WHEN COVARIATE-SHIFTED DATA AUGMENTATION INCREASES TEST ERROR AND HOW TO FIX IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Empirically, data augmentation sometimes improves and sometimes hurts test error, even when only adding points with labels from the true conditional distribution that the hypothesis class is expressive enough to fit. In this paper, we provide precise conditions under which data augmentation can increase test error for minimum norm estimators in linear regression. To mitigate the failure modes of augmentation, we introduce X-regularization, which uses unlabeled data to regularize the parameters towards the non-augmented estimate. We prove that our new estimator never increases test error and in practice exhibits significant improvements for adversarial data augmentation on CIFAR-10.

1 INTRODUCTION

We study *covariate-shifted* data augmentation, where we train on new inputs from some arbitrary distribution with corresponding outputs from the true conditional distribution. Can this form of data augmentation increase test error? On the surface, the answer seems like it should be negative since augmentation provides additional correct information about the true distribution. But the empirical evidence is mixed: In computer vision, while data augmentation by translating or flipping images improves accuracy (Krizhevsky et al., 2012; Yaeger et al., 1996; Ciresan et al., 2011), data augmentation with random or adversarial rotations (Engstrom et al., 2019; Yang et al., 2019) and imperceptible adversarial ℓ_∞ perturbations (Madry et al., 2018) leads to worse standard (non-adversarial) accuracy. Nakkiran (2019) explained this can happen if the hypothesis class is not expressive enough to fit the true function, but in practice, we often work with models which can fit the augmented training data perfectly (e.g. Zhang et al. (2017)).

In this paper, we show that surprisingly, even when the hypothesis class contains the true function, covariate-shifted data augmentation can increase test error. As a simple example, consider the problem of fitting a cubic spline (linear regression with some basis) to a set of points (Figure 1), where the true function is a "staircase". Without data augmentation (dashed blue), one obtains a line that captures the global structure and obtains low (but non-zero) error. With data augmentation using local perturbations (crosses), one incompletely fits the local structure of the high density points, which compromises the global structure on the tail (solid orange). We can show that the ratio between the errors of the two estimators grows as $\Omega(s^2)$, where s is the number of "stairs" by constructing a distribution over inputs such that for a small sample size, all the training points come from the first $s/2$ stairs (see Theorem 4 for details).

Our main result provides a general analysis of the effect of covariate-shifted data augmentation for minimum norm estimators for any linear regression problem. We show in Theorem 1 that whether augmentation helps or hurts hinges on the relation between the span of the augmented points and span of the original points. Intuitively, augmentations that open up new directions may increase increase error if those directions are weighted heavily in the population covariance. However, as the span of the original points grows with additional training data, data augmentation becomes safer. One might expect augmentation to be most helpful in low data settings, but we show this is exactly the regime where it can also be most hurtful, even when all the augmented data is drawn from the correct conditional distribution and the hypothesis class is expressive enough to interpolate it.

To eliminate detrimental effects for any possible true model, we propose a new estimator for covariate-shifted data augmentation that leverages unlabeled data from the test distribution. The estimator adds what we call X-regularization, which pulls the augmented estimator towards the non-augmented estimator, in directions weighted heavily in the risk. We prove that with X-regularization, covariate-shifted

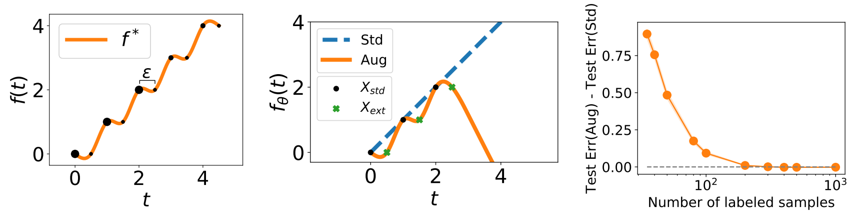


Figure 1: We consider perfectly fitting training points using cubic splines while minimizing *smoothness* of the function. **(Left)** depicts the true function $f^\star$ and the mass P_{xy} on each point (x,y) size of the circles) **(Middle)** With a small number of standard training samples (circles) from the distribution depicted in (a), data augmentation with local perturbations (crosses) causes the estimator to have larger error due to being maximally smooth while also fitting the augmented local perturbations. **(Right)** Difference in test error before (standard) and after data augmentation (augmented)

data augmentation never increases error of any minimum norm estimator for linear regression (Theorem 3). Empirically, for adversarial ℓ_∞ perturbations on CIFAR-10, X-regularization dramatically outperforms vanilla adversarial training methods (Madry et al., 2018) and achieves a relative test error decrease of at least $\approx 20\%$ using the entire and smaller subsets of the labeled training data.

2 SETTING

Well-specified linear regression. We assume the targets $y \in \mathbb{R}$ are drawn from the conditional distribution

$$P_{\text{y}}(\cdot \,|\, x) = \mathcal{N}(x^\top \theta^\star, \sigma), \tag{1}$$

where $x \in \mathbb{R}^d$ are the covariates and $\theta^\star \in \mathbb{R}^d$ are the true parameters. Our goal is to learn a linear predictor $f_\theta(x) = x^\top \theta$. In our main results, we allow the feature vectors to be arbitrary (hence, fixed design) since we focus on the small sample regime and give results in terms of the empirical and population covariance explicitly. The random design case for growing sample sizes however is also discussed in Section 3.3.

Covariate-shifted data augmentation. Let P_{xy} denote the underlying distribution of (x,y) pairs, P_{x} its marginal on $\mathbb{R}^d$ (which is arbitrary and can hence be supported on a fixed given set of points) and $P_{\text{y}}(\cdot \mid x)$ the corresponding conditional distribution of y given input x. Suppose we have n pairs $(x_i, y_i) \sim P_{\text{xy}}$ in the standard training set. We refer to the standard training set by $X_{\text{std}} = [x_1, x_2, \ldots x_n]^\top \in \mathbb{R}^{n \times d}$ and $Y_{\text{std}} = [y_1, y_2, \ldots y_n]^\top \in \mathbb{R}^n$.

Analogously, we consider αn additional training points denoted by $X_{\text{ext}} = [\tilde{x}_1, \tilde{x}_2, \ldots \tilde{x}_{\alpha n}]^\top \in \mathbb{R}^{\alpha n \times d}$ with associated targets $Y_{\text{ext}} = [\tilde{y}_1, \tilde{y}_2, \ldots \tilde{y}_n]^\top \in \mathbb{R}^m$. While the augmented covariates can be arbitrary and hence the distribution of P_{x} of the augmented dataset is potentially different than `Stairs` on the original dataset, the conditional target distribution remains the same for both datasets $\tilde{y}_j \sim P_{\text{y}}(\cdot \,|\, \tilde{x})$.

Minimum norm interpolation estimators. Motivated by the observation that modern machine learning models achieve near zero training loss (on standard and augmented points), we focus on the interpolating regime that has been a focus of a number of recent works (Ma et al., 2018; Belkin et al., 2018; Hastie et al., 2019; Liang & Rakhlin, 2018) . Given covariates X and targets Y as training data, we study the following interpolation estimator:

$$\theta = \operatorname*{argmin}_\theta \Big\{ \theta^\top M \theta : X\theta = Y \Big\}, \tag{2}$$

where M is a positive definite (PD) matrix that incorporates prior knowledge about the true model.

Given an arbitrary PD matrix M, we can rotate the covariates $x \leftarrow M^{-1/2} x$ and parameters $\theta \leftarrow M^{1/2} \theta$. The rotated feature matrix X and target matrix Y are now related via $Y = X\theta + \sigma \mathcal{N}(0, I)$ and the M-norm of the parameters simplifies to $\|\theta\|_2$. Hence, we present our results in terms of the ℓ_2 norm (ridgless regression) although all results hold for arbitrary M–norms via appropriate rotations.

In this work, we compare the performance two estimators: (i) standard estimator $\hat{\theta}_{\text{std}}$ with training data $X, Y = [X_{\text{std}}, Y_{\text{std}}]$ in (2) and (ii) covariate-shifted data augmentation estimator $\hat{\theta}_{\text{aug}}$ where the training data $X = [X_{\text{std}}; X_{\text{ext}}], Y = [Y_{\text{std}}, Y_{\text{ext}}]$ in (2):

$$\begin{aligned} \hat\theta_{\text{std}} &= \underset{\theta}{\operatorname{argmin}} \Big\{ \|\theta\|_2 : X_{\text{std}}\theta = Y_{\text{std}} \Big\}. \\ \hat\theta_{\text{aug}} &= \underset{\theta}{\operatorname{argmin}} \Big\{ \|\theta\|_2 : X_{\text{std}}\theta = Y_{\text{std}}, X_{\text{ext}}\theta = Y_{\text{ext}} \Big\}. \end{aligned} \tag{3}$$

Predictive risk. For both estimators we are interested in the expected predictive risk on a random sample x_{test} from the distribution P_{x} with covariance Σ. For fixed design $X_{\text{std}}, X_{\text{ext}}$, this expected predictive risk can be decomposed into a bias and variance term.

$$\begin{aligned} R(\hat\theta) &= \mathbb{E}[(x_{\text{test}}^\top(\hat\theta - \theta^\star))^2] = \mathbb{E}[(\hat\theta - \theta^\star)^\top \Sigma (\hat\theta - \theta^\star)] \\ &= \underbrace{(\mathbb{E}[\hat\theta] - \theta^\star)^\top \Sigma (\mathbb{E}[\hat\theta] - \theta^\star)}_{\text{Bias } B(\hat\theta)} + \underbrace{\operatorname{tr}(\operatorname{Cov}(\hat\theta)\Sigma)}_{\text{Variance } V(\hat\theta)}. \end{aligned} \tag{4}$$

where the expectation is also taken over the samples Y sampled from $P_{\mathsf{y}}(\cdot \mid x)$. We compare the two estimators defined in Equation (3) in terms of their predictive risk.

3 Predictive risk of covariate-shifted data augmentation

In this section, we study the risk of the minimum norm estimators defined in Equation (3). In particular we compare the predictive bias and variance of a standard estimator and the corresponding covariate-shifted data augmentation estimators (called augmented estimator for brevity).

Bias and variance of minimum norm interpolants. Let Σ be the PD matrix which determines the predictive risk (4), which could but doesn't necessarily have to be the covariance of the underlying distribution P_{x}. Let $\Sigma_{\text{data}} = \frac{1}{n} X^\top X$ and $\Pi_{\text{data}}^\perp = I - \Sigma_{\text{data}}^\dagger \Sigma_{\text{data}}$. For the minimum norm estimators $\hat\theta$ as defined in (3) the bias and variance can be computed as follows.

$$B(\hat\theta) = \theta^{\star\top} \Pi_{\text{data}}^\perp \Sigma \Pi_{\text{data}}^\perp \theta \text{ and } V(\hat\theta) = \frac{\sigma^2}{n} \operatorname{tr}\Big(\Sigma_{\text{data}}^\dagger \Sigma\Big). \tag{5}$$

Throughout the paper we say that augmentation is *safe* if the predictive risk or bias does not increase, and *hurtful* if it does.

Large sample regime. Common intuition from a statistical standpoint would suggest that more data is always good. This is indeed true in the large sample regime when the covariance matrix of the standard training data $\Sigma_{\text{std}} := \frac{1}{n} X_{\text{std}}^\top X_{\text{std}}$ is invertible. In this case, both the standard and augmented estimators are unbiased. Plugging in $\Sigma_{\text{data}} = \Sigma_{\text{std}}$ for the standard estimator, and $\Sigma_{\text{data}} = \Sigma_{\text{aug}} := \frac{1}{n}\Big(X_{\text{std}}^\top X_{\text{std}} + \alpha X_{\text{ext}}^\top X_{\text{ext}}\Big)$ for the augmented estimator in the expression for variance (5) yields that $R(\hat\theta_{\text{aug}}) \leq R(\hat\theta_{\text{std}})$. See Appendix A.1 for a full proof.

However, when Σ_{std} is not invertible, the augmented estimator could have higher predictive risk as exemplified in the spline staircase example (Figure 1). In the following sections, we characterize when the bias (Sec. 3.1) and variance (Sec. 3.2) of the augmented estimator are larger than those of the standard estimator.

3.1 Predictive bias in the small sample regime

In this section, we study the effect of covariate-shifted data augmentation on the predictive bias when Σ_{std} is not invertible. In this case, both the standard estimator and the augmented estimator are biased. The bias is equivalent to the risk when observing noiseless targets. The spline staircase example shows that even with noiseless target observations, data augmentation can increase bias. We now characterize general conditions that leads to bias increase for minimum interpolants in linear regression. Before jumping into the formal characterization we want to give some intuition for how adding noiseless data can hurt using a very simple linear model in $\mathbb{R}^3$.

