation error and error multiplier are not dominant terms in generalization errors, we can only consider bias/variance and survival/contamination. By Proposition 5, the regression testing risk is bounded by

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{aug}}) \lesssim (\lambda_{k+1}^{\text{aug}} \rho_k(\boldsymbol{\Sigma}_{\text{aug}}; n))^2 + \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{\text{aug}}; n)} + \frac{k}{n}\right).$$

However, by the independence of the original data feature components and their augmentations and the boundness assumption on ρ_k, Lemma 2, Lemma 3 and Theorem 5 in [26] shows that there is a matching lower bound such that

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{aug}}) \gtrsim (\lambda_{k+1}^{\text{aug}} \rho_k(\boldsymbol{\Sigma}_{\text{aug}}; n))^2 + \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{\text{aug}}; n)} + \frac{k}{n}\right), \tag{3.133}$$

for some k. On the other hand, by Theorem 9, we know that

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{\text{aug}}) \lesssim \sqrt{(\lambda_{k+1}^{\text{aug}} \rho_k(\boldsymbol{\Sigma}_{\text{aug}}; n))^2 \cdot \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{\text{aug}}; n)} + \frac{k}{n}\right) \log n}, \tag{3.134}$$

for any k. Now combining E. q. (3.133) and (3.134) along with the inequality $x + y \geq 2\sqrt{xy}$ for any $x, y \geq 0$ proves the first statement.

To prove the second statement about $\hat{\theta}_{\text{gn}}$, note that $\hat{\theta}_{\text{gn}} = (\mathbf{X}^\top \mathbf{X} + \sigma^2 n \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \to 0$ almost surely as $\sigma \to \infty$, so

$$\mathrm{MSE}(\hat{\theta}_{\text{gn}}) = \|\boldsymbol{\theta}^* - \hat{\theta}_{\text{gn}}\|_{\boldsymbol{\Sigma}} \xrightarrow{a.s.} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} = 1.$$

On the other hand, by Theorem 9, choose $k = 0$, then

$$\mathrm{SU}(\hat{\theta}_{\text{gn}}) \gtrsim \frac{n \frac{\lambda_t}{\sigma^2}}{n + \frac{\sum \lambda_j}{\sigma^2} + \frac{n \lambda_t}{\sigma^2}}, \quad \mathrm{CN}(\hat{\theta}_{\text{gn}}) \lesssim \frac{1}{\sigma^2} \sqrt{\frac{(\sum \lambda_j^2) n \log n}{(n + \sum \frac{\lambda_j}{\sigma^2})^2}},$$

So,

$$\mathrm{POE}(\hat{\theta}_{\text{gn}}) \leq \frac{\mathrm{CN}(\hat{\theta}_{\text{gn}})}{\mathrm{SU}(\hat{\theta}_{\text{gn}})} \asymp \frac{1}{\lambda_t}\sqrt{\frac{(\sum \lambda_j^2)\log n}{n}} \times \frac{n + \sum \frac{\lambda_j}{\sigma^2} + \frac{n\lambda_t}{\sigma^2}}{n + \sum \frac{\lambda_j}{\sigma^2}}, \tag{3.135}$$

$$\lim_{n\to\infty}\lim_{\sigma\to\infty} \mathrm{POE}(\hat{\theta}_{\text{gn}}) = \lim_{n\to\infty} \frac{1}{\lambda_t}\sqrt{\frac{\log n}{n(1-\gamma^2)}} = 0. \tag{3.136}$$

We can prove the statement for $\hat{\theta}_{\text{rm}}$ similarly. When $\beta \to 1$, $\hat{\theta}_{\text{rm}} = (\mathbf{X}^\top\mathbf{X} + \frac{\beta}{1-\beta}\operatorname{diag}[\mathbf{X}^\top\mathbf{X}])^{-1}\mathbf{X}^\top\mathbf{y} \to 0$ almost surely. So MSE approaches 1 almost surely. But by Corollary 14, we have

$$\lim_{n\to\infty}\lim_{\beta\to 1} \mathrm{POE}(\hat{\theta}_{\text{rm}}) = \lim_{n\to\infty} \sqrt{\frac{p\log n}{n}} = 0. \tag{3.137}$$

Finally, by the closed-form formula of Wasserstein distance between Gaussian distributions,

$$W_2(g(\mathbf{x}), \mathbf{x}) = \|(\mathbf{\Sigma} + \sigma^2\mathbf{I})^{\frac{1}{2}} - \mathbf{\Sigma}^{\frac{1}{2}}\|_F^2 = \Omega(p\sigma^2). \tag{3.138}$$

■

Proof of Proposition 21

Proposition 21 (Non-uniform random mask is easier to tune in classification). *Consider the 1-sparse model $\boldsymbol{\theta}^* = \sqrt{\frac{1}{\lambda_t}}\mathbf{e}_t$. Suppose the approximation error is not dominant in the bounds of Theorem 4 (simple sufficient conditions can be found in Lemma 35 in Appendix 3.8.1) and the assumptions in the two theorems hold. Suppose we apply the non-uniform random mask augmentation and recall the definitions of ψ and ψ_t as in Corollary 45. Then, if $\sqrt{\frac{p}{n}} \ll \frac{\psi}{\psi_t} \ll \frac{p}{n}$, we have*

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{\mathit{rm}}) \xrightarrow{n} 0 \ \textit{while}\ \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\mathit{rm}}) \xrightarrow{n} 1. \tag{3.139}$$

Proof From Corollary 18, we have that the bias scales as

$$\text{Bias} \lesssim \frac{(\psi_t n + \frac{\psi_t p}{\psi})^2}{n^2 + (\psi_t n + \frac{\psi_t p}{\psi})^2} \asymp \frac{(\psi_t n + \frac{\psi_t p}{\psi})^2}{(\psi_t n + \frac{\psi_t p}{\psi})^2} = 1,$$

where the second asymptotic equality uses the assumption that $\frac{\psi_t p}{\psi} \gg n$. Hence the MSE approaches a constant (here we use the fact that the MSE bound is tight when the approximation error is non-dominant, as per [26]). Next we use the bounds in Corollary 45 to find that

$$\text{SU} \asymp \frac{1}{\psi_t + \frac{p\psi_t}{n\psi} + 1} \asymp \frac{1}{\psi_t + \frac{p\psi_t}{n\psi}}, \quad \text{CN} \asymp \sqrt{\frac{np}{(n\psi + p)^2}}.$$

So, if $p \gg n\psi$, we have

$$\frac{SU}{CN} \asymp \frac{1/\psi_t}{(1/\psi)\sqrt{p/n}} = \frac{\psi/\psi_t}{\sqrt{p/n}} \to \infty,$$

and if $p \ll n\psi$, we have

$$\frac{SU}{CN} \asymp \frac{\frac{n\psi}{p\psi_t}}{\sqrt{\frac{n}{p}}} = \frac{\psi/\psi_t}{\sqrt{p/n}} \to \infty.$$

Since we assume we are operating in a regime where the approximation error and error multiplier do not dominate, we can conclude that POE $\to 0$. ■

3.8.5 Derivations of Common Augmented Estimators

Proposition 46 (Common augmentation estimators). *Below are closed-form expression of estimators that solves (3.2) with common data augmentation.*

- Gaussian noise injection with zero-mean noise of covariance $\mathbf{W}$:

$$\hat{\boldsymbol{\theta}}_{aug} = (\mathbf{X}^\top \mathbf{X} + n\mathbf{W})^{-1}\mathbf{X}^\top \mathbf{y}$$

- Unbiased random mask with mask probability β:

$$\hat{\boldsymbol{\theta}}_{aug} = \left(\mathbf{X}^\top\mathbf{X} + \frac{\beta}{1-\beta}\,\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

- Unbiased random cutout with number of cutout features k:

$$\left(\mathbf{X}^\top\mathbf{X} + \frac{p}{p-k}\mathbf{M}\odot\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y},$$

where $\mathbf{M}_{i,j} = \frac{k}{p} - \frac{|j-i|\mathbf{1}_{|j-i|<k-1} + k\mathbf{1}_{|j-i|\geq k-1}}{p-k}$.

- Salt and Pepper (β, μ, σ^2):

$$\hat{\boldsymbol{\theta}}_{aug} = \left(\mathbf{X}^\top\mathbf{X} + \frac{\beta}{1-\beta}\,\mathrm{diag}\left(\mathbf{X}^\top\mathbf{X}\right) + \frac{\beta\sigma^2\mathrm{n}}{(1-\beta)^2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

- Unbiased random rotation with angle α:

$$\hat{\boldsymbol{\theta}}_{aug} = \left(\mathbf{X}^\top\mathbf{X} + \frac{4(1-\cos\alpha)}{p^2}\left(\mathrm{Tr}\left(\mathbf{X}\mathbf{X}^\top\right)\mathbf{I} - \mathbf{X}\mathbf{X}^\top\right)\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Proof To prove all the unbiased augmented estimator formulas, it suffices to derive $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$. Then,

$$\hat{\boldsymbol{\theta}}_{\mathrm{aug}} = (\mathbf{X}^\top\mathbf{X} + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^\top\mathbf{X}\mathbf{y}.$$

Gaussian noise injection $g(\mathbf{x}) = \mathbf{x} + \mathbf{n}$, where $\mathbf{n} \sim \mathcal{N}(0, \mathbf{W})$. Therefore,

$$\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = n^{-1}\sum_i \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}_i) = n^{-1}\sum_i \mathbb{E}_{\mathbf{n}_i}[(\mathbf{x}_i + \mathbf{n}_i)(\mathbf{x} + \mathbf{n}_i)^\top - \mathbf{x}_i\mathbf{x}_i^\top] = \mathbf{W}.$$

Unbiased random mask $g(\mathbf{x}) = (1-\beta)^{-1}\mathbf{b}\odot\mathbf{x}$, where $\mathbf{b}$ has i.i.d. Bernoulli random variable with dropout probability β in its component. The factor $(1-\beta)^{-1}$ is to rescale the estimator to be unbiased. Hence,

$$
\begin{aligned}
\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) &= (1-\beta)^{-2} n^{-1} \sum_i \mathbb{E}_{\mathbf{b}_i}[\mathbf{b}_i \mathbf{b}_i^\top \odot \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{x}_i \mathbf{x}_i^\top] \\
&= n^{-1} \sum_i \left(\frac{\beta}{1-\beta} \mathbf{I} + \mathbf{1}\mathbf{1}^\top - \mathbf{1}\mathbf{1}^\top \right) \odot \mathbf{x}_i \mathbf{x}_i^\top = n^{-1} \frac{\beta}{1-\beta} \operatorname{diag}\left(\mathbf{X}^\top \mathbf{X}\right)
\end{aligned}
$$

Random cutout Define $h(\mathbf{x})$ to be the random cutout of k consecutive features, then the unbiased cutout can be written as $g(\mathbf{x}) = \frac{p}{p-k} h(\mathbf{x})$ as $\mathbb{E}_h h(\mathbf{x}) = \frac{p-k}{k}\mathbf{x}$. Now,

$$
\mathrm{Cov}_h(\mathbf{x}) = \mathbb{E}_h[h(\mathbf{x}) h(\mathbf{x})^\top] - \left(\frac{p-k}{p} \right)^2 \mathbf{x}\mathbf{x}^\top.
$$

