Sample Efficiency of Data Augmentation Consistency Regularization

Shuo Yang*
University of Texas at Austin

Yijun Dong*
University of Texas at Austin

Rachel Ward
University of Texas at Austin

Inderjit S. Dhillon
University of Texas at Austin

Sujay Sanghavi
University of Texas at Austin

Qi Lei
New York University

Abstract

Data augmentation is popular in the training of large neural networks; however, currently, theoretical understanding of the discrepancy between different algorithmic choices of leveraging augmented data remains limited. In this paper, we take a step in this direction – we first present a simple and novel analysis for linear regression with label invariant augmentations, demonstrating that data augmentation consistency (DAC) is intrinsically more efficient than empirical risk minimization on augmented data (DA-ERM). The analysis is then generalized to misspecified augmentations (i.e., augmentations that change the labels), which again demonstrates the merit of DAC over DA-ERM. Further, we extend our analysis to non-linear models (e.g., neural networks) and present generalization bounds. Finally, we perform experiments that make a clean and apples-to-apples comparison (i.e., with no extra modeling or data tweaks) between DAC and DA-ERM using CIFAR-100 and WideResNet; these together demonstrate the superior efficacy of DAC.

1 INTRODUCTION

Modern machine learning models, especially deep learning models, require abundant training samples. Since data collection and human annotation are expensive, data augmentation has become a ubiquitous practice in creating artificial labeled samples and improving generalization performance. This practice is corroborated by the fact that the semantics of images remain the same through simple translations like obscuring, flipping, rotation, color jitter, rescaling (Shorten and Khoshgoftaar, 2019). Conventional algorithms use data augmentation to expand the training data set (Simard et al., 1998; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016; Cubuk et al., 2018).

Data Augmentation Consistency (DAC) regularization, as an alternative, enforces the model to output similar predictions on the original and augmented samples and has contributed to many recent state-of-the-art supervised or semi-supervised algorithms. This idea was first proposed in Bachman et al. (2014) and popularized by Laine and Aila (2016); Sajjadi et al. (2016), and gained more attention recently with the success of FixMatch (Sohn et al., 2020) for semi-supervised few-shot learning as well as AdaMatch (Berthelot et al., 2021) for domain adaptation. DAC can utilize unlabeled samples, as one can augment the training samples and enforce consistent predictions without knowing the true labels. This bypasses the limitation of the conventional algorithms that can only augment labeled samples and add them to the training set (referred to as DA-ERM). However, it is not well-understood whether DAC has additional algorithmic benefits compared to DA-ERM. We are, therefore, seeking a theoretical answer.

Despite the empirical success, the theoretical understanding of data augmentation (DA) remains limited. Existing work (Chen et al., 2020a; Mei et al., 2021; Lyle et al., 2019) focused on establishing that augmenting data saves on the number of labeled samples needed for the same level of accuracy. However, none of these explicitly compare the efficacy (in terms of the number of augmented samples) between different algorithmic choices on *how to use the augmented samples* in an apples-to-apples way.

In this paper, we focus on the following research question:

Is DAC intrinsically more efficient than DA-ERM (even without unlabeled samples)?

*Equal contribution. Correspondence to: yangshuo_ut@utexas.edu, ydong@utexas.edu, ql518@nyu.edu

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

We answer the question affirmatively. We show that DAC is intrinsically more efficient than DA-ERM with a simple and novel analysis for linear regression under label invariant augmentations. We then extend the analysis to misspecified augmentations (i.e., those that change the labels). We further provide generalization bounds under consistency regularization for non-linear models like two-layer neural networks and DNN-based classifiers with expansion-based augmentations. Intuitively, we show DAC is better than DA-ERM in the following sense: 1) DAC enforces stronger invariance in the learned models, yielding smaller estimation error; and 2) DAC better tolerates mis-specified augmentations and incurs smaller approximation error. Our theoretical findings can also explain and guide some technical choices, e.g. why we can use stronger augmentation in consistency regularization but only weaker augmentation when creating pseudo-labels (Sohn et al., 2020).

Specifically, our **main contributions** are:

- **Theoretical comparisons between DAC and DA-ERM.** We first present a simple and novel result for linear regression, which shows that DAC yields a strictly smaller generalization error than DA-ERM using the same augmented data. Further, we demonstrate that with with the flexibility of hyper-parameter tuning, DAC can better handle data augmentation with small misspecification in the labels.
- **Extended analysis for non-linear models.** We derive generalization bounds for DAC under two-layer neural networks, and classification with expansion-based augmentations.
- **Empirical comparisons between DAC and DA-ERM.** We perform experiments that make a clean and apples-to-apples comparison (i.e., with no extra modeling or data tweaks) between DAC and DA-ERM using CIFAR-100 and WideResNet. Our empirical results demonstrate the superior efficacy of DAC.

2 RELATED WORK

Empirical findings. Data augmentation (DA) is an essential ingredient for almost every state-of-the-art supervised learning algorithm since the seminal work of Krizhevsky et al. (2012) (see reference therein (Simard et al., 1998; Simonyan and Zisserman, 2014; He et al., 2016; Cubuk et al., 2018; Kuchnik and Smith, 2018)). It started from adding augmented data to the training samples via (random) perturbations, distortions, scales, crops, rotations, and horizontal flips. More sophisticated variants were subsequently designed; a non-exhaustive list includes Mixup (Zhang et al., 2017), Cutout (DeVries and Taylor, 2017), and Cutmix (Yun et al., 2019). The choice of data augmentation and their combinations require domain knowledge and experts' heuristics, which triggered some automated search algorithms to find the best augmentation strategies (Lim et al., 2019; Cubuk et al., 2019). The effects of different DAs are systematically explored in Tensmeyer and Martinez (2016).

