Journal of Machine Learning Research 25 (2024) 1-85 Submitted 11/22; Revised 1/24; Published 3/24

The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective

Chi-Heng Lin CL3385@GATECH.EDU
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Chiraag Kaushik CKAUSHIK7@GATECH.EDU
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Eva L. Dyer* EVADYER@GATECH.EDU
Coulter Department of Biomedical Engineering
School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Vidya Muthukumar* VMUTHUKUMAR8@GATECH.EDU
School of Electrical & Computer Engineering
School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Editor: Pradeep Ravikumar

Abstract

Data augmentation (DA) is a powerful workhorse for bolstering performance in modern machine learning. Specific augmentations like translations and scaling in computer vision are traditionally believed to improve generalization by generating new (artificial) data from the same distribution. However, this traditional viewpoint does not explain the success of prevalent augmentations in modern machine learning (e.g. randomized masking, cutout, mixup), that greatly alter the training data distribution. In this work, we develop a new theoretical framework to characterize the impact of a general class of DA on underparameterized and overparameterized linear model generalization. Our framework reveals that DA induces *implicit spectral regularization* through a combination of two distinct effects: a) manipulating the relative proportion of eigenvalues of the data covariance matrix in a training-data-dependent manner, and b) uniformly boosting the entire spectrum of the data covariance matrix through ridge regression. These effects, when applied to popular augmentations, give rise to a wide variety of phenomena, including discrepancies in generalization between over-parameterized and under-parameterized regimes and differences between regression and classification tasks. Our framework highlights the nuanced and sometimes surprising impacts of DA on generalization, and serves as a testbed for novel augmentation design.

*. Both senior authors contributed equally.

Keywords: Data augmentation, generalization analysis, overparameterized models, spectral analysis, regression, classification

1. Introduction

Data augmentation (DA), or the transformation of data samples before or during learning, is a workhorse of both supervised (Shorten and Khoshgoftaar, 2019; Iosifidis and Ntoutsi, 2018; Liu et al., 2021b) and self-supervised approaches (Gidaris et al., 2018; Chen et al., 2020b; Grill et al., 2020; Azabou et al., 2021; Zbontar et al., 2021) for machine learning (ML). It is critical to the success of modern ML in multiple domains, e.g., computer vision (Shorten and Khoshgoftaar, 2019), natural language processing (Feng et al., 2021), time series data (Wen et al., 2020), and neuroscience (Lashgari et al., 2020; Azabou et al., 2021; Liu et al., 2021a). This is especially true in settings where data and/or labels are scarce or in other cases where algorithms are prone to overfitting (Zhang et al., 2021). While DA is perhaps one of the most widely used tools for regularization, most augmentations are 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 (Cubuk et al., 2019; Ratner et al., 2017; Balestriero et al., 2022a).

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 (Dao et al., 2019; Chen et al., 2020a). However, state-of-the-art augmentations that are used in practice (e.g. data masking (He et al., 2022), cutout (DeVries and Taylor, 2017), mixup (Zhang et al., 2017)) are stochastic and can significantly alter the distribution of the data (Gontijo-Lopes et al., 2020; He et al., 2022; Yuan et al., 2021). Despite many efforts to explain the success of DA in the literature (Bishop, 1995; Chapelle et al., 2001; Chen et al., 2020a; Dao et al., 2019; Wu et al., 2020), 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 that precisely characterizes the impact of DA on generalization. Our framework enables generalization analysis for: 1. *general stochastic augmentations*, 2. the classical *underparameterized regime* (Hastie et al., 2009) and the modern *overparameterized regime*, 3. *regression* and *classification tasks*, and 4. *strong* and *weak distributional-shift augmentations*. To do this, we borrow and build on finite-sample analysis techniques that simultaneously operate in the underparameterized and overparameterized regime for linear and kernel models (Bartlett et al., 2020; Tsigler and Bartlett, 2020; Muthukumar et al., 2021, 2020).

We find that DA induces two types of implicit, training-data-dependent regularization: manipulation of the spectrum (i.e. eigenvalues) of the data covariance matrix, and the addition of explicit ℓ_2-type regularization to avoid noise overfitting. The first effect of spectral manipulation 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.

Our theory reveals *good, bad, and ugly* sides to DA depending on the setting, nature of task and type of augmentation. We find that on one hand, DA *improves* generalization when it is designed in a targeted manner to reduce variance while preserving bias (for any setting/task) or if the reduction in variance outweighs increase in bias (for classification or underparameterized regression). On the other hand, DA is more *unforgiving* for over-

parameterized regression; here, we find that popular augmentations frequently induce a large increase in both bias and distribution shift between training and test data. We also identify several *ugly* (i.e. subtle/nuanced) features to DA depending on whether the task is regression or classification, the model is underparameterized or overparameterized, and the augmentations are pre-computed or applied on-the-fly.

1.1 Main contributions

Below, we outline and provide a roadmap of the main contributions of this work.

- We propose a new framework for studying non-asymptotic generalization with data augmentation for linear models by building on the recent literature on the theory of overparameterized learning (Bartlett et al., 2020; Tsigler and Bartlett, 2020; Muthukumar et al., 2021). 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 3.1, and sharply characterize the ensuing performance for both regression and classification tasks in Sections 4.3 and 4.4, respectively.

- In Section 5, we apply our theory to provide new interpretations of a broad class of randomized DA strategies used in practice; e.g., random-masking (He et al., 2022), cutout (DeVries and Taylor, 2017), noise injection (Bishop, 1995), and group-invariant augmentations (Chen et al., 2020a). An example is as follows: while the classical noise injection augmentation (Bishop, 1995) causes only a constant shift in the spectrum, data masking (He et al., 2022; Assran et al., 2022), cutout (DeVries and Taylor, 2017) and distribution-preserving augmentations (Chen et al., 2020a) tend to *isotropize* the equivalent data spectrum. We also use our framework as a testbed for new approaches by designing a new augmentation method, inspired by isometries in random feature rotation (Section 7.1). We show that this augmentation achieves smaller bias than the least-squared estimator and variance reduction on the order of the ridge estimator.

- In Section 6 we empirically examine the influence of DA in conjunction with data and model family on generalization. We compare our closed-form expression with augmented stochastic gradient descent (SGD) (Dao et al., 2019; Chen et al., 2020a,b) and pre-computed augmentations (Wu et al., 2020; Shen et al., 2022). In addition to verifying our theoretical insights, our experiments reveal phenomena of independent interest, including surprising distinctions between pre-computed DA and augmented SGD and varying robustness to augmentation hyperparameter tuning between regression and classification tasks.

- We conclude in Section 7 with an extended discussion of the "good, bad and ugly" ideas of DA. In Section 7.1 we discuss how cleverly designed *data-adaptive* covariance modification can reduce both bias and variance, and how a broader class of DA leads to variance reduction that outweighs bias increase for classification and *underparameterized* regression tasks. In Section 7.2 we unpack the suboptimalities of the "isotropizing" effect of DA, particularly in overparameterized regression where bias is especially harmful (Muthukumar et al., 2020; Hastie et al., 2019). Finally, in Section 7.3, we identify strikingly divergent impacts of DA depending on whether the task is regression

or classification, the model is under or overparameterized, and the augmentation is pre-computed or applied to SGD. Our findings here corroborate the empirically observed benefits of DA being primarily applied "on-the-fly", on moderate-dimensional data and classification tasks (Yuan et al., 2021; Dai et al., 2022).

1.2 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 $\mathbf{V}_{a:b} := [\mathbf{v}_a, \mathbf{v}_{a+1}, \ldots, \mathbf{v}_b]$. A similar definition applies to vectors; e.g. for a vector $\mathbf{x}$ we have $\mathbf{x}_{a:b} = [\mathbf{x}_a, \mathbf{x}_{a+1}, \ldots, \mathbf{x}_b]$. The Mahalanobis norm of a vector is defined by $\|\mathbf{x}\|_{\mathbf{H}} = \sqrt{\mathbf{x}^\top\mathbf{H}\mathbf{x}}$. For a matrix $\mathbf{A}$, $\mathrm{diag}(\mathbf{A})$ denotes the diagonal matrix with a diagonal equal to that of $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ 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 which may depend only on σ_x or σ_ε and not on n or p. All asymptotic convergence results are stated in probability.

More specific notation corresponding to our signal model is given in Section 4.1, and some additional notation that is convenient to define for our analysis is postponed to Section 4.2.

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.

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 (Bishop, 1995) and *vicinal risk minimization* (Zhang et al., 2017; Chapelle et al., 2001); 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 (Shorten and Khoshgoftaar, 2019; Iosifidis and Ntoutsi, 2018; Liu et al., 2021b). 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 (Cohen and Welling, 2016; Raj et al., 2017; Mroueh et al., 2015; Bruna and Mallat, 2013; Yang et al., 2019). These studies generally consider cases in which the group structure is explicitly present in the model design via convolutional architectures (Cohen and Welling, 2016; Bruna and Mallat, 2013) or feature maps approximating group-invariant kernels (Raj et al., 2017; Mroueh et al., 2015). The authors of Chen et al. (2020a) 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 (Gontijo-Lopes et al., 2020). They argue that in many cases, it is important to introduce samples which are *out-of-distribution* (OOD) (Sinha et al., 2021; Peng et al., 2022) (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 (He et al., 2022; DeVries and Taylor, 2017). We also consider the modern overparameterized regime (Belkin et al., 2019; Dar et al., 2021). 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: Dao et al. (2019) 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 *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. Li et al. (2019) also propose a kernel classifier based on a notion of invariance to local translations, which produces competitive empirical performance. In another recent analysis, Wu et al. (2020) 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 (Grill et al., 2020; Chen et al., 2020a). Second, they assume that the augmentations would not change the labels generated by the

1. More recent studies of invariant kernel methods, trained to interpolation, suggest that invariance could either improve (Mei et al., 2021) or worsen (Donhauser et al., 2021) generalization depending on the precise setting. Our results for the overparameterized linear model (in particular, Corollary 43) also support this message.

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 (Dao et al., 2019; Chen et al., 2020a), which are arguably more commonly used in practice. Our experiments in Section 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, Hanin and Sun (2021) 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. LeJeune et al. (2019) 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, Shen et al. (2022) 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 (Allen-Zhu and Li, 2020), 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. However, we 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 Shen et al. (2022) only analyze a permutation-style augmentation. We believe that combining our general-purpose framework for DA with a more complex nonlinear model is a promising future direction, and we discuss possible analysis paths for this in Section 8.

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 Dar et al. (2021), 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 (Bartlett et al., 2020; Tsigler and Bartlett, 2020; Hastie et al., 2019; Belkin et al., 2020; Muthukumar et al., 2020). This line of research (most notably, Bartlett et al. (2020); Tsigler and Bartlett (2020)) 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 Tsigler and Bartlett (2020), 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 (Muthukumar et al., 2021; Cao et al., 2021; Wang and Thrampoulidis, 2021; Chatterji and Long, 2021; Shamir, 2022; Deng et al., 2022; Montanari et al., 2019). In particular, Muthukumar et al. (2021); Shamir (2022) 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 Muthukumar et al. (2021) 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 (Chatterji and Long, 2021; Wang and Thrampoulidis, 2021; Cao et al., 2021)), 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 Wu and Xu (2020), which analyzes the regression generalization error of general Tikhonov regularization. However, our work differs from theirs in three key respects. First, the analysis of Wu and Xu (2020) 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.

In addition to explicitly regularized estimators, Wu and Xu (2020) also analyze the ridgeless limit for these regularizers, which can be interpreted as the minimum-Mahalanobis-norm interpolator. In Section 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 (Belkin et al., 2019) 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 (Kobak et al., 2020; Tsigler and Bartlett, 2020), and gradient descent trained to interpolation can sometimes beat ridge regression (Richards et al., 2021). These results show that the scale of the ridge regularization significantly affects model generalization; consequently, recent

2. As remarked on at various points throughout the paper, we believe that the recent work of McRae et al. (2022), which weakens these assumptions further, can also be plugged with our analysis framework; we will explore this in the sequel.

work strives to estimate the optimal scale of ridge regularization using cross-validation techniques (Patil et al., 2021, 2022).

As shown in classical work (Bishop, 1995), 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 6.2 show that random mask (He et al., 2022), cutout (DeVries and Taylor, 2017) 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* (Candes et al., 2018; Romano et al., 2020) and *dropout in parameter space* (Cavazza et al., 2018; Mianjy et al., 2018). 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 (He et al., 2022) augmentation) during the learning process, and has been shown to have a close connection with sparsity regularization (Mianjy et al., 2018). 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. 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 discuss how augmentations affect the ERM procedure. Then, we derive a reduction to ridge regression that paves the way for our analysis in Section 4.

3.1 Empirical risk minimization with data augmentation

Traditionally, high-dimensional ML models are commonly trained to minimize a combination of prediction error on training data and some measure of model complexity. 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 $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} [\ell(\mathbf{X}\boldsymbol{\theta}, \mathbf{y}) + R(\boldsymbol{\theta})]$ 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. 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.

Modern machine learning relies heavily on data augmentation (DA) to achieve state-of-the-art performance (Shorten and Khoshgoftaar, 2019; Zhang et al., 2021). Augmentations are typically applied *on-the-fly* and stochastically to different examples during training (Chen et al., 2020a; Grill et al., 2020; Chen et al., 2020b). This procedure, known as *augmented stochastic gradient descent (aSGD)*, is widely used in practice. Chen et al. (2020a) showsed that aSGD converges to the solution of the following *augmented empirical risk minimization (aERM)* problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_G \left[\|G(\mathbf{X})\boldsymbol{\theta} - \mathbf{y}\|_2^2 \right]. \tag{1}$$

Above, G denotes a stacked data augmentation function applied to each row of the matrix, i.e., $G(\mathbf{X}) = [g_1(\mathbf{x}_1) \ldots, g_n(\mathbf{x}_n)]^T$; we assume that the transformations g_i are stochastic and are drawn i.i.d. from an augmentation function distribution $\mathcal{G}$. For example, the classical *Gaussian noise injection* augmentation (Bishop, 1995) is stochastic and takes the form $g(\mathbf{x}) = \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is an isotropic Gaussian random variable.

To characterize the impact of different augmentations on our solution, we begin by defining the first and second-order statistics of an augmentation distribution. We will show that these quantities appear in the closed-form characterization to the least-squares aERM problem for *any* augmentation function.

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 given data point $\mathbf{x}$ as*

$$\mu_{\mathcal{G}}(\mathbf{x}) := \mathbb{E}_{g\sim\mathcal{G}}[g(\mathbf{x})], \quad \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}) := \mathbb{E}_{g\sim\mathcal{G}}\left[\left(g(\mathbf{x}) - \mu_{\mathcal{G}}(\mathbf{x})\right) \left(g(\mathbf{x}) - \mu_{\mathcal{G}}(\mathbf{x})\right)^\top \right], \tag{2}$$

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

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

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

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

With this notation introduced, we now explain why DA gives rise to implicit regularization. For simplicity, we consider augmentation distributions that are unbiased on average here[4]. Then, we can simplify the objective (1) as:

$$\begin{aligned}
\mathbb{E}_G[\|G(\mathbf{X})\boldsymbol{\theta} - \mathbf{y})\|_2^2] &= \mathbb{E}_G[\| \left(G(\mathbf{X}) - \mu(\mathbf{X})\right)\boldsymbol{\theta} + \mu(\mathbf{X})\boldsymbol{\theta} - \mathbf{y})\|_2^2] \\
&= \|\mu(\mathbf{X})\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \|\boldsymbol{\theta}\|^2_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})} \\
&= \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \|\boldsymbol{\theta}\|^2_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})},
\end{aligned} \tag{3}$$

3. Note that this definition of bias is completely different from the bias-variance decomposition that manifests in regression analysis, i.e., (6).
4. We handle biased augmentation distributions in Section 4.3.2.

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 $\|\boldsymbol{\theta}\|^2_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})}$, 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. As a first step, we unpack the effects of the DA-induced regularizer $\|\boldsymbol{\theta}\|^2_{n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})}$. Note that the objective (3) 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) admits a closed-form solution given by $\hat{\boldsymbol{\theta}}_{\mathrm{aug}} = (\mathbf{X}^\top\mathbf{X} + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mathbf{X}^\top\mathbf{y}$. We use this solution to link the estimator $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ 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 $\mathbf{X}$.

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathrm{aug}} &= (\mathbf{X}^\top\mathbf{X} + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mathbf{X})^\top\mathbf{y} \\
&= \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}(n\mathbf{I}_p + \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\mathbf{X}^\top\mathbf{X}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2})^{-1}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\mathbf{X}^\top\mathbf{y} \\
&= \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}(n\mathbf{I}_p + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y} \qquad (\text{where } \widetilde{\mathbf{X}} := \mathbf{X}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}) \\
&= \mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\hat{\boldsymbol{\theta}}_{\mathrm{ridge}}, \qquad \text{where } \hat{\boldsymbol{\theta}}_{\mathrm{ridge}} := (n\mathbf{I}_p + \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y}. \qquad (4)
\end{aligned}$$

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 $\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$ is equivalent to $\|\hat{\boldsymbol{\theta}}_{\mathrm{ridge}} - \mathrm{Cov}_{\mathcal{G}}{}^{1/2}\boldsymbol{\theta}^*\|^2_{\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\boldsymbol{\Sigma}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}}$. Suppose, for a moment, that $\mathrm{Cov}_{\mathcal{G}}$ were fixed (or independent of $\mathbf{X}$). Then, (4) 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[5] $\lambda = n$, 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) ℓ_2 regularization at a scale that is proportional to the number of training samples ($\lambda_{\mathrm{reg}} = n$),

b) a modification of the original data covariance from $\boldsymbol{\Sigma}$ to $\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}\boldsymbol{\Sigma}\mathrm{Cov}_{\mathcal{G}}{}^{-1/2}$, which can make a sizable impact on the original spectrum.

It is important to note that this equivalence between solutions is only approximate since $\mathrm{Cov}_{\mathcal{G}}$ itself depends on $\mathbf{X}$. We will justify and formalize this approximation in Section 4.2.

3.3 Application to different augmentations used in practice

Understanding the closed-form expression of the aERM estimator above requires exactly characterizing the augmentation covariance operator $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$. In Table 1, we list several common augmentations for which the augmentation covariance can be easily characterized and interpreted, including: White and Correlated Gaussian noise, Unbiased Random Mask, Pepper noise, and Random Cutout. The derivations for these expressions are in Appendix E.

5. This demonstrates that *negative* regularization, which is studied in some recent work (Tsigler and Bartlett, 2020; Kobak et al., 2020) is not possible to achieve through the DA framework.

The expressions for $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ in Table 1 reveal that, in many cases, the augmentation covariance is characterized by an interesting interplay between properties of the training data matrix $\mathbf{X}$ and parameters of the augmentation distribution. For example, in the case of the unbiased random mask augmentation, $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ is a diagonal matrix whose entries depend on the covariance matrix of the training data $\mathbf{X}^\top\mathbf{X}$ and the masking probability β. The salt-and-pepper augmentation has a similar term appear in its augmentation covariance (corresponding to the "salt" part of the augmentation), along with a data-independent term that has the same form as Gaussian noise injection (corresponding to the "pepper" part of the augmentation). We 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 $g(\mathbf{X}) = \mathbf{X} + \mathbf{N}$, $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A}(\mathbf{X}))$.

Table 1: *Examples of common augmentations* for which we can compute a closed-form solution to the aERM objective. Here, $\mathbf{M}$ is a circulant matrix defined in Appendix E.

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

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 7. Table 2 provides the road map of our main results and their applications in this and the next sections.