Note that $\mathbb{E}_h h(\mathbf{x}) h(\mathbf{x})^\top = \mathbf{H} \odot \mathbf{x}\mathbf{x}^\top$, where

$$
\begin{aligned}
\mathbf{H}_{ij} &= \mathrm{P}[\mathbf{x}_i \text{ is not cutout and } \mathbf{x}_j \text{ is not cutout}] \\
&= \mathrm{P}[\text{a random } k \text{ consecutive features does not cover } i \text{ nor } j] \\
&= \frac{p - k - |j-i| \mathbf{1}_{|j-i|<k-1} - k \mathbf{1}_{|j-i| \geq k-1}}{p}.
\end{aligned}
$$

Hence,

$$
\mathrm{Cov}_h(\mathbf{x}) = \left(\mathbf{H} - \left(\frac{p-k}{p} \right)^2 \mathbf{1}\mathbf{1}^\top \right) \odot \mathbf{x}\mathbf{x}^\top,
$$

$$
\left(\mathbf{H} - \left(\frac{p-k}{p} \right)^2 \mathbf{1}\mathbf{1}^\top \right)_{ij} = \frac{p-k}{p} \frac{k}{p} - \frac{|j-i| \mathbf{1}_{|j-i|<k-1} + k \mathbf{1}_{|j-i| \geq k-1}}{p},
$$

and

$$
\begin{aligned}
\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) &= \left(\frac{p}{p-k} \right)^2 \mathrm{Cov}_h(\mathbf{x}) \\
&= \frac{p}{p-k} \left(\frac{k}{p} - \frac{|j-i| \mathbf{1}_{|j-i|<k-1} + k \mathbf{1}_{|j-i| \geq k-1}}{p-k} \right) \odot \mathbf{x}\mathbf{x}^\top \\
&= \frac{p}{p-k} \mathbf{M} \odot \mathbf{x}\mathbf{x}^\top.
\end{aligned}
$$

Salt and pepper This estimator can be derived similarly by combining the derivations of the random mask and the injection of Gaussian noise by writing the augmentation as

$$g(\mathbf{x}) = (1-\beta)^{-1} \left(\mathbf{b} \odot \mathbf{x} + (\mathbf{1} - \mathbf{b}) \odot \mathbf{n}\right),$$

where $\mathbf{b}$ has i.i.d. components of Bernoulli random variables with parameter β and $\mathbf{n} \sim \mathcal{N}(0, \mathbf{I})$.

Random rotation Given a training example $\mathbf{x}$, we will consider rotating $\mathbf{x}$ by a angle α in $\frac{p}{2}$ random plane spanned by two randomly generated orthonormal vectors $\mathbf{u}$ and $\mathbf{v}$. For rotation in each one of the plan, the data transformation can be written by

$$h(\mathbf{x}) = (\mathbf{I} + \sin\alpha(\mathbf{v}\mathbf{u}^\top - \mathbf{u}\mathbf{v}^\top) + (\cos\alpha - 1)(\mathbf{u}\mathbf{u}^\top + \mathbf{v}\mathbf{v}^\top))\mathbf{x}. \tag{3.140}$$

The bias of h is $\Delta = \mathbb{E}_{\mathbf{u},\mathbf{v}}[h(\mathbf{x})] - \mathbf{x}$. We consider the unbiased transform g by subtracting the bias from h where $g(\mathbf{x}) := h(\mathbf{x}) - \Delta$. Since we consider random $\mathbf{u}$ and $\mathbf{v}$, they are distributed uniformly on the sphere of $\mathbf{R}^p$ but orthogonal to each other. The exact joint distribution of $\mathbf{u}$ and $\mathbf{v}$ is intractable, but fortunately when p is large, we know from high dimensional statistics that they are approximately independent vector of $\mathcal{N}(0, \frac{1}{p}\mathbf{I})$. We will thus use this approximation to facilitate our derivation.

Firstly,

$$\mathbb{E}_{\mathbf{u},\mathbf{v}}[h(\mathbf{x})] = \mathbf{x} + \mathbb{E}_{\mathbf{u}} 2(\cos\alpha - 1)\mathbf{u}\mathbf{u}^\top\mathbf{x} = \mathbf{x} + \frac{2}{p}\mathbf{x},$$

so the bias $\Delta = \frac{2}{p}\mathbf{x}$ which is small in high dimensional space. Secondly, subtracting Δ from h, we proceed to calculate the $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \frac{\sum_{i=1}^{n} \mathrm{Cov}_{\mathrm{g_i}}(\mathbf{x}_i)}{n}$ according to Definition 1. After simplification, we have

$$
\begin{aligned}
\mathrm{Cov}_{\mathrm{g_i}}(\mathbf{x}_i) &= \mathbb{E}_g g(\mathbf{x}_i) g(\mathbf{x}_i)^\top \\
&= \mathbb{E}_{\mathbf{u},\mathbf{v}} \Bigg[\sin^2\alpha \left(\mathbf{v}\mathbf{u}^\top - \mathbf{u}\mathbf{v}^\top\right) \mathbf{x}\mathbf{x}^\top (\mathbf{v}\mathbf{u}^\top - \mathbf{u}\mathbf{v}^\top) \\
&+ (\cos\alpha - 1)^2 \left(\mathbf{u}\mathbf{u}^\top + \mathbf{v}\mathbf{v}^\top - \frac{2}{p}\mathbf{I}\right) \mathbf{x}\mathbf{x}^\top \left(\mathbf{u}\mathbf{u}^\top + \mathbf{v}\mathbf{v}^\top - \frac{2}{p}\mathbf{I}\right) \Bigg] \\
&= 2\sin^2\alpha \left(\mathbb{E}_{\mathbf{u},\mathbf{v}} \left[\langle \mathbf{v}, \mathbf{x}\rangle \langle \mathbf{u}, \mathbf{x}\rangle \mathbf{u}\mathbf{v}^\top - \langle \mathbf{u}, \mathbf{x}\rangle^2 \mathbf{v}\mathbf{v}^\top\right]\right) \\
&+ 2(\cos\alpha - 1)^2 \left(\mathbb{E}_{\mathbf{u},\mathbf{v}} \left[\langle \mathbf{u}, \mathbf{x}\rangle^2 \mathbf{v}\mathbf{v}^\top + \langle \mathbf{v}, \mathbf{x}\rangle \langle \mathbf{u}, \mathbf{x}\rangle \mathbf{u}\mathbf{v}^\top - \frac{4}{p} \langle \mathbf{u}, \mathbf{x}\rangle \mathbf{u}\mathbf{x}^\top\right] + \frac{2}{p^2}\mathbf{x}\mathbf{x}^\top\right).
\end{aligned}
$$

By direct calculations, we also have,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{u},\mathbf{v}} \left[\langle \mathbf{u}, \mathbf{x}\rangle^2 \mathbf{v}\mathbf{v}^\top\right] &= \mathbb{E}_{\mathbf{u},\mathbf{v}} \left[\langle \mathbf{u}, \mathbf{x}\rangle^2\right] \mathbb{E}_{\mathbf{u},\mathbf{v}}[\mathbf{v}\mathbf{v}^\top] = \frac{\|\mathbf{x}\|_2^2}{p^2}, \\
\mathbb{E}_{\mathbf{u},\mathbf{v}} \left[\langle \mathbf{v}, \mathbf{x}\rangle \langle \mathbf{u}, \mathbf{x}\rangle \mathbf{u}\mathbf{v}^\top\right] &= \frac{\mathbf{x}\mathbf{x}^\top}{p^2}.
\end{aligned}
$$

Now, plugging in the terms into $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ and multiplying the result by $\frac{p}{2}$ as there are $\frac{p}{2}$ rotations completes the proof. ■

3.8.6 Approximation Error for Dependent Feature Augmentation

In this section, we demonstrate how to bound the approximation error for the augmentation of dependent features using rotation in a random plane and cutout as two examples. At a high level, we partition the augmented covariance operator into diagonal and nondiagonal parts $\mathcal{D}$ and $\mathbf{Q}$ (i.e.,

$\mathrm{Cov}_G(\mathbf{X}) = \mathcal{D} + \mathbf{Q}$) and bound them separately:

$$\begin{aligned}\Delta_G &= \|\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})^{-1/2}(\mathcal{D}+\mathbf{Q})\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})^{-1/2} - \mathbf{I}_p\| \\ &= \|\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})^{-1/2}(\mathcal{D}+\mathbf{Q} - \mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})^{-1/2}\| \\ &\leq \frac{\|\mathcal{D} - \mathbb{E}\mathcal{D}\| + \|\mathbf{Q} - \mathbb{E}\mathbf{Q}\|}{\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))}, \quad \because\ \mathbb{E}\mathcal{D} + \mathbb{E}\mathbf{Q} = \mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}).\end{aligned}$$

Approximation error of random rotations

In this section, we will walk through the steps to bound the approximation error for the random rotation estimator. Specifically, we will prove that

$$\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \frac{4(1-\cos\alpha)}{np}\left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} - \mathbf{X}^\top\mathbf{X}\right), \quad \Delta_G \lesssim \frac{\lambda_1 n + \sum_j \lambda_j}{n\sum_{j>1}\lambda_j}.$$

We follow the bound in E.q. (3.17) from the main text:

$$\Delta_G \lesssim \frac{\|\mathcal{D} - \mathbb{E}\mathcal{D}\| + \|\mathbf{Q} - \mathbb{E}\mathbf{Q}\|}{\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))},$$

where we decompose $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ into diagonal and off-diagonal parts as $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \mathcal{D} + \mathbf{Q}$, $\mathcal{D} = a\left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} + \mathrm{Diag}(\mathbf{X}^\top\mathbf{X})\right)$, $\mathbf{Q} = a\left(\mathbf{X}^\top\mathbf{X} - \mathrm{Diag}(\mathbf{X}^\top\mathbf{X})\right)$, and $a = \frac{4(1-\cos\alpha)}{np} = \Theta(\frac{1}{np})$. Using similar arguments in the proof of Proposition 5 for the independent feature augmentation, the error of the diagonal part can be expressed as a sum of n independent subexponential variables divided by $\Theta(np)$. Then, by the concentration bound in Lemma 23 we have,

$$\|\mathcal{D} - \mathbb{E}\mathcal{D}\| \lesssim \frac{1}{p}\sqrt{\frac{\log n}{n}},$$

with probability $1 - n^{-1}$.

On the other hand, by invoking Lemma 29, we also have,

$$\|\mathbf{Q} - \mathbb{E}\mathbf{Q}\| = \|\mathbf{Q}\| \lesssim \frac{\lambda_1 n + \sum_j \lambda_j}{np},$$

with probability at least $1 - \frac{1}{n}$, using the fact that $\mathbb{E}\mathbf{Q} = 0$. Finally,

$$\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})) = 4(1-\cos\alpha)\frac{\mathrm{Tr}(\boldsymbol{\Sigma}) - \boldsymbol{\Sigma}}{p} \geq 4(1-\cos\alpha)\frac{\sum_{j>1}\lambda_j}{p},$$

so

$$\Delta_G \lesssim \frac{\lambda_1 n + \sum_j \lambda_j}{n\sum_{j>1}\lambda_j},$$

with probability $1 - 2n^{-1}$. Note that Δ_G is $o(1)$ for $\sum_{j>1}\lambda_j \gg \lambda_1$.