Recent practices not only add augmented data to the training set but also enforce similar predictions by adding consistency regularization (Bachman et al., 2014; Laine and Aila, 2016; Sohn et al., 2020). One benefit of DAC is the feasibility of exploiting unlabeled data. Therefore input consistency on augmented data also formed a major component to state-of-the-art algorithms for semi-supervised learning (Laine and Aila, 2016; Sajjadi et al., 2016; Sohn et al., 2020; Xie et al., 2020), self-supervised learning (Chen et al., 2020b), and unsupervised domain adaptation (French et al., 2017; Berthelot et al., 2021).

Theoretical studies. Many interpret the effect of DA as some form of regularization (He et al., 2019). Some work focuses on linear transformations and linear models (Wu et al., 2020) or kernel classifiers (Dao et al., 2019). Convolutional neural networks by design enforce translation equivariance symmetry (Benton et al., 2020; Li et al., 2019); further studies have hard-coded CNN's invariance or equivariance to rotation (Cohen and Welling, 2016; Marcos et al., 2017; Worrall et al., 2017; Zhou et al., 2017), scaling (Sosnovik et al., 2019; Worrall and Welling, 2019) and other types of transformations.

Another line of works view data augmentation as invariant learning by averaging over group actions (Lyle et al., 2019; Chen et al., 2020a; Mei et al., 2021; Bietti et al., 2021; Shao et al., 2022). They consider an ideal setting that is equivalent to ERM with all possible augmented data, bringing a clean mathematical interpretation. In contrast, we are interested in a more realistic setting with limited augmented data. In this setting, it is crucial to utilize the limited data with proper training methods, the difference of which cannot be revealed under previously studied settings.

Some more recent work investigates the feature representation learning procedure with DA for self-supervised learning tasks (Garg and Liang, 2020; Wen and Li, 2021; HaoChen et al., 2021; von Kügelgen et al., 2021). Cai et al. (2021); Wei et al. (2021) studied the effect of data augmentation with label propagation. Data augmentation is also deployed to improve robustness (Rajput et al., 2019), to facilitate domain adaptation and domain generalization (Cai et al., 2021; Sagawa et al., 2019).

3 PROBLEM SETUP AND DATA AUGMENTATION CONSISTENCY

Consider the standard supervised learning problem setup: $\mathbf{x} \in \mathcal{X}$ is input feature, and $y \in \mathcal{Y}$ is its label (or response). Let P be the true distribution of $(\mathbf{x}, y)$ (i.e., the label distribution follows $y \sim P(y|\mathbf{x})$). We have the following def-

inition for label invariant augmentation.

Definition 1 (Label Invariant Augmentation)**.** *For any sample $\mathbf{x} \in \mathcal{X}$, we say that a random transformation $A : \mathcal{X} \to \mathcal{X}$ is a label invariant augmentation if and only if $A(\mathbf{x})$ satisfies $P(y|\mathbf{x}) = P(y|A(\mathbf{x}))$.*

Our work largely relies on label invariant augmentation but also extends to augmentations that incur small misspecification in their labels. Therefore our results apply to the augmentations achieved via certain transformations (e.g., random cropping, rotation), and we do not intend to cover augmentations that can largely alter the semantic meanings (e.g., MixUp (Zhang et al., 2017)).

Now we introduce the learning problem on an augmented dataset. Let $(\mathbf{X}, \mathbf{y}) \in \mathcal{X}^N \times \mathcal{Y}^N$ be a training set consisting of N *i.i.d.* samples. Besides the original $(\mathbf{X}, \mathbf{y})$, each training sample is provided with α augmented samples. The features of the augmented dataset $\widetilde{\mathcal{A}}(\mathbf{x}) \in \mathcal{X}^{(1+\alpha)N}$ is:

$$\widetilde{\mathcal{A}}(\mathbf{X}) = [\mathbf{x}_1; \ldots; \mathbf{x}_N; \mathbf{x}_{1,1}; \ldots; \mathbf{x}_{N,1}; \ldots; \mathbf{x}_{1,\alpha}; \ldots; \mathbf{x}_{N,\alpha}],$$

where $\mathbf{x}_i$ is in the original training set and $\mathbf{x}_{i,j}, \forall j \in [\alpha]$ are the augmentations of $\mathbf{x}_i$. The labels of the augmented samples are kept the same, which can be denoted as $\widetilde{\mathbf{M}}\mathbf{y} \in \mathcal{Y}^{(1+\alpha)N}$, where $\widetilde{\mathbf{M}} \in \mathbb{R}^{(1+\alpha)N \times N}$ is a vertical stack of $(1+\alpha)$ identity mappings.

Data Augmentation Consistency Regularization. Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ be a well-specified function class (e.g., for linear regression problems, $\exists h^* \in \mathcal{H}$, s.t. $h^*(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$) that we hope to learn from. Without loss of generality, we assume that each function $h \in \mathcal{H}$ can be expressed as $h = f_h \circ \phi_h$, where $\phi_h \in \Phi = \{\phi : \mathcal{X} \to \mathcal{W}\}$ is a proper representation mapping and $f_h \in \mathcal{F} = \{f : \mathcal{W} \to \mathcal{Y}\}$ is a predictor on top of the learned representation. We tend to decompose h such that ϕ_h is a powerful feature extraction function whereas f_h can be as simple as a linear combiner. For instance, in a deep neural network, all the layers before the final layer can be viewed as feature extraction ϕ_h, and the predictor f_h is the final linear combination layer.