3.1.1 Simple linear example in $\mathbb{R}^3$ where adding data increases bias

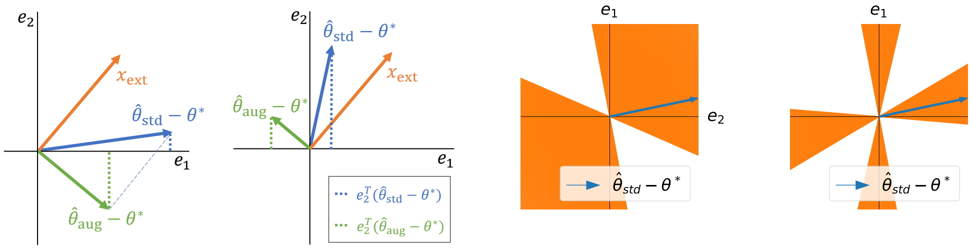
The following concrete example illustrates how the interaction between the span of the standard and augmented training points with the underlying $\theta^\star$ can cause the bias to increase. For simplicity we choose $\Sigma = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ with $\lambda_2 \gg \lambda_1$, $X_{\text{std}} = e_3$ and $X_{\text{ext}} = [e_1; e_2]^\top$ where e_1, e_2, e_3 denote the standard bases in $\mathbb{R}^3$. Plugging these terms into the bias expression in Equation (5) yields

$$B(\hat{\theta}_{\text{std}}) = {\theta_1^\star}^2 \lambda_1 + {\theta_2^\star}^2 \lambda_2 \quad \text{and} \quad B(\hat{\theta}_{\text{aug}}) = (1/4)(\theta_1^\star - \theta_2^\star)^2 \lambda_1 + (1/4)(\theta_1^\star - \theta_2^\star)^2 \lambda_2.$$

Since $\hat{\theta}_{\text{aug}}$ and $\hat{\theta}_{\text{std}}$ are identical and perfectly interpolating on e_3, we restrict our attention to the linear span of e_1, e_2 that is the nullspace of X_{std}. Figure 2 depicts the errors of the two estimators $\hat{\theta}_{\text{std}}$ and $\hat{\theta}_{\text{aug}}$ in said nullspace for different choices of $\theta^\star$. Since, $\lambda_2 \gg \lambda_1$, the bias expression is dominated by the coefficient on λ_2. In particular,

(i) when $\theta_1^\star \gg \theta_2^\star$ as in Fig. 2 (a), augmenting with X_{ext} can be hurtful, that is $B(\hat{\theta}_{\text{aug}}) \gg B(\hat{\theta}_{\text{std}})$. Even though the norm of $\hat{\theta}_{\text{aug}} - \theta^\star$ is smaller than that of $\hat{\theta}_{\text{std}} - \theta^\star$, the increase along e_2 dominates the effect on predictive bias because $\lambda_2 \gg \lambda_1$.

(ii) when $\theta_2^\star \gg \theta_1^\star$ as in Fig. 2 (b), the same X_{ext} causes $B(\hat{\theta}_{\text{aug}})$ to be smaller than $B(\hat{\theta}_{\text{std}})$. Here the augmented error $\hat{\theta}_{\text{aug}} - \theta^\star$ is smaller along e_2 compared to the standard error $\hat{\theta}_{\text{std}} - \theta^\star$, dominating the increase along e_1.


e_2
x_{ext}
$\hat{\theta}_{\text{std}} - \theta^*$
e_1
$\hat{\theta}_{\text{aug}} - \theta^*$
e_2
$\hat{\theta}_{\text{std}} - \theta^*$
$\hat{\theta}_{\text{aug}} - \theta^*$
x_{ext}
e_1
$e_2^T(\hat{\theta}_{\text{std}} - \theta^*)$
$e_2^T(\hat{\theta}_{\text{aug}} - \theta^*)$
e_1
e_2
$\hat{\theta}_{std} - \theta^*$
e_1
e_2
$\hat{\theta}_{std} - \theta^*$

(a) $\theta_1^\star \gg \theta_2^\star$ (b) $\theta_2^\star \gg \theta_2^\star$ (c) $\Sigma = \mathrm{diag}([1,4])$ (d) $\Sigma = \mathrm{diag}([1,25])$

Figure 2: Illustration of the 3-D example described in Sec. 3.1.1. In **(a), (b)** we depict the errors $\hat{\theta}_{\text{aug}} - \theta^\star$ (green solid) and $\hat{\theta}_{\text{std}} - \theta^\star$ (blue solid) projected on the nullspace of $\mathrm{Ker}(\Sigma_{\text{std}})$ that is spanned eigenvectors e_1 and e_2 of Σ with $\lambda_2 \geq \lambda_1$. **(c), (d)** We show the space of safe augmentation directions (orange) that don't increase bias for a given $\theta^\star$ to be cone-shaped where the cone width depends on the alignment of $\theta^\star$ with eigenvectors of Σ and the skew of the eigenvalues of Σ.

In summary, the components of the errors along eigenvectors of Σ with large eigenvalues dominate the predictive bias. The projection onto one of these directions may increase when adding new directions X_{ext}. Even though the norm of the error always decreases with more data points, the same therefore does not have to hold for the predictive bias if it weights the error in these directions more.

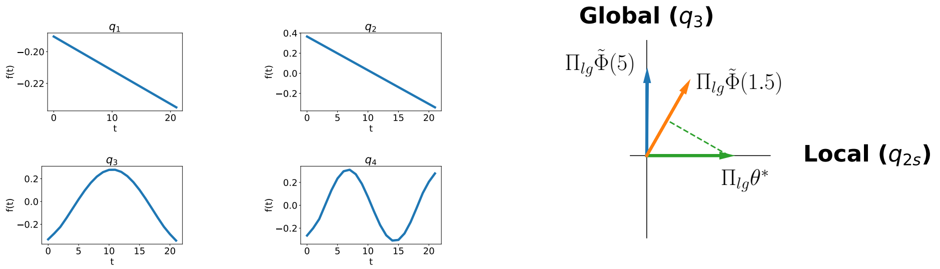
This viewpoint also helps us to explain the observed phenomenon in the spline example in Fig. 1 where the eigenvectors correspond to *local* vs. *global* components. In the next section, we present these ideas formally for a general setting.

3.1.2 General characterizations

As explained above, for the purpose of analyzing the bias difference between the augmented and standard estimators (3), we only need to focus on the nullspace of Σ_{std} with projection matrix $\Pi_{\text{std}}^\perp$. Let $\Sigma_{\text{aug}} = \Sigma_{\text{std}} + \alpha \Sigma_{\text{ext}}$. Let's define an orthogonal decomposition $\mathrm{Ker}(\Sigma_{\text{std}}) = S_{\text{aug}} \oplus \mathrm{Ker}(\Sigma_{\text{aug}})$ where S_{aug} is the subspace orthogonal to $\mathrm{Ker}(\Sigma_{\text{aug}})$. Then, for a given $\theta^\star$ we can always decompose $\Pi_{\text{std}}^\perp \theta^\star = v + w$ where v, w are the (mutually orthogonal) projections of $\theta^\star$ onto S_{aug} and $\mathrm{Ker}(\Sigma_{\text{aug}})$ respectively. The following theorem gives exact conditions and characterizes combinations of $X_{\text{std}}, X_{\text{ext}}, \Sigma$ when augmentation *hurts*, i.e. increases the bias by a positive amount $c = B(\hat{\theta}_{\text{aug}}) - B(\hat{\theta}_{\text{std}})$ that depends on the true model $\theta^\star$. In the following, Π_{aug} and $\Pi_{\text{aug}}^\perp$ denote the projection matrices onto $\mathrm{col}(\Sigma_{\text{aug}})$ and $\mathrm{Ker}(\Sigma_{\text{aug}})$ respectively.

Theorem 1. *The augmented estimator $\hat{\theta}_{aug}$ has higher bias if and only if*

$$v^\top \Sigma v < 2 w^\top \Sigma v, \tag{6}$$



(a) Eigenvectors 1–4 of spline Σ

(b) Nullspace projections onto local vs. global

Figure 3: **(a)** Top 4 eigenvectors $q_1,...,q_{2s}$ of Σ in the splines problem, representing wave functions in the input space. As a smoothing spline, the eigenvalues have much larger weight on "global" eigenvectors that vary smoothly over the domain. **(b)** Projections onto q_3 and q_{2s} in $\text{Ker}(\Sigma_{\text{std}})$ via Π_{lg}, representing global and local eigenvectors respectively. The local perturbation $\Pi_{\text{lg}}\hat{\Phi}(1.5)$ has both local and global components, creating a high-error component in the global direction.

where $v = \Pi_{std}^{\perp}\Pi_{aug}\theta^\star$ and $w = \Pi_{std}^{\perp}\Pi_{aug}^{\perp}\theta^\star$. Furthermore, for a given X_{std}, X_{ext}, Σ, a bias increase of $B(\hat{\theta}_{aug}) - B(\hat{\theta}_{std}) = c > 0$ via augmentation with X_{ext} is possible only if $\theta^\star$ is sufficiently more complex than the interpolant on the original dataset in the ℓ_2 norm, i.e. $\|\theta^\star\|_2 - \|\hat{\theta}_{std}\|_2 > \gamma c$ for some scalar $\gamma > 0$ that depends on X_{std}, X_{ext}, Σ.

The proof of Theorem 1 can be found in Appendix Sec. A.2. We can also use condition (6) to determine augmentations or choices of Σ via the M-norm in (2) where augmentation never hurts for any $\theta^\star$ for given X_{std}, Σ. In particular, if for all directions $w \in \text{Ker}(\Sigma_{\text{aug}})$ and $v \in S_{\text{aug}}$ it holds that $w^\top \Sigma v = 0$, the bias cannot increase for any choice of $\theta^\star$. This holds for example when $\Sigma = I$[1] or when X_{ext} spans the entire nullspace of Σ_{std} (that is Σ_{aug} becomes invertible) and hence $w = 0$.

We now return to the simple 3-D example and illustrate the entire set of safe single augmentation directions in the nullspace of Σ_{std} for different choices of Σ and a fixed $\theta^\star$ in Figure 2 (c), (d). The plots are created using Corollary 1 (a) in Sec. A.4 in the Appendix. In general, the safe augmentation directions lie in a cone around the eigenvectors of Σ, supported by Corollary 1 (b), while the width and alignment of the cone depends on the alignment of $\theta^\star$ with the eigenvectors of Σ.

The second statement in Theorem 1 links back to the intuition in the spline example in the introduction. There, the true staircase function is much more complex than the (good) linear solution that fits most points. Theorem 1 states that this does not only apply to the regression spline setting but that it is indeed necessary for a fixed bias increase that the (rotated) parameter $\theta^\star$ has a higher ℓ_2-norm. In the next section we rigorously define the spline example and use intuitions from the simple 3-D example in Sec. 3.1.1 and our characterizations in Theorem 1 to explain the observed bias increase observed in Fig. 2 in terms of local and global eigenvector functions.

3.1.3 Revisiting the spline staircase

The spline staircase is a linear interpolation problem with noiseless targets $f^\star(x) = \lfloor x \rfloor$. We transform the problem to minimum ℓ_2 norm interpolation using features $\tilde{\Phi}(x) = \Phi(x)M^{-1/2}$ (where $\Phi : \mathbb{R} \mapsto \mathbb{R}^d$ is the cubic spline feature map), so that the results from Section 3.1.2 apply directly. Let Σ be the population covariance of $\tilde{\Phi}$ for a uniform distribution over the discrete domain consisting of s integers and their perturbations (Figure 1). Let $Q = [q_i]_{i=1}^{2s}$ be the eigenvectors of Σ in decreasing order of their corresponding eigenvalues. The visualization in Figure 3(a) shows that q_i are wave functions in the original input space; the "frequency" of the wave increases as i increases.

Suppose the original training set consisted of two points, $X_{\text{std}} = [\tilde{\Phi}(0); \tilde{\Phi}(1)]^\top$. We study the effect of augmenting point x_{ext} in terms of q_i above. We find that $\Pi_{\text{std}}^{\perp} q_1 = \Pi_{\text{std}}^{\perp} q_2 = 0$. For ease of visualization, we consider the 2D space in $\text{Ker}(\Sigma_{\text{std}})$ spanned by $\Pi_{\text{std}}^{\perp} q_3$ (global dimension) and $\Pi_{\text{std}}^{\perp} q_{2s}$ (local

[1] For the general min-norm estimator (2), this is equivalent to the unrotated covariance being equal to M.

dimension), and denote the projection matrix onto this space by Π_{lg}. Note that the same results hold when projecting onto all $\Pi^{\perp}_{\text{std}} q_i$ in Ker(Σ_{std}).