Approximation error of random cutout

In this section, we turn our attention to the bound of the approximation error for random cutout, where k consecutive features are cut out randomly by the augmentation. As the features are dropout dependently, the random cutout belongs to the class of dependent feature augmentation. For simplicity, we consider the unbiased random cutout, where the augmented estimator is rescaled by the factor $\frac{p}{p-k}$ (so $\mu_{\mathcal{G}}(\mathbf{x}) = \mathbf{x}$). The calculations in Section 3.8.5 show that

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \frac{k}{p-k}\,\mathrm{diag}(\boldsymbol{\Sigma}),\;\; \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \frac{p}{p-k}\mathbf{M} \odot \frac{\mathbf{X}^\top\mathbf{X}}{n}, \tag{3.141}$$

where $\mathbf{M}$ is a circulant matrix in which $\mathbf{M}_{i,j} = \frac{k}{p} - \frac{|j-i|\mathbf{1}_{|j-i|<k-1} + k\mathbf{1}_{|j-i|\geq k-1}}{p-k}$ and $\odot$ denotes the element-wise matrix product (Hadamard product). Because $\boldsymbol{\Sigma}$ is diagonal we have,

$$\Delta_G = \frac{p}{k}\|\mathbf{M} \odot (n^{-1}\mathbf{Z}^\top\mathbf{Z} - \mathbf{I}_p)\|,$$

where $\mathbf{Z}$ is a n by p matrix with i.i.d. subgaussian rows that has identity covariance $\mathbf{I}$. Then

$$\Delta_G = \frac{p}{k}\cdot\left(\left\|\widetilde{\mathbf{M}} \odot \mathcal{D} + \frac{k^2}{p(p-k)}n^{-1}\mathbf{Z}^\top\mathbf{Z}\right\|\right) \leq \frac{p}{k}\cdot\left(\underbrace{\left\|\widetilde{\mathbf{M}} \odot \mathcal{D}\right\|}_{L_1} + \underbrace{\left\|\frac{k^2}{p(p-k)}n^{-1}\mathbf{Z}^\top\mathbf{Z}\right\|}_{L_2}\right), \tag{3.142}$$

where $\mathcal{D}$ is an almost diagonal circular matrix with $\mathcal{D}_{ij} = \sum_{l=1}^{n} \frac{\mathbf{Z}_{li}\mathbf{Z}_{lj}}{n} - \delta_{ij}$ if $|i-j| \le k$ and 0 otherwise, while $\widetilde{\mathbf{M}}_{i,j} = \mathbf{M}_{i,j} + \frac{k^2}{p(p-k)}$. Our decomposition strategy here is consistent with our idea in the previous subsection, where we partition the matrix of interest into strong diagonal components and weak off-diagonal components. However, in the random cutout case, approximately $O(k)$ near the diagonal components have a strong covariance with intensity of the order $O(\frac{k}{p})$ while the rest of the order $O(\frac{k^2}{p^2})$; hence, we gather all elements with strong covariance into the "diagonal" part. Now we will bound L_2 and L_1 in a sequence.

Like in the previous section, L_2 can be bounded by invoking the lemma 29 which gives

$$L_2 \lesssim \frac{k^2}{p(p-k)} \frac{n+p}{n},$$

with probability $1 - \frac{c}{n}$ for some constant $c > 0$. For the bounds of L_1, we first bound the elements of $\mathcal{D}$. For $i \neq j$, since $\mathbf{Z}_{ki}^2$ is sub-exponential we have

$$\mathcal{D}_{i,j} \le \sum_{k=1}^{n} \frac{\mathbf{Z}_{ki}\mathbf{Z}_{kj}}{n} \le n^{-1} \sqrt{\sum_{k=1}^{n} \mathbf{Z}_{ki}^2} \sqrt{\sum_{k=1}^{n} \mathbf{Z}_{kj}^2} \le \varepsilon,$$

with probability $\exp(-nC\varepsilon^2)$ for some constant C by Lemma 23, where we have used Cauchy-Schwartz inequality and ε will be determined below. The case where $i = j$ is similar. As there are $O(pk)$ nonzero terms in $\mathcal{D}$, we choose $\varepsilon = \sqrt{\frac{5\log pk}{n}}$. Then, by union bounds over pk terms, we obtain

$$\mathcal{D}_{i,j} \le \sqrt{\frac{5\log pk}{n}}, \ \ \forall i, j$$

with probability at least $1 - \frac{1}{p^3k^3}$. Next, denote $\mathbf{A} := \widetilde{\mathbf{M}} \odot \mathcal{D}$. Note that $|\mathbf{A}_{ij}| \lesssim \frac{k}{p}\varepsilon$ for all $|i-j| \le k$

and 0 otherwise. We will bound the operator norm of $\mathbf{A}$. Consider any $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$,

$$\begin{aligned}\|\mathbf{A}\mathbf{v}\|_2 &= \sqrt{\sum_{i=1}^{k}(\sum_{j=1}^{k}\mathbf{A}_{ij}\mathbf{v}_j)^2} = \sqrt{\sum_{i=1}^{k}(\sum_{j\in i-k:i+k}\mathbf{A}_{i,j}\mathbf{v}_j)^2} \\ &\leq \sqrt{\sum_{i=1}^{k}(\sum_{j\in i-k:i+k}\mathbf{A}_{i,j}^2)(\sum_{j\in i-k:i+k}\mathbf{v}_j^2)} \leq \frac{k}{p}\sqrt{2k\varepsilon^2\sum_{i=1}^{k}\sum_{j\in i-k:i+k}\mathbf{v}_j^2} \\ &= O\left(\frac{k^2}{p}\varepsilon\right),\end{aligned}$$

where we have used the sparsity property that $\mathbf{A}_{ij} = 0$ if $|j-i| > k$. Therefore, $L_1 = \|\mathbf{A}\| \lesssim O(\frac{k^2}{p}\varepsilon) = O\left(\frac{k^2}{p}\sqrt{\frac{5\log pk}{n}}\right)$. Now combining the bounds on L_1 and L_2 and (3.142) we arrive at the result:

$$\Delta_G \lesssim k\sqrt{\frac{\log pk}{n}},$$

with probability at least $1 - \frac{c}{n} - \frac{1}{p^3k^3}$.

Remark 47. *This approximation bound, together with Corollary 35, show that the approximation error is dominated by the bias-variance (survival-contamination) if* ***1.*** *over-parameterized regime ($p \gg n$): p is upper bounded by some polynomial of n and $k \ll \sqrt{\frac{n}{\log p}}$, or* ***2.*** *under-parameterized regime ($p \ll n$): n is upper bounded by some polynomial of p and $k \ll \frac{p}{\sqrt{n}}$.*

CHAPTER 4
MODEL CALIBRATION WITH LOW-RANK OPTIMAL TRANSPORT

In this chapter, we describe our preliminary work on model calibration with optimal transport for domain adaptation. The approach is to find a map between the data in both domains so that we can apply the trained model on the source domain to the data in the target domain through the map. In this way, the method boils down to finding the map between the data domains, and this is where OT comes into play. To solve for the map, we propose a robust distribution alignment technique Latent Optimal Transport (LOT) [33].

4.1 Optimal Transport

Optimal transport (OT) [75, 129, 32] is a distribution alignment technique that learns a transport plan that specifies how to move mass from one distribution to match another. Specifically, consider two sets of data points encoded in matrices, the *source* $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ and the *target* $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_m]$, where $\mathbf{x}_i \in \mathcal{X}$, $\mathbf{y}_j \in \mathcal{Y}$, $\forall i, j$. Assume they are endowed with discrete measures $\mu = \sum_{i=1}^{N} p(\mathbf{x}_i)\delta_{\mathbf{x}_i}$, $\nu = \sum_{j=1}^{M} p(\mathbf{y}_j)\delta_{\mathbf{y}_j}$, respectively. The cost of transporting $\mathbf{x}_i$ to $\mathbf{y}_j$ is $c(\mathbf{x}_i, \mathbf{y}_j)$, where c denotes some cost function. OT considers the most cost-efficient transport by solving the following problem:[1]

$$\mathrm{OT}_{\mathbf{C}}(\mu, \nu) := \min_{\mathbf{P}\mathbf{1}=\mu, \mathbf{P}^T\mathbf{1}=\nu} \langle \mathbf{C}, \mathbf{P} \rangle, \tag{4.1}$$

where $\mathbf{P} := [p(\mathbf{x}_i, \mathbf{y}_j)]_{i,j}$ is the source-to-target transport plan matrix (coupling), and $\mathbf{C} = [c(\mathbf{x}_i, \mathbf{y}_j)]_{i,j}$ is the cost matrix. When $c(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})^p$, where d is a distance function, $\mathcal{W}_p := \mathrm{OT}_{\mathbf{C}}^{1/p}$ defines a distance called the p-Wasserstein distance. The objective in (4.1) is a linear programming problem, where computation speed can be prohibitive if n is large [130]. A common

[1] The problem can be generalized to setting of continuous measures by $\mathrm{OT}_c(\mu, \nu) = \min_{\gamma \in \mathcal{G}} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y)$, $\mathcal{G} = \{\gamma : \int_{\mathcal{Y}} d\gamma(x, y) = \mu, \int_{\mathcal{X}} d\gamma(x, y) = \nu\}$.

speedup is to replace the objective by an entropy-regularized proxy,

$$\mathrm{OT}_{\mathbf{C},\varepsilon}(\mu,\nu) := \min_{\mathbf{P}\mathbf{1}=\mu,\mathbf{P}^T\mathbf{1}=\nu} \langle \mathbf{C},\mathbf{P}\rangle - \varepsilon\mathbf{H}(\mathbf{P}) = \min_{\mathbf{P}\mathbf{1}=\mu,\mathbf{P}^T\mathbf{1}=\nu} \varepsilon\mathrm{KL}(\mathbf{P}||\mathbf{K}), \tag{4.2}$$

where $\mathbf{K}$ is the Gibbs kernel induced by the element-wise exponential of the cost matrix $\mathbf{K} := \exp(-\mathbf{C}/\varepsilon)$, $\mathbf{H}(\mathbf{P}) := -\sum_{ij}\mathbf{P}_{ij}\log(\mathbf{P}_{i,j})$ is the Shannon entropy, and ε is a user-specified hyperparameter that controls the amount of entropic regularization that is introduced. We can alternatively write the objective function as a minimization of $\varepsilon\mathrm{KL}(\mathbf{P}||\mathbf{K})$, where KL denotes the Kullback-Leibler divergence. In practice, the entropy-regularized form is often used over the original objective (4.1) as it admits a fast method called the Sinkhorn algorithm [76, 131]. Hence, we will use OT to refer to the entropy-regularized form unless specified otherwise in the context.