For a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a metric ϱ properly defined on the representation space $\mathcal{W}$, learning with data augmentation consistency (DAC) regularization is:

$$\operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{N} l(h(\mathbf{x}_i), y_i) + \underbrace{\lambda \sum_{i=1}^{N} \sum_{j=1}^{\alpha} \varrho\left(\phi_h(\mathbf{x}_i), \phi_h(\mathbf{x}_{i,j})\right)}_{\textit{DAC regularization}}. \tag{1}$$

Note that the DAC regularization in Equation (1) can be easily implemented empirically as a regularizer. Intuitively, DAC regularization penalizes the representation difference between the original sample $\phi_h(\mathbf{x}_i)$ and the augmented sample $\phi_h(\mathbf{x}_{i,j})$, with the belief that similar samples (i.e., original and augmented samples) should have similar representations. When the data augmentations do not alter the labels, it is reasonable to enforce a strong regularization (i.e., $\lambda \to \infty$) – since the conditional distribution of y does not change. The learned function $\widehat{h}^{dac}$ can then be written as the solution of a constrained optimization problem:

$$\begin{aligned} \widehat{h}^{dac} \triangleq & \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{N} l(h(\mathbf{x}_i), y_i) \\ \text{s.t.} \quad & \phi_h(\mathbf{x}_i) = \phi_h(\mathbf{x}_{i,j}),\ \forall i \in [N], j \in [\alpha]. \end{aligned} \tag{2}$$

In the rest of the paper, we mainly focus on the data augmentations satisfying Definition 1 and our analysis relies on the formulation of Equation (2). When the data augmentations alter the label distributions (i.e., not satisfying Definition 1), it becomes necessary to adopt a finite λ for Equation (1), and such extension is discussed in Section 5.

4 LINEAR MODEL AND LABEL INVARIANT AUGMENTATIONS

In this section, we show the efficacy of DAC regularization with linear regression under label invariant augmentations (Definition 1).

To see the efficacy of DAC regularization (i.e., Equation (2)), we revisit a more commonly adopted training method here – empirical risk minimization on augmented data (DA-ERM):

$$\begin{aligned} \widehat{h}^{da-erm} \triangleq \operatorname*{argmin}_{h \in \mathcal{H}} & \sum_{i=1}^{N} l(h(\mathbf{x}_i), y_i) \\ & + \sum_{i=1}^{N} \sum_{j=1}^{\alpha} l(h(\mathbf{x}_{i,j}), y_i). \end{aligned} \tag{3}$$

Now we show that the DAC regularization (Equation (2)) learns more efficiently than DA-ERM. Consider the following setting: given N observations $\mathbf{X} \in \mathbb{R}^{N \times d}$, the responses $\mathbf{y} \in \mathbb{R}^N$ are generated from a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is zero-mean noise with $\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] = \sigma^2 \mathbf{I}_N$. Recall that $\widetilde{\mathcal{A}}(\mathbf{X})$ is the entire augmented dataset, and $\widetilde{\mathbf{M}}\mathbf{y}$ corresponds to the labels. We focus on the fixed design excess risk of $\boldsymbol{\theta}$ on $\widetilde{\mathcal{A}}(\mathbf{X})$, which is defined as $L(\boldsymbol{\theta}) \triangleq \frac{1}{(1+\alpha)N} \left\| \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* \right\|_2^2$.

Let $\boldsymbol{\Delta} \triangleq \widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}$ and $d_{aug} \triangleq \operatorname{rank}(\boldsymbol{\Delta})$ measure the number of dimensions in the row space of $\mathbf{X}$ perturbed by augmentations (which can be intuitively view as the "strength" of data augmentations where the larger d_{aug} implies the stronger perturbation brought by $\widetilde{\mathcal{A}}(\mathbf{X})$ to $\mathbf{X}$). Assuming that $\widetilde{\mathcal{A}}(\mathbf{X})$ has full column rank (such that the linear regression problem has a unique solution), we have the following result for learning by DAC versus DA-ERM.

Theorem 1 (Informal result on linear regression (formally in Theorem 5))**.** *Learning with DAC regularization,*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[L(\widehat{\boldsymbol{\theta}}^{dac}) - L(\boldsymbol{\theta}^*)\right] = \frac{(d - d_{aug})\sigma^2}{N},$$

while learning with ERM directly on the augmented dataset, there exists $d' \in [0, d_{aug}]$ *such that*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[L(\widehat{\boldsymbol{\theta}}^{da-erm}) - L(\boldsymbol{\theta}^*)\right] = \frac{(d - d_{aug} + d')\sigma^2}{N}.$$

Formally, we have

$$d' \triangleq \frac{\operatorname{tr}\left(\left(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} - \mathbf{P}_{\mathcal{S}}\right)\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top\right)}{1+\alpha},$$

where $\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} \triangleq \widetilde{\mathcal{A}}(\mathbf{X})\widetilde{\mathcal{A}}(\mathbf{X})^\dagger$. $\mathbf{P}_{\mathcal{S}}$ is the projector onto

$$\mathcal{S} \triangleq \left\{\widetilde{\mathbf{M}}\mathbf{X}\boldsymbol{\theta} \mid \forall \boldsymbol{\theta} \in \mathbb{R}^d, s.t. \left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}\right)\boldsymbol{\theta} = 0\right\}.$$

Under standard conditions (e.g., $\mathbf{x}$ is sub-Gaussian and N is not too small), it is not hard to extend Theorem 1 to random design (i.e., the more commonly acknowledged generalization bound) with the same order.

Remark 1 (Why DAC is more effective)**.** *Intuitively, DAC reduces the dimensions from* d *to* $d - d_{aug}$ *by enforcing consistency regularization. DA-ERM, on the other hand, still learns in the original* d*-dimensional space.* d' *characterizes such difference.*

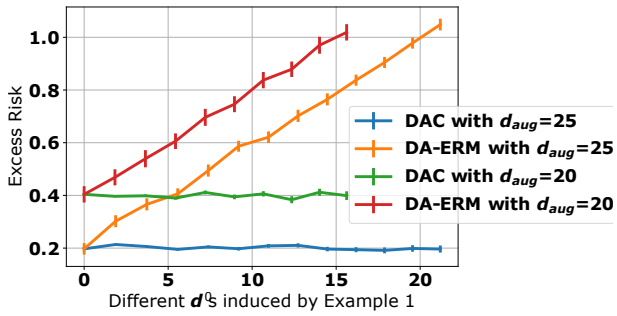

Figure 1: Comparison of DAC regularization and DA-ERM (Example 1). The results precisely match Theorem 1. DA-ERM depends on the d' induced by different augmentations, while the DAC regularization works equally well for all d' and better than the DA-ERM. Further, both DAC and DA-ERM are affected by d_{aug}, the number of dimensions perturbed by $\widetilde{\mathcal{A}}(\mathbf{X})$.