We map the 2D space above to the simple example in Section 3.1.1: global corresponding to e_2 (with large λ_2) and local to e_1. Figure 3 plots the projections $\Pi_{\text{lg}}\theta^\star$ and $\Pi_{\text{lg}}\tilde{\Phi}(X_{\text{ext}})$ for different X_{ext}. When $\theta^\star$ has high frequency variations and is complex, $\Pi_{\text{lg}}\theta^\star$ is aligned with the local dimension. For x_{ext} close to training points, the projection $\Pi_{\text{lg}}\tilde{\Phi}(x_{\text{ext}})$ (orange vector in Figure 3(b)) has both local and global components. Fitting the local component of $\theta^\star$ introduces an error in the global component. For x_{ext} far away from training points, $\Pi_{\text{lg}}\tilde{\Phi}(x_{\text{ext}})$ (blue vector in Figure 3(b)) is almost entirely global and perpendicular to $\theta^\star$, leaving bias unchanged. Thus, augmenting data close to original data cause estimators to fit local components at the cost of the costly global component which changes overall structure of the predictor like in Figure 1(middle). The choice of inductive bias in the M–norm being minimized results in eigenvectors of Σ that correspond to local and global components, dictating this tradeoff.

3.2 Predictive variance in the small sample regime

Recall that in the large sample regime where Σ_{std} is inverible, the variance always decreases with covariance-shifted data augmentation 2. However, when Σ_{std} is not invertible, the variance of the augmented estimator could be larger.

Theorem 2. *The differences in variances of the two estimators can be expressed as follows.*

$$V(\hat{\theta}_{aug}) - V(\hat{\theta}_{std}) = \frac{1}{n}\Big(\underbrace{V_1(\Pi^{\perp}_{std}X_{ext})}_{\textit{Variance increase}} - \underbrace{V_2(\Pi_{std}X_{ext})}_{\textit{Variance decrease}}\Big), \tag{7}$$

where $V_1(U)$ and $V_2(U)$ are scalars that depend on X_{std}, Σ such that $V_1(U), V_2(U) \geq 0$ and $V_1(\mathbf{0}) = 0, V_2(\mathbf{0}) = 0$.

From Theorem 2, we see that when $\Pi_{\text{std}}X_{\text{ext}} = 0$, and X_{ext} is orthogonal to X_{std}, the variance of augmented estimator is larger. However when $\Pi^{\perp}_{\text{std}}X_{\text{ext}} = 0$, the variance of augmented estimator is lower (like in the large sample regime with invertible Σ_{std} where $\Pi^{\perp}_{\text{std}} = 0$). The exact expression the difference in variances appears in Appendix C.1.

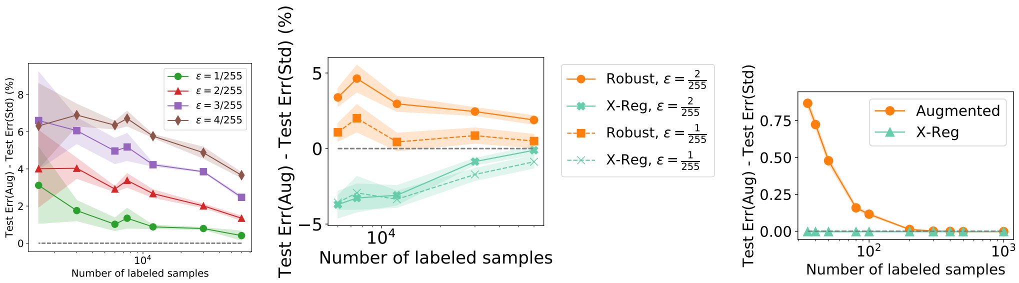
Thus, from Theorems 1 and 2, we see that when X_{ext} is not in the span of X_{std}, covariance-shifted data augmentation *could* increase both the bias and variance.

3.3 Effect of size of the original training set

The analysis of the preceding sections is general and characterizes the risk upon augmentation in terms of Σ_{std} and Σ without making any assumptions on their relation. When the rows in X_{std} correspond to n i.i.d. samples from a distribution with covariance Σ, the empirical covariance $\Sigma_{\text{std}} = \frac{1}{n}X_{\text{std}}^\top X_{\text{std}}$ could be arbitrarily different from Σ in the small sample regime and the risk increase can be severe even in the noiseless case.

As an example, Theorem 4 in the Appendix describes how sampling and augmenting from skewed distributions of the staircase form as depicted in Figure 1 (a) may lead to severe bias increase in the small sample regime (Theorem 4). On the other hand, in the large sample regime $\Sigma_{\text{std}} \to \Sigma$ when Σ_{std} is invertible, we can show in Appendix A.1 that data augmentation never increases the risk. In summary, our analysis on linear regression examples states that as the sample size increases and the span of X_{std} grows, the possible negative effect of data augmentation should decrease.

In order to see whether this trend holds true for more complex models and real world datasets as well, we consider data augmentation via adversarial training (Madry et al., 2018) (AT) with ℓ_∞ perturbations on CIFAR-10 on a WideResnet (in particular, the WRN-40-2 (Zagoruyko & Komodakis, 2016)). Essentially, AT augments with imperceptible perturbations of training images with the corresponding correct target and thereby falls in our framework of covariate-shifted data augmentation. We subsample the full training set and plot the effect of (adversarial) data augmentation on test error as we vary the training set size in Figure 4. Indeed, for CIFAR-10 and a standard deep learning model we observe that the increase in test error upon data augmentation decreases as the number of samples increases. This suggests that the ”tradeoff” between robustness and accuracy commonly observed is a finite sample effect of augmenting a small training set, just like in our analysis of linear regression.



(a) CIFAR-10 (b) CIFAR-10 (c) Splines

Figure 4: **(a), (b)** Effect of covariate-shifted data augmentation via adversarial training (AT) on the error, as a function of the # of original training samples n. Results are for a WRN-40-2 model trained on CIFAR-10. Shaded regions represent 1 STD. With vanilla AT, data augmentation leads to an increase in standard test error, and this harmful effect diminishes as n increases. AT with X-regularization (b) has lower error than vanilla AT. Further, X-regularization has *lower* error than even the standard estimator showing that X-regularization can even take advantage of data augmentation when it hurts standard training. **(c)** X-regularization also eliminates the increase in error upon data augmentation in the spline staircase where points on a line are augmented with local perturbations.

4 LEVERAGING UNLABELED DATA TO ELIMINATE INCREASE IN BIAS

To this point, the paper details ways in which incorporating additional data can make predictors worse. To complement this somewhat negative message, we introduce *X-regularization*, a regularization methodology that leverages unlabeled data to appropriately smooth a predictor, allowing data augmentation while guaranteeing that it (at least) does not decrease accuracy. We propose the most concrete instantiation of this in the context of interpolation in linear regression (Section 4.1), dovetailing with our development in Section 3, then generalize X-regularization to more general problems in Section 4.2. While we do not provide theoretical guarantees in fully general cases, we perform a corresponding empirical evaluation by applying X-regularization in an adversarial training problem with ℓ_∞ perturbations on CIFAR-10, observing that it both mitigates the undesirable drop in standard accuracy that standard adversarial training engenders and simultaneously improves robustness.

4.1 X-REGULARIZATION FOR LINEAR REGRESSION

Our development is most compelling in a stylized setting with noiseless observations from a linear model, $y = x^\top \theta^\star$, though the dimension is (much) larger than the number of observations, as in the now familiar interpolating regime Ma et al. (2018); Belkin et al. (2018). In this case, when the data x has population covariance Σ, the predictive risk of a point θ is $R(\theta) = (\theta - \theta^\star)^\top \Sigma (\theta - \theta^\star)$. Let us suppose as usual that we have pairs $(X_{\text{std}}, Y_{\text{std}})$ and $(X_{\text{ext}}, Y_{\text{ext}})$, and let $\theta_{\text{int-std}}$ interpolate the initial data, satisfying $X_{\text{std}} \theta_{\text{int-std}} = Y_{\text{std}}$. We would like to use $\theta_{\text{int-std}}$ to construct an estimator $\hat{\theta}_{\text{X-aug}}$ that interpolates both the standard data and augmented data while satisfying $R(\hat{\theta}_{\text{X-aug}}) \le R(\theta_{\text{int-std}})$.

To motivate our estimator, recall that while the error of the augmented estimator is smaller in ℓ_2 norm than the standard estimator, the increase in bias upon augmentation occurs because the augmented error could be larger than the standard error in the directions in which Σ is large (recall Figure 2). A natural strategy, then, is to fit the augmented data X_{ext} while staying *close* to $\theta_{\text{int-std}}$ weighted by Σ that determines the population risk.

Thus, we propose the *X-regularized estimator*, which given Σ and an initial interpolant $\theta_{\text{int-std}}$ sets

$$\hat{\theta}_{\text{X-aug}} = \operatorname*{argmin}_{\theta} \left\{ (\theta - \theta_{\text{int-std}})^\top \Sigma (\theta - \theta_{\text{int-std}}) : X_{\text{std}} \theta = Y_{\text{std}}, X_{\text{ext}} \theta = Y_{\text{ext}} \right\}. \tag{8}$$

The assumption that we have Σ deserves some comment. For x following distribution P_{x} with covariance Σ, we have $\|\theta - \theta_{\text{int-std}}\|_\Sigma^2 = \mathbb{E}_{P_{\mathsf{x}}}[(x^\top \theta - x^\top \theta_{\text{int-std}})^2]$, so unlabeled data from P_{x} can provide an arbitrarily accurate estimate of Σ. As we discuss presently, such unlabeled data is available in many settings. With the definition (8) of the X-regularized estimator, we have the following.

Theorem 3. *Assume the noiseless linear model $y = x^\top \theta^\star$. Let $\theta_{\text{int-std}}$ be an arbitrary interpolant of the standard data, i.e. $X_{std}\theta_{\text{int-std}} = Y_{std}$. Let $\hat{\theta}_{\text{X-aug}}$ be the X-regularized interpolant* (8)*. Then*

$$R\big(\hat{\theta}_{\text{X-aug}}\big) \leq R(\theta_{\text{int-std}}).$$

See Appendix D for proof. To provide some graphical intuition for the result, consider the spline interpolant $\hat{\theta}_{\text{std}}$ Fig. 1 illustrates. In this case, the marginal distribution P_{x} on x puts higher mass on points on the diagonal $\mathcal{T}_{\text{line}} := \{(t, f(t) = t)\}$. The X-regularized estimator $\hat{\theta}_{\text{X-aug}}$ thus both interpolates observed perturbations—as desired—while matching $\hat{\theta}_{\text{std}}$ wherever there is *not* augmented data, so that it is linear on $\mathcal{T}_{\text{line}}$. Thus, $\hat{\theta}_{\text{X-aug}}$ incorporates the local structure X_{ext} identifies, while the X-regularization (8) means there is no compromise of the global structure of $\hat{\theta}_{\text{std}}$.

4.2 General X-regularization

Theorem 3 holds for *any* interpolant on the training data, highlighting the generality of X-regularization and it is natural to go beyond the linear regime. Revisiting the estimator (8), we see roughly that it performs two tasks: achieving small error on the available data $(X_{\text{std}}, Y_{\text{std}})$ and $(X_{\text{ext}}, Y_{\text{ext}})$, while keeping the predictions that $\hat{\theta}_{\text{X-aug}}$ makes close to those of $\theta_{\text{int-std}}$, weighted by the distribution of the (unlabeled) data x. We can leverage unlabeled data to perform the general form of X-regularization. To do so, we slightly generalize our setting. Now, we assume we have a domain $\mathcal{X}$, target set $\mathcal{Y}$ and vector-valued prediction functions $f_\theta : \mathcal{X} \to \mathbb{R}^k$ indexed by parameter θ; we also have a loss $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ measuring the error in a prediction $f(x)$ for a true label y and some distance-like measure $\text{dist} : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$ that measures similarity in predictions. (In the case of regression, both of these are simply the squared error.) Now, suppose that we have a collection of m unlabeled samples $\widetilde{x}_i \overset{i.i.d}{\sim} P_{\text{x}}$, a model $\hat{\theta}_{\text{std}}$ trained on the original data $(X_{\text{std}}, Y_{\text{std}})$, and an augmented data set $(X_{\text{ext}}, Y_{\text{ext}})$, jointly consisting of N samples. Then the general X-regularized estimator is

$$\hat{\theta}_{\text{X-aug}} := \operatorname*{argmin}_{\theta} \left\{ N^{-1} \sum_{(x,y) \in [X_{\text{data}}, Y_{\text{data}}]} \ell(f_\theta(x), y) + \lambda m^{-1} \sum_{i=1}^{m} \text{dist}\Big(f_\theta(\tilde{x}), f_{\hat{\theta}_{\text{std}}}(\widetilde{x}_i)\Big) \right\}, \tag{9}$$

where λ is a regularization multiplier and $X_{\text{data}} = [X_{\text{std}}; X_{\text{ext}}], Y_{\text{data}} = [Y_{\text{std}}; Y_{\text{ext}}]$. In general, the optimal value of λ depends on the quality of $\hat{\theta}_{\text{std}}$. For the squared loss (regression), the estimator (9) is a Lagrangian form of the estimator (8), where the empirical expectation over P_{x} replaces its population counterpart. We investigate empirical performance of the estimator (8) in the next section. We can view X-regularization as a generalization of the classical self-training (Rosenberg et al., 2005), where in addition to unlabeled data from the right distribution, we have additional labeled data from a covariate-shifted distribution.