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 $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{z}$, where we assume, without loss of generality, that $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ 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 ($p > n$); however, 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 McRae et al. (2022), where features are not assumed to be sub-Gaussian, but we leave this extension to future work.

4.1.1 Error Metrics

In this work, we will focus on the squared loss training objective (1) 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 (Hui and Belkin, 2020; Muthukumar et al., 2021; Wang and Thrampoulidis, 2021; Chatterji and Long, 2021) has shown that the squared loss can achieve competitive results when compared with the cross-entropy loss in classification tasks[6]. For the regression task, we use the *mean squared error (MSE)*, defined for an estimator $\hat{\boldsymbol{\theta}}$ as:

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

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:

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

4.1.2 Spectral quantities of interest

Recent works studying overparameterized regression and classification tasks (Bartlett et al., 2020; Tsigler and Bartlett, 2020; Muthukumar et al., 2021; Zou et al., 2021) 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, (Bartlett et al., 2020)) *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:*

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

6. 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 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.* 15, 17, 16, 34 Compositions: *Cor.* 18	Cutout: *Cor.* 17, 42, 45 Group Invariant: *Cor.* 43
Interplay with Signal Model	*Corollary* 16	*Corollary* 45
Comparisons Between Under- & Over-parameterized Regimes	*Corollary* 15, 42, 43	
Comparisons between Regression & Classification	*Proposition* 46, 47	

Using this notation, the risk for the minimum-norm least squares estimate from Bartlett et al. (2020); Tsigler and Bartlett (2020) can be sharply characterized as

$$\text{MSE} \asymp \underbrace{\|\boldsymbol{\theta}^* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Bias}} + \underbrace{\|\hat{\boldsymbol{\theta}} - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Variance}}, \text{ where}$$

$$\text{Bias} \lesssim \|\boldsymbol{\theta}^*_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + \|\boldsymbol{\theta}^*_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 \lambda_{k+1}^2 \rho_k(\boldsymbol{\Sigma};0)^2, \quad \text{Variance} \asymp \frac{k}{n} + \frac{n}{R_k(\boldsymbol{\Sigma};0)}.$$

Here, $k \le \min(n,p)$ is an index that partitions the spectrum of the data covariance $\boldsymbol{\Sigma}$ 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 up to universal constant factors for certain types of signal: either random (Tsigler and Bartlett, 2020) or sparse (Muthukumar et al., 2021).

Intuitively, this characterization implies a two-fold requirement on the data spectrum for good generalization (in the sense of statistical consistency: 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).

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 3.2. 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 $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ with an idealized estimator $\bar{\boldsymbol{\theta}}_{\mathrm{aug}}$ that uses the *expected* augmentation covariance (over the original data distribution). The two estimators are formally defined below:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{aug}} &= (\mu_{\mathcal{G}}(\mathbf{X})^\top \mu_{\mathcal{G}}(\mathbf{X}) + n\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu_{\mathcal{G}}(\mathbf{X})^\top \mathbf{y},\\ \bar{\boldsymbol{\theta}}_{\mathrm{aug}} &= (\mu_{\mathcal{G}}(\mathbf{X})^\top \mu_{\mathcal{G}}(\mathbf{X}) + n\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1}\mu_{\mathcal{G}}(\mathbf{X})^\top \mathbf{y},\end{aligned} \tag{5}$$

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

$$\mathrm{MSE} \lesssim \underbrace{\|\boldsymbol{\theta}^* - \mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\mathrm{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Bias}} + \underbrace{\|\bar{\boldsymbol{\theta}}_{\mathrm{aug}} - \mathbb{E}_\varepsilon[\bar{\boldsymbol{\theta}}_{\mathrm{aug}}|\mathbf{X}]\|_{\boldsymbol{\Sigma}}^2}_{\text{Variance}} + \underbrace{\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\boldsymbol{\Sigma}}^2}_{\text{Approximation Error}}. \tag{6}$$

The bias and variance terms can be analyzed with relative ease through an extension of the techniques of Bartlett et al. (2020); Tsigler and Bartlett (2020) to general positive-semidefinite regularizers that are not dependent on the training data[7] $\mathbf{X}$, as we outlined in Section 3.2. We provide a novel analysis of the approximation error term in Section 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 4.3.1, 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 1 confirms that the approximation error is indeed negligible. In this plot, we show the decomposition corresponding to the terms in (6) for random mask augmentation with different masking probabilities denoted by β. We can see that the approximation error is small compared with the other error components.

That the approximation error is negligible is a 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 $n, p \to \infty$.

7. For this case, a related contribution lies in the work of Wu and Xu (2020). Note that Wu and Xu (2020) 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 equivalent version of Assumption 1 for the proportional regime $p \propto n$, but do not pursue this path in this paper.

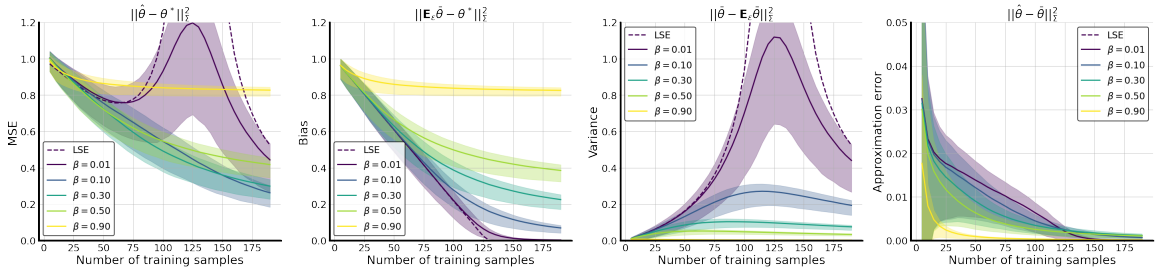


Figure 1: *Decomposition of MSE into the bias, variance, and approximation error as in Theorem 4.* A random masking augmentation is applied with different dropout probability β and the bias, variance, and approximation error are computed as a function of the number of training samples. The approximation error is small compared to the bias and variance and goes to zero quickly with more training data.

Assumption 1 *Let the data dimension p grows with n at a polynomial rate $p \asymp n^\alpha$ for some $\alpha > 0$. 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 $\Delta_G \to 0$ as $n \to \infty$ almost surely, where*

$$\Delta_G := \left\| \frac{1}{n} \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} \sum_{i=1}^{n} \mathrm{Cov}_{\mathcal{G}}(\mathbf{x}_i) \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}_p \right\|. \tag{7}$$

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 fast as $n, p \to \infty$. We will show in Section 4.5 that a wide class of popular 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.

4.3 Regression analysis

With the connection of DA to ridge regression established in Section 3.2 and the deterministic approximation method established in Section 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 4.3.1, and biased-in-average augmented estimators are studied in Section 4.3.2. The applications of the general theorem in this section will be discussed in detail in Section 7.

4.3.1 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 Appendix B. 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-*

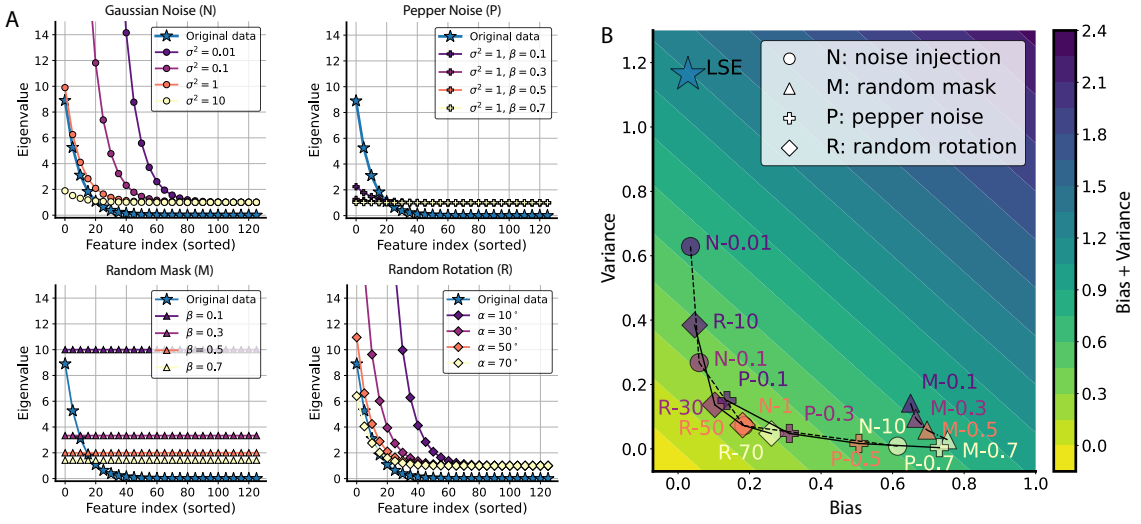


Figure 2: *Visualizing the augmented data spectrum and generalization for different forms of DA.* On the left in (A), we visualize the regularized augmented spectrum in Equation (9)), clockwise for Gaussian noise, pepper noise, random mask, and our novel random rotation introduced in Section 5.5. On the right in (B), we show their corresponding generalization, where the number indicated for each data point denotes the strength of its augmentation parameter. The LSE (star) represents the baseline of least-squared estimator without any augmentation.

transformed data covariance $\mathbf{\Sigma}_{aug}$*, by*

$$\bar{\mathbf{\Sigma}} := \mathbb{E}_{\mathbf{x}}[(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])^\top], \tag{8}$$

$$\mathbf{\Sigma}_{aug} := \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\bar{\mathbf{\Sigma}}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}. \tag{9}$$

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*

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

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 that* $\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}$ *if the augmentation is unbiased in average.*

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

Theorem 4 (High probability bound for MSE with unbiased DA) *Consider an unbiased data augmentation* g *and its corresponding estimator* $\hat{\boldsymbol{\theta}}_{aug}$*, where* Δ_G *is defined in Eq. 7 and* κ *is the condition number of* $\mathbf{\Sigma}_{aug}$*. Assume for some integers* k_1*,* k_2*, the condition numbers for the matrices* $\mathcal{A}_{k_1}(\mathbf{X}_{aug}; n)$*,* $\mathcal{A}_{k_2}(\mathbf{X}_{aug}; n)$ *(defined in Section 1.2) are bounded by* L_1 *and* L_2 *respectively with probability* $1 - \delta'$*, and that* $\Delta_G \leq c'$ *for some constant* $c' < 1$*.*

Then , with probability $1-\delta'-4n^{-1}$*, the test mean-squared error is bounded by*

$$\begin{aligned}
\text{MSE} &\lesssim \text{Bias} + \text{Variance} + \text{ApproximationError}, \qquad (10)\\
\frac{\text{Bias}}{L_1^4} &\lesssim \left(\left\| \mathbf{P}_{k_1+1:p}^{\mathbf{\Sigma}_{aug}} \theta^*_{aug} \right\|^2_{\mathbf{\Sigma}_{aug}} + \left\| \mathbf{P}_{1:k_1}^{\mathbf{\Sigma}_{aug}} \theta^*_{aug} \right\|^2_{\mathbf{\Sigma}_{aug}^{-1}} \frac{(\rho_{k_1}^{aug})^2}{(\lambda_{k_1+1}^{aug})^{-2} + (\lambda_1^{aug})^{-2}(\rho_{k_1}^{aug})^2} \right),\\
\frac{\text{Variance}}{L_2^2} &\lesssim \left(\frac{k_2}{n} + \frac{n}{R_k^{aug}} \right) \log n, \quad \text{Approx.Error} \lesssim \kappa^{\frac{1}{2}} \Delta_G \left(\|\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}} + \sqrt{\text{Bias} + \text{Variance}} \right).
\end{aligned}$$

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

Theorem 4 illustrates the critical role that the spectrum of the augmentation-transformed data covariance $\mathbf{\Sigma}_{\text{aug}}$ plays in generalization. In particular, we find that, up to an approximation error term, the generalization error is characterized by the effective ranks ρ_k^{aug} and R_k^{aug}(rather than the original effective ranks of the covariance, as in Tsigler and Bartlett (2020)). Intuitively, we expect an increase in the bias as ρ^{aug} increases and variance reduction as R^{aug} increases.

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 Tsigler and Bartlett (2020). 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 $n^{-1}\text{Cov}_{\mathcal{G}}(\mathbf{X})$ to $\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]$ is sufficiently fast with respect to n, i.e. Δ_G is sufficiently small. This condition is formalized in the lemma below.

Lemma 5 (Condition on bias/variance dominating error approximation) *Suppose the conditions of Theorem 4 hold. If*

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

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

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

Proof The lemma follows from Theorem 4 with the observation that

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

■

4.3.2 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*

$$\xi(\mathbf{x}) := \mu_g(\mathbf{x}) - \mathbf{x}, \quad \mathrm{Cov}_\xi := \mathbb{E}_{\mathbf{x}}\left[\xi(\mathbf{x})\xi(\mathbf{x})^\top\right]. \tag{11}$$

In a similar spirit to Δ_G, we define $\Delta_\xi := \left\|\frac{1}{n}\sum_{i=1}^n(\mu_g(\mathbf{x}_i) - \mathbf{x}_i)(\mu_g(\mathbf{x}_i) - \mathbf{x}_i)^\top - \mathrm{Cov}_\xi\right\|$.

Since $\xi(\mathbf{x})$ is not zero for a biased augmentation, the closed-form expression for the aERM estimator $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ becomes more complicated and we lose the exact equivalence to an ridge regression in (4). 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 $\mu(\mathbf{x}) = \bar{\boldsymbol{\Sigma}}^{1/2}\bar{\mathbf{z}}$, where $\bar{\boldsymbol{\Sigma}}$ 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 $\bar{\boldsymbol{\Sigma}} := \mathbb{E}_{\mathbf{x}}[(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])(\mu_{\mathcal{G}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}[\mu_{\mathcal{G}}(\mathbf{x})])^\top]$. Now we are ready to state our theorem for biased augmentations. The proof is deferred to Appendix B.3.

Theorem 7 (Bounds on the MSE for Biased Augmentations) *Consider the estimator $\hat{\boldsymbol{\theta}}_{aug}$ obtained by solving the aERM in (1). Let $\mathrm{MSE}^o(\hat{\boldsymbol{\theta}}_{aug})$ denote the unbiased MSE bound in Eq. (10) of Theorem 4, and Δ_G defined in Eq. 7. Suppose the assumptions in Theorem 4 hold for the mean augmentation $\mu(\mathbf{x})$ and that $\Delta_G \le c < 1$. 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

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

Our upper bound for the MSE in the biased augmentation case is a generalization of the bound in Tsigler and Bartlett (2020) 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 $\boldsymbol{\Sigma}$ but our training covariates have covariance $\bar{\boldsymbol{\Sigma}}$ 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-on-average, i.e., $\mu_{\mathcal{G}}(\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x}$, since $\mathbf{\Sigma} = \bar{\mathbf{\Sigma}}$, $\Delta_\xi = 0$ and $\mathrm{Cov}_\xi = 0$. Thus, we directly recover Theorem 4 in this case. Whether Theorem 7 is tight in general is an interesting open question for future work.

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 7.

4.4.1 Classification analysis setup

We adopt the random signed model from Muthukumar et al. (2021), noting that we expect similar analysis to be possible for the Gaussian-mixture-model setting of Chatterji and Long (2021); Wang and Thrampoulidis (2021) (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

$$y_i = \begin{cases} \mathrm{sgn}(\mathbf{x}_i^\top \boldsymbol{\theta}^*) & \text{with probability } 1 - \nu^* \\ -\mathrm{sgn}(\mathbf{x}_i^\top \boldsymbol{\theta}^*) & \text{with probability } \nu^* \end{cases}. \tag{12}$$

Just as in Muthukumar et al. (2021), we make a *1-sparse* assumption on the true signal $\boldsymbol{\theta}^* = \frac{1}{\sqrt{\lambda_t}} \mathbf{e}_t$. We denote $\mathbf{x}_{\mathrm{sig}} := \mathbf{x}_t$ to emphasize the signal feature. Motivated by recent results which demonstrate the effectiveness of training with the squared loss for classification tasks (Hui and Belkin, 2020; Muthukumar et al., 2021), 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. (1)).

Muthukumar et al. (2021) 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 37, the contamination term scales identically to the variance from regression analysis). The definitions of these quantities are given below.

Definition 8 (Survival and contamination (Muthukumar et al., 2021)) *Given an estimator $\hat{\theta}$, its survival (SU) and contamination (CN) are defined as*

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

For Gaussian data, Muthukumar et al. (2021) derived the following closed-form expression for the Probability-of-Error (POE):

$$\text{POE}(\hat{\boldsymbol{\theta}}) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{\text{SU}(\hat{\boldsymbol{\theta}})}{\text{CN}(\hat{\boldsymbol{\theta}})}. \tag{13}$$

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 and are augmented independently*[8]*, 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.

4.4.2 Classification analysis for unbiased-on-average augmentations

Now, we present our main theorem for the classification task under the setting in Assumption 3. The proof of this theorem is deferred to Appendix C.

Theorem 9 (Bounds on Probability of Classification Error) *Let* $t \le n$ *be the index (arranged according to the eigenvalues of* $\boldsymbol{\Sigma}_{aug}$*) of the non-zero coordinate of* $\boldsymbol{\theta}^*$, $\widetilde{\boldsymbol{\Sigma}}_{aug}$ *be the leave-one-out modified spectrum corresponding to index* t, *and* $\widetilde{\mathbf{X}}_{aug}$ *be the leave-one-column-out data matrix corresponding to column* t. *Suppose there exists a* $t \le k \le n$ *such that with probability at least* $1-\delta$, *the condition numbers of* $n\mathbf{I}+\widetilde{\mathbf{X}}^{aug}_{k+1:p}(\widetilde{\mathbf{X}}^{aug}_{k+1:p})^\top$, $n\mathbf{I}+\mathbf{X}^{aug}_{k+1:p}(\mathbf{X}^{aug}_{k+1:p})^\top$, *and* $\widetilde{\mathbf{X}}_{k+1:p}\boldsymbol{\Sigma}_{k+1:p}\widetilde{\mathbf{X}}^T_{k+1:p}$ *are at most* L. *Then as long as* $\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\text{SU})$ *and* $\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\text{CN})$,

$$\text{POE}(\hat{\theta}) \lesssim \frac{\text{CN}}{\text{SU}}\left(1 + \sigma_z\sqrt{\log\frac{\text{SU}}{\text{CN}}}\right), \tag{14}$$

8. 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 McRae et al. (2022), we believe that the independence between signal and noise features can also be relaxed.

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

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

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

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

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

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* $\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}}=O(\text{SU})$ *and* $\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}}=O(\text{CN})$ *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 40.*

Based on the upper and lower bounds provided for SU and CN, we see that these quantities depend crucially on the effective ranks of the induced covariance matrix $\boldsymbol{\Sigma}_{aug}$. In particular, we note that SU is large when ρ_k^{aug} is small and CN is small when R_k^{aug} is large; good generalization relies on having a careful balance of these two factors to ensure that the ratio of CN to SU is small. For favorable classification performance, Theorem 9 also requires $t\le n$. 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 $\boldsymbol{\Sigma}_{aug}$. Such requirements have also been used in past analyses of both regression (Tsigler and Bartlett, 2020) and classification (Muthukumar et al., 2021).

4.4.3 Classification analysis for general biased-on-average 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 C.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 (1) 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*

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

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.

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. Note that the sign preservation is only required in expectation and not for every realization of the augmentation, i.e., we only require $\mathbb{E}_g[g(\mathbf{x})_t]$ 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 Wu et al. (2020).