Optimal Transport via Factored Couplings: Factored Coupling (FC) is proposed in [88] to reduce the sample complexity of OT in high dimensions. Specifically, it adds an additional constraint to (4.1) by enforcing the transport plan to be of the following factored form,

$$p(\mathbf{x}_i,\mathbf{y}_j) = \sum_{l=1}^{k} p(\mathbf{z}_l)p(\mathbf{x}_i|\mathbf{z}_l)p(\mathbf{y}_j|\mathbf{z}_l). \tag{4.3}$$

This has a nice interpretation: $\mathbf{z}_l$ serves as a common "anchor" that transportation from $\mathbf{x}_i$ to $\mathbf{y}_j$ must pass through. It turns out that FC is closely related to the Wasserstein barycenter problem [132, 133, 134], $\min_\nu \sum_{i=1}^{N} \mathcal{W}_2^2(\mu_i,\nu)$, where ν is the Procrustes mean to distributions μ_i, $i = 1,\dots,N$ with respect to the squared 2-Wasserstein distance. A crucial insight from [88] is that for $N = 2$, the barycenter ν could approximate the optimal anchors to a transport plan of the form (4.3) that minimizes the objective in (4.1).

4.2 Latent Optimal Transport

4.2.1 Problem formulation

Now we introduce our robust transport method, Latent Optimal Transport (LOT). Consider data matrices $\mathbf{X}$ and $\mathbf{Y}$ and their measures μ, ν. We introduce "anchors" through which points must flow, thus constraining the transportation. The anchors are stacked in data matrices $\mathbf{Z}_x := [\mathbf{z}_1^x, \ldots, \mathbf{z}_{k_x}^x]$, $\mathbf{Z}_y := [\mathbf{z}_1^y, \ldots, \mathbf{z}_{k_y}^y]$. We denote the measures of the source and target anchors as $\mu_z = \sum_{m=1}^{k_x} p(\mathbf{z}_m^x)\delta_{\mathbf{z}_m^x}$ and $\nu_z = \sum_{n=1}^{k_y} p(\mathbf{z}_n^y)\delta_{\mathbf{z}_n^y}$. For any set $\mathcal{A}$, we further denote $\Delta_{\mathcal{A}}^k := \left\{\sum_{i=1}^k \omega_i \delta_{\mathbf{a}_i} : \sum_{i=1}^k \omega_i = 1, \omega_i \geq 0, \mathbf{a}_i \in \mathcal{A}, \forall i\right\}$ as the set of probability measures on $\mathcal{A}$ that has discrete support of size up to k. Hence $\mu_z \in \Delta_{\mathcal{Z}_x}^{k_x}$, $\nu_z \in \Delta_{\mathcal{Z}_y}^{k_y}$, where $\mathcal{Z}_x$ (resp. $\mathcal{Z}_y$) is the space of source (resp. target) anchors. If we interpret the conditional probability $p(a|b)$ as the strength of transportation from b to a, then, using the chain rule, the concurrence probability $p(\mathbf{x}_i, \mathbf{y}_j)$ of $\mathbf{x}_i$ and $\mathbf{y}_j$ can be written as,

$$p(\mathbf{x}_i, \mathbf{y}_j) = \sum_{m,n} p(\mathbf{x}_i)p(\mathbf{z}_m^x|\mathbf{x}_i)p(\mathbf{z}_n^y|\mathbf{z}_m^x)p(\mathbf{y}_j|\mathbf{z}_n^y) = \sum_{m,n} p(\mathbf{x}_i, \mathbf{z}_m^x)\frac{p(\mathbf{z}_m^x, \mathbf{z}_n^y)}{p(\mathbf{z}_m^x)p(\mathbf{z}_n^y)}p(\mathbf{z}_n^y, \mathbf{y}_j). \quad (4.4)$$

When encoding these probabilities using a transport matrix $\mathbf{P} := [p(\mathbf{x}_i, \mathbf{y}_j)]_{i,j}$, the factorized form (4.4) can be written as,

$$\mathbf{P} = \mathbf{P}_x \mathrm{diag}(\mathbf{u}_z^{-1})\mathbf{P}_z \mathrm{diag}(\mathbf{v}_z^{-1})\mathbf{P}_y, \quad (4.5)$$

where $\mathbf{P}_x$ encodes transport from source space to source anchor space (i.e., $p(\mathbf{x}_i, \mathbf{z}_m^x)$), $\mathbf{P}_z$ encodes transport from source anchor space to target anchor space, $\mathbf{P}_y$ encodes transport from target anchor space to target space , and $\mathbf{u}_z := [p(\mathbf{z}_1^x), \cdots, p(\mathbf{z}_{k_x}^x)]$, $\mathbf{v}_z := [p(\mathbf{z}_1^y), \cdots, p(\mathbf{z}_{k_y}^y)]$ encode the latent distributions of anchors. To learn each of these transport plans, we must first designate the ground metric used to define the cost in each of the three stages. The cost matrices $\mathbf{C}_x, \mathbf{C}_y$ determine how points will be transported to their respective anchors and thus dictate how the data structure will be extracted. We will elaborate on the choice of costs in Section 4.2.2.

We now formalize our proposed approach to transport in the following definition.

Definition 48. *Let $\mathbf{C}_x$, $\mathbf{C}_y$ denote the cost matrices between the source/target and their representative anchors, and let $\mathbf{C}_z$ denote the cost matrix between anchors. We define the latent optimal transport (`LOT`) problem as,*

$$\mathrm{OT}^{\mathrm{L}}(\mu,\nu) := \inf_{\mu_z \in \Delta^{k_x}_{\mathcal{Z}_x}, \nu_z \in \Delta^{k_y}_{\mathcal{Z}_y}} \big\{ \mathrm{OT}_{\mathbf{C}_\mathrm{x}}(\mu,\mu_\mathrm{z}) + \mathrm{OT}_{\mathbf{C}_\mathrm{z}}(\mu_\mathrm{z},\nu_\mathrm{z}) + \mathrm{OT}_{\mathbf{C}_\mathrm{y}}(\nu_\mathrm{z},\nu) \big\},$$

where $\mathcal{Z}_x$ and $\mathcal{Z}_y$ are the latent spaces of the source and target anchors, respectively.[2]

The intuition behind Def. 48 is that we use $\mathrm{OT}_{\mathbf{C}_\mathrm{x}}(\mu,\mu_\mathrm{z})$ and $\mathrm{OT}_{\mathbf{C}_\mathrm{y}}(\nu_\mathrm{z},\nu)$ to capture group structure in each space, and then $\mathrm{OT}_{\mathbf{C}_\mathrm{z}}(\mu_\mathrm{z},\nu_\mathrm{z})$ to align the source and target by determining the transportation across anchors. Hence, `LOT` can be interpreted as an optimization of joint clustering and alignment. The flexibility of cost matrices allows `LOT` to capture different structures and induce different transport plans. In Section 4.3, we further show that `LOT` can be regarded as a relaxation of an OT problem.

Next, we show some properties of `LOT` that highlight its similarity to a metric.

Proposition 49. *Suppose the latent spaces $\mathcal{Z}_x = \mathcal{Z}_y$ are the same as the original data spaces $\mathcal{X} = \mathcal{Y}$, and the cost matrices are defined by $\mathbf{C}_x[a,b] = \mathbf{C}_z[a,b] = \mathbf{C}_y[a,b] = d(a,b)^p$, where $p \geq 1$ and d is some distance function. If we define the latent Wasserstein discrepancy as $\mathcal{W}_p^L := (\mathrm{OT}^{\mathrm{L}})^{1/p}$, then there exist $\kappa > 0$ such that, for any μ, ν and ζ having latent distributions of support sizes up to k, the discrepancy satisfies,*

- $\mathcal{W}_p^L(\mu,\nu) \geq 0\,,\ \mathcal{W}_p^L(\mu,\nu) = \mathcal{W}_p^L(\nu,\mu),\ \mathcal{W}_p^L(\mu,\nu) \leq \kappa\left(\mathcal{W}_p^L(\mu,\zeta) + \mathcal{W}_p^L(\zeta,\nu)\right).$

The low-rank nature of `LOT` has a biasing effect that results in $\mathcal{W}_p^L(\mu,\mu) > 0$ for a general μ. We can debias it by defining its variant $\tilde{\mathcal{W}}_p^L(\mu,\nu) := \left(\left(\mathcal{W}_p^L(\mu,\nu)\right)^p - \min_{z_k \in \Phi_x} \mathcal{W}_p^p(\mu, z_k) - \min_{z'_k \in \Phi_y} \mathcal{W}_p^p(\nu, z'_k) \right)^{1/p}$, where $\Phi_x = \Delta^{k_x}_{\mathcal{Z}_x}$, $\Phi_y = \Delta^{k_y}_{\mathcal{Z}_y}$. The following property connects $\tilde{\mathcal{W}}_p^L(\mu,\nu)$ to k-means clustering.

[2] This definition extends naturally to continuous measures by replacing cost matrix $\mathbf{C}$ with cost function c.

Corollary 50. *Under the assumptions of Proposition 49, if $p = 2$ and $k_x = k_y = k$, then $\forall \mu, \nu$, we have $\tilde{\mathcal{W}}_2^L(\mu, \nu) \geq 0$. Furthermore, $\tilde{\mathcal{W}}_2^L(\mu, \nu) > 0$ if their k-means centroids or sizes of their k-means clusters differ.*

4.2.2 Establishing a ground metric

In what follows, we will focus on the Euclidean space $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. Instead of considering every source-to-target distance to build our transportation cost, we can use anchors as proxies for each point. A well-established way of encoding the distance that each point needs to travel to get to its nearest anchor, is to define the cost as: $\mathbf{C}_x = d_{\mathbf{M}_x}, \mathbf{C}_z = d_{\mathbf{M}_z}, \mathbf{C}_y = d_{\mathbf{M}_y}$, where $d_{\mathbf{M}}$ denotes the Mahalanobis distance: $d^2_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) := (\mathbf{x} - \mathbf{y})^T \mathbf{M}(\mathbf{x} - \mathbf{y})$ and $\mathbf{M}$ is some positive semidefinite matrix. The Mahalanobis distance generalizes the squared Euclidean distance and allows us to consider different costs based on correlations between features. The framework of Mahalanobis distance benefits from efficient metric learning techniques [135]; recent research also establishes connections between it and robust OT [136, 137]. When a simple L2-distance is used ($\mathbf{M} = \mathbf{I}$), we will denote this specific variant as LOT-L2.