Now we take a closer look at $d' \triangleq \frac{\operatorname{tr}\left(\left(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} - \mathbf{P}_{\mathcal{S}}\right)\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top\right)}{1+\alpha}$ characterizing the discrepancy between DAC and DA-ERM. We first observe that $\sigma^2 \cdot \widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top$ is the noise covariance matrix of the augmented dataset. $\operatorname{tr}\left(\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top\right)$ represents the variance of $\widehat{\boldsymbol{\theta}}^{dac}$, while $\operatorname{tr}\left(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top\right)$ denotes the variance of $\widehat{\boldsymbol{\theta}}^{da-erm}$. Therefore, $d' \propto \operatorname{tr}\left(\left(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} - \mathbf{P}_{\mathcal{S}}\right)\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top\right)$ measures the excess variance of $\widehat{\boldsymbol{\theta}}^{da-erm}$ in comparison to $\widehat{\boldsymbol{\theta}}^{dac}$. When $\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} \neq \mathbf{P}_{\mathcal{S}}$ (a common scenario as instantiated in Example 1), DAC is strictly better than DA-ERM.

Example 1. *Consider a 30-dimensional linear regression. The original training set contains 50 samples. The inputs* $\mathbf{x}_i$*s are generated independently from* $\mathcal{N}(0, \mathbf{I}_{30})$ *and we set* $\boldsymbol{\theta}^* = [\boldsymbol{\theta}^*_c; \mathbf{0}]$ *with* $\boldsymbol{\theta}^*_c \sim \mathcal{N}(0, \mathbf{I}_5)$ *and* $\mathbf{0} \in \mathbb{R}^{25}$*. The noise variance* σ *is set to* 1*. We partition* $\mathbf{x}$ *into 3 parts* $[x_{c1}, x_{e1}, x_{e2}]$ *and take the following augmentations:* $A([x_{c1}; x_{e1}; x_{e2}]) = [x_{c1}; 2x_{e1}; -x_{e2}], x_{c1} \in \mathbb{R}^{d_{c1}}, x_{e1} \in \mathbb{R}^{d_{e1}}, x_{e2} \in \mathbb{R}^{d_{e2}}$*, where* $d_{c1} + d_{e1} + d_{e2} = 30$*.*

Notice that the augmentation perturbs x_{e1} *and* x_{e2} *and leaving* x_{c1} *unchanged, we therefore have* $d_{aug} = 30 - d_{c1}$*. By changing* d_{c1} *and* d_{e1}*, we can have different augmentations with different* d_{aug}, d'*. The results for* $d_{aug} \in \{20, 25\}$ *and various* d'*s are presented in Figure 1. The excess risks precisely match Theorem 1. It confirms that the DAC regularization is strictly better than DA-ERM for a wide variety of augmentations.*

5 BEYOND LABEL INVARIANT AUGMENTATION

In this section, we extend our analysis to misspecified augmentations by relaxing the label invariance assumption (such that $P(y|\mathbf{x}) \neq P(y|A(\mathbf{x}))$). With an illustrative linear regression problem, we show that DAC also brings advantages over DA-ERM for misspecified augmentations.

We first recall the linear regression setup: given a set of N *i.i.d.* samples $(\mathbf{X}, \mathbf{y})$ that follows $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ are zero-mean independent noise with $\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] = \sigma^2\mathbf{I}_N$, we aim to learn the unknown ground truth $\boldsymbol{\theta}^*$. For randomly generated misspecified augmentations $\widetilde{\mathcal{A}}(\mathbf{X})$ that alter the labels (*i.e.*, $\widetilde{\mathcal{A}}(\mathbf{X})\,\boldsymbol{\theta}^* \neq \widetilde{\mathbf{M}}\mathbf{X}\boldsymbol{\theta}^*$), a proper consistency constraint is $\|\phi_h(\mathbf{x}_i) - \phi_h(\mathbf{x}_{i,j})\|_2 \leq C_{mis}$ (where $\mathbf{x}_{i,j}$ is an augmentation of $\mathbf{x}_i$, noticing that $C_{mis} = 0$ corresponds to label invariant augmentations in Definition 1). For $C_{mis} > 0$, the constrained optimization is equivalent to:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{dac} = &\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{N}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \\ &+ \frac{\lambda}{(1+\alpha)N}\left\|\left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}\right)\boldsymbol{\theta}\right\|_2^2\end{aligned} \tag{4}$$

for some finite $0 < \lambda < \infty$. We compare $\widehat{\boldsymbol{\theta}}^{dac}$ to the solution learned with ERM on augmented data (as in Equation (3)):

$$\widehat{\boldsymbol{\theta}}^{da-erm} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{(1+\alpha)N}\left\|\widetilde{\mathcal{A}}(\mathbf{X})\,\boldsymbol{\theta} - \widetilde{\mathbf{M}}\mathbf{y}\right\|_2^2.$$