4.5 Classes of augmentations for which our theory applies

In this section, we delineate important classes of augmentations for which our theory provides a sharp characterization of their impact on generalization. More formally, we show under these classes of augmentations that the approximation error term of Theorem 4 is negligible with respect to the bias/variance terms and our analysis is tight, i.e. Lemma 5 holds and Theorem 4 is tight up to constant/logarithmic factors. Recall that Lemma 5 requires the (normalized) error of the "sample" augmentation covariance, denoted by Δ_G, to be sufficiently small with respect to the sum of the bias and variance terms. This section shows that the value of Δ_G inherently depends on the extent of correlation between the augmentations across features (where the correlation is defined for fixed data, and only with respect to the stochasticity in the augmentations). At a high level, we show that: a) Δ_G is negligible as long as the correlations between the feature augmentations are weak enough, and that b) this sufficiently weak level of correlation is indeed the case for several popular classes of augmentations.

We first analyze the simplest case of *uncorrelated feature augmentations*, and then generalize our analysis to *regionally correlated feature augmentations* and augmentations with *"small" off-diagonal component.*

Uncorrelated-feature augmentations: Many augmentations involve independently augmenting each of the features (or, more generally, applying an augmentation which is uncorrelated across features). This class subsumes many prevailing augmentations like random mask and salt-and-pepper noise. Because the augmentation covariance $\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})$ is diagonal for such augmentations, we can show that Δ_G is small.

Proposition 12 (Uncorrelated Feature Augmentations) *Let the augmentation g be composed of p uncorrelated feature augmentation maps, i.e., $g(\mathbf{x}) = \begin{bmatrix} g_1(x_1) & \dots & g_d(x_d) \end{bmatrix}$*

where $\{g_i(\cdot)\}_{i\in[p]}$ *are uncorrelated (with respect to the randomness in the augmentation). If the variance of each feature augmentation* $\mathrm{Var}_{g_i}(g_i(x_i))$ *(which is a random variable due to the randomness in* x_i*) is sub-exponential with sub-exponential norm* σ_i^2 *and mean* $\bar{\sigma}_i^2$ *for all* $i \in [p]$*, then we have*

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

with probability at least $1 - \frac{1}{n}$.

Proposition 12 is proved in Appendix B.4 and gives a bound on Δ_G of the order $O(\sqrt{\frac{\log n}{n}})$. However, one might wonder whether the approximation error still vanishes for stochastic augmentations that include dependencies between features, i.e. the random variables $\{g_i(x_i)\}_{i=1}^p$ are not necessarily uncorrelated for a fixed value of $\mathbf{x}$. To address this question, the following subsection describes two techniques to bound Δ_G for two important types of such "feature-dependent" augmentations.

Regionally correlated feature augmentations: First, we consider a popular class of augmentations that are correlated to a limited extent across features. This encompasses many "patch"-based augmentations used in image applications, such as the PatchShuffle augmentation of Kang et al. (2017)[9]. To define this type of augmentation, we categorize the features into k groups denoted by $B_1, \ldots, B_k \subset [p]$. In image applications, each group B_j can represent a local region of an image. We assume that only the augmentations within a group can be correlated, meaning that $\mathbb{E}[g_{j_1}(x_{j_1})g_{j_2}(x_{j_2})] \neq 0$ only if $j_1, j_2 \in B_j$ for some $j \in [k]$, i.e. if the features belong to the same group. We then overload notation and write the augmentation in block form as $g(\mathbf{x}) = [g_1(\mathbf{x}_1), g_2(\mathbf{x}_2), \ldots, g_k(\mathbf{x}_k)]$, where $g_j \sim \mathcal{G}_j$. In this notation, each *sub-feature* $\mathbf{x}_j$ has smaller dimensionality $\mathbf{x}_j \in \mathbb{R}^{|B_j|}$ (and we have $\sum_{j\in[k]} |B_j| = p$ by definition) and the augmentation covariance for the jth block is denoted $\mathrm{Cov}_{\mathcal{G}_j}(\mathbf{x}_j)$. Note that our assumption on correlations being only within the blocks $\{B_j\}_{j\in[k]}$ implies that the covariance matrix $\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})$ will have a block-diagonal structure for any data point $\mathbf{x}$.

We have the following proposition for regionally correlated feature augmentations. The proof is contained in Appendix B.5.

Proposition 13 *Consider a correlated-feature augmentation of the form described above. Further, assume that the smallest eigenvalue of* $\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})$ *is lower bounded by* σ *for every* k*, and* g_k *is component-wise bounded, i.e.,* $\|g_k(\mathbf{x}_k)\|_\infty \leq M$ *for any* k*. Then, we have*

$$\Delta_G \lesssim \frac{M^2 \max_k |B_k|}{\sigma} \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - \frac{1}{p}$.

9. We derive $\mathrm{Cov}_{\mathcal{G}}$ for this augmentation in Appendix E.

Augmentations with a "small" off-diagonal component: At this stage, it is natural to ask whether any guarantees are possible for augmentations that do not enjoy the properties of independence or weak correlation. While we do not provide a guarantee for arbitrary augmentations on high-dimensional data, 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 more complex dependencies between features. Specifically, we state and prove the following result.

Proposition 14 *Consider the decomposition* $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \mathbf{D} + \mathbf{Q}$*, where* $\mathbf{D}$ *is a diagonal matrix representing the independent feature augmentation part. Then, we have*

$$\Delta_G \lesssim \frac{\|\mathbf{D} - \mathbb{E}\mathbf{D}\| + \|\mathbf{Q} - \mathbb{E}\mathbf{Q}\|}{\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))}. \tag{15}$$

The proof of Proposition 14 is provided in Section B.6, along with further discussion on the approximation error for dependent feature augmentations. We use Eq. (15) to show that even if the quantity $\|\mathbf{Q} - \mathbb{E}\mathbf{Q}\|$ is large (due to dependencies among features in the augmentations), it can be mitigated by the denominator of Eq. (15) for augmentations for which $\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))$ is large. We use this in Appendix F to characterize the approximation error for two examples of augmentations that induce global dependencies between features: a) the new *random-rotation* augmentation that we introduced in Section 5.5, b) the cutout augmentation which is popular in deep learning practice (DeVries and Taylor, 2017).

5. Case Studies: Applying Our Theory to Study Different Classes of DA

In this section, we will use the meta-theorems established in Section 4.3 and 4.4 to get further insight into the impact of DA on generalization. In particular, we present and interpret generalization guarantees for commonly used augmentations including: *Gaussian noise injection, randomized mask, cutout, salt-and-pepper noise, and our newly proposed random-rotation augmentation.*

5.1 Gaussian noise injection

As a preliminary example, we note that Theorem 4 generalizes and recovers the existing bounds on the ridge and ridgeless estimators (Bartlett et al., 2020; Tsigler and Bartlett, 2020). This is consistent with classical results (Bishop, 1995) that show an equivalence between augmented ERM with Gaussian noise injection and ridge regularization. Specifically, an application of the theorem to Gaussian noise injection with variance σ^2 recovers existing bounds for ridge estimators with regularization parameter $\lambda = n\sigma^2$, where the number of samples n controls the amount of regularization applied to the estimator. For completeness, we include this bound in Appendix B.7.

5.2 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 (He et al., 2022; Konda et al., 2015),[10] and is a simplified version of the popular cutout augmentation (DeVries and Taylor, 2017). The following corollary characterizes the generalization error arising from the randomized mask augmentation in regression tasks.

Corollary 15 (Regression bounds for unbiased randomized mask) *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 Theorem 4. Then, for any set $\mathcal{K} \subset \{1, 2, \ldots, p\}$ consisting of k_1 elements and some choice of $k_2 \in [0, n]$, 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 $1 - \delta' - n^{-1}$.

Note that $\psi = \frac{\beta}{1-\beta}$ increases monotonically in the mask probability β, Corollary 15 shows that bias increases with the mask intensity β, while the variance decreases. Figure 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 (Muthukumar et al., 2020; Hastie et al., 2019; Bartlett et al., 2020), 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 also note 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 Tsigler and Bartlett (2020) imply tightness of our bound in this range).

We also derive guarantees for regression with the biased variant of random masking in Corollary 34 in Appendix B.7. In Appendix C.4, we provide bounds for unbiased random mask in the classification setting (note that Theorem 11 implies the biased and unbiased case behave similarly for classification).

5.2.1 Feature-adaptive random masking

We consider, as in Corollary 16, the case of a nonuniform random masking augmentation in which the features that encode signal are masked with a lower probability than the remaining

10. We note that a superficially similar implicit regularization mechanism is at play in *dropout* (Bouthillier et al., 2015), 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.

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 B.7 and the analogous classification result to Corollary 45 in Appendix C.4.

Corollary 16 (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 $\psi_1 \le \psi_0$, we have with probability at least $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$*

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

$$\text{Approx.Error} \lesssim \sqrt{\frac{\psi_1}{\psi_0}} \sigma_z^2 \sqrt{\frac{\log n}{n}} \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}.$$

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

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

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 16 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).

5.3 Random cutout

Next, we consider the popularly used *cutout* augmentation (DeVries and Taylor, 2017), 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 B.7.

Corollary 17 (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*

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

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

5.4 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 B.7.

Corollary 18 (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:*

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

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* $\hat{\boldsymbol{\theta}}_{rm}(\beta)$*):*

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

Corollary 18 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.

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

5.5 A new "random-rotation" augmentation

Our framework can also serve as a testbed for designing new augmentations that have desired properties in terms of how they effect the spectrum. As an example, we introduce a novel augmentation that performs multiple rotations in random planes. Specifically, for an input $\mathbf{x} \in \mathbb{R}^p$ and user specified rotation angle α, 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 $\mathbf{U}_i = [\mathbf{u}_{2i-1}, \mathbf{u}_{2i}]$ and $i = 1, 2, \ldots, \frac{p}{2}$.
3. Rotate $\mathbf{x}$ by an angle α in each of these planes $\mathbf{U}_i$, $i = 1, 2, \ldots, \frac{p}{2}$.

Note that in an implementation of aSGD, we would pick an independent orthonormal basis for each iteration and each training example in Step 1. 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

$$\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \frac{4(1-\cos\alpha)}{np} \left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} - \mathbf{X}^\top\mathbf{X} \right).$$

The full derivation is deferred to Appendix E.

We can use our theory to study the generalization error for our proposed augmentation and compare it with ridge regression. Interestingly, this augmentation enjoys good generalization performance, *regardless of the signal model*. This result is summarized through the following Corollary.

Corollary 19 (Generalization of random-rotation augmentation) *Let $\hat{\boldsymbol{\theta}}_{rot}$ denote the estimator induced by the random-rotation augmentation with angle parameter α. 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* $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{rot}) \lesssim \mathrm{Var}(\hat{\boldsymbol{\theta}}_{ridge,\lambda})$. *Let $\hat{\boldsymbol{\theta}}_{lse}$ 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*

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

The proof of the bias and variance expressions are provided in Appendix C.4, and the proof of the approximation error is provided in Appendix F (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.

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.

6.1 Convergence of aSGD to the closed-form aERM solution

In this paper, we mathematically study a-ERM (the solution in Equation (1)); 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. 1 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) $\times$ (augmentation size) per backward pass. Fig. 3 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 η_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,...,$\frac{\text{n}}{\text{B}}$ do

    for i=1,...,B in the batch $\mathcal{B}_k$ do

      Draw H augmentations $g_{ij} \sim \mathcal{G}$, $j = 1, \ldots,$H

    end

    $\hat{\boldsymbol{\theta}}_{t+1} \leftarrow \hat{\boldsymbol{\theta}}_t - \eta_t \sum_{i=1}^{B} \sum_{j=1}^{H} \nabla_{\boldsymbol{\theta}} (\langle \boldsymbol{\theta}, g_{ij}(\mathbf{x}_i) \rangle - y_i)_2^2 |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_t}$

  end

end

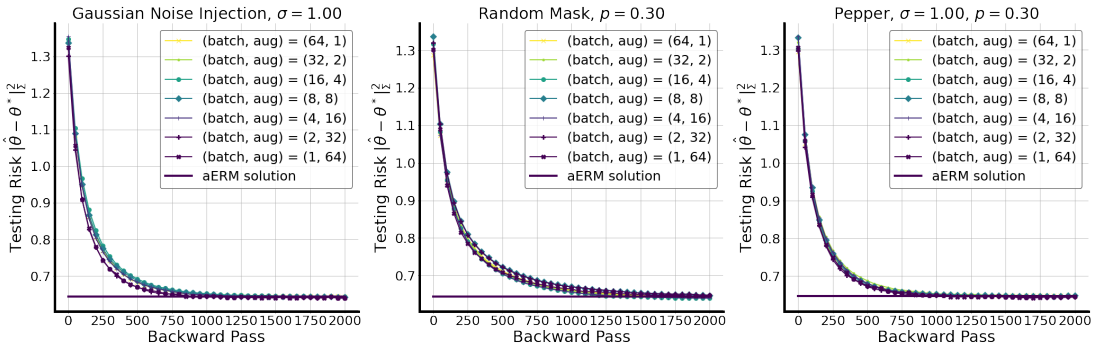


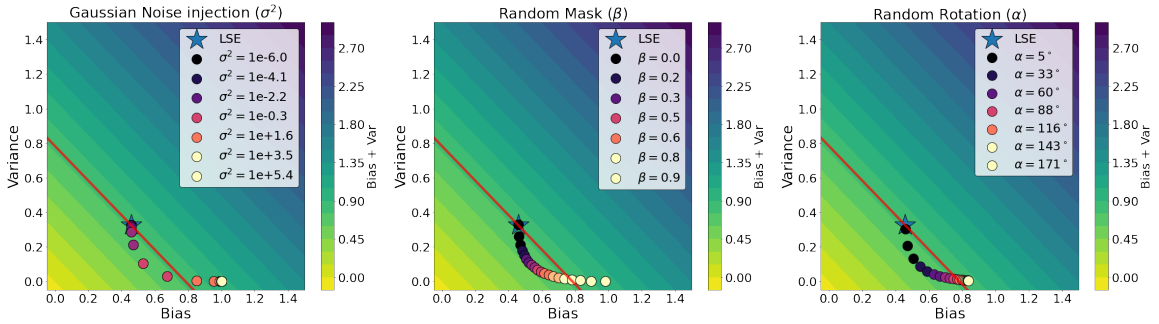
Figure 3: *Convergence of augmented stochastic gradient descent (a-SGD, Algorithm 1)* as a function of the number of backward passes to the closed-form solution of the a-ERM objective (Equation (1)). The result shows fairly stable convergence across different batch sizes and augmentation copies per sample.

6.2 Comparisons of different types of augmentations

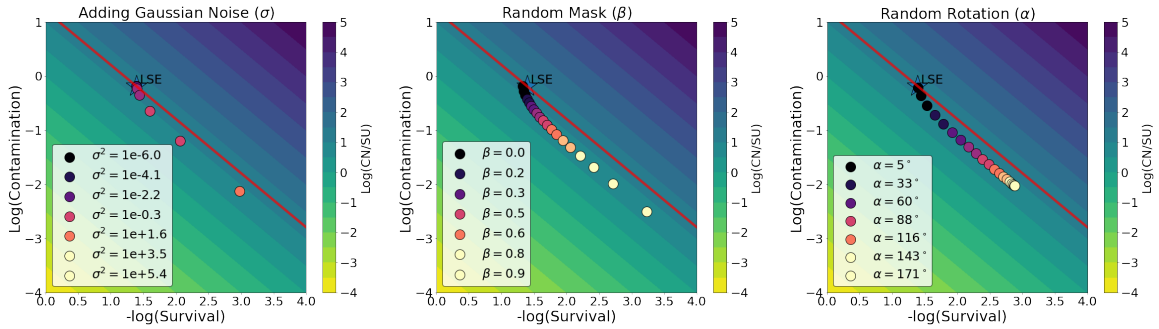
In this section, we compare the generalization of: 1) Gaussian noise injection (Bishop, 1995), 2) random mask (He et al., 2022), and 3) random rotation (which we introduced in Section 5.5). As in Section 6.1, we consider the random isotropic signal $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. 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 $\boldsymbol{\Sigma}$ and two choices of spectrum: 1. isotropic (i.e. $\boldsymbol{\Sigma} = \mathbf{I}_p$) and 2. decaying spectrum where $\Sigma_{ii} \propto \gamma^i$ with $\gamma = 0.95$.

Figure 4 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 4:

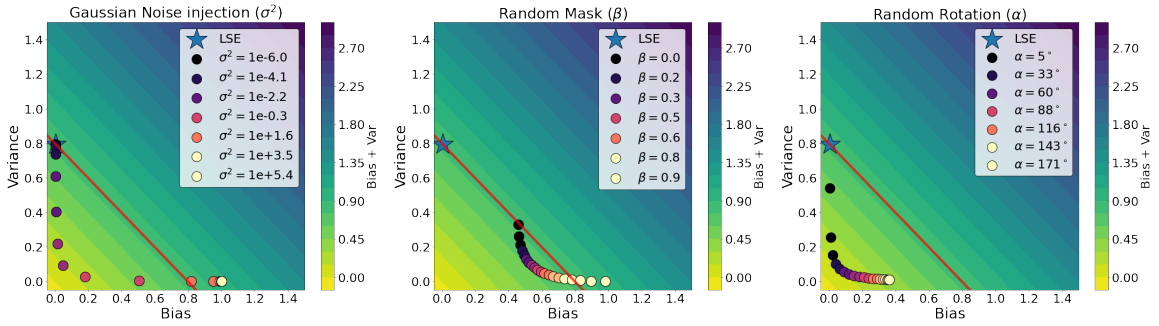
1. For isotropic data, all three augmentations achieve similar results in terms of generalization, while for the case of decaying spectrum, Gaussian injection and random rotation outperform random mask when their respective hyperparameters are optimally tuned.
2. For regression, Gaussian injection requires careful hyperparameter tuning in the range $[0, 1.8]$, while random mask and random rotation are fairly robust in performance in the entire tested hyperparameter range. 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 relatively robust generalization with respect to their hyperparameters. This verifies our theoretical observations in Propositions 46 and 47.
4. Our novel random rotation augmentation achieves the best of both worlds across different data distributions and tasks, achieving comparable generalization to noise injection when optimally tuned, while also being robust with respect to hyperparameter choice (like random mask). This observation is consistent with the theoretical prediction of Corollary 19.



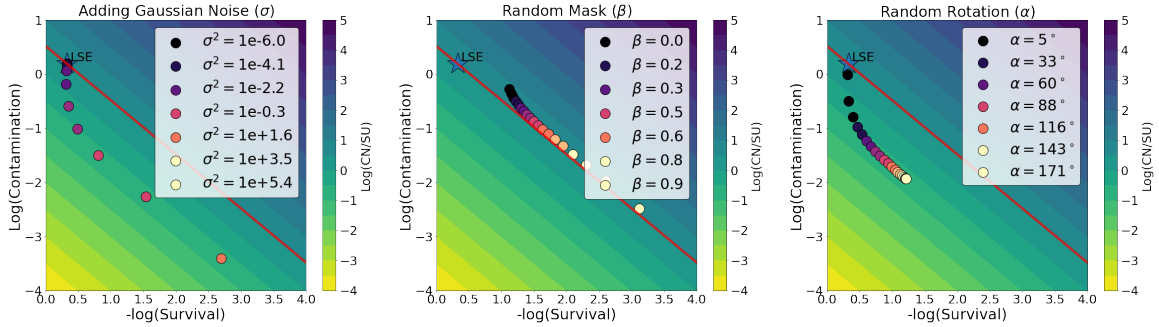
(a) Bias and variance distribution comparison in uniform covariate spectrum.