4.3 Empirical performance of X-regularization

Using X-regularization for linear regression as in Equation (8) for the spline staircase problem eliminates the increase in error as implied in Theorem 3. General X-regularization (Equation (9)) however can be applied broadly to any loss and function class like neural networks. We return to augmenting CIFAR-10 with ℓ_∞ adversarial examples that leads to increases error and test whether X-regularizationcan mitigate this effect. We use the same WRN-40-2 and compare the test accuracies of the AT with and without X-regularization. We obtain the unlabeled data required for X-regularization by following the procedure employed in (Carmon et al., 2019) and source 500K unlabeled images from 80 Million TinyImages. The parameter λ in equation (9) is chosen via hyperparameter search. The exact training procedure can be found in Appendix E.

We compare the standard test accuracies obtained by standard training, vanilla adversarial training and X-regularized adversarial training for ℓ_∞ perturbations in Figure 4(b). We observe that X-regularization mitigates the risk increase upon augmentation with adversarial examples, even outperforming the standard estimator. This benefit is more pronounced in the small sample regime, since the standard estimator becomes more accurate with increasing labeled sample size. Finally, adversarial training is typically used to improve the test robustness of a model. We find that X-regularized adversarial training exhibits robust accuracies within $1-2\%$ of vanilla adversarial training, for small training sets, and matches vanilla adversarial training as the standard training set increases (see Figure 6 in Appendix). Therefore, empirically, we find that X-regularization applied on retains the benefits of adversarial training while mitigating its harmful effects on the standard accuracy.

5 Related work and discussion

Robustness via semisupervised learning (RSL). Unlabeled data in the context of adversarial training (that we call RSL) has also been explored in several recent works (Carmon et al., 2019; Najafi et al., 2019; Uesato et al., 2019). Morally, RSL differs from our work in that the metric of interest in both the theoretical and empirical study is adversarial robustness, while we focus on the standard (unperturbed) error metric. RSL views unlabeled data as ”extra samples” to break the sample complexity barrier of robustness (Schmidt et al., 2018), while we use unlabeled data to regularize around the standard non-augmented estimator. Operationally, we differ from RSL by *not* adversarially perturbing the unlabeled data while training. Hence empirically, our X-regularization has lower standard error but higher robust error than RSL. Our theoretical results complement RSL work by showing that in addition to robustness, unlabeled data can provably also lower standard error.

Decreasing standard error of adversarial training. Mitigating the undesirable increase in standard error upon adversarial training has been a recent topic of interest. Better neural network architectures found via Neural Architecture Search (Zoph & Le, 2016) and improved training methods such as via mixup interpolated training (Lamb et al., 2019) have shown some success in improving the standard error of adversarially trained networks. Orthogonal to such approaches which focus on the optimization/training aspect of neural networks, our proposed X-regularization is statistically motivated (by studying a convex well-specified problem) and we leverage additional unlabeled data to obtain gains. We posit that these approaches could be used in conjunction to see further gains.

References

M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning (ICML)*, 2018.

Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

S. Chen, E. Dobriban, and J. H. Lee. Invariance reduces variance: Understanding data augmentation in deep learning and beyond. *arXiv preprint arXiv:1907.10905*, 2019.

D. C. Ciresan, U. Meier, J. M., L. M. Gambardella, and J. Schmidhuber. High-performance neural networks for visual object classification. *arXiv*, 2011.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research (JMLR)*, 17(83):1–5, 2016.

L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, pp. 1802–1811, 2019.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA: Springer series in statistics New York, NY, USA:, 2001 2001.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

D. Kleinman and M. Athans. The design of suboptimal linear time-varying systems. *IEEE Transactions on Automatic Control*, 13:150–159, 1968.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1097–1105, 2012.

A. Lamb, V. Verma, J. Kannala, and Y. Bengio. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. *arXiv*, 2019.

T. Liang and A. Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *International Conference on Machine Learning (ICML)*, 2018.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

A. Najafi, S. Maeda, M. Koyama, and T. Miyato. Robustness to adversarial perturbations in learning from incomplete data. *arXiv preprint arXiv:1905.13021*, 2019.

P. Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. In *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision*, 2005.

L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5014–5026, 2018.

J. Uesato, J. Alayrac, P. Huang, R. Stanforth, A. Fawzi, and P. Kohli. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

L. Yaeger, R. Lyon, and B. Webb. Effective training of a neural network character classifier for word recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 807–813, 1996.

F. Yang, Z. Wang, and C. Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. *arXiv preprint arXiv:1906.11235*, 2019.

S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

B. Zoph and Q. V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

A Appendix

A.1 Underparameterized regime—standard ERM.

In this section consider the regime where Σ_{std} is invertible. We refer to this as the underparameterized regime—this is true asymptotically for finite d, and for some distributions $\mathbb{P}_x$ in the high dimensional asymptotic regime with $\frac{d}{n} \to \gamma < 1$ (Hastie et al., 2019). In this case, both the standard estimator $\hat{\theta}_{\text{std}}$ and the augmented estimator $\hat{\theta}_{\text{aug}}$ are unbiased. The variance of $\hat{\theta}_{\text{aug}}$ is *always* lower than $\hat{\theta}_{\text{std}}$. We define the normalized covariance matrices (for constant eigenvalues as n grows) $\Sigma_{\text{std}} := \frac{1}{n} X_{\text{std}}^\top X_{\text{std}}$ and $\Sigma_{\text{ext}} := \frac{1}{\alpha n} X_{\text{ext}}^\top X_{\text{ext}}$. Formally, we have the following.

Proposition 1. *When Σ_{std} is invertible, the predictive risk of the minimum-augmented estimator $\hat{\theta}_{aug}$ is always smaller than the predictive risk of the standard estimator.*

$$R(\hat{\theta}_{aug}) - R(\hat{\theta}_{std}) = \frac{\sigma^2}{n} \operatorname{tr}\Big((\Sigma_{std} + \alpha \Sigma_{ext})^{-1} - \Sigma_{std}^{-1}) \Sigma \Big) \leq 0. \tag{10}$$

In particular, the predictive risk of the data augmented estimator $R(\hat{\theta}_{aug})$ is never greater than the predictive risk of the standard estimator $R(\hat{\theta}_{std})$, showing that data augmentation never hurts.

Proof. Plugging in $\Sigma_{\text{data}} = \Sigma_{\text{std}} + \alpha\Sigma_{\text{ext}}$ (for the augmented estimator) and $\Sigma_{\text{data}} = \Sigma_{\text{std}}$ (for the standard estimator) into Equations (**??**), (5) yields zero and variance difference of $V(\hat{\theta}_{\text{std}}) - V(\hat{\theta}_{\text{aug}}) = \operatorname{tr}\Big((\Sigma_{\text{std}}^{-1} - (\Sigma_{\text{std}} + \alpha\Sigma_{\text{ext}})^{-1})\Sigma\Big)$. Since Σ_{std} is invertible and $\Sigma_{\text{ext}} \succeq 0$, we have $\Sigma_{\text{std}}^{-1} - (\Sigma_{\text{std}} + \alpha\Sigma_{\text{ext}})^{-1} \succeq 0$. Multiplying with another PSD matrix Σ gives us

$$\begin{aligned} &\operatorname{tr}\Big((\Sigma_{\text{std}}^{-1} - (\Sigma_{\text{std}} + \alpha\Sigma_{\text{ext}})^{-1})\Sigma\Big) \geq 0 \\ &\qquad\qquad \implies V(\hat{\theta}_{\text{std}}) - V(\hat{\theta}_{\text{aug}}) \geq 0. \end{aligned} \tag{11}$$

Therefore, in terms of the predictive risk, we have $R(\hat{\theta}_{\text{std}}) = B(\hat{\theta}_{\text{std}}) + V(\hat{\theta}_{\text{std}}) \geq R(\hat{\theta}_{\text{aug}})$, and the data augmented estimator never has worse performance in this regime. □

Note that when Σ_{std} is invertible, minimum norm interpolation reduces to unregularized empirical risk minimization (ERM). In recent work, Chen et al. (2019) showed that a particular form of data augmentation (using invariance groups) never hurts performance of unregularized ERM in general. Essentially this is saying that if we have enough data to cover the space, augmentation never hurts but decreases the variances.

A.2 Proof of Theorem 1

We prove the two statements of Theorem 1 separately.

A.2.1 Proof of inequality (6)

Inequality (6) follows from

$$\begin{aligned} B(\hat{\theta}_{\text{aug}}) - B(\hat{\theta}_{\text{std}}) &= (\theta^\star - \hat{\theta}_{\text{aug}})^\top \Sigma (\theta^\star - \hat{\theta}_{\text{aug}}) - (\theta^\star - \hat{\theta}_{\text{std}})^\top \Sigma (\theta^\star - \hat{\theta}_{\text{std}}) \\ &= (\Pi_{\text{aug}}^\perp \theta^\star)^\top \Sigma \Pi_{\text{aug}}^\perp \theta^\star - (\Pi_{\text{std}}^\perp \theta^\star)^\top \Sigma \Pi_{\text{std}}^\perp \theta^\star \\ &= w^\top \Sigma w - (w+v)^\top \Sigma (w+v) \\ &= 2w^\top \Sigma v - v^\top \Sigma v \end{aligned} \tag{12}$$

by decomposition of $\Pi_{\text{std}}^\perp \theta^\star = v + w$ where $v = \Pi_{\text{std}}^\perp \Pi_{\text{aug}} \theta^\star$ and $w = \Pi_{\text{std}}^\perp \Pi_{\text{aug}}^\perp \theta^\star$.

A.2.2 Proof of lower bound on $\|\theta^\star\|_2^2 - \|\hat{\theta}_{\text{STD}}\|_2^2$

The proofs of Theorem 1 (b) and (c) are based on the following two lemmas that follow from simple linear algebra but are nonetheless used for characterization purposes as well.

Lemma 1. *If a PSD matrix Σ has non-equal eigenvalues, one can find two unit vectors w,v for which the following holds*

$$w^\top v = 0 \qquad \textit{and} \qquad w^\top \Sigma v \neq 0 \tag{13}$$

Note that neither w nor v can be eigenvectors of Σ in order for both conditions in equation (13) to hold. Notice that it doesn't matter if the Σ inner product smaller or bigger than zero.

Note that the bias difference scales with $\|\theta^\star\|^2$. Combining the two former lemmas allows an explicit construction of $\theta^\star$ for which augmentation hurts bias.

Lemma 2. *Assume Σ, X_{std}, X_{ext} are fixed. Condition (13) holds for two directions $v \in col(\Pi_{std}^\perp \Pi_{aug})$ and $w \in col(\Pi_{std}^\perp \Pi_{aug}^\perp)$ iff there exists a $\theta^\star$ such that $B(\hat{\theta}_{aug}) - B(\hat{\theta}_{std}) \geq c > 0$ for some $c > 0$. Furthermore, the ℓ_2 norm of $\theta^\star$ needs to satisfy the following lower bounds with $c_1 := \|\hat{\theta}_{aug}\|^2 - \|\hat{\theta}_{std}\|^2$*

$$\begin{aligned} \|\theta^\star\|^2 - \|\hat{\theta}_{aug}\|^2 &\geq \beta_1 c_1 + \beta_2 \frac{c^2}{c_1} \\ \|\theta^\star\|^2 - \|\hat{\theta}_{std}\|^2 &\geq (\beta_1 + 1) c_1 + \beta_2 \frac{c^2}{c_1} \end{aligned} \tag{14}$$

where β_i are constants that depend on X_{std}, X_{ext}, Σ.

Theorem 1(b) follows directly from the second statement of Lemma 2 by minimizing the bound (14) with respect to c_1 which is a free parameter to be chosen during construction of $\theta^\star$ (see proof of Lemma (2). The minimum is attained for $c_1 = 2\sqrt{(\beta_1+1)(\beta_2 c^2)}$. We hence conclude that $\theta^\star$ needs to be sufficiently more complex than a good standard solution, i.e. $\|\theta^\star\|_2^2 - \|\hat{\theta}_{\text{std}}\|_2^2 > \alpha c$ where $\alpha > 0$ is a constant that depends on the $X_{\text{std}}, X_{\text{ext}}$.

A.3 Proof of Lemma 2

We first construct $\Sigma_{\text{std}}, \Sigma_{\text{ext}}$ using w, v from which we can reconstruct $X_{\text{std}}, X_{\text{ext}}$ use any standard decomposition method to obtain $\Sigma = \frac{1}{n} X^\top X$ for any desired $|X| = n$ where $|X|$ is the number of rows of X. Given $\Sigma_{\text{std}}, \Sigma_{\text{ext}}$ we construct $\theta^\star$ for which the inequality (6) in Theorem 1 (b) holds.