Let $\mathbf{\Sigma_X} \triangleq \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ and $\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})} \triangleq \frac{1}{(1+\alpha)N}\widetilde{\mathcal{A}}(\mathbf{X})^\top\widetilde{\mathcal{A}}(\mathbf{X})$. With $\mathbf{S} = \frac{1}{1+\alpha}\widetilde{\mathbf{M}}^\top\widetilde{\mathcal{A}}(\mathbf{X})$, $\mathbf{\Delta} \triangleq \widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}$, and its reweighted analog $\widetilde{\mathbf{\Delta}} \triangleq \left(\widetilde{\mathbf{M}}\mathbf{X}\right)\widetilde{\mathcal{A}}(\mathbf{X})^\dagger\mathbf{\Delta}$, we further introduce positive semidefinite matrices: $\mathbf{\Sigma_S} \triangleq \frac{1}{N}\mathbf{S}^\top\mathbf{S}$, $\mathbf{\Sigma_\Delta} \triangleq \frac{1}{(1+\alpha)N}\mathbf{\Delta}^\top\mathbf{\Delta}$, and $\mathbf{\Sigma}_{\widetilde{\mathbf{\Delta}}} \triangleq \frac{1}{(1+\alpha)N}\widetilde{\mathbf{\Delta}}^\top\widetilde{\mathbf{\Delta}}$. For demonstration purpose, we consider fixed $\mathbf{X}$ and $\widetilde{\mathcal{A}}(\mathbf{X})$, with respect to which we introduce distortion factors $c_X, c_S > 0$ as the minimum constants that satisfy $\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})} \preccurlyeq c_X\mathbf{\Sigma_X}$ and $\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})} \preccurlyeq c_S\mathbf{\Sigma_S}$ (notice that such c_X, c_S exist almost surely when $\mathbf{X}$ and $\widetilde{\mathcal{A}}(\mathbf{X})$ are drawn from absolutely continuous marginal distributions).

Recall $d_{aug} \triangleq \operatorname{rank}(\mathbf{\Delta})$ from Section 4. Let $\mathbf{P_\Delta} \triangleq \mathbf{\Delta}^\dagger\mathbf{\Delta}$ denote the rank-d_{aug} orthogonal projector onto $\operatorname{Range}\left(\mathbf{\Delta}^\top\right)$. Then, for $L(\boldsymbol{\theta}) = \frac{1}{N}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$, we have the following result:

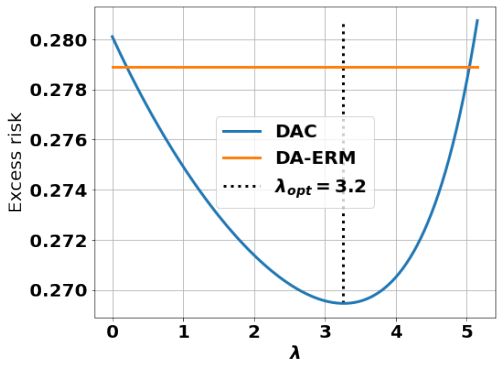

Figure 2: Comparison of DAC with different λ (optimal choice at $\lambda_{opt} = 3.2$) and DA-ERM in Example 2, where $d_{aug} = 24$ and $\alpha = 1$. The results demonstrate that, with a proper λ, DAC can outperform DA-ERM under misspecified augmentations.

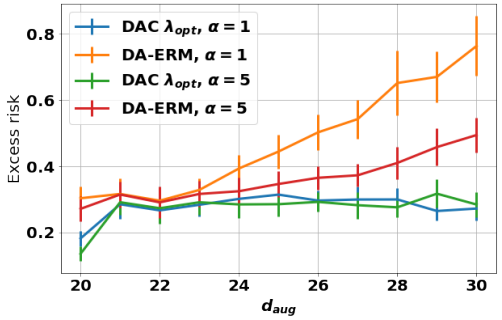

Figure 3: Comparison of DAC with the optimal λ and DA-ERM in Example 2 for different augmentation strength d_{aug}. $d_{aug} = 20$ corresponds to the label-invariance augmentations, whereas increasing d_{aug} leads to more misspecification.

Theorem 2. *Learning with DAC regularization (Equation (4)), we have that, at the optimal λ*[1]*,*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[L(\widehat{\boldsymbol{\theta}}^{dac}) - L(\boldsymbol{\theta}^*)\right] \le \frac{\sigma^2(d - d_{aug})}{N} + \|\mathbf{P_\Delta}\boldsymbol{\theta}^*\|_{\mathbf{\Sigma_\Delta}}\sqrt{\frac{\sigma^2}{N}\operatorname{tr}\left(\mathbf{\Sigma_X}\mathbf{\Sigma}_{\mathbf{\Delta}}^\dagger\right)},$$

whereas learning with DA-ERM (Equation (3)),

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[L(\widehat{\boldsymbol{\theta}}^{da-erm}) - L(\boldsymbol{\theta}^*)\right] \ge \frac{\sigma^2 d}{N c_X c_S} + \|\mathbf{P_\Delta}\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}_{\widetilde{\mathbf{\Delta}}}}^2.$$

Here, $\mathbf{P_\Delta}\boldsymbol{\theta}^*$ measures the misspecification in $\boldsymbol{\theta}^*$ by the augmentations $\widetilde{\mathcal{A}}(\mathbf{X})$.

One advantage of DAC regularization derives from its flexibility in choosing regularization parameter λ. With a proper λ (*e.g.*, see Figure 2) that matches misspecification $C_{mis}^2 = \frac{1}{(1+\alpha)N}\left\|\left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}\right)\boldsymbol{\theta}^*\right\|_2^2 = \|\mathbf{P_\Delta}\boldsymbol{\theta}^*\|_{\mathbf{\Sigma_\Delta}}^2$, DAC effectively reduces the function class from $\mathbb{R}^d$ to $\left\{\boldsymbol{\theta} \mid \|\mathbf{P_\Delta}\boldsymbol{\theta}\|_{\mathbf{\Sigma_\Delta}} \le C_{mis}\right\}$ and therefore improves the sample efficiency.

Another advantage of DAC is that, in contrast to DA-ERM, the consistency regularization term in Equation (4) refrains from learning the original labels with misspecified augmentations $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widetilde{\mathbf{M}}\mathbf{y}\right] \ne \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*$ when a suitable C_{mis} is identified implicitly via λ. This allows DAC to learn from fewer but stronger (potentially more severely misspecified) augmentations (*e.g.*, Figure 3). Specifically, as $N \to \infty$, the excess risk of DAC with the optimal λ converges to zero by learning from unbiased labels $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathbf{y}] = \mathbf{X}\boldsymbol{\theta}^*$, whereas DA-ERM suffers from a bias term $\|\mathbf{P_\Delta}\boldsymbol{\theta}^*\|_{\mathbf{\Sigma}_{\widetilde{\mathbf{\Delta}}}}^2 > 0$ due to the bias from misspecified augmentations.