(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 for decaying covariate spectrum, $\gamma = 0.95$

Figure 4: *Visualizing the generalization error for different augmentations, across 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.

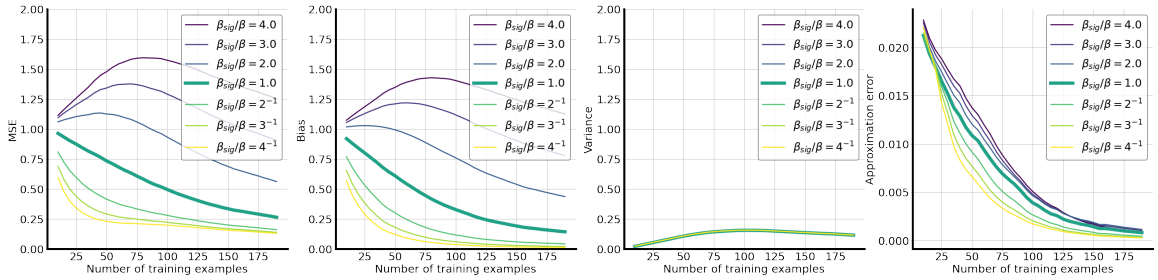


Figure 5: *Bias and variance decomposition for non-uniform random masking.* We vary the relative mask intensities (β_{sig}/β) across the signal and noise features. The result suggests that noise features can be augmented more heavily in comparison to the signal features.

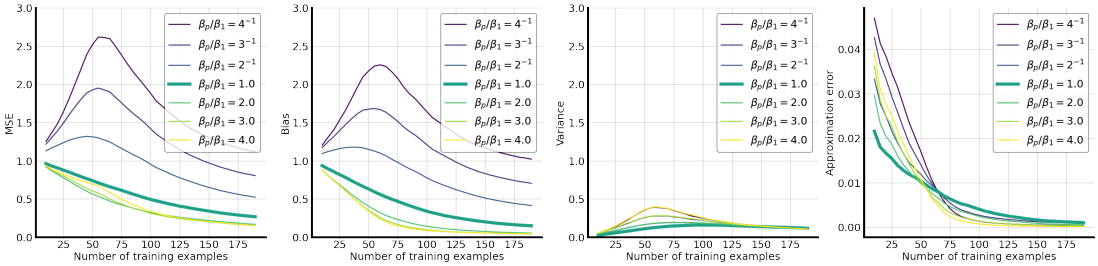


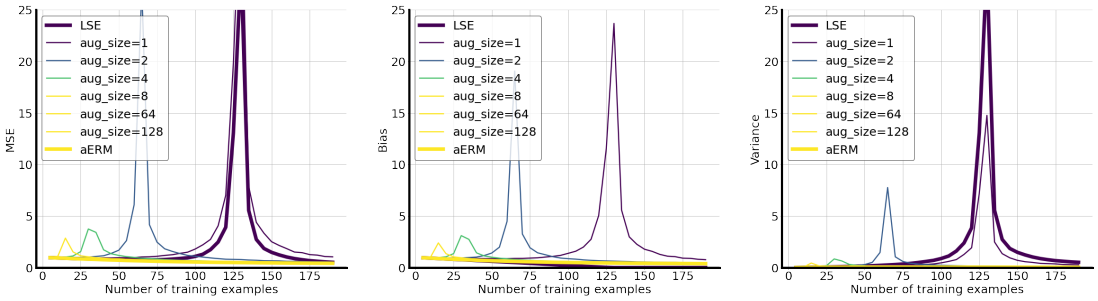
Figure 6: *Bias and variance decomposition of random masking for a bi-level spectrum.* We investigate the bi-level random mask strategies in data with decaying spectrum $\propto 0.95^i$. The first half of features are masked with probability β_1 while the rest are with β_p. We vary the ratio between the intensity β_p/β_1. We observe that augmenting more for features with higher variance benefits generalization.

6.3 Studying the interactions between the original covariance and augmentations

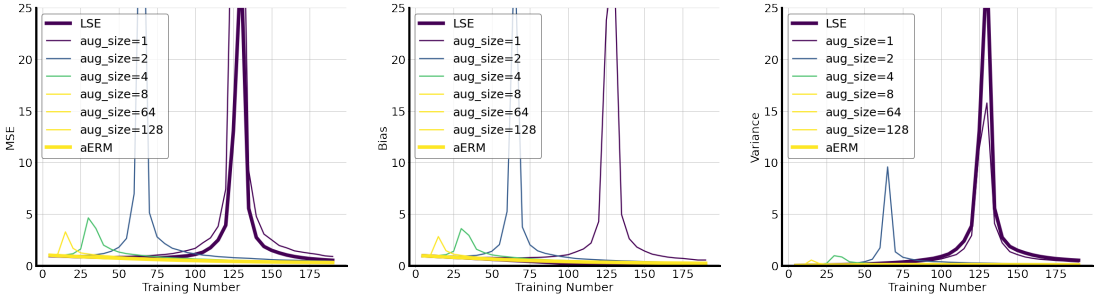
In this section, we try to understand the impact of the true model $\boldsymbol{\theta}^*$ and the data covariance $\boldsymbol{\Sigma}$ on the efficacy of different augmentations, focusing on the nonuniform random mask introduced in Section 7.1. We set the ambient dimension to $p = 128$ and consider the noise standard deviation $\sigma_\epsilon = 0.5$.

Effect of the true model: We study the impact of nonuniform masking on the 1-sparse model $\boldsymbol{\theta}^* = \mathbf{e}_1$, as depicted in Section 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. 5 and verify our analysis in Corollary 16 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 16, which depends only on the probability of the noise mask β.

Effect of the covariance spectrum: Next, to understand the impact of the covariance 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



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



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

Figure 7: *Pre-computed augmentations versus aERM.* 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.

Fig. 6. We observe from this figure that it is more beneficial to augment more for features with smaller eigenvalues.

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. 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 $\boldsymbol{\Sigma} = \mathbf{I}_{128}$. In Figs. 7 (a)-(b), we observe the well-known double descent peaks (Belkin et al., 2020; Nakkiran et al., 2020) 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.

7. The good, the bad and the ugly sides of data augmentation

In this section, we will dive into specific implications of our theory and experiments, and list key properties through which augmentations can be good (improve learning), bad (hurt learning), or "ugly" (have unexpected/surprising outcomes).

7.1 The good: when DA helps generalization

1) Data-adaptive spectral modification: Section 3.2 provided an interpretable and succinct characterization of the impact of any augmentation: namely, that it modifies the entire spectrum of the data covariance. We show that this modification to the covariance can be translated into modifications to the two *effective ranks*, as defined in Bartlett et al. (2020), and used to derive generalization bounds that reveal the impact of a given augmentation. This modification can itself be data-adaptive and lead to rich types of Tikhonov-style spectral regularization. Our analysis of several augmentations (including but not exclusive to the examples listed in Table 1) reveals that DA can have a much richer impact on the covariance spectrum, leading to unique benefits in generalization. For example, our theoretical and empirical analysis of the *non-uniform* random mask (Cor. 16) and random-rotation augmentation (Cor. 19) reveals that it is possible to generate data-adaptive regularizers through DA that reduce variance without a harmful increase in bias.

2) Variance reduction: Our analysis in Section 3.2 implies that any stochastic augmentation will reduce variance. Thus, in situations where the bias (of the original, unaugmented estimator) is already minimal or has a minimal impact on task performance, we expect many types of DA to be beneficial through this form of variance reduction. For example, Section 6.2 revealed that the random masking augmentation leads to stable improvements in classification performance for any choice of masking rate. We also see in Figure 4 that Gaussian noise injection, random mask and random rotation *all* improve overall generalization for both classification and moderately overparameterized regression ($p = 2n$), where bias is either minimal or does not significantly impact task performance. The improvement is maximal for the random rotation augmentation and Gaussian noise injection, both of which incur less bias than the random masking augmentation. An interesting result of our experiments is that our proposed random-rotation augmentation achieves a good generalization performance for a wide range of hyperparameters (rotation angles).

3) Reducing effective overparameterization and mitigating "double descent" behavior: The variance reduction effect is especially beneficial in overparameterized scenarios where d exceeds n but not by much, which is a scenario that often arises in ML practice. Here, the variance of the original estimator would be very large, leading to the "peak" observed in the double descent curve (Belkin et al., 2019). In these regimes, DA reduces the effective dimension of the data and effectively creates a synthetic underparameterized regime. As shown in Figure 7, this benefit takes place for both the pre-computed augmentation

implementation and the aERM implementation. These results support the perspective that augmentations can act as "virtual samples" (Balestriero et al., 2022b).

7.2 The bad: when DA hurts generalization

1) Augmentations can erase helpful data structure by "isotropizing" the spectrum An important ingredient for generalization of ridge estimators in high-dimensional settings is low-dimensional structure in the data. One key data structure that is often used to restrict the solution space, is to assume that the data have low-rank structure, or that the data covariance has a sufficiently fast rate of decay in its eigenvalues with the bulk of the signal energy contained in the top eigenvectors (Muthukumar et al., 2020; Tsigler and Bartlett, 2020; Muthukumar et al., 2021). Our theory reveals that some popular augmentations, such as random-masking and group-invariant augmentations, begin to *erase* all of this helpful data structure by *isotropizing* the "equivalent" data covariance in expectation. Putting this in the language of Section 7.2.1, such isotropization has the benefit of variance reduction but this comes at a cost of increased bias. While we believe this isotropization effect might be specific to high-dimensional linear models and may not occur even for nonlinear kernel methods, it is an important factor that our theory identifies as, overall, pessimistic for generalization.

2) Increase in bias could offset variance reduction: Our theory demonstrates that augmentations can have the undesired effect of increasing the bias of the aERM estimator. This increase in bias is particularly acute when the data are high-dimensional (Hastie et al., 2019; Muthukumar et al., 2020). Figure 4 illustrates that in this regime, certain augmentations like Gaussian noise injection and random mask can lead to poor performance due to high bias. Moreover, since bias increases with increased augmentation intensity, these augmentations can even hurt performance if applied too strongly!

Another class of augmentations for which we show counterintuitive effects is the class of *group-invariant augmentations*, i.e. augmentations that are created with the ostensible aim of inducing invariance in prediction within a specific algebraic group (for example, for the rotation or translation group, augmentations would consist of rotations or translations of images). In previous work Chen et al. (2020a), the authors show that such group-invariant augmentations always improves generalization through variance reduction; however, they primarily considered the underparameterized regime where bias is much less significant. We show in Appendix C.4, such augmentations may generalize poorly in the overparameterized regime.

3) Augmentations can induce distribution shift between training and test data Finally, we show that augmentations that are *biased-on-average*, meaning that $\mathbb{E}[g(\mathbf{x})] \neq \mathbf{x}$, can induce an undesirable distribution shift between training and test data that is harmful, particularly for regression tasks. For example, comparing the regression error bounds for the unbiased variant of randomized mask (Corollary 15) and its biased variant (Corollary 34), reveals that the biased variant incurs an additional penalty due to distribution shift. On the other hand, as predicted in Corollary 11, the impact of bias in augmentation on classification tasks might not be as pronounced. Thus, our results highlight the importance of debiasing augmentations when applied to regression tasks.

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

The previous two subsections highlight ways in which augmentations can both help and harm learning. Now we will discuss a few ways that augmentations can give rise to what we call "ugly" behavior, impacting performance in curious and unexpected ways.

1) Differences in under and overparameterized settings Our results also highlight ways in which augmentations will impact models differently in the under vs. overparameterized regimes. Corollary 15 shows that when applying random masking for regression tasks, 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 (recall that p is the data dimension and n is the number of training examples). From this, we can draw the following insights: 1. the variance is vanishing in both regimes, 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 $p \gg n$. Said another way, the isotropization effect described in Section 7.2.2 can be beneficial in the underparameterized regime, as the contribution of bias is relatively minimal, but harmful in the overparameterized regime where the contribution of bias can be substantial. This supports the benefits of group-invariant augmentations shown by Chen et al. (2020a) in the underparameterized regime.

2) Differences between augmentations that are precomputed or generated on-the-fly. Our experiments also demonstrated interesting subtleties between precomputing and on-the-fly-generated augmentations (Fig. 7). While a small number of pre-computed augmentations induces a similar double-descent behavior to the original LSE, the aERM error gracefully decreases without any interpolation peak. We also note that the double descent MSE peak shifts for different numbers of pre-computed samples, appearing approximately at the location $n = \frac{p}{k}$, where p and k denote the data augmentation and number of pre-computed aumgentations per sample, respectively. The complete mitigation of double descent by aERM can be explained by our theory and is directly connected to the beneficial effect of variance reduction in mitigating double descent (previously observed for ridge regularization (Hastie et al., 2019)). On the other hand, we believe pre-computed augmentation has a different effect of adding "virtual samples", therefore leading to the observed effect of shifting the effective interpolation threshold to the right. Proving this rigorously will require novel random matrix theory tools due to the synthetic but correlated nature of the virtual samples. We defer mathematical analysis of this phenomenon to future work.

3) The effect of weak DA Our analysis framework also allows us to understand a counterintuitive phenomenon that emerges for "weak" augmentations. 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 aERM estimator when the augmentation intensity ξ approaches zero is given by

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

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_∞. 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 $\text{Cov}_\infty = n^{-1}\text{diag}(\mathbf{X}^T\mathbf{X}) \approx \mathbf{\Sigma}$. Our empirical results in Figure 8 confirm this effect as well, where we find a gap between the LSE and even very weak augmentations. We discuss this effect further in Appendix G.1.

8. Conclusions and Future Work

In this paper, we established a new framework to analyze the generalization error for linear models with data augmentation in underparameterized and overparameterized regimes. We characterized 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.

While we do not formally study nonlinear models in this paper, we believe our analysis provides powerful tools that we could build on to handle the nonlinear case in future work. Our approach extends most naturally to the case of *kernel methods* (Schölkopf and Smola (2002)), *random features or last-layer retuning* (Mei et al. (2021)), and the *neural tangent kernel regime* (Jacot et al. (2018)). In these cases, the primary technical challenge is understanding the effect of the augmentation covariance $\text{Cov}_{\mathcal{G}}(X)$, which can be very different than in our analysis, as the feature map in kernel methods is typically nonlinear in the data. Nevertheless, we believe our generalization analysis can be applied in a plug-and-play manner with such covariance calculations, by combining the insights of our work with tools established, e.g., in McRae et al. (2022).

While our current analysis focuses on the effect of augmentations on supervised learning, understanding how augmentation impacts self-supervised and contrastive learning is an important area for future work. In these approaches, the choice of augmentations can have even more harmful effects on learning and in some cases, cause representational collapse Cabannes et al. (2023). Thus, we hope that our results on the implicit spectral manipulation induced by DA can also be applied to study SSL in the future.

Acknowledgements

We would like to thank Mehdi Azabou and Max Dabagia for feedback on the work at various stages and many helpful discussions. This work was funded through NSF IIS-2212182, NSF IIS-2039741, a NSF Graduate Research Fellowship (DGE-2039655), NIH 1R01EB029852, and the support from the Canadian Institute for Advanced Research (CIFAR) through the Global Scholars Program (ELD).

Appendix

Contents

Appendix A. 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 20 (A useful identity for the ridge estimator (Tsigler and Bartlett, 2020)) *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:*

$$(\mathbf{I}_k + \mathbf{V}^\top\mathbf{X}^\top\mathbf{P}_k^{-1}\mathbf{X}\mathbf{V})\mathbf{V}^\top\hat{\theta} = \mathbf{V}^\top\mathbf{X}^\top\mathbf{P}_k^{-1}\mathbf{y}, \tag{16}$$

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 21 (Bernstein-type inequality for sum of sub-exponential variables) *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent zero-mean sub-exponential random variables with sub-exponential norm at most σ_x^2. Then for every $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ 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 22 (Concentration of regularized truncated empirical covariance, Lemma 21 in Tsigler and Bartlett (2020)) *Suppose $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p] \in \mathbb{R}^{n\times p}$ is a matrix with independent isotropic sub-gaussian rows with norm σ. Consider $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ 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 $1 - \delta - 2\exp(-ct)$

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

Lemma 23 (Concentration of leave-one-out empirical covariance) *Under the same notations and assumptions in Lemma 22, 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 22 and the observations of $\mu_1(\mathbf{A}_{-t}) \leq \mu_1(\mathbf{A}_0))$ and $\mathbf{A}_{-t} \succeq \mathbf{A}_k$. ■

Lemma 24 (**Concentration of matrix with independent sub-gaussian rows, Theorem 5.39 in Vershynin (2010)**) *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 25 (**Concentration of the sum of squared norms, Lemma 17 in Tsigler and Bartlett (2020)**) *Suppose* $\mathbf{Z} \in \mathbb{R}^{n \times p}$ *is a matrix with independent isotropic sub-gaussian rows with norm* σ*. Consider* $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ *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 26 (**Applications of Hanson-Wright inequality as done in Muthukumar et al. (2021)**) *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*

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

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

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

Lemma 27 (**Operator norm bound of matrix with sub-gaussian rows (Tsigler and Bartlett, 2020)**) *Suppose* $\{\mathbf{z}_i\}_{i=1}^n$ *is a sequence of independent sub-gaussian vectors in* $\mathbb{R}^p$ *with* $\|\mathbf{z}_i\| \leq \sigma$*. Consider* $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ *for some positive non-decreasing sequence* $\{\lambda_i\}_{i=1}^p$*. Denote* $\mathbf{X}$ *to be the matrix with rows* $\mathbf{\Sigma}^{1/2}\mathbf{z}_i$*. Then for some absolute constant* c*, for any* $t > 0$ *with probability at least* $1 - 4e^{-t/c}$

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

Appendix B. Proofs of Regression Results

In this section, we will include essential lemmas in B.1 to prove the main theorems for regression analysis in the sections B.2 and B.3. Then, we will use these theorems to prove the propositions and corollaries in sections B.4 and B.7, respectively.