When LOT moves source points through anchors, the anchors impose a type of bottleneck, and this results in a loss of information that makes it difficult to estimate the corresponding point in the target space. In cases where accurate point-to-point alignment is desired, we propose an alternative strategy for defining the cost matrix $\mathbf{C}_z$. The idea is to represent an anchor as the distribution of points assigned to it. Specifically, we represent $\mathbf{z}^x, \mathbf{z}^y$ as measures in $\mathbb{R}^d$: $\tilde{\mathbf{z}}^x = \sum_{i=1}^N \mathbf{P}_x(\mathbf{x}_i | \mathbf{z}^x) \delta_{\mathbf{x}_i}$, $\tilde{\mathbf{z}}^y = \sum_{j=1}^M \mathbf{P}_y(\mathbf{y}_j | \mathbf{z}^y) \delta_{\mathbf{y}_j}$. Then we measure the cost between anchors as the squared Wasserstein distance between their respective distributions,

$$\mathbf{C}_z := [\mathcal{W}_2^2(\mathbf{P}_x(\cdot | \mathbf{z}^x_m), \mathbf{P}_y(\cdot | \mathbf{z}^y_n))]_{m,n}. \tag{4.6}$$

Besides the quantity itself, the transport plan returned by calculating $\mathbf{C}_z$ is also very important as it provides accurate point-to-point maps. Since the cost matrix is now a function of $\mathbf{P}_x$ and $\mathbf{P}_y$,

Algorithm 3: Latent Optimal Transport - `LOT`

```
Input Data matrices $\mathbf{X}$, $\mathbf{Y}$; metric costs $\mathbf{M}_x, \mathbf{M}_y, \mathbf{M}_z$; entropy regularization parameters
  $\varepsilon_x$, $\varepsilon_y$, $\varepsilon_z$; initial anchors $\mathbf{Z}_x, \mathbf{Z}_y$.
while not converging do
    $(\mathrm{vec}(\mathbf{Z}_x), \mathrm{vec}(\mathbf{Z}_y)) \leftarrow$ Eqn. (4.8) $\mathbf{K}_x = \{\exp(-\|\mathbf{X}[i] - \mathbf{Z}_x[j]\|^2_{\mathbf{M}_x}/\varepsilon_x)\}_{i,j}$
    $\mathbf{K}_y = \{\exp(-\|\mathbf{Y}[i] - \mathbf{Z}_y[j]\|^2_{\mathbf{M}_y}/\varepsilon_y)\}_{j,q}$
    $\mathbf{K}_z = \{\exp(-\|\mathbf{Z}_x[i] - \mathbf{Z}_y[j]\|^2_{\mathbf{M}_z}/\varepsilon_z)\}_{p,q}$
    $\mathbf{P}_x, \mathbf{P}_y, \mathbf{P}_z \leftarrow$ UpdatePlan $(\mathbf{K}_x, \mathbf{K}_y, \mathbf{K}_z)$
Return $\mathbf{P}_x$, $\mathbf{P}_y$, $\mathbf{P}_z$, $\mathbf{Z}_x$, $\mathbf{Z}_y$
UpdatePlan $(\mathbf{K}_x, \mathbf{K}_y, \mathbf{K}_z)$
Initialize $\alpha_x \leftarrow \mathbf{1}_N; \beta_x \leftarrow \mathbf{1}_{k_1}; \alpha_y \leftarrow \mathbf{1}_{k_2};$     $\beta_y \leftarrow \mathbf{1}_M; \alpha_z \leftarrow \mathbf{1}_{k_1}; \beta_z \leftarrow \mathbf{1}_{k_2}$
while not converging do
    $\alpha_x \leftarrow \mu \oslash \mathbf{K}_x \beta_x$; $\beta_y \leftarrow \nu \oslash \mathbf{K}_y^T \alpha_y$
    $\mathbf{u}_z \leftarrow ((\alpha_z \odot \mathbf{K}_z \beta_z) \odot (\beta_x \odot \mathbf{K}_x^T \alpha_x))^{\frac{1}{2}}$
    $\beta_x \leftarrow \mathbf{u}_z \oslash \mathbf{K}_x^T \alpha_x$; $\alpha_z \leftarrow \mathbf{u}_z \oslash \mathbf{K}_z \beta_z$
    $\mathbf{v}_z \leftarrow ((\alpha_y \odot \mathbf{K}_y \beta_y) \odot (\beta_z \odot \mathbf{K}_z^T \alpha_z))^{\frac{1}{2}}$
    $\beta_z \leftarrow \mathbf{v}_z \oslash \mathbf{K}_z^T \alpha_z$; $\alpha_y \leftarrow \mathbf{v}_z \oslash \mathbf{K}_y \beta_y$
Return $\mathbf{P}_i = \mathrm{diag}(\alpha_i)\mathbf{K}_i\mathrm{diag}(\beta_i)$, $i \in \{x, y, z\}$
```

we use an additional alternating scheme to solve the problem: we alternate between updating $\mathbf{C}_z$ while keeping $\mathbf{P}_x$ and $\mathbf{P}_y$ fixed, and then updating $\mathbf{P}_x, \mathbf{P}_y, \mathbf{P}_z$ while keeping $\mathbf{C}_z$ fixed. An efficient algorithm is presented in [33] to reduce the computation complexity. This variant, `LOT-WA` , can yield better performance in downstream tasks that require precise alignment at the cost of additional computation.

4.2.3 Algorithm

In the rest of this section, we will develop our main approach for solving the problem in Def. 48. We provide an outline of the algorithm in Algorithm 2 and an implementation of the algorithm in Python at: `http://nerdslab.github.io/latentOT`.

(1) Optimizing $\mathbf{P}_x, \mathbf{P}_y$ ***and*** $\mathbf{P}_z$***:*** To begin, we assume that the anchors and cost matrices $\mathbf{C}_x, \mathbf{C}_z, \mathbf{C}_y$ are already specified. Let $\mathbf{K}_x, \mathbf{K}_z, \mathbf{K}_y$ be the Gibbs kernels induced from the cost matrices

$\mathbf{C}_x, \mathbf{C}_z, \mathbf{C}_y$ as in (4.2). The optimization problem can be written as,

$$\min_{\mathbf{u}_z, \mathbf{v}_z, \mathbf{P}_x, \mathbf{P}_z, \mathbf{P}_y} \sum_{i \in \{x,y,z\}} \varepsilon_i \mathrm{KL}(\mathbf{P}_i \| \mathbf{K}_i),$$
$$\text{subject to: } \mathbf{P}_x \mathbf{1} = \mu, \mathbf{P}_x^T \mathbf{1} = \mathbf{u}_z, \mathbf{P}_z \mathbf{1} = \mathbf{u}_z, \mathbf{P}_z^T \mathbf{1} = \mathbf{v}_z, \mathbf{P}_y \mathbf{1} = \mathbf{v}_z, \mathbf{P}_y^T \mathbf{1} = \nu. \tag{4.7}$$

This is a Bregman projection problem with affine constraints. An iterative projection procedure can thus be applied to solve the problem [138]. We present the procedure as UPDATEPLAN in Algorithm 2, where $\mathbf{P}_x, \mathbf{P}_z, \mathbf{P}_y$ are successively projected onto the constrained sets of fixed marginal distributions.

(2) Optimizing the anchor locations: Now we consider the case where we are free to select the anchor locations in $\mathbb{R}^d$. We consider the class of Mahalanobis costs described in Section 4.2.2. Let $\mathbf{M}_x$, $\mathbf{M}_z$, $\mathbf{M}_y$ be the Mahalanobis matrices correspond to $\mathbf{C}_x$, $\mathbf{C}_z$, and $\mathbf{C}_y$, respectively.

Given the transport plans generated after solving (4.7), we can derive the the first-order stationary condition of OT^L with respect to $\mathbf{Z}_x$ and $\mathbf{Z}_y$. Let

$$\mathbf{A} = \begin{bmatrix} D(\mathbf{u}_z) \otimes (\mathbf{M}_x + \mathbf{M}_z) & \mathbf{P}_z \otimes \mathbf{M}_z \\ -\mathbf{P}_z^T \otimes \mathbf{M}_z & D(\mathbf{v}_z) \otimes (\mathbf{M}_y + \mathbf{M}_z) \end{bmatrix}$$

The update formula is given by

$$\begin{bmatrix} \mathrm{vec}(\mathbf{Z}_x^*) \\ \mathrm{vec}(\mathbf{Z}_y^*) \end{bmatrix} = \mathbf{A}^{-1} \times \begin{bmatrix} (\mathbf{P}_x^T \otimes \mathbf{M}_x)\mathrm{vec}(\mathbf{X}) \\ (\mathbf{P}_y \otimes \mathbf{M}_y)\mathrm{vec}(\mathbf{Y}) \end{bmatrix}, \tag{4.8}$$

where $\mathrm{vec}(\cdot)$ denotes the operator converting a matrix to a column vector, and $D(\cdot)$ denotes the operator converting a vector to a diagonal matrix. Pseudo-code for the combined scheme can be found in Algorithm 2.

(3) Robust estimation of data transport: LOT provides robust transport in the target domain by aligning the data through anchors, which can facilitate regression, and classification in downstream applications. We denote the centroids of the source and target by $\mathbf{Q}_x = \text{diag}(\mathbf{u}_z^{-1})\mathbf{P}_x^T\mathbf{X}^T$, $\mathbf{Q}_y = \text{diag}(\mathbf{v}_z^{-1})\mathbf{P}_y\mathbf{Y}^T$. We propose the estimator $\hat{\mathbf{X}} := \sum_{m,n} p(\mathbf{z}_m^x, \mathbf{z}_n^y|\mathbf{x})(\mathbf{Q}_m^y - \mathbf{Q}_n^x) = \text{diag}(\mu^{-1})\mathbf{P}_x\text{diag}((\mathbf{P}_z\mathbf{1})^{-1})\mathbf{P}_z(\mathbf{Q}_y - \mathbf{Q}_x)$. In contrast to factored coupling [88], where $\mathbf{Z}_x = \mathbf{Z}_y$, LOT is robust even when the source and target have different structures (see Table 4.1 MNIST-DU, Figure 4.2).

4.3 Theoretical Analysis

LOT as a relaxation of OT: We now ask how the optimal value of our original rank-constrained objective in (4.7) is related to the transportation cost defined in entropy-regularized OT. It turns out their objectives are connected by an inequality described below.

Proposition 51. *Let* $\mathbf{P}$ *be a transport plan of the form in (4.5). Assume that* $\mathbf{K}$ *is some Gibbs kernel that satisfies,*

$$\mathbf{K}_x\mathbf{K}_z\mathbf{K}_y \leq \mathbf{K}, \tag{4.9}$$

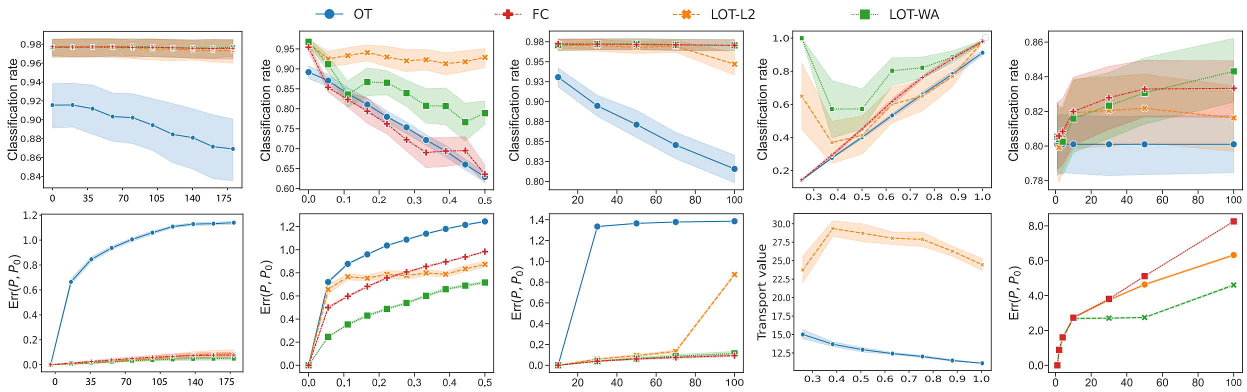
where the inequality is over each entry. Then we have,

$$\varepsilon \mathit{KL}(\mathbf{P}\|\mathbf{K}) \leq \varepsilon(\mathit{KL}(\mathbf{P}_x\|\mathbf{K}_x) + \mathit{KL}(\mathbf{P}_z\|\mathbf{K}_z) + \mathit{KL}(\mathbf{P}_y\|\mathbf{K}_y)) + \varepsilon(\mathbf{H}(\mathbf{u}_z) + \mathbf{H}(\mathbf{v}_z)),$$

where $\mathbf{H}(\mathbf{a}) := -\sum_i \mathbf{a}_i \log \mathbf{a}_i$ *denotes the entropy.*

The proposition shows that an OT objective, corresponding to a kernel $\mathbf{K}$ (resp. $\mathbf{C}$), can be upper bounded by three sub-OT problems defined by subsequent kernels $\mathbf{K}_x, \mathbf{K}_z, \mathbf{K}_y$ (resp. $\mathbf{C}_x, \mathbf{C}_z, \mathbf{C}_y$) that satisfies (4.9) (resp. $\exp(-\mathbf{C}_x/\varepsilon)\exp(-\mathbf{C}_z/\varepsilon)\exp(-\mathbf{C}_y/\varepsilon) \leq \exp(-\mathbf{C}/\varepsilon)$).