Example 2. *As in Example 1, we consider a linear regression problem of dimension $d = 30$ with $\alpha \ge 1$ misspecified augmentations on $N = 50$ i.i.d. training samples drawn from $\mathcal{N}(\boldsymbol{0}, \mathbf{I}_d)$. We aim to learn $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_c^*; \mathbf{0}] \in \mathbb{R}^d$ (where $\boldsymbol{\theta}_c^* \in \{-1, +1\}^{d_c}$, $d_c = 10$) under label noise $\sigma = 0.1$. The misspecified augmentations mimic the effect of color jitter by adding* i.i.d. *Gaussian noise entry-wisely to the last d_{aug} feature coordinates: $\widetilde{\mathcal{A}}(\mathbf{X}) = [\mathbf{X}; \mathbf{X}']$ where $\mathbf{X}'_{ij} = \mathbf{X}_{ij} + \mathcal{N}(0, 0.1)$ for all $i \in [N]$, $d - d_{aug} + 1 \le j \le d$ – such that $d_{aug} = \operatorname{rank}(\mathbf{\Delta})$ with probability 1. The $(d - d_{aug} + 1), \ldots, d_c$-th coordinates of $\boldsymbol{\theta}^*$ are misspecified by the augmentations.*

As previously discussed on Theorem 2, DAC is more robust than DA-ERM to misspecified augmentations, and therefore can learn with fewer (smaller α) and stronger (larger d_{aug}) augmentations. In addition, DAC generally achieves better generalization than DA-ERM with limited samples.

[1] A positive (semi)definite matrix $\mathbf{\Sigma}$ induces a (semi)norm: $\|\mathbf{u}\|_{\mathbf{\Sigma}} = \left(\mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}\right)^{1/2}$ for all conformable $\mathbf{u}$.

6 BEYOND LINEAR MODEL

In this section, we extend our analysis of DAC regularization to non-linear models, including the two-layer neural networks, and DNN-based classifiers with expansion-based augmentations.

Further, in addition to the popular in-distribution setting where we consider a unique distribution P for both training and testing, DAC regularization is also known to improve out-of-distribution generalization for settings like domain adaptation. We defer detailed discussion on such advantage of DAC regularization for linear regression in the domain adaptation setting to Appendix D.

6.1 Two-layer Neural Network

We first generalize our analysis to an illustrative nonlinear model – two-layer ReLU network. With $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, we consider a ground truth distribution $P(y|\mathbf{x})$ induced by $y = \left(\mathbf{x}^\top \mathbf{B}^*\right)_+ \mathbf{w}^* + \epsilon$. For the unknown ground truth function $h^*(\mathbf{x}) \triangleq \left(\mathbf{x}^\top \mathbf{B}^*\right)_+ \mathbf{w}^*$, $(\cdot)_+ \triangleq \max(0, \cdot)$ denotes the element-wisely ReLU function; $\mathbf{B}^* = \left[\mathbf{b}_1^* \dots \mathbf{b}_k^* \dots \mathbf{b}_q^*\right] \in \mathbb{R}^{d\times q}$ consists of $\mathbf{b}_k^* \in \mathbb{S}^{d-1}$ for all $k \in [q]$; and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is *i.i.d.* Gaussian noise. In terms of the function class $\mathcal{H}$, for some constant $C_w \geq \|\mathbf{w}^*\|_1$, let

$$\mathcal{H} = \big\{h(\mathbf{x}) = (\mathbf{x}^\top \mathbf{B})_+ \mathbf{w} \;\big|\; \mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_q] \in \mathbb{R}^{d\times q}, \\ \|\mathbf{b}_k\|_2 = 1 \;\forall\, j \in [q], \|\mathbf{w}\|_1 \leq C_w\big\},$$

such that $h^* \in \mathcal{H}$. For regression, we again consider square loss $l(h(\mathbf{x}), y) = \frac{1}{2}(h(\mathbf{x}) - y)^2$ and learn with DAC on the first layer: $\left(\mathbf{x}_i^\top \mathbf{B}\right)_+ = \left(\mathbf{x}_{i,j}^\top \mathbf{B}\right)_+$.

Let $\mathbf{\Delta} \triangleq \widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}$, and $\mathbf{P}_{\mathbf{\Delta}}^\perp$ be the projector onto the null space of $\mathbf{\Delta}$. Under mild regularity conditions (*i.e.*, αN being sufficiently large, $\mathbf{x}$ being subgaussian, and distribution of $\mathbf{\Delta}$ being absolutely continuous, as specified in Appendix B), regression over two-layer ReLU networks with the DAC regularization generalizes as following:

Theorem 3 (Informal result on two-layer neural network with DAC (formally in Theorem 6))**.** *Conditioned on $\mathbf{X}$ and $\mathbf{\Delta}$, with $L(h) = \frac{1}{N}\|h(\mathbf{X}) - h^*(\mathbf{X})\|_2^2$ and $\sqrt{\frac{1}{N}\sum_{i=1}^N \left\|\mathbf{P}_{\mathbf{\Delta}}^\perp \mathbf{x}_i\right\|_2^2} \leq C_{\mathcal{N}}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $\boldsymbol{\epsilon}$,*

$$L\left(\widehat{h}^{dac}\right) - L\left(h^*\right) \lesssim \sigma C_w C_{\mathcal{N}} \left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right).$$

Recall $d_{aug} = \operatorname{rank}(\mathbf{\Delta})$. With a sufficiently large N (as specified in Appendix B), we have $C_{\mathcal{N}} \lesssim \sqrt{d - d_{aug}}$ with high probability[2]. Meanwhile, applying DA-ERM directly on the augmented samples achieves no better than

$$L(\widehat{h}^{da-erm}) - L(h^*) \lesssim \sigma C_w \max\left(\sqrt{\tfrac{d}{(\alpha+1)N}}, \sqrt{\tfrac{d-d_{aug}}{N}}\right),$$

where the first term corresponds to the generalization bound for a d-dimensional regression with $(\alpha+1)N$ *i.i.d.* samples (in contrast to augmented samples that are potentially dependent); and the second term follows as the augmentations $\widetilde{\mathcal{A}}(\mathbf{X})$ keep a $(d - d_{aug})$-dimensional subspace (*i.e.*, the null space of $\mathbf{\Delta} = \widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}$) intact, in which DA-ERM can only rely on the N original samples for learning. In specific, the first term will dominate the max with limited augmented data (i.e., α being small).