B.1 Regression Lemmas

Lemma 28 (**Sharpened bias of ridge regression, extension of Tsigler and Bartlett (2020)**)

$$\frac{\text{Bias}}{C_x L_1^4} \lesssim \left\|\mathbf{P}^{\mathbf{\Sigma}}_{k_1+1:p}\theta^*\right\|^2_{\mathbf{\Sigma}} + \left\|\mathbf{P}^{\mathbf{\Sigma}}_{1:k_1}\theta^*\right\|^2_{\mathbf{\Sigma}^{-1}} \frac{\rho^2_{k_1}(\mathbf{\Sigma};n)}{(\lambda_{k_1+1})^{-2}+(\lambda_1)^{-2}\rho^2_{k_1}(\mathbf{\Sigma};n)} \tag{17}$$

Remark 29 *The reason we modify the bound from Tsigler and Bartlett (2020) is twofold: 1. We consider non-diagonal covariance matrix $\mathbf{\Sigma}$. 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 Tsigler and Bartlett (2020), 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 Tsigler and Bartlett (2020), 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 $\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$. 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 20 with $\mathbf{V} = \mathbf{V}_1$ to obtain

$$(\mathbf{I}_k + \mathbf{V}_1^\top\mathbf{X}^\top\mathcal{A}_k(\mathbf{\Sigma};\lambda)^{-1}\mathbf{X}\mathbf{V}_1)\mathbf{V}_1^\top\hat{\theta} = \mathbf{V}_1^\top\mathbf{X}^\top\mathcal{A}_k(\mathbf{\Sigma};\lambda)^{-1}\mathbf{y}, \tag{18}$$

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

$$\text{Bias} \le 2\|\mathbf{V}_1\mathbf{V}_1^\top(\mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*)\|^2_{\mathbf{\Sigma}} + 2\|\mathbf{V}_2\mathbf{V}_2^\top(\mathbb{E}_\varepsilon[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^*)\|^2_{\mathbf{\Sigma}}, \tag{19}$$

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 20 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

$$\begin{aligned}&= (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{-1/2} \mathbf{X}^\top \mathbf{A}_k^{-1} \mathbf{X} \mathbf{V}_2 \mathbf{V}_2^\top \boldsymbol{\theta}^* - (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\theta}^* \\ &\leq \|\mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}} \mu_n(\mathbf{A}_k)^{-1} \sqrt{\mu_1 \left(\mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{X}^\top \mathbf{X} \boldsymbol{\Sigma}^{-1/2} \mathbf{V}_1 \mathbf{V}_1^\top \right)} \|\mathbf{X} \mathbf{V}_2 \mathbf{V}_2^\top \boldsymbol{\theta}^*\| \\ &\quad + \|\mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}} \|\mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}^{-1}}. \qquad (20)\end{aligned}$$

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

$$\geq \lambda_1^{-1} \|\mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 + (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{V}_1 \mathbf{V}_1^\top \mathbf{X}^\top \mathbf{A}_k^{-1} \mathbf{X} \mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \qquad (21)$$

in which the second term is

$$\begin{aligned}&= (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{X}^\top \mathbf{A}_k^{-1} \mathbf{X} \boldsymbol{\Sigma}^{-1/2} \mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{V}_1 \mathbf{V}_1^\top (\mathbf{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &\geq \|\mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \|\mathbf{V}_1 \mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{X}^\top \mathbf{A}_k^{-1} \mathbf{X} \boldsymbol{\Sigma}^{-1/2} \mathbf{V}_1 \mathbf{V}_1^\top\| \\ &\geq \|\mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \mu_k(\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{X}^\top \mathbf{A}_k^{-1} \mathbf{X} \boldsymbol{\Sigma}^{-1/2} \mathbf{V}_1) \\ &\geq \|\mathbf{V}_1 \mathbf{V}_1^\top (\mathbb{E}_\varepsilon \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2 \mu_1(\mathbf{A}_k)^{-1} \mu_k(\mathbf{V}_1^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{X}^\top \mathbf{X} \boldsymbol{\Sigma}^{-1/2} \mathbf{V}_1). \qquad (22)\end{aligned}$$

Therefore, combining e.q. (20), (21) and (22), 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 Tsigler and Bartlett (2020) except we use projection matrices to accommodate for the non-diagonal

covariance:

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

T_2 is bounded by

$$\mu_n^{-2}(\mathbf{A}_k)\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2. \tag{23}$$

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,

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

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

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

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

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

where for the diagonal covariance $\boldsymbol{\Sigma}$, the first two terms are sharpened with additional λ_1^{-2} in the denominators as compared to Tsigler and Bartlett (2020). As in Tsigler and Bartlett (2020), 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 24, $\mu_j(\mathbf{A}_k)$ by Lemma 22, $\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2\|_2^2$ and $\|\mathbf{X}\mathbf{V}_2\mathbf{V}_2^\top\boldsymbol{\Sigma}\mathbf{V}_2\mathbf{V}_2^\top\mathbf{X}^\top\|$ by Lemma 25. The details can be found in the proof of MSE bound of Tsigler and Bartlett (2020). ■

Lemma 30 (**Variance bound of ridge regression for non-diagonal covariance data (Tsigler and Bartlett, 2020)**) *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* $\lambda_1 \geq \lambda_2 \ldots \lambda_p$. *Given a ridge estimator* $\hat{\theta} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ *and* $\lambda \geq 0$, *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* $1-\delta-n^{-1}$,

$$\frac{\text{Variance}}{\sigma_\varepsilon^2 L_2^2 \tilde{C}_x} \lesssim \left(\frac{k_2}{n} + \frac{n}{R_{k_2}(\boldsymbol{\Sigma};n)}\right)\log n. \tag{24}$$

Lemma 31 (**Generalization bound of ridge regression for non-diagonal covariance data, extension of Tsigler and Bartlett (2020)**) *Consider the regression task with the model setting in Section 3 where the input variable* $\mathbf{x}$ *has possibly non-diagonal covariance* $\boldsymbol{\Sigma}$ *with eigenvalues* $\lambda_1 \geq \lambda_2 \ldots$. *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* $1-n^{-1}$,

$$\begin{aligned}\text{MSE} \lesssim & \underbrace{C_x L_1^4\left(\left\|\mathbf{P}_{k_1+1:p}^{\boldsymbol{\Sigma}}\theta^*\right\|_{\boldsymbol{\Sigma}}^2 + \left\|\mathbf{P}_{1:k_1}^{\boldsymbol{\Sigma}}\theta^*\right\|_{\boldsymbol{\Sigma}^{-1}}^2 \frac{\rho_{k_1}^2(\boldsymbol{\Sigma};n)}{(\lambda_{k_1+1})^{-2}+(\lambda_1)^{-2}\rho_{k_1}^2(\boldsymbol{\Sigma};n)}\right)}_{\text{Bias}} \\ & + \underbrace{\sigma_\varepsilon^2 L_2^2\tilde{C}_x\left(\frac{k_2}{n}+\frac{n}{R_{k_2}(\boldsymbol{\Sigma};n)}\right)\log n}_{\text{Variance}}\end{aligned} \tag{25}$$

Proof The statement is a direct combination of Lemma 28, 30 and the bias-variance decomposition of MSE from Tsigler and Bartlett (2020). ■

Lemma 32 (Bounds on the approximation error for regression) *Denote*

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

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

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

Then the approximation error is bounded by,

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

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

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\boldsymbol{\Sigma}} &= \|(\mathbf{X}^\top\mathbf{X} + n\mathbf{D})^{-1}\mathbf{X}^\top\mathbf{y} - (\mathbf{X}^\top\mathbf{X} + n\bar{\mathbf{D}})^{-1}\mathbf{X}^\top\mathbf{y}\|_{\boldsymbol{\Sigma}} \\
&= \|(\mathbf{X}^\top\mathbf{X} + n\mathbf{D})^{-1}(\mathbf{X}^\top\mathbf{X} + n\bar{\mathbf{D}} - \mathbf{X}^\top\mathbf{X} - n\mathbf{D})(\mathbf{X}^\top\mathbf{X} + n\bar{\mathbf{D}})^{-1}\mathbf{X}^\top\mathbf{y}\|_{\boldsymbol{\Sigma}} \\
&= n\|\boldsymbol{\Sigma}^{\frac{1}{2}}\bar{\mathbf{D}}^{-\frac{1}{2}}\bar{\mathbf{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathbf{D})^{-1}\bar{\mathbf{D}}^{\frac{1}{2}}\boldsymbol{\Delta}\bar{\mathbf{D}}^{\frac{1}{2}}\bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_2, \\
&\lesssim n\|\boldsymbol{\Sigma}^{\frac{1}{2}}\bar{\mathbf{D}}^{-\frac{1}{2}}\|\|\bar{\mathbf{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathbf{D})^{-1}\bar{\mathbf{D}}^{\frac{1}{2}}\|\|\boldsymbol{\Delta}\|\|\bar{\mathbf{D}}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|\|\bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_2 \\
&\lesssim n\kappa^{\frac{1}{2}}\Delta_G\|\bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\boldsymbol{\Sigma}}\|\bar{\mathbf{D}}^{\frac{1}{2}}(\mathbf{X}^\top\mathbf{X} + n\mathbf{D})^{-1}\bar{\mathbf{D}}^{\frac{1}{2}}\| \qquad (26)
\end{aligned}$$

By (30), $\|\bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\boldsymbol{\Sigma}}$ can be bounded as,

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

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

Now, observe

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

However,

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

where $\tilde{\mathbf{X}}$ has sub-gaussian rows with covariance $\mathbf{\Sigma}_{\text{aug}}$. 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,

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

and finally we have,

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

■

B.2 Proof of Theorem 4

Theorem 4 (High probability bound for MSE with unbiased DA) *Consider an unbiased data augmentation g and its corresponding estimator $\hat{\boldsymbol{\theta}}_{aug}$, where Δ_G is defined in Eq. 7 and κ is the condition number of $\mathbf{\Sigma}_{aug}$. Assume for some integers k_1, k_2, the condition numbers for the matrices $\mathcal{A}_{k_1}(\mathbf{X}_{aug}; n)$, $\mathcal{A}_{k_2}(\mathbf{X}_{aug}; n)$ (defined in Section 1.2) 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 , with probability $1 - \delta' - 4n^{-1}$, the test mean-squared error is bounded by*

$$\begin{aligned}
\text{MSE} &\lesssim \text{Bias} + \text{Variance} + \text{ApproximationError}, \qquad (10)\\
\frac{\text{Bias}}{L_1^4} &\lesssim \left(\left\|\mathbf{P}^{\mathbf{\Sigma}_{aug}}_{k_1+1:p}\theta^*_{aug}\right\|^2_{\mathbf{\Sigma}_{aug}} + \left\|\mathbf{P}^{\mathbf{\Sigma}_{aug}}_{1:k_1}\theta^*_{aug}\right\|^2_{\mathbf{\Sigma}^{-1}_{aug}} \frac{(\rho^{aug}_{k_1})^2}{(\lambda^{aug}_{k_1+1})^{-2} + (\lambda^{aug}_1)^{-2}(\rho^{aug}_{k_1})^2}\right),\\
\frac{\text{Variance}}{L_2^2} &\lesssim \left(\frac{k_2}{n} + \frac{n}{R^{aug}_k}\right)\log n, \quad \text{Approx.Error} \lesssim \kappa^{\frac{1}{2}}\Delta_G\left(\|\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}} + \sqrt{\text{Bias} + \text{Variance}}\right).
\end{aligned}$$

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

Proof

$$\text{MSE} = \mathbb{E}_{\mathbf{x}}[(\mathbf{x}^\top(\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*))^2|\mathbf{X}, \varepsilon] = \|\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|^2_{\mathbf{\Sigma}}. \qquad (27)$$

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

$$\text{MSE} \le 2\|\bar{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|^2_{\mathbf{\Sigma}} + 2\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|^2_{\mathbf{\Sigma}} \qquad (28)$$

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

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}_{\text{aug}} = (\mathbf{X}^\top\mathbf{X} + n\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1}\mathbf{X}^\top\mathbf{y}\\
&= \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}(n\mathbf{I}_p + \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\mathbf{X}^\top\mathbf{X}\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2})^{-1}\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\mathbf{X}^\top\mathbf{y}\\
&= \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}(n\mathbf{I}_p + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{y} \qquad (\tilde{\mathbf{X}} := \mathbf{X}\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2})\\
&= \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\hat{\theta}_{\text{ridge}}, \qquad (\hat{\boldsymbol{\theta}}_{\text{ridge}} := (n\mathbf{I}_p + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{y}). \qquad (29)
\end{aligned}$$

So the MSE becomes $\|\hat{\boldsymbol{\theta}}_{\text{ridge}} - \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{1/2}\boldsymbol{\theta}^*\|^2_{\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}}$. These observations have shown an approximate equivalence to a ridge estimator with data matrix $\tilde{\mathbf{X}}$, which has data covariance $= \mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{-1/2}$, ridge intensity $\lambda = n$, and true model parameter $\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})]^{1/2}\boldsymbol{\theta}^*$. Hence, we can apply Lemma 31 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,

$$\|\mathbb{E}_{\varepsilon}[\bar{\boldsymbol{\theta}}_{\text{aug}}] - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}} \lesssim$$
$$C_x L_1^4 \left(\left\| \mathbf{P}^{\boldsymbol{\Sigma}_{\text{aug}}}_{k_1+1:p}\theta^*_{\text{aug}} \right\|^2_{\boldsymbol{\Sigma}_{\text{aug}}} + \left\| \mathbf{P}^{\boldsymbol{\Sigma}_{\text{aug}}}_{1:k_1}\theta^*_{\text{aug}} \right\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{aug}}} \frac{\rho^2_{k_1}(\boldsymbol{\Sigma}_{\text{aug}}; n)}{(\lambda^{\text{aug}}_{k_1+1})^{-2} + (\lambda^{\text{aug}}_1)^{-2}\rho^2_{k_1}(\boldsymbol{\Sigma}_{\text{aug}}; n)} \right), \quad (30)$$
$$\|\mathbb{E}_{\varepsilon}[\bar{\boldsymbol{\theta}}_{\text{aug}}] - \bar{\boldsymbol{\theta}}_{\text{aug}}\|^2_{\boldsymbol{\Sigma}} \lesssim \sigma^2_{\varepsilon} t L_2^2 \tilde{C}_x \left(\frac{k_2}{n} + \frac{n}{R_{k_2}(\boldsymbol{\Sigma}_{\text{aug}}; n)} \right). \quad (31)$$

For the second error term $\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \bar{\boldsymbol{\theta}}_{\text{aug}}\|^2_{\boldsymbol{\Sigma}}$, we apply Lemma 32.

■

B.3 Proof of Theorem 7

Theorem 7 (Bounds on the MSE for Biased Augmentations) *Consider the estimator $\hat{\boldsymbol{\theta}}_{aug}$ obtained by solving the aERM in (1). Let $\text{MSE}^o(\hat{\boldsymbol{\theta}}_{aug})$ denote the unbiased MSE bound in Eq. (10) of Theorem 4, and Δ_G defined in Eq. 7. 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,*

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

where

$$R_1 = 1 + \|\boldsymbol{\Sigma}^{\frac{1}{2}}\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\| \text{ and}$$
$$R_2 = \sqrt{\|\bar{\boldsymbol{\Sigma}}(\mathbb{E}_{\mathbf{x}}[\text{Cov}_{\mathcal{G}}(\mathbf{x})])^{-1}\|} \left(1 + \frac{\Delta_G}{1-c}\right) \left(\sqrt{\Delta_\xi}\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\text{Cov}_\xi} \right)$$
$$\times \left(\sqrt{\frac{1}{\lambda^{aug}_k}} + \sqrt{\frac{\lambda^{aug}_{k+1}(1+\rho^{aug}_k)}{(\lambda^{aug}_1\rho^{aug}_0)^2}} \right).$$

Proof

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

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

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

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

$$L_1 + L_2 \leq (1 + \|\boldsymbol{\Sigma}^{\frac{1}{2}} \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \mathbf{I}_p\|) \|\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}}. \tag{32}$$

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,

$$\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}} \lesssim \underbrace{\|\tilde{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}}}_{L_3} + \underbrace{\|\hat{\boldsymbol{\theta}}_{\text{aug}} - \tilde{\boldsymbol{\theta}}_{\text{aug}}\|_{\bar{\boldsymbol{\Sigma}}}}_{L_4}. \tag{33}$$

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:

$$\|\tilde{\boldsymbol{\theta}}_{\text{aug}} - \boldsymbol{\theta}^*\|_{\bar{\boldsymbol{\Sigma}}} \leq \sqrt{\text{MSE}^o}, \tag{34}$$

where MSE^o is the bound in E.q. (10). 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,

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

Observe that

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

where $\tilde{\mathbf{X}}$ has sub-gaussian rows with covariance $\mathbf{\Sigma}_{\text{aug}}$ as defined in E.q. (9).

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

$$\begin{aligned} &\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top + n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{1:k}\| = \|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k}(\mathbf{I}_p + \tilde{\mathbf{X}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k})^{-1}\| \\ \leq\ & \mu_n(\mathbf{A}_k)^{-1}\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\|\|\tilde{\mathbf{X}}_{1:k}(\mathbf{I}_p + \tilde{\mathbf{X}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{X}}_{1:k})^{-1}\| \\ =\ & \mu_n(\mathbf{A}_k)^{-1}\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\|\|\tilde{\mathbf{Z}}_{1:k}(\mathbf{\Sigma}_{\text{aug},1:\text{k}}^{-1} + \tilde{\mathbf{Z}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{Z}}_{1:k})^{-1}\mathbf{\Sigma}_{\text{aug, 1:k}}^{-\frac{1}{2}}\| \\ \leq\ & \mu_n(\mathbf{A}_k)^{-1}\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}\|\|\mathbf{\Sigma}_{\text{aug},1:\text{k}}^{-\frac{1}{2}}\|\|\tilde{\mathbf{Z}}_{1:k}(\mathbf{\Sigma}_{\text{aug},1:\text{k}}^{-1} + \tilde{\mathbf{Z}}_{1:k}^\top\mathbf{A}_k^{-1}\tilde{\mathbf{Z}}_{1:k})^{-1}\|, \end{aligned} \tag{35}$$

where $\tilde{\mathbf{Z}}$ has sub-gaussian rows with isotropic covariance $\mathbf{I}_p$. Now applying Lemma 24, we have, with probability $1 - 5n^{-3}$,

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

Combining the above and E.q. (35) with Lemma 22, we have with probability $1 - \delta - 2n^{-3}$ that

$$L_9 = \|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\bar{\mathbf{C}}^{-\frac{1}{2}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top + n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}_{1:k}\| \lesssim \sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{\lambda_k^{\text{aug}}n}}, \tag{36}$$

where λ_k^{aug} is the k-th eigenvalue of $\mathbf{\Sigma}_{\text{aug}}$. On the other hand, by Lemma 22 and 27,

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

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

$$L_7 \leq L_9 + L_{10} \lesssim \sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{n}}\left(\sqrt{\frac{1}{\lambda_k^{\text{aug}}}} + \sqrt{\frac{\lambda_{k+1}^{\text{aug}}(1+\rho_k(\mathbf{\Sigma}_{\text{aug}};n))}{(\lambda_1^{\text{aug}}\rho_0(\mathbf{\Sigma}_{\text{aug}};n))^2}}\right). \tag{37}$$

Next, we bound L_8:

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

The term L_{11} is identical to (37) 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 22

$$\mu_p\left(\left(\bar{\mathbf{C}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\right)^{-1}\right)=\mu_p\left(\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}+n\mathbf{I}_p\right)\right)\geq n \tag{38}$$

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

$$\begin{aligned}
&\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\mu(\mathbf{X})(\mu(\mathbf{X})^\top\mu(\mathbf{X})+\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}-\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\mu(\mathbf{X})(\mu(\mathbf{X})^\top\mu(\mathbf{X})+n\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}^{\frac{1}{2}}\|\\
\lesssim\;&\frac{\Delta_G}{1-\Delta_G}\sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{n}}\left(\sqrt{\frac{1}{\lambda_k^{\mathrm{aug}}}}+\sqrt{\frac{\lambda_{k+1}^{\mathrm{aug}}(1+\rho_k(\mathbf{\Sigma}_{\mathrm{aug}};n))}{(\lambda_1^{\mathrm{aug}}\rho_0(\mathbf{\Sigma}_{\mathrm{aug}};n))^2}}\right),
\end{aligned} \tag{39}$$

by our assumption that $\Delta_G\leq c$ for some $c<1$. E.q. (37) and (39) now imply

$$\begin{aligned}
L_5=\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}(\mu(\mathbf{X})^\top\mu(\mathbf{X})+\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}))^{-1}\mu(\mathbf{X})^\top\|&\leq L_7+L_8\\
\lesssim\sqrt{\frac{\|\bar{\mathbf{\Sigma}}\bar{\mathbf{C}}^{-1}\|}{n}}\left(\sqrt{\frac{1}{\lambda_k^{\mathrm{aug}}}}+\sqrt{\frac{\lambda_{k+1}^{\mathrm{aug}}(1+\rho_k(\mathbf{\Sigma}_{\mathrm{aug}};n))}{(\lambda_1^{\mathrm{aug}}\rho_0(\mathbf{\Sigma}_{\mathrm{aug}};n))^2}}\right)\cdot\left(1+\frac{\Delta_G}{1-c}\right).&
\end{aligned} \tag{40}$$

On the other hand,

$$\begin{aligned}
L_6=\|\mathbf{y}-\tilde{\mathbf{y}}\|\;&=\|(\mu(\mathbf{X})-\mathbf{X})\boldsymbol{\theta}^*\|=\sqrt{n}\|\boldsymbol{\theta}^*\|_{n^{-1}(\mu(\mathbf{X})-\mathbf{X})(\mu(\mathbf{X})-\mathbf{X})^\top}\\
&\leq\sqrt{n}\left(\|\boldsymbol{\theta}^*\|\sqrt{\|n^{-1}(\mu(\mathbf{X})-\mathbf{X})(\mu(\mathbf{X})-\mathbf{X})^\top-\mathrm{Cov}_\xi\|}+\|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right)\\
&\leq\sqrt{n}\left(\sqrt{\Delta_\xi}\|\boldsymbol{\theta}^*\|+\|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right),
\end{aligned} \tag{41}$$

where Cov_ξ is defined in Definition 6.