Let us compare the upper bound given by Proposition 51 with Def. 48 and ignore the entropy terms; we recognize that it is precisely the entropy-regularized objective of LOT . In other words,



(a) Rotation angle (b) Outlier rate (c) Dimensionality (d) Cluster mismatch (e) Transport rank

Figure 4.1: **Results on Gaussian mixture models.** In (a), we apply a rotation between the source and target, in (b) we add outliers, in (c) we vary the ambient dimension, in (d) the target is set to have 8 components, and we vary the number of components in the source to simulate source-target mismatch, in (e) we fix the rank to 10 and vary the number of factors (anchors) used in the approximation. Throughout, we simulate data according to a GMM and evaluate performance by measuring the classification accuracy (top) and computing the deviation between the transport plans before and after the perturbations with respect to the Fröbenius norm (bottom).

with suitable cost matrices satisfying (4.9), LOT could be interpreted as a relaxation of an OT problem in a decomposed form. We then ask what $\mathbf{C}_x$, $\mathbf{C}_z$, $\mathbf{C}_y$ should be to satisfy (4.9). In cases where cost $\mathbf{C}$ is defined by the L^p-norm to the power p, the following corollary shows that the same form suffices.

Corollary 52. *Let $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_p^p$. Consider an optimal transport problem $\mathrm{OT}_{\mathbf{C},\varepsilon}$ with cost $\mathbf{C}[i, j] = d(\mathbf{x}_i, \mathbf{y}_j)$, where $p \geq 1$. Then for a sufficiently small ε, the latent optimal transport OT^L with cost matrices, $\mathbf{C}_x[i, m] = 3^{p-1}d(\mathbf{x}_i - \mathbf{z}^x_m), \mathbf{C}_z[m, n] = 3^{p-1}d(\mathbf{z}^x_m - \mathbf{z}^y_n), \mathbf{C}_y[n, j] = 3^{p-1}d(\mathbf{z}^y_n - \mathbf{y}_j)$ minimizes an upper bound of the entropy-regularized OT objective in (4.2).*

Corollary (52) provides natural costs for LOT to be posed as a relaxation to a OT problem with L^p norm. More generally, finding the optimal cost functions that obey (4.9) and minimize the gap in the inequality in Proposition 51 is outside the scope of this work but would be an interesting topic for future investigation.

Sampling complexity: Below we analyze LOT from a statistical point of view. Specifically, we bound the sampling rate of OT^{L} in Def. 48 when the true distributions μ and ν are estimated by their empirical distributions.

Proposition 53. *Suppose X and Y have distributions μ and ν supported on a compact region Ω in $\mathbb{R}^d$, the cost functions $c_x(\cdot,\cdot)$ and $c_y(\cdot,\cdot)$ are defined as the squared Euclidean distance, and $\hat{\mu}$, $\hat{\nu}$ are empirical distributions of n and m i.i.d. samples from μ and ν, respectively. If the spaces for latent distributions are equal to $\mathcal{Z}_x = \mathcal{Z}_y = \mathbb{R}^d$, and there are k_x and k_y anchors in the source and target, respectively, then with probability at least $1-\delta$,*

$$\mathrm{Err} \;\leq\; C\sqrt{\frac{k_{max}^3 d \log k_{max} + \log(2/\delta)}{N}}, \tag{4.10}$$

where $\mathrm{Err} = |\mathrm{OT}^L(\mu,\nu) - \mathrm{OT}^L(\hat{\mu},\hat{\nu})|$, $k_{max} = \max\{k_x, k_y\}$, $N = \min\{n, m\}$ and $C \geq 0$ is some constant not depending on N.

As shown in [139], the general sampling rate of a plug-in OT scales with $N^{\frac{1}{d}}$, suffering from the "curse of dimensionality". On the other hand, as evidence from [88], structural optimal transport paves ways to overcome the issue. In particular, LOT achieves $N^{-\frac{1}{2}}$ scaling by regularizing the transport rank.

4.4 Experiments

In this section, we conduct empirical investigations.

E1) Testing robustness to various data perturbations: To better understand how different types of domain shift impact the transport plans generated by our approach, we considered different transformations between the source and target. To create synthetic data for this task, we generated multiple clusters/components using a k-dimensional Gaussian with random mean and covariance sampled from a Wishart distribution, randomly projected to a 5-dimensional subspace. The source and target are generated independently: we randomly sample a fixed number of points according to the true distribution for each cluster. We compared the performance of the LOT variants proposed in Section 4.2.2: LOT-L2 (orange curves) and LOT-WA (green curves) with baselines OT (blue curves) and rank regularized factored coupling (FC) [140] (red curves) in terms of their (i) classification rates and (ii) deviation from the original transport plan without perturbations, which we compute as

$\text{Err}(\mathbf{P} - \mathbf{P}_0) = \|\mathbf{P} - \mathbf{P}_0\|_F / \|\mathbf{P}_0\|_F$, where $\mathbf{P}_0$ is the transport plan obtained before perturbations. The results are averaged over 20 runs, and a 75% confidence interval is used. See Appendix **??** for further details.

When compared with OT, both our method and FC provide more stable class recovery, even with significant amounts of perturbations (Figure 4.1). When we examine the error term in the transport plan, we observe that, in most cases, the OT plan deviates rapidly, even for small amounts of perturbations. Both FC and LOT appear to have similar performances across rotations while OT's performance decreases quickly. In experiment (b), we found that both LOT variants provide substantial improvements on classification subject to outliers, implying the applicability of LOT for noisy data. In experiment (c), we study LOT in the high-dimensional setting; we find that LOT-WA behaves similarly to FC with some degradation in performance after the dimension increases beyond 70. Next, in experiment (d), we fix the number of components in the target to be 10, while varying the number in the source from 4 to 10. In contrast to the outlier experiment in (b), LOT-WA shows more resilience to mismatches between the source and target. At the bottom of plot (d), we show the 2-Wasserstein distance (blue) and latent Wasserstein discrepancy (orange) defined in Proposition 49. This shows that the latent Wasserstein discrepancy does indeed provide an upper bound on the 2-Wasserstein distance. Finally, we look at the effect of transport rank on LOT and FC in (e). The plot shows that the slope for LOT is flatter than FC while maintaining similar performances.

E2) Domain adaptation application: In our next experiment, we used LOT to correct for domain shift in a neural network that is trained on one dataset but underperforms on a new but similar dataset (Table 4.1, Figure 4.2). MNIST and USPS are two handwritten digits datasets that are semantically similar but that have different pixel-level distributions and thus introduce domain shift (Figure 4.2a). We train a multi-layer perceptron (MLP) on the training set of the MNIST dataset, freeze the network, and use it for the remaining experiments. The classifier achieves 100% training accuracy and a 98% validation accuracy on MNIST but only achieves 79.3% accuracy on the USPS validation set. We project MNIST's training samples in the classifier's output space (logits) and consider the 10D projection to be the target distribution. Similarly, we project images

Table 4.1: **Results for concept drift and domain adaptation for handwritten digits.** The classification accuracy and L2-error are computed after transport for MNIST to USPS (left) and coarse dropout (right). Our method is compared with the accuracy before alignment (Original), entropy-regularized OT, k-means plus OT (KOT), and subspace alignment (SA).

	MNIST-USPS	MNIST-DU	
	Accuracy	Accuracy	L2 error
Original	79.3	72.6	0.72
OT	76.9	61.5	0.71
KOT	79.4	60.9	0.73
SA	81.3	72.3	-
FC	84.1	67.2	0.59
LOT-WA	**86.2**	**77.7**	**0.56**

from the USPS dataset in the network's output space to get our source distribution. We study the performance of LOT in correcting the classifier's outputs and compare with FC , k-means OT (KOT) [88], and subspace alignment (SA) [141].

In Table 4.1, we summarize the results of our comparisons on the domain adaptation task (MNIST-USPS). Our results suggest that both FC and LOT perform pretty well on this task, with LOT beating FC by 2% in terms of their final classification accuracy. We also show that LOT does better than naive KOT . In Figure 4.2a, we use Isomap to project the distribution of USPS images as well as the alignment results for LOT, FC , and OT. For both LOT and FC , we also display the anchors; note that for LOT , we have two different sets of anchors (source, red; target, blue). This example highlights the alignment of the anchors in our approach and contrasts it with that of FC .

Taking inspiration from studies in self-supervised learning [142, 143] that use different transformations of an input image (e.g., masking parts of the image) to build invariances into a network's representations, here we ask how augmentations of the images introduce domain shift and whether our approach can correct/identify it. To test this, we apply coarse dropout on test samples in MNIST and feed them to the classifier to get a new source distribution. We do this in a balanced (all digits in source and target) and an unbalanced setting (2, 4, 8 removed from source, all digits in target). The