Comparing the two, we see that DAC tends to be more efficient than DA-ERM, and such advantage is enhanced with strong but limited data augmentations (i.e., large d_{aug} and small α). For instance, with $\alpha = 1$ and $d_{aug} = d - 1$, the generalization error of DA-ERM scales as $\sqrt{d/N}$, while DAC yields a dimension-free $\sqrt{1/N}$ error.

As a synopsis for the regression cases in Section 4, Section 5, and Section 6.1 generally, the effect of DAC regularization can be casted as a dimension reduction by d_{aug} – dimension of the subspace perturbed by data augmentations where features contain scarce label information.

6.2 Classification with Expansion-based Augmentations

A natural generalization of the dimension reduction viewpoint on DAC regularization in the regression setting is the complexity reduction for general function classes. Here we demonstrate the power of DAC on function class reduction in a DNN-based classification setting.

Concretely, we consider a multi-class classification problem: given a probability space $\mathcal{X}$ with marginal distribution $P(\mathbf{x})$ and K classes $\mathcal{Y} = [K]$, let $h^* : \mathcal{X} \to [K]$ be the ground truth classifier, partitioning $\mathcal{X}$ into K disjoint sets $\{\mathcal{X}_k\}_{k\in[K]}$ such that $P(y|\mathbf{x}) = \mathbf{1}\{y = h^*(\mathbf{x})\} = \mathbf{1}\{\mathbf{x} \in \mathcal{X}_y\}$. In the classification setting, we replace Definition 1 with *expansion-based data augmentations* introduced in Wei et al. (2021); Cai et al. (2021).

Definition 2 (Expansion-based augmentations (formally in Definition 4))**.** *With respect to an augmentation function $\mathcal{A} : \mathcal{X} \to 2^{\mathcal{X}}$, let $NB(S) \triangleq \cup_{\mathbf{x}\in S}\left\{\mathbf{x}' \in \mathcal{X} \;\middle|\; \mathcal{A}(\mathbf{x}) \cap \mathcal{A}(\mathbf{x}') \neq \emptyset\right\}$ be the neighborhood of $S \subseteq \mathcal{X}$. For any $c > 1$, we say that $\mathcal{A}$ induces c-expansion-based data augmentations if (a) $\{\mathbf{x}\} \subsetneq \mathcal{A}(\mathbf{x}) \subseteq \{\mathbf{x}' \in \mathcal{X} \mid h^*(\mathbf{x}) = h^*(\mathbf{x}')\}$ for all $\mathbf{x} \in \mathcal{X}$; and (b) for all $k \in [K]$, given any $S \subseteq \mathcal{X}$ with $P(S \cap \mathcal{X}_k) \leq \frac{1}{2}$, $P(NB(S) \cap \mathcal{X}_k) \geq \min\{c \cdot P(S \cap \mathcal{X}_k), 1\}$.*

Particularly, Definition 2(a) enforces that the ground truth classifier h^* is invariant throughout each neighborhood.

[2]Here we only account for the randomness in $\mathbf{X}$ but not that in $\mathbf{\Delta}|\mathbf{X}$ which characterizes d_{aug} for conciseness. We refer the readers to Appendix B for a formal tail bound on $C_{\mathcal{N}}$.

Meanwhile, the expansion factor c in Definition 2(b) serves as a quantification of augmentation strength – a larger c implies a stronger augmentation $\mathcal{A}$.

We aim to learn $h(\mathbf{x}) \triangleq \operatorname{argmax}_{k\in[K]} f(\mathbf{x})_k$ with loss $l_{01}(h(\mathbf{x}), y) = \mathbf{1}\{h(\mathbf{x}) \neq y\}$ from $\mathcal{H}$ induced by the class of p-layer fully connected neural networks with maximum width q, $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}^K \mid f = f_{2p-1} \circ \cdots \circ f_1,\}$ where $f_{2\iota-1}(\mathbf{x}) = \mathbf{W}_\iota \mathbf{x}$, $f_{2\iota}(\boldsymbol{\epsilon}) = \varphi(\boldsymbol{\epsilon})$, $\mathbf{W}_\iota \in \mathbb{R}^{d_\iota \times d_{\iota-1}}$ $\forall \iota \in [p]$, $q \triangleq \max_{\iota\in[p]} d_\iota$, and φ is the activation function.