Construct $\Sigma_{\text{std}}, \Sigma_{\text{ext}}$ Let's construct $\Sigma_{\text{std}}, \Sigma_{\text{ext}}$ using w, v. Wlog we can make them simultaneously diagonalizable. So we construct a set of eigenvectors that is the same for both matrices and different eigenvalues. Let the eigenvectors include w, v. Then if we set the corresponding eigenvalues $\lambda_w(\Sigma_{\text{ext}}) = 0, \lambda_v(\Sigma_{\text{ext}}) > 0$ and $\lambda_w(\Sigma_{\text{std}}) = 0, \lambda_v(\Sigma_{\text{std}}) = 0$ (hence $w \in \text{col}(\Pi_{\text{aug}}^\perp)$ and $v \in \text{col}(\Pi_{\text{std}}^\perp \Pi_{\text{aug}})$ and $\lambda_w(\Sigma_{\text{aug}}) = 0$) and we can design a $\theta^\star$ that creates tradeoff as follows.

Construct $\theta^\star$ We now construct a $\theta^\star$ such that inequality (6) holds. First choose some arbitrary $\hat{\theta}_{\text{std}} \in \text{col}(\Sigma_{\text{std}})$. One can decompose the space $\text{Ker}(\Sigma_{\text{std}})$ into the direct sum of two orthogonal subspaces

$$\begin{aligned} \text{Ker}(\Sigma_{\text{std}}) &= \text{Ker}(\Sigma_{\text{aug}}) \oplus \text{col}(\Pi_{\text{std}}^\perp \Pi_{\text{aug}}) \\ &= \text{col}(\Pi_{\text{std}}^\perp \Pi_{\text{aug}}^\perp) \oplus \text{col}(\Pi_{\text{std}}^\perp \Pi_{\text{aug}}) \end{aligned}$$

and hence using the minimum-norm property, we can always decompose the (rotated) augmented $\hat{\theta}_{\text{aug}} \in \text{col}(\Pi_{\text{aug}}^\perp)$ and true parameter

$$\begin{aligned} \hat{\theta}_{\text{aug}} &= \hat{\theta}_{\text{std}} + \sum_{i \in \text{ext}} \zeta_i v_i \\ \theta^\star &= \hat{\theta}_{\text{aug}} + \sum_{j \in \text{rest}} \xi_j w_j. \end{aligned}$$

Now if we require the tradeoff to be some $c > 0$, rewrite the bias condition using the identity (6)

$$\begin{aligned} & B(\hat{\theta}_{\text{aug}}) - B(\hat{\theta}_{\text{std}}) = c \\ \Longleftrightarrow\ & (\sum_{i \in \text{ext}} \zeta_i v_i)^\top \Sigma (\sum_{i \in \text{ext}} \zeta_i v_i) + c = -2(\sum_{j \in \text{rest}} \xi_j w_j) \Sigma (\sum_{i \in \text{ext}} \zeta_i v_i) \\ \Longleftrightarrow\ & (\sum_{i \in \text{ext}} \zeta_i v_i)^\top \Sigma (\sum_{i \in \text{ext}} \zeta_i v_i) + c = -2 \sum_{j,i} \xi_j \zeta_i w_j^\top \Sigma v_i \end{aligned} \tag{15}$$

The left hand side of equation (A.3) is always positive, hence it is necessary that there exists at least one pair i,j such that $w_j^\top \Sigma v_i \neq 0$. By construction we know that this is the case and we assume wlog that $\xi_j \zeta_i w_j^\top \Sigma v_i < 0$ (since if positive reverse the signs). For this particular choice of i (or more) we set $\zeta_i \neq 0$ such that $\|\hat{\theta}_{\text{aug}} - \hat{\theta}_{\text{std}}\|^2 = \|\zeta\|^2 = c_1 > 0$, i.e. that the augmented estimator is not equal to the standard estimator (else obviously there can be no difference in bias and equation cannot be satisfied for any desired bias increase $c > 0$).

The choice of ξ_j minimizing $\sum_j \xi_j^2$ is in the direction of the vector $x = W^\top \Sigma V \zeta$ with $x_j = \sum_i \zeta_i w_j^\top \Sigma v_i$. Defining $c_0 = (\sum_{i \in \text{ext}} \zeta_i v_i)^\top \Sigma (\sum_{i \in \text{ext}} \zeta_i v_i)$ for convenience and then setting

$$\xi = -\frac{c_0 + c}{2\|x\|_2^2} x \tag{16}$$

which is well-defined as $x \neq 0$ and yields a $\theta^\star$ such that augmentation hurts. It is thus necessary for $B(\hat{\theta}_{\text{aug}}) - B(\hat{\theta}_{\text{std}}) = c$ that

$$\begin{aligned}\sum_j \xi_j^2 &= \frac{(c_0+c)^2}{4\|W^\top \Sigma V \zeta\|^2} = \frac{(\zeta^\top V^\top \Sigma V \zeta + c)^2}{4\zeta^\top V^\top \Sigma W W^\top \Sigma V \zeta} \\ &\geq \frac{(\zeta^\top V^\top \Sigma V \zeta)^2}{4\zeta^\top V^\top \Sigma W W^\top \Sigma V \zeta} + \frac{c^2}{4\zeta^\top V^\top \Sigma W W^\top \Sigma V \zeta} \\ &\geq \frac{c_1}{4} \frac{\lambda_{\min}^2(V^\top \Sigma V)}{\lambda_{\max}^2(W^\top \Sigma V)} + \frac{c^2}{4 c_1 \lambda_{\max}^2(W^\top \Sigma V)}\end{aligned}$$

Notice that by construction of V we have $\lambda_{\min}(V^\top \Sigma V) > 0$ and $\lambda_{\max}(W^\top \Sigma V) > 0$

Note due to construction we have $\|\theta^\star\|_2^2 = \|\hat{\theta}_{\text{std}}\|_2^2 + \sum_i \zeta_i^2 + \sum_j \xi_j^2$ and plugging in the choice of ξ_j in equation (16) we have

$$\|\theta^\star\|_2^2 - \|\hat{\theta}_{\text{std}}\|_2^2 \geq c_1 \left[1 + \frac{\lambda_{\min}^2(V^\top \Sigma V)}{4\lambda_{\max}^2(W^\top \Sigma V)}\right] + \frac{c^2}{4\lambda_{\max}^2(W^\top \Sigma V)} \frac{1}{c_1}$$

where we define $V \in \mathbb{R}^{d \times m}$ where m is the dimension of $\text{col}(\Pi_{\text{std}}^\perp \Pi_{\text{aug}}^\perp)$ and $\zeta \in \mathbb{R}^m$. Setting $\beta_1 = \left[1 + \frac{\lambda_{\min}^2(V^\top \Sigma V)}{4\lambda_{\max}^2(W^\top \Sigma V)}\right]$, $\beta_2 = \frac{1}{4\lambda_{\max}^2(W^\top \Sigma V)}$ yields the result.

Some more observations. For $\|\theta^\star\|_2$ larger than the minimum, we can allow $\|\hat{\theta}_{\text{aug}} - \hat{\theta}_{\text{std}}\|_2$ to be smaller than the c_1 that achieves the last inequality.

A.3.1 Proof of Lemma 1

Let λ_i be the m non-zero eigenvalues of Σ and u_i be the corresponding eigenvectors. Then choose v to be any combination of the eigenvectors $v = U\beta$ where $U = [u_1, ..., u_m]$ where at least $\beta_i, \beta_j \neq 0$ for $\lambda_i \neq \lambda_j$. We next construct $w = U\alpha$ by choosing α as follows such that the inequality in (13) holds:

$$\begin{aligned}\alpha_i &= \frac{\beta_j}{\beta_i^2 + \beta_j^2} \\ \alpha_j &= \frac{-\beta_i}{\beta_i^2 + \beta_j^2}\end{aligned}$$

Then we have that $\alpha^\top \beta = 0$ and hence $w^\top v = 0$ and simultaneously

$$\begin{aligned}w^\top \Sigma v &= \lambda_i \beta_i \alpha_i + \lambda_2 \beta_j \alpha_j \\ &= (\lambda_i - \lambda_j) \frac{\beta_i \beta_j}{\beta_i^2 + \beta_j^2} \neq 0\end{aligned}$$

which concludes the proof of the lemma.

A.4 Characterization Corollary 1

Given fixed $X_{\text{std}}, \Sigma, \theta^\star$, the following corollary characterizes which single augmentation directions do and do not lead to higher prediction error for the augmented estimator.

Corollary 1. *The following characterizations hold for augmentation directions that do not cause the bias of the augmented estimator to be higher than the original estimator*

(a) (in terms of ratios of inner products) *For a given $\theta^\star$, data augmentation does not increase the bias of the augmented estimator for a single augmentation direction x_{ext} if*

$$\frac{x_{ext}^\top \Pi_{std}^\perp \Sigma \Pi_{std}^\perp x_{ext}}{x_{ext}^\top \Pi_{std}^\perp x_{ext}} - 2\frac{(\Pi_{std}^\perp x_{ext})^\top \Sigma \Pi_{std}^\perp \theta^\star}{x_{ext}^\top \Pi_{std}^\perp \theta^\star} \leq 0 \tag{17}$$

(b) (in terms of eigenvectors) *Data augmentation does not increase bias for any $\theta^\star$ if $\Pi_{std}^\perp x_{ext}$ is an eigenvector of Σ. However if one augments in the direction of a mixture of eigenvectors of Σ with different eigenvalues, there exists a $\theta^\star$ such that augmentation hurts.*

The form in Equation (17) compares ratios of inner products of $\Pi_{\text{std}}^\perp x_{\text{ext}}$ and $\Pi_{\text{std}}^\perp \theta^\star$ in two spaces: the one in the numerator is weighted by Σ whereas the denominator is the standard inner product. Thus, if Σ scales and rotates rather inhomogeneously, then augmenting with x_{ext} may hurt bias. Here again, if $\Sigma = \gamma I$ for $\gamma > 0$, then the condition must hold.

A.4.1 Proof of Corollary 1 (a) - safe single augmentation directions dependent on $\theta^\star$

The proof of Corollary 1 (a) is a direct result of condition (6) when restricted to one augmentation point. Let the one augmentation data point be x_{ext}. When $X_{\text{ext}} = x_{\text{ext}}^\top$, then $\bar{X}_{\text{ext}} = X_{\text{ext}} \Pi_{\text{std}}^\perp = x_{\text{ext}}^\top \Pi_{\text{std}}^\perp$. Then by definition from Lemma **??**,

$$\begin{aligned} a &= (\bar{X}_{\text{ext}}{}^\top \bar{X}_{\text{ext}})^\dagger \bar{X}_{\text{ext}}{}^\top \bar{X}_{\text{ext}} \theta^\star \\ &= ((\Pi_{\text{std}}^\perp x_{\text{ext}})(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top)^\dagger ((\Pi_{\text{std}}^\perp x_{\text{ext}})(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top)\theta^\star \\ &= \frac{1}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \left(\left(\frac{\Pi_{\text{std}}^\perp x_{\text{ext}}}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|} \right) \left(\frac{\Pi_{\text{std}}^\perp x_{\text{ext}}}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|} \right)^\top \right) ((\Pi_{\text{std}}^\perp x_{\text{ext}})(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top)\theta^\star \\ &= \frac{1}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \left(\Pi_{\text{std}}^\perp x_{\text{ext}} \left(\frac{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top (\Pi_{\text{std}}^\perp x_{\text{ext}})}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \right) \right) (\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \theta^\star \\ &= \frac{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \theta^\star}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \Pi_{\text{std}}^\perp x_{\text{ext}} \end{aligned}$$

Substituting this a into condition (**??**),

$$\begin{aligned} \left(\frac{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \theta^\star}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \right)^2 x_{\text{ext}}^\top \Pi_{\text{std}}^\perp \Sigma \Pi_{\text{std}}^\perp x_{\text{ext}} &\le 2 \left(\frac{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \theta^\star}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \right) (\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \Sigma \Pi_{\text{std}}^\perp \theta^\star \\ \iff x_{\text{ext}}^\top \Pi_{\text{std}}^\perp \Sigma \Pi_{\text{std}}^\perp x_{\text{ext}} &\le 2 \left(\frac{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2}{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \theta^\star} \right) (\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \Sigma \Pi_{\text{std}}^\perp \theta^\star \\ \iff \frac{x_{\text{ext}}^\top \Pi_{\text{std}}^\perp \Sigma \Pi_{\text{std}}^\perp x_{\text{ext}}}{x_{\text{ext}}^\top \Pi_{\text{std}}^\perp x_{\text{ext}}} &\le 2 \frac{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \Sigma \Pi_{\text{std}}^\perp \theta^\star}{x_{\text{ext}}^\top \Pi_{\text{std}}^\perp \theta^\star}. \end{aligned}$$

Notice that the bias difference scales as

$$B(\hat{\theta}_{\text{aug}}) - B(\hat{\theta}_{\text{std}}) = \frac{[(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \theta^\star]^2}{\|\Pi_{\text{std}}^\perp x_{\text{ext}}\|^2} \left[\frac{x_{\text{ext}}^\top \Pi_{\text{std}}^\perp \Sigma \Pi_{\text{std}}^\perp x_{\text{ext}}}{x_{\text{ext}}^\top \Pi_{\text{std}}^\perp x_{\text{ext}}} - 2 \frac{(\Pi_{\text{std}}^\perp x_{\text{ext}})^\top \Sigma \Pi_{\text{std}}^\perp \theta^\star}{x_{\text{ext}}^\top \Pi_{\text{std}}^\perp \theta^\star} \right]$$

and hence does scale with the ℓ_2 norm of $\theta^\star$. Safe augmentation directions for specific choices of $\theta^\star$ and Σ are illustrated in Figure 2.