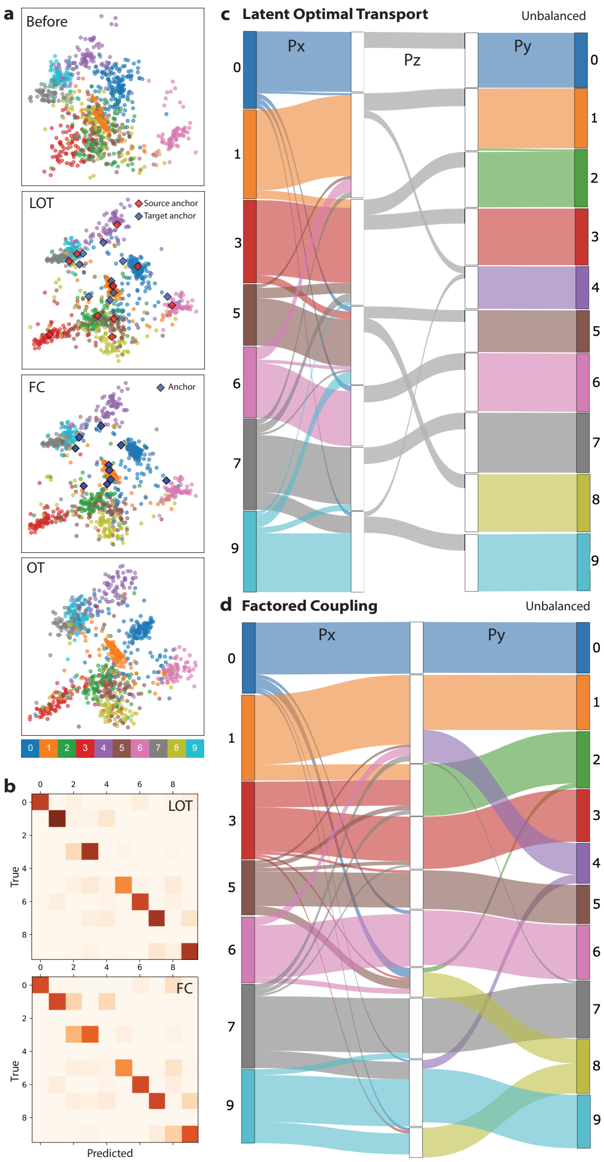


Figure 4.2: **Visualization of results on handwritten digits and examples of domain shift.** (a) 2D projections of representations formed in deep neural network before (top) and after different alignment methods (LOT , FC , OT). (b) confusion matrices for LOT (top) and FC (bottom) after alignment. The transport plans are visualized for LOT (c) and FC (d) in the unbalanced case.

results of the unbalanced dropout are summarized in Table 4.1 (MNIST-DU). In this case, we have the features of the testing samples pre-transformation, and thus, we can compare the transported features to the ground truth features in terms of their point-to-point error (L2 distance). In the unbalanced case, we observe even more significant gaps between FC and LOT, as the source and target datasets have different structures. To quantify these different class-level errors, we compare the confusion matrices for the classifier's output after alignment (Figure 4.2b). By examining the columns corresponding to the removed digits, we see that `FC` is more likely to misclassify these images. Our results suggest that `LOT` has comparable performance with `FC` in a balanced setting and outperforms `FC` in an unbalanced case.

The decomposition in both `LOT` and `FC` allows us to visualize transport between the source, anchors, and the target (Figure 4.2c-d). This visualization highlights the interpretability of the transport plans learned via our approach, with the middle transport plan $\mathbf{P}_z$ providing a concise map of interactions between class manifolds in the unbalanced setting. With `LOT` (Figure 4.2c), we find that each source anchor is mapped to the correct target anchor, with some minor interactions with the target anchors corresponding to the removed digits. In comparison, `FC` (Figure 4.2d) has more spurious interactions between source, anchors, and target.

CHAPTER 5
CONCLUSION

5.1 Summary of Contributions

In this thesis, we establish frameworks and algorithms to understand and leverage data manipulation to improve model generalization. We establish a framework to understand the data augmentation (DA) in the first part of our proposal. Despite DA's simplicity, when and what DA leads to generalization improvement for a machine learning model remains elusive in theory; hence, we study DA theoretically in linear regression, fully characterizing its generalization for different data distributions and augmentation in our preliminary works. Furthermore, we propose to find the optimal augmentation based on our analytical bound. In the second part of the thesis, we develop an data-manipulation algorithm to improve model generalization for the domain adaptation problem. Our method is an extension to the optimal transport (0T) technique . Although OT is effective for domain adaptation by finding connections between data in the two domains, it is often vulnerable to noise and outliers. We thus design a robust low-rank transport to improve upon the ordinary OT in our preliminary works. We also provide substantial theoretical and empirical analysis to validate our algorithm.

Our work paves way formodel generalization improvement through pure data manipulations, and the key contributions can be summarized as follows:

- Establish simple framework to understand DA for linear regression and classifications.
- Show that DA has L2 regularization and change of spectrum effects.
- Provide analytical testbed for novel DA invention.
- Develop Low-rank transportation, more robust to outliers and transformations, which can be used to calibrate models for domain adaptations.

- The proposed OT extension is theoretically grounded with low sampling complexity and is an interpretable relaxation of OT.

5.2 Future Works

For the future work of data manipulation for machine learning, we list two main research directions.

The first direction is to extend our results to the nonlinear model classes. However, we would like to outline two main challenges. The first is the estimator derived from the learning objective does not admit a closed-form solution. Hence, to study its generalization, we will have to analyze the algorithms used to optimize the objective, so a joint consideration of learning dynamic and the generalization analysis of the resultant estimators must be adopted. The second challenge is that the optimization is nonconvex which might admits non-unique solution. Fully characterization of the possible estimators will require additional advanced analytical technique. To bridge the gap between linear model and artificial neural network (ANN), we think the random feature model is a good candidate. The reason is that, intuitively, the model still preserves many aspect in linear model when viewed in the feature space. Further, the connection of ANN and random feature model has also been established through the theory of neural tangent kernel, which is an interesting direction to pursue.

A second interesting direction is to consider the adaptive data manipulation strategy where the augmentation adapts along with the learning process. There is a connection where adaptive gaussian noise injection augmentation can induce sparsity to the learned model parameters. This opens the question whether other adaptive augmentation strategy will lead to different properties of models. The adaptive process will also call for a algorithmic design that might be based on feedback loop or adversarial learning.

Lastly, in this thesis we show a data manipulation complements the model architecture design that facilitates efficient machine learning algorithms. We would like to see how this technique can be applied to more general domains or even the latent space of neural networks.

REFERENCES

[1] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[2] V. Iosifidis and E. Ntoutsi, “Dealing with bias via data augmentation in supervised learning scenarios,” *Jo Bates Paul D. Clough Robert Jäschke*, vol. 24, 2018.

[3] T. Liu, J. Fan, Y. Luo, N. Tang, G. Li, and X. Du, “Adaptive data augmentation for supervised learning over missing data,” *Proceedings of the VLDB Endowment*, vol. 14, no. 7, pp. 1202–1214, 2021.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[5] J.-B. Grill *et al.*, “Bootstrap your own latent: A new approach to self-supervised learning,” *arXiv preprint arXiv:2006.07733*, 2020.

[6] M. Azabou *et al.*, “Mine your own view: Self-supervised learning through across-sample prediction,” *arXiv preprint arXiv:2102.10106*, 2021.

[7] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 12 310–12 320.

[8] S. Y. Feng *et al.*, “A survey of data augmentation approaches for nlp,” *arXiv preprint arXiv:2105.03075*, 2021.

[9] Q. Wen *et al.*, “Time series data augmentation for deep learning: A survey,” *arXiv preprint arXiv:2002.12478*, 2020.

[10] E. Lashgari, D. Liang, and U. Maoz, “Data augmentation for deep-learning-based electroencephalography,” *Journal of Neuroscience Methods*, vol. 346, p. 108 885, 2020.

[11] R. Liu *et al.*, “Drop, swap, and generate: A self-supervised approach for generating neural activity,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 587–10 599, 2021.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning (still) requires rethinking generalization,” *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[13] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. R. Le, “Practical data augmentation with no separate search,” *arXiv preprint arXiv:1909.13719*, 2019.

[14] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, “Learning to compose domain-specific transformations for data augmentation,” *Advances in neural information processing systems*, vol. 30, p. 3239, 2017.

[15] T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Ré, “A kernel theory of modern data augmentation,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 1528–1537.

[16] S. Chen, E. Dobriban, and J. H. Lee, “A group-theoretic framework for data augmentation,” *Journal of Machine Learning Research*, vol. 21, no. 245, pp. 1–71, 2020.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[18] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[20] R. Gontijo-Lopes, S. J. Smullin, E. D. Cubuk, and E. Dyer, “Affinity and diversity: Quantifying mechanisms of data augmentation,” *arXiv preprint arXiv:2002.08973*, 2020.

[21] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, “A simple baseline for semi-supervised semantic segmentation with strong data augmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8229–8238.

[22] M. Assran *et al.*, “Masked siamese networks for label-efficient learning,” *arXiv preprint arXiv:2204.07141*, 2022.

[23] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[24] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[25] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, “Benign overfitting in linear regression,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[26] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *arXiv preprint arXiv:2009.14286*, 2020.

[27] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai, "Classification vs regression in overparameterized regimes: Does the loss function matter?" *Journal of Machine Learning Research*, vol. 22, no. 222, pp. 1–69, 2021.

[28] K. Wang and C. Thrampoulidis, "Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization," *arXiv preprint arXiv:2011.09148*, 2021.

[29] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 67–83, 2020.

[30] N. Courty, R. Flamary, and D. Tuia, "Domain adaptation with regularized optimal transport," in *Machine Learning and Knowledge Discovery in Databases*, Springer, 2014, pp. 274–289.

[31] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1853–1865, 2017.

[32] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[33] C.-H. Lin, M. Azabou, and E. L. Dyer, "Making transport more robust and interpretable by moving data through a small number of anchor points," *Proceedings of machine learning research*, vol. 139, p. 6631, 2021.

[34] J. Gama, I. vecZliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[35] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[36] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *Journal of Big Data*, vol. 4, no. 1, pp. 1–42, 2017.

[37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[38] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

[39] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[40] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 1019–1034, 2014.

[41] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[42] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning*, PMLR, 2013, pp. 222–230.

[43] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[44] H. Lu *et al.*, "When unsupervised domain adaptation meets tensor representations," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 599–608.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5927–5935.

[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[49] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[50] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[51] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, 2019, pp. 4780–4789.

[52] T. Cohen and M. Welling, “Group equivariant convolutional networks,” in *International conference on machine learning*, PMLR, 2016, pp. 2990–2999.

[53] R. Kondor and S. Trivedi, “On the generalization of equivariance and convolution in neural networks to the action of compact groups,” in *International Conference on Machine Learning*, PMLR, 2018, pp. 2747–2755.

[54] X. Wang, K. Wang, and S. Lian, “A survey on face data augmentation for the training of deep neural networks,” *Neural computing and applications*, pp. 1–29, 2020.

[55] B. K. Iwana and S. Uchida, “An empirical survey of data augmentation for time series classification with neural networks,” *Plos one*, vol. 16, no. 7, e0254841, 2021.

[56] M. Bayer, M.-A. Kaufhold, and C. Reuter, “A survey on data augmentation for text classification,” *arXiv preprint arXiv:2107.03158*, 2021.

[57] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[58] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation policies from data,” *arXiv preprint arXiv:1805.09501*, 2018.

[59] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[60] R. Hataya, J. Zdenek, K. Yoshizoe, and H. Nakayama, “Faster autoaugment: Learning augmentation strategies using backpropagation,” in *European Conference on Computer Vision*, Springer, 2020, pp. 1–16.

[61] I. Misra and L. v. d. Maaten, “Self-supervised learning of pretext-invariant representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[62] B. Wallace and B. Hariharan, “Extending and analyzing self-supervised learning across domains,” in *European Conference on Computer Vision*, Springer, 2020, pp. 717–734.

[63] S. Rajput, Z. Feng, Z. Charles, P.-L. Loh, and D. Papailiopoulos, “Does data augmentation lead to positive margin?” In *International Conference on Machine Learning*, PMLR, 2019, pp. 5321–5330.

[64] S. Wu, H. Zhang, G. Valiant, and C. Ré, “On the generalization effects of linear transformations in data augmentation,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 10 410–10 420.