Combining E.q. (40) and (41), we obtain the following:

$$L_4 = \|\hat{\boldsymbol{\theta}}_{\text{aug}} - \tilde{\boldsymbol{\theta}}_{\text{aug}}\|_{\bar{\boldsymbol{\Sigma}}} = L_5 \cdot L_6 \lesssim \sqrt{\|\bar{\boldsymbol{\Sigma}}\bar{\mathbf{C}}^{-1}\|}\left(1 + \frac{\Delta_G}{1-c}\right)\left(\sqrt{\Delta_\xi}\|\boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|_{\mathrm{Cov}_\xi}\right) \cdot \left(\sqrt{\frac{1}{\lambda_k^{\text{aug}}}} + \sqrt{\frac{\lambda_{k+1}^{\text{aug}}(1+\rho_k(\boldsymbol{\Sigma}_{\text{aug}};n))}{(\lambda_1^{\text{aug}}\rho_0(\boldsymbol{\Sigma}_{\text{aug}};n))^2}}\right) \tag{42}$$

Finally, putting together the results of Eq. (32), (33), (34) and (42) completes the proof. ■

B.4 Proof of Proposition 12

Proposition 12 (Uncorrelated Feature Augmentations) *Let the augmentation g be composed of p uncorrelated feature augmentation maps, i.e., $g(\mathbf{x}) = \begin{bmatrix} g_1(x_1) & \dots & g_d(x_d) \end{bmatrix}$ where $\{g_i(\cdot)\}_{i\in[p]}$ are uncorrelated (with respect to the randomness in the augmentation). If the variance of each feature augmentation $\mathrm{Var}_{g_i}(g_i(x_i))$ (which is a random variable due to the randomness in x_i) is sub-exponential with sub-exponential norm σ_i^2 and mean $\bar{\sigma}_i^2$ for all $i \in [p]$, then we have*

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

with probability at least $1 - \frac{1}{n}$.

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

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

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

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 $\mathbf{D} = n^{-1}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^{-\frac{1}{2}}$, we have $\mathbf{Q} = \mathbf{D} - \mathbf{I}$. For any $i \in \{1, 2, \ldots, p\}$, $\mathbf{D}_{ii} = n^{-1}\sum_{j=1}^{n}\frac{\mathrm{Var}_{g_i}(\mathbf{x}_{ji})}{\mathbb{E}_{\mathbf{x}}[\mathrm{Var}_{g_i}(\mathbf{x})]}$, 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 21 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,

$$\mathbf{Q}_{ii} = \frac{1}{\bar{\sigma}_i{}^2}\left(n^{-1}\sum_{j=1}^{n}\mathrm{Var}_{g_i}(\mathbf{x}_j) - \mathbb{E}_{\mathbf{x}}[\mathrm{Var}_{g_i}(\mathbf{x})]\right) \leq c'\frac{\sigma_i^2}{\bar{\sigma}_i^2}\sqrt{\frac{\log n}{n}}. \tag{43}$$

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.

■

B.5 Proof of Proposition 13

Proposition 13 *Consider a correlated-feature augmentation of the form described above. Further, assume that the smallest eigenvalue of $\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})$ is lower bounded by σ for every k, and g_k is component-wise bounded, i.e., $\|g_k(\mathbf{x}_k)\|_\infty \leq M$ for any k. Then, we have*

$$\Delta_G \lesssim \frac{M^2\max_k|B_k|}{\sigma}\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - \frac{1}{p}$.

Proof We begin by bounding Δ_{G_k}, which is the component of Δ_G corresponding to the kth block of $\mathrm{Cov}_{\mathcal{G}}$. Specifically, we have

$$\begin{aligned}
\Delta_{G_k} :=& \left\|\frac{1}{n}\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]^{-\frac{1}{2}}\sum_{i=1}^{n}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i)\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]^{-\frac{1}{2}} - \mathbf{I}_{|B_k|}\right\| \\
=& \left\|\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]^{-\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]\right)\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]^{-\frac{1}{2}}\right\| \\
\leq& \sigma^{-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]\right\|,
\end{aligned}$$

where the last inequality uses the assumption that $\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})) \geq \sigma$. Note that $\left\|\frac{1}{n}\sum_{i=1}^n \mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]\right\|$ is the norm of a sum of n independent, zero-mean, bounded random matrices. In particular, note that

$$\begin{aligned}
\lambda_{\max}(\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})) &\leq \|\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i)\| + \mathbb{E}_{\mathbf{x}}\|\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i)\| \\
&\leq 2\mathbb{E}\|(g(\mathbf{x}_i) - \mu_g(\mathbf{x}_i))(g(\mathbf{x}_i) - \mu_g(\mathbf{x}_i))^\top\| \\
&\leq 2\mathbb{E}\|g(\mathbf{x}_i) - \mu_g(\mathbf{x}_i)\|_2^2 \\
&\leq 2|B_k|\mathbb{E}\|g(\mathbf{x}_i) - \mu_g(\mathbf{x}_i)\|_\infty^2 \\
&\leq 8|B_k|M^2
\end{aligned}$$

where the last inequality follows from the assumption that $\|g_k(\mathbf{x}_k)\|_\infty \leq M$ almost surely. Moreover, we have

$$\begin{aligned}
\left\|\sum_{i=1}^n \mathbb{E}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]^2\right\| &\leq \sum_{i=1}^n \mathbb{E}\|\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})\|^2 \\
&\leq 64|B_k|^2M^4n
\end{aligned}$$

We can then apply the Matrix Bernstein inequality, e.g., (Tropp, 2012, Theorem 1.4) with $t = 32|B_k|M^2\sqrt{n\log p}$, to conclude that

$$\Delta_{G_k} \lesssim \frac{|B_k|M^2}{\sigma}\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - \frac{1}{p^2}$. Finally, applying a union bound over each of the B_k (i.e. at most p events) yields the result. ■

B.6 Proof of Proposition 14

Proposition 14 *Consider the decomposition* $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \mathbf{D} + \mathbf{Q}$*, where* $\mathbf{D}$ *is a diagonal matrix representing the independent feature augmentation part. Then, we have*

$$\Delta_G \lesssim \frac{\|\mathbf{D} - \mathbb{E}\mathbf{D}\| + \|\mathbf{Q} - \mathbb{E}\mathbf{Q}\|}{\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))}. \tag{15}$$

Proof This proof proceeds by partition the augmented covariance operator into diagonal and nondiagonal parts $\mathbf{D}$ and $\mathbf{Q}$ (i.e., $\mathrm{Cov}_G(\mathbf{X}) = \mathbf{D} + \mathbf{Q}$). We then bound the terms separately as below:

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

■

B.7 Proofs of Corollaries

Corollary 33 (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*

$$\text{MSE} \lesssim \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \lambda_{k+1}^2 \rho_k^2(\boldsymbol{\Sigma}; n\sigma^2) + R_k^{-1}(\boldsymbol{\Sigma}; n\sigma^2) + kn^{-1}.$$

Proof Since this belongs to the independent feature augmentation class, we can apply Corollary 12. Below are the quantities in the corollary.

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

hence,

$$\rho_k^{\text{aug}} = \rho_k(\boldsymbol{\Sigma}_{\text{aug}}; n) = \frac{n + \sum_{i=k+1}^{p} \lambda_i^{\text{aug}}}{n \lambda_{k+1}^{\text{aug}}} = \frac{n\sigma^2 + \sum_{i=k+1}^{p} \lambda_i}{n \lambda_{k+1}} = \rho_k(\boldsymbol{\Sigma}; n\sigma^2),$$

$$R_k^{\text{aug}} = R_k(\boldsymbol{\Sigma}_{\text{aug}}; n) = \frac{\left(n + \sum_{i=k+1}^{p} \lambda_i^{\text{aug}}\right)^2}{\sum_{i=k+1}^{p} (\lambda_{k+1}^{\text{aug}})^2} = \frac{\left(n\sigma^2 + \sum_{i=k+1}^{p} \lambda_i\right)^2}{n \sum_{i=k+1}^{p} \lambda_{k+1}^2} = R_k(\boldsymbol{\Sigma}; n\sigma^2).$$

Note that $R_k(\boldsymbol{\Sigma}; n\sigma^2)$ and $\rho_k(\boldsymbol{\Sigma}_{\text{aug}}; n\sigma^2)$ are the effective dimensions of the original spectrum for ridge regression with regularization parameter $n\sigma^2$, as defined in Tsigler and Bartlett (2020). Finally, the approximation error term is zero because $\Delta_G = 0$. ■

Corollary 15 (Regression bounds for unbiased randomized mask) *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 Theorem 4. Then, for any set $\mathcal{K} \subset \{1, 2, \ldots, p\}$ consisting of k_1 elements and some choice of $k_2 \in [0, n]$, the regression MSE is upper-bounded by*

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

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

Proof Random mask belongs to independent feature augmentation class, so we can apply Proposition 12. We calculate the quantities used in the corollary.

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

The effective ranks of the augmentation modified spectrum are

$$\rho_k^{\mathrm{aug}} = \frac{\psi n + p - k}{n}, \tag{44}$$

$$R_k^{\mathrm{aug}} = \frac{(\psi n + p - k)^2}{p - k}. \tag{45}$$

Now, we apply Proposition 12. 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

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

■

Corollary 17 (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*

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

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

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

while for random mask

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

Furthermore, the approximation is negligible when $k \ll \min(\sqrt{\frac{n}{\log p}}, \frac{p}{\sqrt{n}})$ as shown in Appendix F.2. Now, setting $\psi = \frac{k}{d-k}$ gives $\beta = \frac{k}{p}$. ■

Corollary 16 (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 $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$*

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

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

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

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 12 with:

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

Now as in the proof of Proposition 15, 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 18 (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 $\hat{\boldsymbol{\theta}}_{rm}(\beta)$):

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

Proof Proposition 12 is applicable to salt/pepper augmentation. The related quantities in the proposition are:

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

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,

$$\lambda_{pepper}(\beta,\sigma^2)^{-1} = \lambda_{rm}(\beta)^{-1} + \beta^{-1}\lambda_{gn}(\sigma^2)^{-1}. \tag{46}$$

Hence, the statement of MSE limit is clear as we take $\sigma \to 0$ in (46) along with the fact that $\lambda_{\mathrm{gn}} \to \infty$. Now we prove the bias statement. By Proposition 12,

$$\hat{\boldsymbol{\theta}}_{\mathrm{pepper}}(\beta,\sigma) \lesssim \|\boldsymbol{\theta}^*_{k+1:p}\|^2_{\boldsymbol{\Sigma}_{k+1:p}} + \left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2\Sigma^{-1}_{\pi(1:k_1)}} (\lambda^{\mathrm{aug}}_{k+1}\rho^{\mathrm{aug}}_k)^2. \tag{47}$$

In particular,

$$\left\|\theta^*_{\pi(1:k_1)}\right\|^2_{\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})]^2\Sigma^{-1}_{\pi(1:k_1)}} = \sum_{i\le k} \frac{\left(\psi\lambda_i + \frac{\psi\sigma^2}{1-\beta}\right)^2}{\lambda_i}(\boldsymbol{\theta}^*_i)^2, \tag{48}$$

$$\lambda^{\mathrm{aug}}_{k+1}\rho^{\mathrm{aug}}_k = \frac{n + \sum_{i>k}\frac{\lambda_i}{\psi(\lambda_i+\frac{\sigma^2}{1-\beta})}}{n} \le \frac{n + \sum_{i>k}\frac{\lambda_i}{\psi\frac{\sigma^2}{1-\beta}}}{n}. \tag{49}$$

Now the result follows by combining Eq. (47), (48) and (49).

The variance statement can be proved using similar calculations. From Corollary 12, 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 34 (Generalization of 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-β). Define*

$\psi = \frac{\beta}{1-\beta} \in [0, \infty)$. *Assume the assumptions in Corollary 15 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 15.*

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

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

Secondly, observing $\delta(\mathbf{x}) = -\beta\mathbf{x}$, $\mathrm{Cov}_\delta = \beta^2\boldsymbol{\Sigma}$, so concentration bound in Lemma 27 gives that

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

■

Appendix C. Proofs of Classification Results

C.1 Classification Lemmas

Lemma 35 (**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 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*

$$\mathrm{POE}(\hat{\theta}) \lesssim \frac{\mathrm{CN}}{\mathrm{SU}}\left(1 + \sigma_z\sqrt{\log\frac{\mathrm{SU}}{\mathrm{CN}}}\right). \tag{50}$$

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

$$\mathrm{POE}(\hat{\theta}) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})}{\mathrm{CN}(\hat{\boldsymbol{\theta}})} \leq \frac{\mathrm{CN}(\hat{\boldsymbol{\theta}})}{\mathrm{SU}(\hat{\boldsymbol{\theta}})}. \tag{51}$$

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 36. Assume without loss

of generality that $\mathbf{x}_{sig} = \mathbf{x}_t = \mathbf{x}_1$.

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

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 $\mathbf{u} = \frac{1}{\mathbf{x}_{\mathrm{sig}}}[\sqrt{\lambda_2}\hat{\boldsymbol{\theta}}_2, \sqrt{\lambda_3}\hat{\boldsymbol{\theta}}_3, \ldots, \sqrt{\lambda_p}\hat{\boldsymbol{\theta}}_p]$, which, by assumption, is independent of $\mathbf{z}'$. Then,

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

where the last inequality follows from the assumption that $\mathbf{z}_{sig}$ has bounded density and a small ball probability bound from (Vershynin, 2010, 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 Muthukumar et al. (2021) and the bound $\tan^{-1}(x) \geq \frac{\pi}{2} - \frac{1}{x}$, for all $x > 0$. ■

Lemma 36 (**Survival of ridge estimator for dependent features**) *Consider the classification task under the model and assumption described in Section 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:*

$$\frac{\lambda_t(1-2\nu^*)\left(1-\frac{k}{n}\right)}{L\left(\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda) + \lambda_t L\right)} \lesssim \mathrm{SU}(\hat{\boldsymbol{\theta}}) \lesssim \frac{L\lambda_t(1-2\nu^*)}{\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda) + L^{-1}\lambda_t\left(1-\frac{k}{n}\right)}. \tag{52}$$

Proof Our bound is a generalization to Theorem 22 in Muthukumar et al. (2021) 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 Muthukumar et al. (2021), our proof begins with writing the SU in terms of a quadratic form of signal vector and applying Hanson-Wright inequality, Lemma 26, by invoking the independence between the signal and noise. The result is that, with probability $1 - \exp(-\sqrt{n})$,

$$\mathrm{SU} \gtrsim \frac{\lambda_t \cdot \left((1-2\nu^*)\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) - 2c_1 \left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)}{1 + \lambda_t\left(\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) + c_1 \left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)} \text{ and} \tag{53}$$

$$\mathrm{SU} \lesssim \frac{\lambda_t \cdot \left((1-2\nu^*)\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) + 2c_1 \left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)}{1 + \lambda_t\left(\operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) - c_1 \left\|\mathbf{A}_{-t}^{-1}\right\| \cdot n^{3/4}\right)}, \tag{54}$$

Now observe $\|\mathbf{A}_{-t}^{-1}\| = \mu_n(\mathbf{A}_{-t})^{-1}$, so by Lemma 23, we have

$$\|\mathbf{A}_{-t}^{-1}\| \lesssim \frac{L}{n\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)}. \tag{55}$$

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

$$\lambda_1\rho_0(\mathbf{\Sigma};\lambda) = n^{-1}\sum_{i=1}^{k}\lambda_i + \lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda) \leq \lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)(1 + \frac{k\lambda_1}{nc}). \tag{56}$$

Also, using the same Lemma and (56),

$$\begin{aligned}\frac{(1-k/n)(1+\frac{k\lambda_1}{nc})^{-1}}{L\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)} \lesssim \frac{1-k/n}{L\lambda_1\rho_0(\mathbf{\Sigma};\lambda)} \lesssim \frac{n-k}{\mu_{k+1}(\mathbf{A}_{-t})} \lesssim \operatorname{tr}\left(\mathbf{A}_{-t}^{-1}\right) = \sum_{i=1}^{n}\frac{1}{\mu_i(\mathbf{A}_{-t})} \lesssim \frac{n}{\mu_n(\mathbf{A}_{-t})} \\ \lesssim \frac{L}{\lambda_{k+1}\rho_k(\mathbf{\Sigma};\lambda)}.\end{aligned} \tag{57}$$

Finally, plugging in the bounds in (57) and (55) into (53) completes the proof. ■

Lemma 37 (Contamination of ridge estimator for dependent features) *Consider the classification task under the model and assumption described in Section 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* $\tilde{\mathbf{X}} = [\mathbf{X}_{:1}, \ldots, \mathbf{X}_{:t-1}, \mathbf{X}_{:t+1}, \ldots, \mathbf{X}_{:p}]$*, respectively. Let* $\hat{\boldsymbol{\theta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}$ *be a ridge 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:*

$$\sqrt{\frac{\tilde{\lambda}_{k+1}\rho_k(\tilde{\mathbf{\Sigma}}^2;0)}{L'^2\lambda_1^2(1+\rho_0(\mathbf{\Sigma};\lambda))^2}} \lesssim \mathrm{CN}(\hat{\boldsymbol{\theta}}) \lesssim \sqrt{(1+\mathrm{SU}(\hat{\boldsymbol{\theta}})^2)L^2\left(\frac{k}{n} + \frac{n}{R_k(\tilde{\mathbf{\Sigma}};\lambda)}\right)\log n}. \tag{58}$$

Proof We begin with the same argument as in Lemma 28 in Muthukumar et al. (2021) to write the CN as a quadratic form of signal vector. For notation convenience, we denote the columns of $\mathbf{X}$ to be $\mathbf{X}_{:i},\ i \in \{1, 2, \ldots, p\}$, and define the leave-one-out quantities $\tilde{\mathbf{X}} := [\mathbf{X}_{:1}, \ldots, \mathbf{X}_{:t-1}, \mathbf{X}_{:t+1}, \ldots, \mathbf{X}_{:p}]$, $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{t-1}, \lambda_{t+1}, \ldots, \lambda_p)$, 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

$$\mathbf{y}^\top \widetilde{\mathbf{C}}\mathbf{y} = \|\hat{\boldsymbol{\theta}}_{-t}(\mathbf{y})\|_{\tilde{\mathbf{\Sigma}}}^2, \quad \mathbf{z}^\top \widetilde{\mathbf{C}}\mathbf{z} = \|\hat{\boldsymbol{\theta}}_{-t}(\mathbf{z})\|_{\tilde{\mathbf{\Sigma}}}^2, \tag{59}$$