A.4.2 Proof of Corollary 1 (b) - safe and hurtful augmentation directions independent of $\theta^\star$

The proof of this statement follows directly from the iff statement in Lemma 2. When one adds one data point x_{ext}, $\text{col}(\Pi_{\text{std}}^\perp \Pi_{\text{aug}}) = \text{span}(X_{\text{ext}})$ and $u = \Pi_{\text{std}}^\perp x_{\text{ext}}$. Using Corollary 1 (a), it follows that if v is an eigenvector of Σ with eigenvalue $\lambda > 0$ in the nullspace of Σ_{std}, we have

$$v^\top \Sigma v - 2 \frac{u^\top \Sigma \Pi_{\text{std}}^\perp \theta^\star}{u^\top \Pi_{\text{std}}^\perp \theta^\star} = -\lambda < 0$$

for any $\theta^\star$. Hence by Lemma 2 the bias doesn't increase by augmenting with eigenvectors of Σ for any $\theta^\star$.

When the single augmentation direction v is not an eigenvector of Σ, by Lemma 1 one can find w such that $w^\top \Sigma v \neq 0$. The proof in Lemma 1 gives an explicit construction for w such that condition (13) holds and the result then follows directly by Lemma 2

B Details for spline example

B.1 Spline problem data distribution

We consider a finite input domain $\mathcal{T} = \{0, \epsilon, 1, 1+\epsilon, ..., s-1, s-1+\epsilon\}$ for some integer s. Let $\mathcal{T}_{\text{line}} \subset \mathcal{T} = \{0,1,...,s-1\}$. We define $\mathbb{P}_x$ such that $\mathbb{P}_x(\mathcal{T}_{\text{line}}) = 1-\delta$ for some small δ such that points not in $\mathcal{T}_{\text{line}}$ have low probability. We define the underlying function $f^\star : \mathbb{R} \mapsto \mathbb{R}$ as $f^\star(t) = \lfloor t \rfloor$. This function takes a staircase shape, and is linear when restricted to $\mathcal{T}_{\text{line}}$.

We describe the data distribution in terms of the one-dimensional input t, and by the one-to-one correspondence with $x = \Phi(t)$, this also defines the distribution of spline features $x \in \mathcal{X}$. Let s be the total number of "stairs" in the staircase problem. Define $\delta \in [0,1]$ to be the probability of sampling a perturbation point, i.e. $t \in \mathcal{T}_{\text{line}}^c$, which we will choose to be close to zero. The size of the perturbations is $\epsilon = \frac{1}{2}$, and $\lfloor t+\epsilon \rfloor = t$ for any $t \in \mathcal{T}_{\text{line}}$.

Let $w \in \Delta_s$ be a distribution over $\mathcal{T}_{\text{line}}$ where Δ_s is the probability simplex of dimension s. We define the data distribution with the following generative process for one sample t. First, sample a point i from $\mathcal{T}_{\text{line}}$ according to the categorical distribution described by w, such that $i \sim \text{Categorical}(w)$. Second, sample t by perturbing i with probability δ such that

$$t = \begin{cases} i & \text{w.p. } 1-\delta \\ i+\epsilon & \text{w.p. } \delta. \end{cases}$$

The sampled t is in $\mathcal{T}_{\text{line}}$ with probability $1-\delta$ and $\mathcal{T}_{\text{line}}^c$ with probability δ, where we choose δ to be small.

Augmentation For each element t_i in the training set, we augment with $\tilde{T}_i = [u \overset{u.a.r}{\sim} B(t_i)]$, an input chosen uniformly at random from $B(t_i) = \{\lfloor t_i \rfloor, \lfloor t_i \rfloor + \epsilon\}$. Recall that in our work, we consider data augmentation where the targets associated with the augmented points are from the ground truth oracle. Notice that by definition, $f^\star(\tilde{t}_i) = f^\star(t_i)$ for all $\tilde{t} \in B(t_i)$, and thus we can set the augmented targets to be $\tilde{y}_i = y_i$. This is similar to random data augmentation in images (Yaeger et al., 1996; Krizhevsky et al., 2012), where inputs are perturbed in a way that preserves the label.

In addition, in order to exaggerate the difference between augmented and standard estimators for small sample sizes, we set w such that the first $s_0 < s$ stairs have the majority of probability mass. To achieve this, we set the unnormalized probabilities of w as

$$\hat{w}_j = \begin{cases} 1/s_0 & j < s_0 \\ 0.01 & j \geq s_0 \end{cases}$$

and define w by normalizing $w = \hat{w}/\sum_j \hat{w}_j$. For our examples, we fix $s_0 = 5$.

B.2 Spline model

We parameterize the spline predictors as $f_\theta(t) = \theta^\top \Phi(t)$ where $\Phi : \mathbb{R} \to \mathbb{R}^d$ is the cubic B-spline feature mapping (Friedman et al., 2001 2001) and the norm of $f_\theta(t)$ can be expressed as $\theta^\top M \theta$ for a matrix M that penalizes a large second derivative norm. Notice that the splines problem is a linear regression problem from $\mathbb{R}^d$ to $\mathbb{R}$ in the feature domain $\Phi(t)$, allowing direct application of Theorem 1 with $[M]_{ij} = \int \Phi_i''(u) \Phi_j''(u) \mathrm{d}u$. As a linear regression problem, we define the finite domain as $\mathcal{X} = \{\Phi(t) : t \in \mathcal{T}\}$ containing $2s$ possible elements in $\mathbb{R}^d$. There is a one-to-one correspondence between t and $\Phi(t)$, such that Φ^{-1} is well-defined. We define the features that correspond to inputs in $\mathcal{T}_{\text{line}}$ as $\mathcal{X}_{\text{line}} = \{x : \Phi^{-1}(x) \in \mathcal{T}_{\text{line}}\}$. Using this feature mapping, there exists a $\theta^\star$ such that $f_{\theta^\star}(t) = f^\star(t)$ for $t \in \mathcal{T}$.

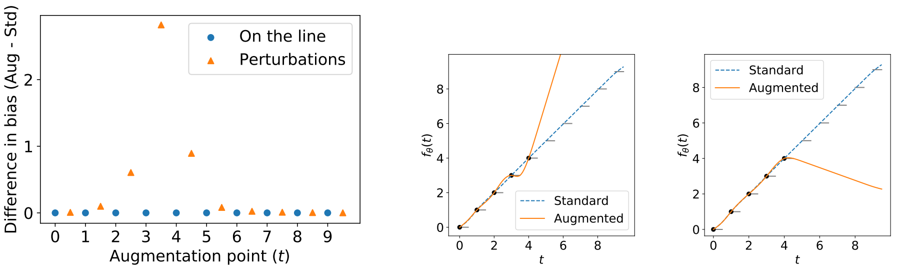
Our hypothesis class is the family of cubic B-splines as defined in (Friedman et al., 2001 2001). Cubic B-splines are piecewise cubic functions, where the endpoints of each cubic function are called the knots. In our example, we fix the knots to be $[0, \epsilon, 1, ..., s-1, s-1+\epsilon]$, which places a knot on every point in $\mathcal{T}$. This ensures that for some $\theta^\star$,

$$f_{\theta^\star}(t) = {\theta^\star}^\top \Phi(t) = f^\star(t) = \lfloor t \rfloor \tag{18}$$

such that $f^\star$ is in the hypothesis class.

We solve the minimum norm problem

$$\hat{\theta}_{\text{std}} = \underset{\theta}{\operatorname{argmin}} \{\theta^\top M \theta : X_{\text{std}} \theta = Y_{\text{std}}\} \tag{19}$$



(a)

(b) Augment with $x = \Phi(3.5)$

(c) Augment with $x = \Phi(4.5)$

Figure 5: Visualization of the effect of single augmentation points in the noiseless spline problem given an initial dataset $X_{\text{std}} = \{\Phi(t) : t \in \{0,1,2,3,4\}\}$. The standard estimator defined by X_{std} is linear. **(a)** Plot of the difference in predictive risk for all possible single augmentation points. Augmenting with points on $\mathcal{X}_{\text{line}}$ does not affect the bias, but augmenting with any element of $\{\Phi(t) : t \in \{2.5,3.5,4.5\}\}$ hurts the bias of the augmented estimator dramatically. **(b), (c)** Augmenting with $\Phi(3.5)$ or $\Phi(4.5)$ hurts the bias by changing the direction of extrapolation.

and the corresponding augmented problem for the augmented estimator. Here, $M_{i,j} = \int \Phi''(t)_i \Phi''(t)_j dt$ measures smoothness in terms of the second derivative.

We implement the optimization of the standard and robust objectives using the basis described in (Friedman et al., 2001 2001). The penalty matrix M computes second-order finite differences of the parameters θ. In Figure 1, we solve the min-norm objective directly using CVXPY (Diamond & Boyd, 2016).

B.3 Evaluating Corollary 1 for splines

To check our (single-point) characterization in Theorem 1 against possible augmentation points in the splines problem, we use the rotated spline features $\Phi(X)M^{-1/2}$. Note that in our case, M is not full rank. We add a small identity matrix ($(1\text{e}{-}10)I$) to M to make it invertible.

Now assume our original data $X_{\text{std}} = \{\Phi(t) : t \in \{0,1,2,3,4\}\}$ where $t \in \mathcal{T}_{\text{line}}$ as defined below (on integer points) and now we examine all possible single augmentation points corresponding to all points in $\mathcal{T}$ in Figure 5 (a) and plot the calculated predictive risk difference. Figure 5 shows that augmenting with an additional point from $\{\Phi(t) : t \in \mathcal{T}_{\text{line}}\}$ does not affect the bias, but adding any perturbation point in $\{\Phi(t) : t \in \{2.5,3.5,4.5\}\}$ where $t \notin \mathcal{T}_{\text{line}}$ increases the error significantly by changing the direction in which the estimator extrapolates. Particularly, *local* augmentations hurt while other augmentations do not significantly affect the bias of the augmented estimator.

B.4 Data augmentation can be quite painful for splines

Suppose a dataset $T = [t_1 \ldots, t_n]$ is provided. Let the domain of the problem be $\cup_{t=0}^{s-1}[t, t+\epsilon]$. Considering only s which is a multiple of 2, define the data distribution through the following generative process: first sample $t_0 \in \mathcal{T}_{\text{line}}$ from the following distribution

$$p(t) = \begin{cases} \frac{1-\gamma}{s/2} & t < s/2, t \in \mathcal{T}_{\text{line}} \\ \frac{\gamma}{s/2} & t \geq s/2, t \in \mathcal{T}_{\text{line}}. \end{cases} \tag{20}$$

for $\gamma \in [0,1)$. Then with probability $1-\delta$, return $t = t_0$. Otherwise with probability δ, return $t \sim \text{Uniform}([t_0, t_0+\epsilon])$. Consider a modified augmented estimator for the splines problem, where for each point t_i we augment with the entire interval $[\lfloor t_i \rfloor, \lfloor t_i \rfloor + \epsilon]$, where $\epsilon \in [0,1)$ is a constant and the target is $y_i = \lfloor t_i \rfloor$ on the whole interval. Additionally, suppose that the ratio $s/n = O(1)$ between the number of stairs s and the number of samples n is constant. Note that in the feature domain, this corresponds to $d/n = O(1)$ where $d = 2s+2$ is the dimensionality of the spline basis as in Friedman et al. (2001 2001).

Define the squared loss to be limited with knowledge of the number of stairs s, such that we have a risk function bounded with respect to s:

$$R(\hat{\theta})=\mathbb{E}_t[\min\{(s-1)^2,(\Phi(t)^\top\hat{\theta}-\Phi(t)^\top\theta^\star)^2\}\,|\,T]. \tag{21}$$

In this simplified setting, we can show that the risk of the augmented estimator grows while the risk of the standard estimator decays to 0.