[65] H. Inoue, “Data augmentation by pairing samples for images classification,” *arXiv preprint arXiv:1801.02929*, 2018.

[66] L. Sun, C. Xia, W. Yin, T. Liang, P. S. Yu, and L. He, “Mixup-transformer: Dynamic data augmentation for nlp tasks,” *arXiv preprint arXiv:2010.02394*, 2020.

[67] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, “Understanding and mitigating the tradeoff between robustness and accuracy,” *arXiv preprint arXiv:2002.10716*, 2020.

[68] S. M. Xie, A. Raghunathan, F. Yang, J. C. Duchi, and P. Liang, “When covariate-shifted data augmentation increases test error and how to fix it,” 2019.

[69] H. Javadi, R. Balestriero, and R. Baraniuk, “A hessian based complexity measure for deep networks,” *arXiv preprint arXiv:1905.11639*, 2019.

[70] S. Liu, D. Papailiopoulos, and D. Achlioptas, “Bad global minima exist and sgd can reach them,” *arXiv preprint arXiv:1906.02613*, 2019.

[71] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[72] D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade, “Benign overfitting of constant-stepsize sgd for linear regression,” *arXiv preprint arXiv:2103.12692*, 2021.

[73] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, “Surprises in high-dimensional ridgeless least squares interpolation,” *arXiv preprint arXiv:1903.08560*, 2019.

[74] S. Mei and A. Montanari, “The generalization error of random features regression: Precise asymptotics and the double descent curve,” *Communications on Pure and Applied Mathematics*, 2019.

[75] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[76] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2292–2300.

[77] S. Martin Arjovsky and L. Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[78] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” *arXiv preprint arXiv:1711.01558*, 2017.

[79] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, “From word embeddings to document distances,” in *Proceedings of the Thirty-Forth International Conference on Artificial Intelligence and Statistics*, 2015, pp. 957–966.

[80] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher, “Earth mover’s distances on discrete surfaces,” *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–12, 2014.

[81] J. Solomon *et al.*, “Convolutional wasserstein distances: Efficient optimal transportation on geometric domains,” *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–11, 2015.

[82] N. Bonneel, G. Peyré, and M. Cuturi, “Wasserstein barycentric coordinates: Histogram regression using optimal transport.,” *ACM Trans. Graph.*, vol. 35, no. 4, pp. 71–1, 2016.

[83] A. Gramfort, G. Peyré, and M. Cuturi, “Fast optimal transport averaging of neuroimaging data,” in *Information Processing in Medical Imaging*, Springer, 2015, pp. 261–272.

[84] J. Lee, M. Dabagia, E. Dyer, and C. Rozell, “Hierarchical optimal transport for multimodal distribution alignment,” in *Advances in Neural Information Processing Systems*, 2019, pp. 13 474–13 484.

[85] M. Yurochkin, S. Claici, E. Chien, F. Mirzazadeh, and J. M. Solomon, “Hierarchical optimal transport for document representation,” in *Advances in Neural Information Processing Systems*, 2019, pp. 1601–1611.

[86] H. Xu, D. Luo, R. Henao, S. Shah, and L. Carin, “Learning autoencoders with relational regularization,” 2020.

[87] D. Alvarez-Melis, T. Jaakkola, and S. Jegelka, “Structured optimal transport,” in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1771–1780.

[88] A. Forrow, J.-C. Hütter, M. Nitzan, P. Rigollet, G. Schiebinger, and J. Weed, “Statistical optimal transport via factored couplings,” in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2454–2465.

[89] J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed, “Massively scalable sinkhorn distances via the nyström method,” in *Advances in Neural Information Processing Systems*, 2019, pp. 4427–4437.

[90] M. Scetbon, M. Cuturi, and G. Peyré, “Low-rank sinkhorn factorization,” *arXiv preprint arXiv:2103.04737*, 2021.

[91] C. M. Bishop, “Training with noise is equivalent to tikhonov regularization,” *Neural computation*, vol. 7, no. 1, pp. 108–116, 1995.

[92] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal risk minimization," *Advances in neural information processing systems*, pp. 416–422, 2001.

[93] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, 2020.

[94] Y. Dai, B. Price, H. Zhang, and C. Shen, "Boosting robustness of image matting with context assembling and strong data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 707–11 716.

[95] R. Shen, S. Bubeck, and S. Gunasekar, "Data augmentation as feature manipulation: A story of desert cows and grass cows," *arXiv preprint arXiv:2203.01572*, 2022.

[96] A. Raj, A. Kumar, Y. Mroueh, T. Fletcher, and B. Schölkopf, "Local group invariant representations via orbit embeddings," in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 1225–1235.

[97] Y. Mroueh, S. Voinea, and T. A. Poggio, "Learning with group invariant features: A kernel perspective.," *Advances in neural information processing systems*, vol. 28, 2015.

[98] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[99] F. Yang, Z. Wang, and C. Heinze-Deml, "Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[100] S. Mei, T. Misiakiewicz, and A. Montanari, "Learning with invariances in random features and kernel models," in *Conference on Learning Theory*, PMLR, 2021, pp. 3351–3418.

[101] K. Donhauser, M. Wu, and F. Yang, "How rotational invariance of common kernels prevents generalization in high dimensions," in *International Conference on Machine Learning*, PMLR, 2021, pp. 2804–2814.

[102] A. Sinha, K. Ayush, J. Song, B. Uzkent, H. Jin, and S. Ermon, "Negative data augmentation," *arXiv preprint arXiv:2102.05113*, 2021.

[103] P. Peng, J. Lu, T. Xie, S. Tao, H. Wang, and H. Zhang, "Open-set fault diagnosis via supervised contrastive learning with negative out-of-distribution data augmentation," *IEEE Transactions on Industrial Informatics*, 2022.

[104] Y. Dar, V. Muthukumar, and R. G. Baraniuk, "A farewell to the bias-variance trade-off? an overview of the theory of overparameterized machine learning," *arXiv preprint arXiv:2109.02355*, 2021.

[105] Z. Li *et al.*, “Enhanced convolutional neural tangent kernels,” *arXiv preprint arXiv:1911.00809*, 2019.

[106] B. Hanin and Y. Sun, “How data augmentation affects optimization for linear regression,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8095–8105, 2021.

[107] D. LeJeune, R. Balestriero, H. Javadi, and R. G. Baraniuk, “Implicit rugosity regularization via data augmentation,” *arXiv preprint arXiv:1905.11639*, 2019.

[108] Z. Allen-Zhu and Y. Li, “Towards understanding ensemble, knowledge distillation and self-distillation in deep learning,” *arXiv preprint arXiv:2012.09816*, 2020.

[109] A. D. McRae, S. Karnik, M. Davenport, and V. K. Muthukumar, “Harmless interpolation in regression and classification with structured features,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 5853–5875.

[110] Y. Cao, Q. Gu, and M. Belkin, “Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8407–8418, 2021.

[111] N. S. Chatterji and P. M. Long, “Finite-sample analysis of interpolating linear classifiers in the overparameterized regime.,” *J. Mach. Learn. Res.*, vol. 22, pp. 129–1, 2021.

[112] O. Shamir, “The implicit bias of benign overfitting,” *arXiv preprint arXiv:2201.11489*, 2022.

[113] Z. Deng, A. Kammoun, and C. Thrampoulidis, “A model of double descent for high-dimensional binary linear classification,” *Information and Inference: A Journal of the IMA*, vol. 11, no. 2, pp. 435–495, 2022.

[114] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, “The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime,” *arXiv preprint arXiv:1911.01544*, 2019.

[115] D. Wu and J. Xu, “On the optimal weighted

$$\backslash$$

ell_2 $regularizationinoverparameterizedlinearregression$,” *arXiv preprint arXiv:2006.05800*, 2020.

[116] D. Kobak, J. Lomond, and B. Sanchez, “The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization.,” *J. Mach. Learn. Res.*, vol. 21, pp. 169–1, 2020.

[117] D. Richards, E. Dobriban, and P. Rebeschini, “Comparing classes of estimators: When does gradient descent beat ridge regression in linear models?” *arXiv preprint arXiv:2108.11872*, 2021.

[118] P. Patil, Y. Wei, A. Rinaldo, and R. Tibshirani, “Uniform consistency of cross-validation estimators for high-dimensional ridge regression,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 3178–3186.

[119] P. Patil, A. K. Kuchibhotla, Y. Wei, and A. Rinaldo, “Mitigating multiple descents: A model-agnostic framework for risk monotonization,” *arXiv preprint arXiv:2205.12937*, 2022.

[120] E. Candes, Y. Fan, L. Janson, and J. Lv, “Panning for gold:‘model-x’knockoffs for high dimensional controlled variable selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 551–577, 2018.

[121] Y. Romano, M. Sesia, and E. Candès, “Deep knockoffs,” *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1861–1872, 2020.

[122] J. Cavazza, P. Morerio, B. Haeffele, C. Lane, V. Murino, and R. Vidal, “Dropout as a low-rank regularizer for matrix factorization,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2018, pp. 435–444.

[123] P. Mianjy, R. Arora, and R. Vidal, “On the implicit bias of dropout,” in *International Conference on Machine Learning*, PMLR, 2018, pp. 3540–3548.

[124] L. Hui and M. Belkin, “Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks,” *arXiv preprint arXiv:2006.07322*, 2020.

[125] K. Konda, X. Bouthillier, R. Memisevic, and P. Vincent, “Dropout as data augmentation,” *stat*, vol. 1050, p. 29, 2015.

[126] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, “Dropout as data augmentation,” *arXiv preprint arXiv:1506.08700*, 2015.

[127] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma, “Optimal regularization can mitigate double descent,” *arXiv preprint arXiv:2003.01897*, 2020.

[128] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*, 2010.

[129] F. Santambrogio, “Optimal transport for applied mathematicians,” *Birkäuser, NY*, vol. 55, no. 58-63, p. 94, 2015.

[130] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *2009 IEEE 12th International Conference on Computer Vision*, IEEE, 2009, pp. 460–467.

[131] J. Altschuler, J. Niles-Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via sinkhorn iteration," in *Advances in neural information processing systems*, 2017, pp. 1964–1974.

[132] M. Agueh and G. Carlier, "Barycenters in the wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[133] M. Cuturi and A. Doucet, "Fast computation of wasserstein barycenters," pp. 685–693, 2014.

[134] M. Cuturi and G. Peyré, "A smoothed dual approach for variational wasserstein problems," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 320–343, 2016.

[135] M. Cuturi and D. Avis, "Ground metric learning," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 533–564, 2014.

[136] F.-P. Paty and M. Cuturi, "Subspace robust wasserstein distances," 2019.

[137] S. Dhouib, I. Redko, T. Kerdoncuff, R. Emonet, and M. Sebban, "A swiss army knife for minimax optimal transport," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[138] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, A1111–A1138, 2015.

[139] J. Weed, F. Bach, *et al.*, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance," *Bernoulli*, vol. 25, no. 4A, pp. 2620–2648, 2019.

[140] S. Graf and H. Luschgy, *Foundations of quantization for probability distributions*. Springer, 2007.

[141] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *IEEE International Conference on Computer Vision*, 2013, pp. 2960–2967.

[142] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[143] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.