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 Tsigler and Bartlett (2020),

$$\|\hat{\boldsymbol{\theta}}_{-t}(\varepsilon)\|_{\tilde{\mathbf{\Sigma}}}^2 \leq \frac{\varepsilon^\top \tilde{\mathbf{A}}_k^{-1}\tilde{\mathbf{X}}_{0:k}\tilde{\mathbf{\Sigma}}_{0:k}^{-1}\tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{A}}_k^{-1}\varepsilon}{\mu_n\left(\tilde{\mathbf{A}}_k^{-1}\right)^2 \mu_k\left(\tilde{\mathbf{\Sigma}}_{0:k}^{-1/2}\tilde{\mathbf{X}}_{0:k}^\top \tilde{\mathbf{X}}_{0:k}\tilde{\mathbf{\Sigma}}_{0:k}^{-1/2}\right)^2} + \varepsilon^\top \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{X}}_{k:\infty}\tilde{\mathbf{\Sigma}}_{k:\infty}\tilde{\mathbf{X}}_{k:\infty}^\top \tilde{\mathbf{A}}^{-1}\varepsilon, \tag{60}$$

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 26, for some constant c_1, with probability $1 - 2n^{-1}$,

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

where the last follows from the concentration of sum of sub-Gaussian variables. On the other hand, by Lemma 24, for some constant $c_2 > 0$,

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

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 26, Lemma 22, and Lemma 25,

we have for some constant $c_4 > 0$,

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

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

Combining all above, we deduce that

$$\mathrm{CN}(\hat{\boldsymbol{\theta}})^2 \lesssim (1+\mathrm{SU}(\hat{\boldsymbol{\theta}})^2)L^2\left(\frac{k}{n}+\frac{n}{R_k(\tilde{\mathbf{\Sigma}};\lambda)}\right)\log n. \tag{61}$$

For the lower bound of $\mathrm{CN}(\hat{\boldsymbol{\theta}})^2$, as shown in Muthukumar et al. (2021),

$$\mathrm{CN}(\hat{\boldsymbol{\theta}})^2 = \mathbf{y}^\top\mathbf{C}\mathbf{y} \geq \mu_n(\mathbf{C})\|\mathbf{y}\|_2^2 = n\mu_n(\mathbf{C}), \tag{62}$$

where $\mathbf{C} = (\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{X}}^T(\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-1}$. Now, by Lemma 27, we have

$$\mu_1(\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-2} \lesssim \frac{1}{(\lambda_1 n+\sum_{j=1}^p\lambda_j+\lambda)^2} \lesssim \frac{1}{\lambda_1^2 n^2(1+\rho_0(\mathbf{\Sigma};\lambda))^2}. \tag{63}$$

Also, by the boundness assumption on the condition number of $\tilde{\mathbf{X}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{X}}^\top$ and Lemma 22 we have

$$\mu_n(\tilde{\mathbf{X}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{X}}^\top) \gtrsim \frac{n}{L'}\tilde{\lambda}_{k+1}\rho_k(\tilde{\mathbf{\Sigma}}^2;\lambda), \tag{64}$$

with probability $1-\delta-n^{-1}$. Finally, the lower bound in the theorem is established by combining eq. (63) and (64):

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

■

Lemma 38 (**Probability of classification error of ridge estimator for dependent features**) *Consider the classification task under the model and assumption described in Section 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-one-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 $\tilde{\mathbf{X}} = [\mathbf{X}_{:1},\ldots,\mathbf{X}_{:t-1},\mathbf{X}_{:t+1},\ldots,\mathbf{X}_{:p}]$, 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:

$$\mathrm{POE}(\hat{\theta}) \lesssim \frac{\mathrm{CN}(\hat{\boldsymbol{\theta}})}{\mathrm{SU}(\hat{\boldsymbol{\theta}})}\left(1+\sigma_z\sqrt{\log\frac{\mathrm{SU}(\hat{\boldsymbol{\theta}})}{\mathrm{CN}(\hat{\boldsymbol{\theta}})}}\right), \tag{65}$$

$$\frac{\lambda_t(1-2\nu^*)\left(1-\frac{k}{n}\right)}{L\left(\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda)+\lambda_t L\right)} \lesssim \underbrace{\mathrm{SU}(\hat{\boldsymbol{\theta}})}_{\textit{Survival}} \lesssim \frac{L\lambda_t(1-2\nu^*)}{\lambda_{k+1}\rho_k(\boldsymbol{\Sigma};\lambda)+L^{-1}\lambda_t\left(1-\frac{k}{n}\right)}, \tag{66}$$

$$\sqrt{\frac{\tilde{\lambda}_{k+1}\rho_k(\tilde{\boldsymbol{\Sigma}}^2;0)}{L'^2\lambda_1^2(1+\rho_0(\boldsymbol{\Sigma};\lambda))^2}} \lesssim \underbrace{\mathrm{CN}(\hat{\boldsymbol{\theta}})}_{\textit{Contamination}} \lesssim \sqrt{(1+\mathrm{SU}(\hat{\boldsymbol{\theta}})^2)L^2\left(\frac{k}{n}+\frac{n}{R_k(\tilde{\boldsymbol{\Sigma}};\lambda)}\right)\log n}. \tag{67}$$

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

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

Proof This is a direct combination of Lemma 35, 36, and 37. ■

Lemma 39 (Bounds on the survival-to-contamination ratio between $\hat{\boldsymbol{\theta}}_{\mathbf{aug}}$ and $\bar{\boldsymbol{\theta}}_{\mathbf{aug}}$) *Consider an estimator* $\hat{\boldsymbol{\theta}}_{aug}$ *that solves the objective (1). Denote its averaged approximation* $\bar{\boldsymbol{\theta}}_{aug}$ *as in (5). Suppose* $\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug}))$ *and* $\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug}))$. *Then, the probability of classification error of* $\hat{\boldsymbol{\theta}}_{aug}$ *can be bounded by:*

$$\frac{1}{\mathrm{EM}}\frac{\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})}{\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})} \le \frac{\mathrm{SU}(\hat{\boldsymbol{\theta}}_{aug})}{\mathrm{CN}(\hat{\boldsymbol{\theta}}_{aug})} \le \mathrm{EM}\frac{\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})}{\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})}, \tag{68}$$

where $EM := \exp\left(\left(1+\frac{\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}}}{\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})}\right)\left(1+\frac{\|\hat{\boldsymbol{\theta}}_{aug}-\bar{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}}}{\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})}\right)-1\right) \in [1,\infty]$ *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, $\frac{\mathrm{SU}(\theta)}{\mathrm{CN}(\theta)} = \exp(f(\boldsymbol{\Sigma}^{1/2}\hat{\theta}))$ By the mean value theorem we have

$$f(\boldsymbol{\Sigma}^{1/2}\hat{\theta}) = f(\boldsymbol{\Sigma}^{1/2}\bar{\theta}) + \nabla f(\boldsymbol{\Sigma}^{1/2}\eta)\boldsymbol{\Sigma}^{1/2}(\hat{\theta}-\bar{\theta}), \tag{69}$$

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,

$$\|\nabla f(\theta)\| = \frac{1}{\|\mathbf{U}^\top\theta\|\|\mathbf{V}^\top\theta\|}\left\|\frac{\|\mathbf{U}^\top\theta\|\mathbf{V}\mathbf{V}^\top\theta}{\|\mathbf{V}^\top\theta\|} - \frac{\|\mathbf{V}^\top\theta\|\mathbf{U}\mathbf{U}^\top\theta}{\|\mathbf{U}^\top\theta\|}\right\| \tag{70}$$

$$= \frac{1}{\|\mathbf{U}^\top\theta\|\|\mathbf{V}^\top\theta\|}\sqrt{\frac{\|\mathbf{U}^\top\theta\|^2}{\|\mathbf{V}^\top\theta\|^2}\|\mathbf{V}\mathbf{V}^\top\theta\|^2 + \frac{\|\mathbf{V}^\top\theta\|^2}{\|\mathbf{U}^\top\theta\|^2}\|\mathbf{U}\mathbf{U}^\top\theta\|^2} \tag{71}$$

$$= \frac{\|\theta\|}{\|\mathbf{U}^T\theta\|\|\mathbf{V}^T\boldsymbol{\theta}\|}. \tag{72}$$

Hence,

$$\|\nabla f(\mathbf{\Sigma}^{1/2}\eta)\|\|\hat{\theta}-\bar{\theta}\|_{\Sigma} \leq \frac{(\|\mathbf{\Sigma}^{1/2}\bar{\theta}\|+t\|\mathbf{\Sigma}^{1/2}(\hat{\theta}-\bar{\theta})\|)\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{(\|\mathbf{U}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\|-t\|\mathbf{U}^T\mathbf{\Sigma}^{1/2}(\hat{\theta}-\bar{\theta})\|)(\|\mathbf{V}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\|-t\|\mathbf{V}^T\mathbf{\Sigma}^{1/2}(\hat{\theta}-\bar{\theta})\|)}$$
$$\leq \frac{(\|\bar{\theta}\|_{\mathbf{\Sigma}}+t\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}})\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{(\|\mathbf{U}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\|-t\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}})(\|\mathbf{V}^T\mathbf{\Sigma}^{1/2}\bar{\theta}\|-t\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}})}, \tag{73}$$

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. (73) is bounded by

$$\lesssim \left(\frac{1}{\mathrm{SU}(\bar{\theta})}+\frac{1}{\mathrm{CN}(\bar{\theta})}+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})\mathrm{SU}(\bar{\theta})}\right)\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}$$
$$= \left(1+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})}\right)\left(1+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\theta})}\right)-1. \tag{74}$$

Hence,

$$f(\mathbf{\Sigma}^{1/2}\hat{\theta}) \geq f(\mathbf{\Sigma}^{1/2}\bar{\theta})-\left(1+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})}\right)\left(1+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\theta})}\right)+1, \tag{75}$$

and

$$\frac{\mathrm{SU}(\hat{\theta})}{\mathrm{CN}(\hat{\theta})} \geq \frac{\mathrm{SU}(\bar{\theta})}{\mathrm{CN}(\bar{\theta})}\exp\left(1-\left(1+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{CN}(\bar{\theta})}\right)\left(1+\frac{\|\hat{\theta}-\bar{\theta}\|_{\mathbf{\Sigma}}}{\mathrm{SU}(\bar{\theta})}\right)\right) := \frac{\mathrm{SU}(\bar{\theta})}{\mathrm{CN}(\bar{\theta})}\frac{1}{\mathrm{EM}}. \tag{76}$$

The upper bound follows by an identical argument. ■

Lemma 40 *Let $\hat{\boldsymbol{\theta}}_{aug}$ and $\bar{\boldsymbol{\theta}}_{aug}$ be defined as in (5) 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,

$$\|\bar{\boldsymbol{\theta}}_{aug}-\hat{\boldsymbol{\theta}}_{aug}\|_{\mathbf{\Sigma}}^2 \leq \kappa\Delta_G^2\left(\mathrm{SU}(\bar{\boldsymbol{\theta}}_{aug})^2+\mathrm{CN}(\bar{\boldsymbol{\theta}}_{aug})^2\right). \tag{77}$$

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

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

where, by the assumption $\Delta_G < c$, one can prove $\mu_p((\mathbf{X}^\top\mathbf{X}+\mathbf{D})\,\bar{\mathbf{D}}^{-1})^2 \gtrsim n^2$ similarly as in Lemma 32. Finally, recalling Definition 8, we observe that

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

■

Remark 41 *Comparing with Lemma 32, 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.*

C.2 Proof of Theorem 9

Theorem 9 (Bounds on Probability of Classification Error) *Let $t \le n$ be the index (arranged according to the eigenvalues of $\boldsymbol{\Sigma}_{aug}$) of the non-zero coordinate of $\boldsymbol{\theta}^*$, $\widetilde{\boldsymbol{\Sigma}}_{aug}$ be the leave-one-out modified spectrum corresponding to index t, and $\widetilde{\mathbf{X}}_{aug}$ be the leave-one-column-out data matrix corresponding to column t. Suppose there exists a $t \le k \le n$ such that with probability at least $1-\delta$, the condition numbers of $n\mathbf{I}+\widetilde{\mathbf{X}}^{aug}_{k+1:p}(\widetilde{\mathbf{X}}^{aug}_{k+1:p})^\top$, $n\mathbf{I}+\mathbf{X}^{aug}_{k+1:p}(\mathbf{X}^{aug}_{k+1:p})^\top$, and $\widetilde{\mathbf{X}}_{k+1:p}\boldsymbol{\Sigma}_{k+1:p}\widetilde{\mathbf{X}}^T_{k+1:p}$ are at most L. Then as long as $\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\text{SU})$ and $\|\bar{\boldsymbol{\theta}}_{aug} - \hat{\boldsymbol{\theta}}_{aug}\|_{\boldsymbol{\Sigma}} = O(\text{CN})$,*

$$\text{POE}(\hat{\theta}) \;\lesssim\; \frac{\text{CN}}{\text{SU}}\left(1+\sigma_z\sqrt{\log\frac{\text{SU}}{\text{CN}}}\right), \tag{14}$$

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

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

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

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

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

where c is a universal constant.

Proof We can prove the theorem by carefully walking through the proofs of Lemma 35, 36, 37, and 39 and noting that the error multiplier defined in Lemma 39 is on the order of a constant under the assumptions made in this theorem. ■

C.3 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 (1) 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*

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

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 35, we know that the POE can be written as a function of the SU and CN of $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$. Next, recall that from the analysis in Section 3.2, the biased estimator is given by

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

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 $\hat{\boldsymbol{\theta}}_{\mathrm{aug}}$ by using the bounds in Theorem 9 with $\mathbf{X}$ and $\boldsymbol{\Sigma}$ replaced by $\mu(\mathbf{X})$ and $\bar{\boldsymbol{\Sigma}}$, respectively. ■

C.4 Proofs of Corollaries

Corollary 42 (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 $1 - \delta - \exp(-\sqrt{n}) - 5n^{-1}$*

$$\mathrm{POE} \lesssim Q^{-1}(1 + \sqrt{\log Q}) \text{ where} \tag{78}$$

$$Q = (1 - 2\nu)\sqrt{\frac{n}{p \log n}} \left(1 + \frac{n}{n\psi + p}\right)^{-1}. \tag{79}$$

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

$$\mathrm{POE} \asymp \frac{1}{2} - \frac{1}{\pi} \tan^{-1} Q \tag{80}$$

with the same probability.

Proof We first note the following key quantities:

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

and the effective ranks of the augmentation modified spectrum are

$$\rho_k^{\mathrm{aug}} = \frac{\psi n + p - k}{n}, \tag{81}$$

$$R_k^{\mathrm{aug}} = \frac{(\psi n + p - k)^2}{p - k}. \tag{82}$$

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

$$\mathrm{SU} \asymp (1 - 2\nu)\frac{n}{n\psi + n + p}, \tag{83}$$

$$\sqrt{\frac{np}{(n\psi + p)^2}} \lesssim \mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\frac{np\log n}{(n\psi + p)^2}} \tag{84}$$

$$\tag{85}$$

It remains to check when the conditions $\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\mathbf{\Sigma}} = O(SU)$ and $\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\mathbf{\Sigma}} = O(CN)$ 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 40, we have

$$\|\hat{\boldsymbol{\theta}}_{\mathrm{aug}} - \bar{\boldsymbol{\theta}}_{\mathrm{aug}}\|_{\mathbf{\Sigma}} \lesssim \kappa^{1/2}\Delta_G(SU + CN) \tag{86}$$

$$\lesssim \sigma_{\mathbf{z}}^2\sqrt{\frac{\log n}{n}}\sqrt{\frac{n}{p}} \tag{87}$$

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

Corollary 43 (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

$$\mathrm{POE} \asymp \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{\mathrm{SU}}{\mathrm{CN}},\ \textit{where} \tag{88}$$

$$\mathrm{SU} \asymp (1 - 2\nu)\frac{n}{2n + p},\quad \sqrt{\frac{np}{(n + p)^2}} \lesssim \mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\frac{np\log n}{(n + p)^2}}. \tag{89}$$

Proof By definition and the assumption of group invariance,

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

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 $\boldsymbol{\Sigma}_{\text{aug}}$.

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 $\boldsymbol{\Sigma}_{\text{aug}} = \frac{1}{2}\boldsymbol{\Sigma}$. Furthermore, $\mathrm{Cov}_G(\mathbf{X}) = \frac{1}{2}\boldsymbol{\Sigma}$ 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 augmentation) *Let $\hat{\boldsymbol{\theta}}_{rot}$ denote the estimator induced by the random-rotation augmentation with angle parameter α. 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* $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{rot}) \lesssim \mathrm{Var}(\hat{\boldsymbol{\theta}}_{ridge,\lambda})$. *Let $\hat{\boldsymbol{\theta}}_{lse}$ 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*

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

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

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

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

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

where the last equality is by Corollary 33 with $\lambda = 0$. The variance part can be proved similarly. The approximation error bound is proved in Appendix F. ■

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 $\tilde{\mathbf{\Sigma}}$ be the leave-one-out spectrum corresponding to index t. Then, with probability at least $1 - \exp(\sqrt{n}) - 5n^{-1}$,*

$$\mathrm{SU} \asymp (1 - 2\nu^*) \frac{\lambda_t}{\lambda_{k+1}\rho_k(\mathbf{\Sigma}; n\sigma^2) + \lambda_t}, \tag{90}$$

$$\mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\left(\frac{k}{n} + \frac{n}{R_k(\tilde{\mathbf{\Sigma}}; n\sigma^2)}\right)\log n}, \tag{91}$$

$$\tag{92}$$

and $\mathrm{EM} = 1$.