Theorem 4. *Let the setting be defined as above. Then with the choice of* $\delta=\frac{\log(s^3)-\log(s^3-1)}{s}$ *and* $\gamma=c/s$ *for a constant* $c\in[0,1)$*, the ratio between risks is lower bounded as*

$$\frac{R(\hat{\theta}_{aug})}{R(\hat{\theta}_{std})}=\Omega(s^2) \tag{22}$$

which goes to infinity as $s\to\infty$*. Furthermore,* $R(\hat{\theta}_{std})\to 0$ *as* $s\to\infty$*.*

Proof. We first lower bound the risk of the augmented estimator. We lower bound by consider only the case where all the points are sampled from $\{t:t<s/2\}$, which occurs with probability $(1-\gamma)^n$. Let $t^\star=\max_i\lfloor t_i\rfloor$ be the largest "stair" value seen in the training set. Note that the min-norm augmented estimator will extrapolate with zero derivative for $t\geq\max_i\lfloor t_i\rfloor$. This is because on the interval $[t^\star,t^\star+\epsilon]$, the augmented estimator is forced to have zero derivative, and the solution minimizing the second derivative of the prediction continues with zero derivative for all $t\geq t^\star$. The best case min-norm solution in the case where all points are sampled from $\{t:t<s/2\}$ is that $t^\star=s/2-1$. As a result, for the $s/2$ stairs where $t>(s/2-1)$, the augmented estimator incurs large error:

$$\begin{aligned}
R(\hat{\theta}_{\text{aug}})&\geq(1-\gamma)^n\sum_{t=1}^{s/2}t^2\cdot\frac{\gamma}{s/2}\\
&=(1-\gamma)^n\frac{\gamma}{s/2}\cdot\frac{1}{12}(s^3+2s^2+s)\\
&=\frac{1}{6}\gamma(1-\gamma)^n(s^2+2s+1)\\
&\geq\frac{1}{6}\gamma(1-\gamma n)(s^2+2s+1)\\
&=\Omega(\frac{c-c^2}{s}(s^2+2s+1))\\
&=\Omega(s).
\end{aligned}$$

where in the first line, we note that the error on each interval is the same and the probability of each interval is $(1-\delta)\frac{\gamma}{s/2}+\epsilon\frac{\delta}{\epsilon}\cdot\frac{\gamma}{s/2}=\frac{\gamma}{s/2}$.

Next we upper bound the risk of the standard estimator. We first focus on the case where all points are sampled from $\mathcal{T}_{\text{line}}$, which occurs with probability $(1-\delta)^n$. In this case, the standard estimator is linear and fits the points on $\mathcal{T}_{\text{line}}$ with zero error, while incurring error for all points not in $\mathcal{T}_{\text{line}}$. Note that the probability density of sampling a point not in $\mathcal{T}_{\text{line}}$ is either $\frac{\delta}{\epsilon}\cdot\frac{1-\gamma}{s/2}$ or $\frac{\delta}{\epsilon}\cdot\frac{\gamma}{s/2}$, which we upper bound as $\frac{\delta}{\epsilon}\cdot\frac{1}{s/2}$. The contribution to the $R(\hat{\theta}_{\text{std}})$ for this case is upper bounded as

$$\begin{aligned}(1-\delta)^n\sum_{t=1}^{s-1}\frac{\delta}{\epsilon}\cdot\frac{1}{s/2}\int_0^{\epsilon}u^2du &= (1-e^{-\delta n})\frac{\delta}{\epsilon}\cdot\frac{1}{s/2}\sum_{t=1}^{s-1}\int_0^{\epsilon}u^2du\\ &= O(1-\frac{s^3-1}{s^3})\frac{\delta}{\epsilon}\cdot\frac{1}{s/2}\sum_{t=1}^{s-1}\int_0^{\epsilon}u^2du\\ &= O(\frac{1}{s^3})\frac{\delta}{\epsilon}\cdot\frac{1}{s/2}\sum_{t=1}^{s-1}\int_0^{\epsilon}u^2du\\ &= O(\frac{1}{s^3})\frac{\delta}{\epsilon}\cdot\frac{1}{s/2}O(s\epsilon^3)\\ &= O(\frac{\log(s^3)-\log(s^3-1)}{s^4})\\ &= O(1/s^4)\end{aligned}$$

since $\log(s^3)-\log(s^3-1)\leq 8/7$ for $s\geq 2$. For all other cases, we upper bound the error with the worst possible standard estimator, which has risk $R(\hat{\theta}_{\text{std}})\leq(s-1)^2$ by the bounded risk. Then the contribution to $R(\hat{\theta}_{\text{std}})$ for this case is upper bounded as

$$\begin{aligned}(1-(1-\delta)^n)(s-1)^2 &\leq (1-e^{-\delta n})(s-1)^2\\ &= O(\frac{1}{s^3})(s-1)^2\\ &= O(1/s)\end{aligned}$$

Thus overall, $R(\hat{\theta}_{\text{std}})=O(1/s)$ and the ratio $\frac{R(\hat{\theta}_{\text{aug}})}{R(\hat{\theta}_{\text{std}})}=\Omega(s^2)$. □

C Analysis of variance in the overparameterized regime

Previous work/standard results consider the underparameterized regime where both the standard and augmented estimators are unbiased. In thsi case, the variance (and equivalently the predictive risk) of the augmented estimator is never larger than that of the standard estimator. In contrast, we saw in Section 3.1 that in the overparameterized regime, the bias of an augmented estimator can be larger than the bias of the standard estimator, and this could lead to larger predictive risk upon augmentation. However, does the cariance show similar trends as the underparamterized regime or could augmented estimators also have larger variance than their standard counterparts?

Just like in Section 3.1, we consider minimum P-norm interpolation, where the covariates are rotated by $P^{\frac{-1}{2}}$, so that the minimum norm interpolant takes the form in Equation 3. Theorem **??** describes the effect of data augmentation on the variance.

C.1 Analysis of variance

In the previous subsections, we focused on the analysis of the predictive bias which is equivalent to the predictive risk of the estimator $R(\hat{\theta})$ when the observed targets Y are noiseless. In this section, we consider the case where the noise is non-zero, and compute the variances of the two estimators of interest: the standard estimator $\hat{\theta}_{\text{std}}$ and data augmented estimator $\hat{\theta}_{\text{aug}}$.

Theorem 5 (Variance). *The difference in the variances of a standard and augmented estimator can be expressed as follows.*

$$n(V(\hat{\theta}_{aug})-V(\hat{\theta}_{std}))=\underbrace{\mathrm{tr}\big(\Sigma\bar{X}_{ext}^{\dagger}(\bar{X}_{ext}^{\dagger})^{\top}\big)}_{T_1\text{: Variance increase}}-\underbrace{\mathrm{tr}\big(\Sigma\Sigma_{std}^{\dagger}X_{ext}^{\top}(I+X_{ext}\Sigma_{std}^{\dagger}X_{ext}^{\top})^{-1}X_{ext}\Sigma_{std}^{\dagger}\big)}_{T_2\text{: Variance reduction}}, \tag{23}$$

where $\bar{X}_{ext}\overset{\text{def}}{=}\Pi_{std}^{\perp}X_{ext}$, is the component of X_{ext} in the null space of Σ_{std}.

Proof and some corollaries appear in Appendix C.1. From Theorem 5, we see that (i) if X_{ext} is entirely in the span of Σ_{std} making $\bar{X}_{\text{ext}}=0$, $T_1=0$ making $V(\hat{\theta}_{\text{aug}})\leq V(\hat{\theta}_{\text{std}})$ (ii) on the other extreme, if X_{ext} is entirely in the null space with $\Sigma_{\text{std}}^{\dagger}X_{\text{ext}}=0$, $T_2=0$ and hence $V(\hat{\theta}_{\text{aug}})\geq V(\hat{\theta}_{\text{std}})$.

Note that both terms T1 and T2 are traces of PSD matrices and hence non-negative.

Unlike the underparameterized regime, the variance of $\hat{\theta}_{\text{aug}}$ could be larger than that of $\hat{\theta}_{\text{std}}$. X_{ext} could open up new dimensions to be estimated in $\hat{\theta}_{\text{aug}}$, which were otherwise zero in $\hat{\theta}_{\text{std}}$ by virtue of minimizing the norm. The precise effect can be expressed as T1. On the other hand, any augmentation in the column space of Σ_{std} provides a noise canceling effect similarly to the underparameterized regime and reduces the variance, captured in term T2. The overall effect of augmentation on variance is an interplay of these two competing factors.

Proof. Recall from (4) that the $V(\hat{\theta})=\text{tr}(\text{Cov}(\hat{\theta}\mid X_{\text{std}}X_{\text{ext}})\Sigma)$. For the minimum norm interpolation estimators $\hat{\theta}_{\text{std}}$ and $\hat{\theta}_{\text{aug}}$ (3), we have

$$\text{Cov}(\hat{\theta}_{\text{std}}\mid X_{\text{std}},X_{\text{ext}})=\frac{1}{n^2}\text{Cov}(\Sigma_{\text{std}}^{\dagger}X_{\text{std}}^{\top}\epsilon)=\frac{\sigma^2}{n^2}\Sigma_{\text{std}}^{\dagger}X_{\text{std}}^{\top}X_{\text{std}}\Sigma_{\text{std}}^{\dagger}=\frac{\sigma^2}{n}\Sigma_{\text{std}}^{\dagger}. \tag{24}$$

Similarly, we have $\text{Cov}(\hat{\theta}_{\text{aug}}\mid X_{\text{std}},X_{\text{ext}})=\frac{\sigma^2}{n}(\Sigma_{\text{std}}+\alpha\Sigma_{\text{ext}})^{\dagger}$. This gives the following expressions for the variances of the estimators.

$$V(\hat{\theta}_{\text{std}})=\frac{\sigma^2}{n}\text{tr}\left(\Sigma_{\text{std}}^{\dagger}\Sigma\right).$$
$$V(\hat{\theta}_{\text{aug}})=\frac{\sigma^2}{n}\text{tr}\left((\Sigma_{\text{std}}+\alpha\Sigma_{\text{ext}})^{\dagger}\Sigma\right).$$

In order to compare $V(\hat{\theta}_{\text{std}})$ and $V(\hat{\theta}_{\text{aug}})$, we need to compare $\Sigma_{\text{std}}^{\dagger}$ and $(\Sigma_{\text{std}}+\alpha\Sigma_{\text{ext}})^{\dagger}$. In order to do this, we leverage the result from [CITE] on the pseudo-inverse of the sum of two symmetric matrices.

$$\begin{aligned}(\Sigma_{\text{std}}+\alpha\Sigma_{\text{ext}})^{\dagger}&=(\Sigma_{\text{std}}+\frac{1}{n}X_{\text{ext}}^{\top}X_{\text{ext}})^{\dagger}\\&=\Sigma_{\text{std}}^{\dagger}-\frac{1}{n}\Sigma_{\text{std}}^{\dagger}X_{\text{ext}}^{\top}(I+X_{\text{ext}}\Sigma_{\text{std}}^{+}X_{\text{ext}}^{\top})^{-1}X_{\text{ext}}\Sigma_{\text{std}}^{\dagger}+\frac{1}{n}\bar{X}_{\text{ext}}^{\dagger}(\bar{X}_{\text{ext}}^{\dagger})^{\top},\end{aligned}$$

where $\bar{X}_{\text{ext}}\overset{\text{def}}{=}(I-\Sigma_{\text{std}}\Sigma_{\text{std}}^{\dagger})X_{\text{ext}}$, is the component of X_{ext} in the null space of Σ_{std}. Therefore, Multiplying each term by Σ and using linearity of trace, we get the required expression (5). □

We now interpret the above condition to compare the variances of the two estimators, and provide some corollaries of Theorem 5.

Corollary 2 (Nullspace augmentation)**.** *If the augmented points X_{ext} are entirely in the null space of the covariance of the original training points Σ_{std}, the variance of the augmented estimator is never lesser than the variance of the standard estimator, i.e.*

$$\Sigma_{std}^{\top}X_{ext}=0\implies V(\hat{\theta}_{aug})\geq V(\hat{\theta}_{std}). \tag{25}$$

Proof. When $\Sigma_{\text{std}}^{\top}X_{\text{ext}}=0$, term T2 in (5) is zero. Since T1 is non-negative by virtue of being the trace of a PSD matrix, we have $V(\hat{\theta}_{\text{aug}})\geq V(\hat{\theta}_{\text{std}})$. □

Therefore, augmenting solely in the null space of Σ_{std} leads to larger variance. This is because, augmenting in the null space introduces new dimensions that have to be estimated from the data. In contrast, for the standard estimator, these dimensions would be identically zero.

In the next corollary, we consider the case where $\Sigma_{\text{std}}^{\top}X_{\text{ext}}\neq 0$, and study the effect of α—the proportion of augmented points.

Corollary 3 (Amount of augmentation)**.** *Suppose we fix the covariance of the augmented points Σ_{ext}, such that $\Sigma_{std}^{\top}X_{ext}\neq 0$, and vary the number of augmented samples αn. For a large enough α,* *the variance of the augmented estimator is never larger than the variance of the standard estimator. Formally, let $\tilde{X}_{ext} = \frac{1}{\sqrt{\alpha}} X_{ext}$ such that $\Sigma_{ext} = \frac{1}{n} \tilde{X}_{ext}^\top \tilde{X}_{ext}$. For a fixed $\tilde{X}_{ext}$ (and hence Σ_{ext}), we have*

$$\alpha \geq \left\{ \frac{\operatorname{tr}\left(\Sigma \tilde{X}_{ext_\perp}^\dagger (\tilde{X}_{ext_\perp}^\dagger)^\top\right)}{\operatorname{tr}\left(\tilde{X}_{ext} \Sigma_{std}^\dagger \Sigma \Sigma_{std}^\dagger \tilde{X}_{ext}^\top\right)} \left(1 + \lambda_{max}(\tilde{X}_{ext} \Sigma_{std}^\dagger \tilde{X}_{ext}^\top)\right), 1 \right\} \implies V(\hat{\theta}_{aug}) \leq V(\hat{\theta}_{std}). \tag{26}$$

Proof. We first express the variance increase term $T1$ as function of α.

$$T1(\alpha) = \frac{1}{\alpha^2} \operatorname{tr}\left(\Sigma \tilde{X}_{\text{ext}_\perp}^\dagger (\tilde{X}_{\text{ext}_\perp}^\dagger)^\top\right). \tag{27}$$

We now bound the variance reduction term as a function of α. We apply the cyclic trace property and a standard bound on the trace of a product of symmetric matrices (Kleinman & Athans, 1968).

$$\begin{aligned} T2(\alpha) &= \alpha \operatorname{tr}\left(\Sigma \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top (I + \alpha \tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top)^{-1} \tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger\right) \\ &= \alpha \operatorname{tr}\left(\tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \Sigma \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top (I + \alpha \tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top)^{-1}\right) \\ &\geq \alpha \lambda_{\text{min}}((I + \alpha \tilde{X}_{\text{ext}} \bar{X}_{\text{ext}}^\dagger \tilde{X}_{\text{ext}}^\top)^{-1}) \operatorname{tr}\left(\tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \Sigma \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top\right). \end{aligned}$$

By simple algebra, we have

$$\begin{aligned} \lambda_{\text{min}}((I + \alpha \tilde{X}_{\text{ext}} \bar{X}_{\text{ext}}^\dagger \tilde{X}_{\text{ext}}^\top)^{-1}) &= \frac{1}{\lambda_{\text{max}}(I + \alpha \tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top)} \geq \frac{1}{1 + \alpha \lambda_{\text{max}}(\tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top)} \\ \implies T2(\alpha) &\geq \frac{\operatorname{tr}\left(\tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \Sigma \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top\right)}{\frac{1}{\alpha} + \lambda_{\text{max}}(\tilde{X}_{\text{ext}} \Sigma_{\text{std}}^\dagger \tilde{X}_{\text{ext}}^\top)} \end{aligned} \tag{28}$$

Expressing the difference $V(\hat{\theta}_{\text{aug}}) - V(\hat{\theta}_{\text{std}}) = T1(\alpha) - T2(\alpha)$ using (27) and (28) and rearranging the terms via simple linear algebra gives the result. □

To summarize, in the overparameterized regime where the standard training data is not invertible, data augmentation could lead to higher variance due to opening up new dimensions in which the parameters have to be estimated from the data.

D LEVERAGING UNLABELED DATA TO ELIMINATE INCREASE IN BIAS

In this section, we prove Theorem 3, which we reproduce here.

Theorem 3. *Assume the noiseless linear model $y = x^\top \theta^\star$. Let $\theta_{\text{int-std}}$ be an arbitrary interpolant of the standard data, i.e. $X_{std} \theta_{\text{int-std}} = Y_{std}$. Let $\hat{\theta}_{\text{X-aug}}$ be the X-regularized interpolant* (8). *Then*

$$R(\hat{\theta}_{\text{X-aug}}) \leq R(\theta_{\text{int-std}}).$$

Proof. Let $\{u_i\}$ be an orthonormal basis of the kernel $\text{Ker}(\Sigma_{\text{std}} + \Sigma_{\text{ext}})$ and $\{v_i\}$ be an orthonormal basis for $\text{Ker}(\Sigma_{\text{std}}) \setminus \text{span}(\{u_i\})$. Let U and V be the linear operators defined by $Uw = \sum_i u_i w_i$ and $Vw = \sum_i v_i w_i$, respectively, noting that $U^\top V = 0$. Defining $\Pi_{\text{std}}^\perp := (I - \Sigma_{\text{std}}^\dagger \Sigma_{\text{std}})$ to be the projection onto the null space of X_{std}, we see that there are unique vectors ρ, α such that

$$\theta^\star = (I - \Pi_{\text{std}}^\perp) \theta^\star + U\rho + V\alpha. \tag{29a}$$

As $\theta_{\text{int-std}}$ interpolates the standard data, we also have

$$\theta_{\text{int-std}} = (I - \Pi_{\text{std}}^\perp) \theta^\star + Uw + Vz, \tag{29b}$$

as $X_{\text{std}} Uw = X_{\text{std}} Vz = 0$, and finally,

$$\hat{\theta}_{\text{X-aug}} = (I - \Pi_{\text{std}}^\perp) \theta^\star + U\rho + V\lambda \tag{29c}$$

where we note the common ρ between Eqs. (29a) and (29c).

Using the representations (29) we may provide an alternative formulation for the augmented estimator (**??**), using this to prove the theorem. Indeed, writing $\theta_{\text{int-std}} - \hat{\theta}_{\text{X-aug}} = U(w-\rho) + V(z-\lambda)$, we immediately have that the estimator has the form (29c), with the choice

$$\lambda = \operatorname*{argmin}_{\lambda} \left\{ (U(w-\rho)+V(z-\lambda))^\top \Sigma (U(w-\rho)+V(z-\lambda)) \right\}.$$

The optimality conditions for this quadratic imply that

$$V^\top \Sigma V(\lambda - z) = V^\top \Sigma U(w-\rho). \tag{30}$$

Now, recall that the predictive risk of a vector θ is $R(\theta) = (\theta - \theta^\star)^\top \Sigma (\theta - \theta^\star) = \|\theta - \theta^\star\|_\Sigma^2$, using Mahalanobis norm notation. In particular, a few quadratic expansions yield

$$\begin{aligned}
&R(\theta_{\text{int-std}}) - R(\hat{\theta}_{\text{X-aug}}) \\
&= \|U(w-\rho)+V(z-\alpha)\|_\Sigma^2 - \|V(\lambda-\alpha)\|_\Sigma^2 \\
&= \|U(w-\rho)+Vz\|_\Sigma^2 + \|V\alpha\|_\Sigma^2 - 2(U(w-\rho)+Vz)^\top \Sigma V\alpha - \|V\lambda\|_\Sigma^2 - \|V\alpha\|_\Sigma^2 + 2(V\lambda)^\top \Sigma V\alpha \\
&\overset{(i)}{=} \|U(w-\rho)+Vz\|_\Sigma^2 - 2(V\lambda)^\top \Sigma V\alpha - \|V\lambda\|_\Sigma^2 + 2(V\lambda)^\top V\alpha \\
&= \|U(w-\rho)+Vz\|_\Sigma^2 - \|V\lambda\|_\Sigma^2, \qquad (31)
\end{aligned}$$

where step (i) used that $(U(w-\rho))^\top \Sigma V = (V(\lambda - z))^\top \Sigma V$ from the optimality conditions (30).

Finally, we consider the rightmost term in equality (31). Again using the optimality conditions (30), we have

$$\|V\lambda\|_\Sigma^2 = \lambda^\top V^\top \Sigma^{1/2} \Sigma^{1/2} (U(w-\rho)+Vz) \leq \|V\lambda\|_\Sigma \|U(w-\rho)+Vz\|_\Sigma$$

by Cauchy-Schwarz. Revisiting equality (31), we obtain

$$\begin{aligned}
R(\theta_{\text{int-std}}) - R(\hat{\theta}_{\text{X-aug}}) &= \|U(w-\rho)+Vz\|_\Sigma^2 - \frac{\|V\lambda\|_\Sigma^4}{\|V\lambda\|_\Sigma^2} \\
&\geq \|U(w-\rho)+Vz\|_\Sigma^2 - \frac{\|V\lambda\|_\Sigma^2 \|U(w-\rho)+Vz\|_\Sigma^2}{\|V\lambda\|_\Sigma^2} = 0,
\end{aligned}$$

as desired. □

E Experimental Details

E.1 Adversarial augmentation

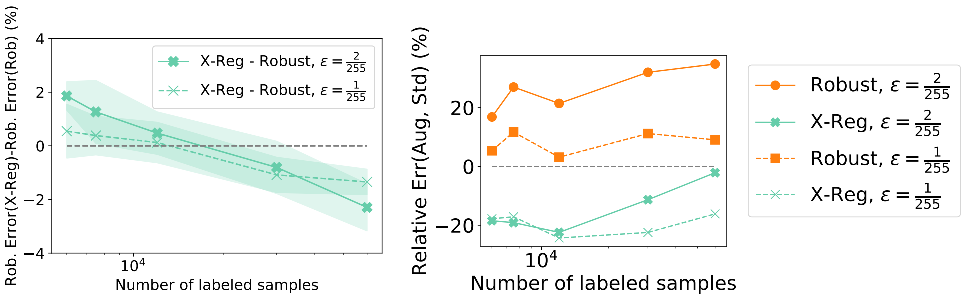
We augment with $\ell_i nfty$ adversarial perturbations of various sizes. In each epoch, we find the augmented examples via Projected Gradient Ascent on the multiclass logistic loss (cross-entropy loss) of the incorrect class. Augmenting in this fashion is essentially adversarial training procedure of (Madry et al., 2018), with equal weight on both the "clean" and adversarial examples.

E.2 Subsampling experiments

We train Wide ResNet 40-2 models (Zagoruyko & Komodakis, 2016) while varying the number of samples in CIFAR-10. We sub-sample CIFAR-10 by factors of $\{1,2,5,8,10,20,40\}$ in Figure 4(a) and $\{1,2,5,8,10\}$ in Figure 4(b). For sub-sample factors 1 to 20, we report results averaged from 2 trials for each model. For sub-sample factors greater than 20, we average over 5 trials.

E.3 X-regularization

We instantiate the general X-regularization estimator that is defined in Equation 9. We use the multiclass logistic loss as the classification loss ℓ and also as the "distance" like measure dist.



(a) Robust error, CIFAR-10

(b) Relative standard error, CIFAR-10

Figure 6: **(a)** Difference in robust test error between our X-regularized model and the robust model for CIFAR-10. X-regularization keeps the robust accuracy within 2% of the robust model for small subsamples and even improves over the robust model for larger subsamples of CIFAR-10. **(b)** Relative difference in standard error between augmented estimators (our X-regularized model and the robust model) and the standard estimator on CIFAR-10. We achieve up to 20% better standard error than the standard model for small subsamples.

E.3.1 COMPARISON TO ROBUST SELF-TRAINING.

X-regularized adversarial training is quite related to robust self-training and it's close variants studied in (Carmon et al., 2019; Uesato et al., 2019; Najafi et al., 2019). However, there is one key difference. Note that no robust loss is minimized on the unlabeled samples $\tilde{x}$; i.e., the perturbations of the unlabeled samples are not added to during X-regularized adversarial training. This distinguishing factor allows X-regularized adversarial training to completely eliminate the increase in standard error. Robust self-training improves over vanilla adversarial training, but still suffers a drop in accuracy.

E.3.2 EVALUATING ROBUST ACCURACY

We evaluate the robustness of models to the strong PGD-attack with 40 steps and 5 restarts. In Figure 4(b), we used a simple heuristic to set the regularization strength λ in the general X-regularization problem (9) to be $\lambda = \min(0.9, \beta)/(1 - \min(0.9, \beta)$ where $\beta \in [0, 1]$ is the fraction of the original CIFAR-10 dataset sampled. Intuitively, we give more weight to the unlabeled data when the original dataset is larger, meaning that the standard estimator produces more accurate pseudo-labels.

Figure 6 shows that the robust accuracy of our X-regularized model stays within 2% of the robust model (trained using PGD adversarial training) for all subsamples, and even improves upon the robust model on the full dataset.

Note that we cannot directly compare the empirical performance of X-regularized adversarial training on CIFAR-10with other methods to obtain robust models that are modifications of vanilla adversarial training. We use a smaller model due to computational constraints enforced by adversarial training. Since the model is small, we could only fit adversarially augmented examples with small $\epsilon = 2/255$, while existing baselines use $\epsilon = 8/255$. Note that even for $\epsilon = 2/255$, adversarial data augmentation leads to an increase in error. We show that X-regularization can fix this. While ensuring models are robust is an important goal in itself, in this work, we view adversarial training through the lens of covariate-shifted data augmentation and study how to use augmented data without increasing test error. We show that X-regularization preserves the other benefits of some kinds of data augmentation like increased robustness to adversarial examples.