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

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

and the effective ranks are given by

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

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 $1 - \delta - \exp(\sqrt{n}) - 5n^{-1}$,*

$$\mathrm{SU} \asymp \frac{1}{\psi_1 + \frac{p\psi_t}{n\psi} + 1} \tag{93}$$

$$\mathrm{CN} \lesssim \sqrt{(1 + \mathrm{SU}^2)\frac{np\log n}{(n\psi + p)^2}} \tag{94}$$

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:

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

and the effective ranks of the augmentation modified spectrum are

$$\rho_k^{\mathrm{aug}} = \frac{\psi n + p - k}{n}, \tag{95}$$

$$R_k^{\mathrm{aug}} = \frac{(\psi n + p - k)^2}{p - k}. \tag{96}$$

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

Appendix D. Comparisons between Regression and Classification

D.1 Proof of Proposition 46

Proposition 46 (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 5 in Appendix A) and the assumptions in the two theorems hold, then,*

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

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

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{aug}) \lesssim \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{aug})\sqrt{\log n}. \tag{97}$$

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 $\boldsymbol{\Sigma}_{ii} = \gamma^i,\ \forall i \in \{1, 2, \ldots, p\}$, where γ is some constant satisfying $0 < \gamma < 1$. Let $\hat{\boldsymbol{\theta}}_{gn}$ and $\hat{\boldsymbol{\theta}}_{rm}$ be the corresponding estimators, then

$$\lim_{n\to\infty}\lim_{\sigma\to\infty}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{gn}) = 0 \quad \textit{while} \quad \lim_{n\to\infty}\lim_{\sigma\to\infty}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{gn}) = 1. \tag{98}$$

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

$$\lim_{n\to\infty}\lim_{\beta\to 1}\mathrm{POE}(\hat{\boldsymbol{\theta}}_{rm}) = 0 \quad \textit{while} \quad \lim_{n\to\infty}\lim_{\beta\to 1}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{rm}) = 1. \tag{99}$$

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

$$\frac{W_2^2(g(\mathbf{x}), \mathbf{x})}{p} \xrightarrow{n,\sigma} \infty, \tag{100}$$

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 12, the regression testing risk is bounded by

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

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 Tsigler and Bartlett (2020) shows that there is a matching lower bound such that

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\mathrm{aug}}) \gtrsim (\lambda_{k+1}^{\mathrm{aug}} \rho_k(\boldsymbol{\Sigma}_{\mathrm{aug}}; n))^2 + \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{\mathrm{aug}}; n)} + \frac{k}{n} \right), \tag{101}$$

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

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{\mathrm{aug}}) \lesssim \sqrt{(\lambda_{k+1}^{\mathrm{aug}} \rho_k(\boldsymbol{\Sigma}_{\mathrm{aug}}; n))^2 \cdot \left(\frac{n}{R_k(\boldsymbol{\Sigma}_{\mathrm{aug}}; n)} + \frac{k}{n} \right) \log n}, \tag{102}$$

for any k. Now combining E. q. (101) and (102) 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 $\hat{\theta}_{\mathrm{gn}}$, note that $\hat{\theta}_{\mathrm{gn}} = (\mathbf{X}^\top \mathbf{X} + \sigma^2 n \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \to 0$ almost surely as $\sigma \to \infty$, so

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

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

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

So,

$$\mathrm{POE}(\hat{\theta}_{\mathrm{gn}}) \leq \frac{\mathrm{CN}(\hat{\theta}_{\mathrm{gn}})}{\mathrm{SU}(\hat{\theta}_{\mathrm{gn}})} \asymp \frac{1}{\lambda_t} \sqrt{\frac{(\sum \lambda_j^2) \log n}{n}} \times \frac{n + \sum \frac{\lambda_j}{\sigma^2} + \frac{n\lambda_t}{\sigma^2}}{n + \sum \frac{\lambda_j}{\sigma^2}}, \tag{103}$$

$$\lim_{n \to \infty} \lim_{\sigma \to \infty} \mathrm{POE}(\hat{\theta}_{\mathrm{gn}}) = \lim_{n \to \infty} \frac{1}{\lambda_t} \sqrt{\frac{\log n}{n(1 - \gamma^2)}} = 0. \tag{104}$$

We can prove the statement for $\hat{\theta}_{\mathrm{rm}}$ similarly. When $\beta \to 1$, $\hat{\theta}_{\mathrm{rm}} = (\mathbf{X}^\top \mathbf{X} + \frac{\beta}{1-\beta} \operatorname{diag}[\mathbf{X}^\top \mathbf{X}])^{-1} \mathbf{X}^\top \mathbf{y} \to 0$ almost surely. So MSE approaches 1 almost surely. But by Corollary 42, we have

$$\lim_{n \to \infty} \lim_{\beta \to 1} \mathrm{POE}(\hat{\theta}_{\mathrm{rm}}) = \lim_{n \to \infty} \sqrt{\frac{p \log n}{n}} = 0. \tag{105}$$

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

$$W_2(g(\mathbf{x}), \mathbf{x}) = \|(\boldsymbol{\Sigma} + \sigma^2 \mathbf{I})^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}}\|_F^2 = \Omega(p\sigma^2). \tag{106}$$

■

D.2 Classification/regression separation for non-uniform random mask

Proposition 47 (**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 5 in Appendix A) 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*

$$\mathrm{POE}(\hat{\boldsymbol{\theta}}_{rm}) \xrightarrow{n} 0 \quad \textit{while} \quad \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{rm}) \xrightarrow{n} 1. \tag{107}$$

Proof From Corollary 16, we have that the bias scales as

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

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 Tsigler and Bartlett (2020)). Next we use the bounds in Corollary 45 to find that

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

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 $\to 0$. ■

Appendix E. Derivations of Common Augmented Estimators

Proposition 48 (Common augmentation estimators) *Below are closed-form expression of estimators that solves (1) 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,

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

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

$$h(\mathbf{x}) = (\mathbf{I} + \sin\alpha(\mathbf{v}\mathbf{u}^\top - \mathbf{u}\mathbf{v}^\top) + (\cos\alpha - 1)(\mathbf{u}\mathbf{u}^\top + \mathbf{v}\mathbf{v}^\top))\mathbf{x}. \tag{108}$$

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. ■

PatchShuffle regularization (Kang et al. (2017)) This augmentation is an example of a patch-based method where the original feature vector $\mathbf{x}$ is partitioned into sub-vectors $\tilde{\mathbf{x}}$ each with dimension b. The augmentation function to each sub-vector is given by $g(\tilde{\mathbf{x}}) = (1-r)\tilde{\mathbf{x}} + r\Pi(\tilde{\mathbf{x}})$ where $r \sim \mathrm{Bernoulli}(1-\beta)$ (chosen independently for each patch), and Π is a uniform random permutation to the features in $\tilde{\mathbf{x}}$.

Given a patch vector feature vector $\mathbf{x} \in \mathbb{R}^b$, we will show that $\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})$ and $\mathbb{E}_x \mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})$ are given as

$$\begin{aligned}
\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}) &= \beta(1-\beta)\mathbf{x}\mathbf{x}^\top + \left[\frac{1-\beta}{b(b-1)}\left(\left(\mathbf{1}^\top\mathbf{x}\right)^2 - \mathbf{1}^\top\mathbf{x}^{\odot 2}\right) - \left(\frac{1-\beta}{b}\right)^2\left(\mathbf{1}^\top\mathbf{x}\right)^2\right]\mathbf{1}\mathbf{1}^\top \\
&+ \left(\frac{1-\beta}{1-b}\mathbf{1}^\top\mathbf{x}^{\odot 2} - \frac{1-\beta}{b(1-b)}\left(\mathbf{1}^\top\mathbf{x}\right)^2\right)\mathbf{I}_b - \frac{\beta(1-\beta)}{b}\mathbf{1}^\top\mathbf{x}\left(\mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top\right), \\
\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})] &= \beta(1-\beta)\mathbf{\Sigma} - \frac{1-\beta}{b}\left(\mathbf{1}^\top\lambda\right)\mathbf{I}_b - \left(\frac{1-\beta}{b}\right)^2\mathbf{1}^\top\lambda\mathbf{1}\mathbf{1}^\top \\
&- \frac{\beta(1-\beta)}{b}\left(\lambda\mathbf{1}^\top + \mathbf{1}\lambda^\top\right),
\end{aligned}$$

where $\mathbf{x}^{\odot 2}$ denotes the entrywise (Hadamard) product of $\mathbf{x}$ with itself.

First, note that by definition, $\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x}) = \mathbb{E}[g(\mathbf{x})g(\mathbf{x})^\top] - \mathbb{E}[g(\mathbf{x})]\mathbb{E}[g(\mathbf{x})]^\top$. The first term is

$$\begin{aligned}
&\mathbb{E}[g(\mathbf{x})g(\mathbf{x})^\top] \\
=\ &\mathbb{E}[(1-r)^2]\mathbf{x}\mathbf{x}^\top + \mathbb{E}[(1-r)r]\mathbb{E}[\mathbf{x}\Pi(\mathbf{x})^\top] + \mathbb{E}[(1-r)r]\mathbb{E}[\Pi(\mathbf{x})\mathbf{x}^\top] + \mathbb{E}[r^2]\mathbb{E}[\Pi(\mathbf{x})\Pi(\mathbf{x})^\top] \\
=\ &\beta\mathbf{x}\mathbf{x}^\top + (1-\beta)\mathbb{E}[\Pi(\mathbf{x})\Pi(\mathbf{x})^\top] \\
=\ &\beta\mathbf{x}\mathbf{x}^\top + (1-\beta)\left[\left(\frac{(\mathbf{1}^\top\mathbf{x})^2 - \mathbf{1}^\top\mathbf{x}^{\odot 2}}{b(b-1)}\right)\mathbf{1}\mathbf{1}^\top + \left(\frac{1}{b-1}\mathbf{1}^\top\mathbf{x}^{\odot 2} - \frac{(\mathbf{1}^\top\mathbf{x})^2}{b(b-1)}\right)\mathbf{I}\right].
\end{aligned}$$

The last equation follows since the diagonal elements of $\mathbb{E}[\Pi(\mathbf{x})\Pi(\mathbf{x})^\top]$ are all equal to $\frac{1}{b}\sum_i \mathbf{x}_i^2$, while the (i,j) off-diagonal element is $\frac{\sum_i\sum_{j\neq i}\mathbf{x}_i\mathbf{x}_j}{b(b-1)} = \frac{(\sum_i \mathbf{x}_i)^2 - \sum_i \mathbf{x}_i^2}{b(b-1)}$. For the second term, $\mathbb{E}[g(\mathbf{x})] = \beta\mathbf{x} + (1-\beta)\frac{\mathbf{1}^\top\mathbf{x}}{b}\mathbf{1}$. Combining the above expressions gives the result of $\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})$. Finally, notice that $\mathbb{E}[\left(\mathbf{1}^\top\mathbf{x}\right)^2] = \mathbb{E}[\mathbf{1}^\top\mathbf{x}^{\odot 2}] = \mathbf{1}^\top\boldsymbol{\lambda}$ and $\mathbb{E}[\mathbf{1}^\top\mathbf{x}\mathbf{x}\mathbf{1}^\top] = \boldsymbol{\lambda}\mathbf{1}^\top$, where $\boldsymbol{\lambda}$ denotes the spectrum of $\boldsymbol{\Sigma}$. This completes the calculation of $\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}_k}(\mathbf{x})]$.

Appendix F. Approximation Error for Dependent Feature Augmentation

In this section, we demonstrate how to bound the approximation error for the augmentation of dependent features which satisfy neither the independent-augmentation nor regionally-correlated augmentation assumptions. We use rotation in a random plane (Section 5.5 and cutout DeVries and Taylor (2017) as our two examples. We will build on Proposition 14, which we restate here for convenience.

Proposition 14 *Consider the decomposition* $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \mathbf{D} + \mathbf{Q}$*, where* $\mathbf{D}$ *is a diagonal matrix representing the independent feature augmentation part. Then, we have*

$$\Delta_G \lesssim \frac{\|\mathbf{D} - \mathbb{E}\mathbf{D}\| + \|\mathbf{Q} - \mathbb{E}\mathbf{Q}\|}{\mu_p(\mathbb{E}_{\mathbf{x}}\mathrm{Cov}_{\mathcal{G}}(\mathbf{x}))}. \tag{15}$$

F.1 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

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

We follow the bound in E.q. (15) from the main text:

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

where we decompose $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X})$ into diagonal and off-diagonal parts as $\mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \mathbf{D} + \mathbf{Q}$, $\mathbf{D} = a\left(\mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} + \mathrm{Diag}(\mathbf{X}^\top\mathbf{X})\right)$, $\mathbf{Q} = a\left(\mathbf{X}^\top\mathbf{X} - \mathrm{Diag}(\mathbf{X}^\top\mathbf{X})\right)$, and $a = \frac{4(1-\cos\alpha)}{np} = \Theta(\frac{1}{np})$. Using similar arguments in the proof of Proposition 12 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 21 we have,

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

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

On the other hand, by invoking Lemma 27, 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,

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

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 $\sum_{j>1}\lambda_j \gg \lambda_1$.

F.2 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 E show that

$$\mathbb{E}_{\mathbf{x}}[\mathrm{Cov}_{\mathcal{G}}(\mathbf{x})] = \frac{k}{p-k}\,\mathrm{diag}(\mathbf{\Sigma}), \quad \mathrm{Cov}_{\mathcal{G}}(\mathbf{X}) = \frac{p}{p-k}\mathbf{M} \odot \frac{\mathbf{X}^\top\mathbf{X}}{n}, \tag{109}$$

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 $\mathbf{\Sigma}$ 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

$$\Delta_G = \frac{p}{k}\cdot\left(\left\|\widetilde{\mathbf{M}} \odot \mathbf{D} + \frac{k^2}{p(p-k)}n^{-1}\mathbf{Z}^\top\mathbf{Z}\right\|\right) \leq \frac{p}{k}\cdot\left(\underbrace{\left\|\widetilde{\mathbf{M}} \odot \mathbf{D}\right\|}_{L_1} + \underbrace{\left\|\frac{k^2}{p(p-k)}n^{-1}\mathbf{Z}^\top\mathbf{Z}\right\|}_{L_2}\right), \tag{110}$$

where $\mathbf{D}$ is an almost diagonal circular matrix with $\mathbf{D}_{ij} = \sum_{l=1}^n \frac{\mathbf{Z}_{li}\mathbf{Z}_{lj}}{n} - \delta_{ij}$ if $|i-j| \leq k$ 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 27 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 $\mathbf{D}$. For $i \neq j$, since $\mathbf{Z}_{ki}^2$ is sub-exponential we have

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

with probability $\exp(-nC\varepsilon^2)$ for some constant C by Lemma 21, 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 $\mathbf{D}$, we choose $\varepsilon = \sqrt{\frac{5\log pk}{n}}$. Then, by union bounds over pk terms, we obtain

$$\mathbf{D}_{i,j} \leq \sqrt{\frac{5\log pk}{n}}, \quad \forall i,j$$

with probability at least $1-\frac{1}{p^3k^3}$. Next, denote $\mathbf{A} := \widetilde{\mathbf{M}} \odot \mathbf{D}$. Note that $|\mathbf{A}_{ij}| \lesssim \frac{k}{p}\varepsilon$ for all $|i-j| \leq k$ 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 (110) 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 49 *This approximation bound, together with Corollary 5, 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 $k \ll \frac{p}{\sqrt{n}}$.*

Appendix G. Additional experiments

G.1 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

$$\hat{\theta}_{aug} \xrightarrow{\xi \to 0} \mathrm{Cov}_\infty^{-1} \mathbf{X}^\top \left(\mathbf{X} \, \mathrm{Cov}_\infty^{-1} \mathbf{X}^\top \right)^\dagger \mathbf{y}. \tag{111}$$

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

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\mathrm{Cov}_\infty} \quad \text{s.t. } \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \tag{112}$$

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. 8 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 of the form $\boldsymbol{\Sigma}_{ii} \propto \gamma^i$, where γ is chosen such that $\mu_p(\boldsymbol{\Sigma}) = 0.2\mu_1(\boldsymbol{\Sigma})$. It is clear from the plot that both aSGD and aERM converges to the M-LSE solution of Eq. 111). 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

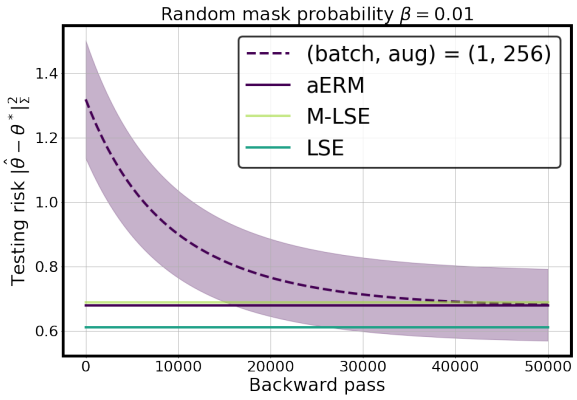


Figure 8: **aSGD convergence to aERM for small random mask.** We simulate the convergence of aSGD for random mask with dropout probability 0.01. We compare its converging estimator with the aERM limit in Eq. 111).

convergence of aSGD to aERM more generally) is beyond the scope of this work and would be interesting to explore in the future.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022.

Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021.

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. *Advances in Neural Information Processing Systems*, 35:37878–37891, 2022a.

Randall Balestriero, Ishan Misra, and Yann LeCun. A data-augmentation is worth a thousand samples: Exact quantification from analytical augmented sample moments. *arXiv preprint arXiv:2202.08325*, 2022b.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Xavier Bouthillier, Kishore Konda, Pascal Vincent, and Roland Memisevic. Dropout as data augmentation. *arXiv preprint arXiv:1506.08700*, 2015.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International Conference on Machine Learning*, pages 3252–3298. PMLR, 2023.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a low-rank regularizer for matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2018.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, pages 416–422, 2001.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *J. Mach. Learn. Res.*, 22:129–1, 2021.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020b.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Randaugment Le. Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

Yutong Dai, Brian Price, He Zhang, and Chunhua 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*, pages 11707–11716, 2022.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pages 1528–1537. PMLR, 2019.

Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*, 2021.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Boris Hanin and Yi Sun. How data augmentation affects optimization for linear regression. *Advances in Neural Information Processing Systems*, 34:8095–8105, 2021.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization. *arXiv preprint arXiv:1707.07103*, 2017.

Dmitry Kobak, Jonathan Lomond, and Benoit 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.*, 21:169–1, 2020.

Kishore Konda, Xavier Bouthillier, Roland Memisevic, and Pascal Vincent. Dropout as data augmentation. *stat*, 1050:29, 2015.

Elnaz Lashgari, Dehua Liang, and Uri Maoz. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885, 2020.

Daniel LeJeune, Randall Balestriero, Hamid Javadi, and Richard G Baraniuk. Implicit rugosity regularization via data augmentation. *arXiv preprint arXiv:1905.11639*, 2019.

Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Ran Liu, Mehdi Azabou, Max Dabagia, Chi-Heng Lin, Mohammad Gheshlaghi Azar, Keith Hengen, Michal Valko, and Eva Dyer. Drop, swap, and generate: A self-supervised approach for generating neural activity. *Advances in Neural Information Processing Systems*, 34: 10587–10599, 2021a.

Tongyu Liu, Ju Fan, Yinqing Luo, Nan Tang, Guoliang Li, and Xiaoyong Du. Adaptive data augmentation for supervised learning over missing data. *Proceedings of the VLDB Endowment*, 14(7):1202–1214, 2021b.

Andrew D McRae, Santhosh Karnik, Mark Davenport, and Vidya K Muthukumar. Harmless interpolation in regression and classification with structured features. In *International Conference on Artificial Intelligence and Statistics*, pages 5853–5875. PMLR, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pages 3351–3418. PMLR, 2021.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. In *International Conference on Machine Learning*, pages 3540–3548. PMLR, 2018.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. *Advances in neural information processing systems*, 28, 2015.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021. URL `http://jmlr.org/papers/v22/20-603.html`.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR, 2021.

Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022.

Peng Peng, Jiaxun Lu, Tingyu Xie, Shuting Tao, Hongwei Wang, and Heming Zhang. Open-set fault diagnosis via supervised contrastive learning with negative out-of-distribution data augmentation. *IEEE Transactions on Industrial Informatics*, 2022.

Anant Raj, Abhishek Kumar, Youssef Mroueh, Tom Fletcher, and Bernhard Schölkopf. Local group invariant representations via orbit embeddings. In *Artificial Intelligence and Statistics*, pages 1225–1235. PMLR, 2017.

Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, 30, 2017.

Dominic Richards, Edgar Dobriban, and Patrick Rebeschini. Comparing classes of estimators: When does gradient descent beat ridge regression in linear models? *arXiv preprint arXiv:2108.11872*, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Ohad Shamir. The implicit bias of benign overfitting. *arXiv preprint arXiv:2201.11489*, 2022.

Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation: a story of desert cows and grass cows. *arXiv preprint arXiv:2203.01572*, 2022.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. *arXiv preprint arXiv:2102.05113*, 2021.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.