Over a general probability space $\mathcal{X}$, DAC with expansion-based augmentations requires stronger conditions than merely consistent classification over $\mathcal{A}(\mathbf{x}_i)$ for all labeled training samples $i \in [N]$. Instead, we enforce a large robust margin $m_{\mathcal{A}}(f, \mathbf{x}^u)$ (adapted from Wei et al. (2021), see Appendix C) over an finite set of unlabeled samples $\mathbf{X}^u$ that is independent of $\mathbf{X}$ and drawn *i.i.d.* from $P(\mathbf{x})$. Intuitively, $m_{\mathcal{A}}(f, \mathbf{x}^u)$ measures the maximum allowed perturbation in all parameters of f such that predictions remain consistent throughout $\mathcal{A}(\mathbf{x}^u)$ (*e.g.*, $m_{\mathcal{A}}(f, \mathbf{x}^u) > 0$ is equivalent to enforcing consistent classification outputs). For any $0 < \tau \leq \max_{f\in\mathcal{F}} \inf_{\mathbf{x}^u\in\mathcal{X}} m_{\mathcal{A}}(f, \mathbf{x}^u)$, the DAC regularization reduces the function class $\mathcal{H}$ to

$$\mathcal{H}_{dac} \triangleq \{h \in \mathcal{H} \mid m_{\mathcal{A}}(f, \mathbf{x}^u) > \tau \quad \forall\, \mathbf{x}^u \in \mathbf{X}^u\}.$$

Then for $\widehat{h}^{dac} = \operatorname{argmin}_{h\in\mathcal{H}_{dac}} \frac{1}{N}\sum_{i=1}^N l_{01}(h(\mathbf{x}_i), y_i)$, we have the following.

Theorem 4 (formally in Theorem 8)**.** *Given an augmentation function $\mathcal{A}$ that induces c-expansion-based data augmentations (Definition 2) such that*

$$\mu \triangleq \sup_{h\in\mathcal{H}_{dac}} \mathbb{P}_P\left[\exists\, \mathbf{x}' \in \mathcal{A}(\mathbf{x}) : h(\mathbf{x}) \neq h(\mathbf{x}')\right] \leq \frac{c-1}{4},$$

for any $\delta \in (0, 1)$, with probability at least $1-\delta$, we have

$$\mu \leq \widetilde{O}\left(\frac{\sum_{\iota=1}^p \sqrt{q}\,\|\mathbf{W}_\iota\|_F}{\tau\sqrt{|\mathbf{X}^u|}} + \sqrt{\frac{p\log|\mathbf{X}^u|}{|\mathbf{X}^u|}}\right)$$

such that

$$L_{01}\left(\widehat{h}^{dac}\right) - L_{01}(h^*) \lesssim \sqrt{\frac{K\log(N)}{N} + \frac{\mu}{\min\{c-1, 1\}}} + \sqrt{\frac{\log(1/\delta)}{N}}.$$

In particular, DAC regularization leverages the unlabeled samples $\mathbf{X}^u$ and effectively decouples the labeled sample complexity $N = \widetilde{O}(K)$ from the complexity of the function class $\mathcal{H}$ (characterized by $\{\mathbf{W}_\iota\}_{\iota\in[p]}$ and q and encapsulated in μ) via the reduced function class $\mathcal{H}_{dac}$. Notably, Theorem 4 is reminiscent of Wei et al. (2021) Theorem 3.6, 3.7, and Cai et al. (2021) Theorem 2.1, 2.2, 2.3. We unified the existing theories under our function class reduction viewpoint to demonstrate its generality.

7 EXPERIMENTS

In this section, we empirically verify that training with DAC learns more efficiently than DA-ERM. The dataset is derived from CIFAR-100, where we randomly select 10,000 labeled data as the training set (i.e., 100 labeled samples per class). During the training time, given a training batch, we generate augmentations by RandAugment (Cubuk et al., 2020). We set the number of augmentations per sample to 7 unless otherwise mentioned.

The experiments focus on comparisons of 1) training with consistency regularization (DAC), and 2) empirical risk minimization on the augmented dataset (DA-ERM). We use the same network architecture (a WideResNet-28-2 (Zagoruyko and Komodakis, 2016)) and the same training settings (e.g., optimizer, learning rate schedule, etc) for both methods. We defer the detailed experiment settings to Appendix F. Our test set is the standard CIFAR-100 test set, and we report the average and standard deviation of the testing accuracy of 5 independent runs. The consistency regularizer is implemented as the l_2 distance of the model's predictions on the original and augmented samples.

Efficacy of DAC regularization. We first show that the DAC regularization learns more efficiently than DA-ERM. The results are listed in Table 1. In practice, the augmentations almost always alter the label distribution, we therefore follow the discussion in section 5 and adopt a finite λ (i.e., the multiplicative coefficient before the DAC regularization, see Equation (1)). With proper choice of λ, training with DAC significantly improves over DA-ERM.

DAC regularization helps more with limited augmentations. Our theoretical results suggest that the DAC regularization learns efficiently with a limited number of augmentations. While keeping the number of labeled samples to be 10,000, we evaluate the performance of the DAC regularization and DA-ERM with different numbers of augmentations. The number of augmentations for each training sample ranges from 1 to 15, and the results are listed in Table 2. The DAC regularization offers a more significant improvement when the number of augmentations is small. This clearly demonstrates that the DAC regularization learns more efficiently than DA-ERM.

DAC regularization helps more when data is scarce. We conduct experiments with different numbers of labeled samples, ranging from 1,000 (i.e., 10 images per class) to 20,000 samples (i.e., 200 images per class). We generate 3 augmentations for each of the samples during the training time, and the results are presented in Table 3. Notice that the DAC regularization gives a bigger improvement over DA-ERM when the labeled samples are scarce. This matches the intuition that when there are sufficient training samples, data augmentation is less necessary. Therefore, the difference between different ways of utilizing the aug-

DA-ERM	DAC Regularization				
	$\lambda = 0$	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$
69.40 ± 0.05	62.82 ± 0.21	68.63 ± 0.11	$\mathbf{70.56 \pm 0.07}$	$\mathbf{70.52 \pm 0.14}$	68.65 ± 0.27

Table 1: Testing accuracy of DA-ERM and DAC with different λ's (regularization coeff.).

Number of Augmentations	1	3	7	15
DA-ERM	67.92 ± 0.08	69.04 ± 0.05	69.25 ± 0.16	69.30 ± 0.11
DAC ($\lambda = 10$)	$\mathbf{70.06 \pm 0.08}$	$\mathbf{70.77 \pm 0.20}$	$\mathbf{70.74 \pm 0.11}$	$\mathbf{70.31 \pm 0.12}$

Table 2: Testing accuracy of DA-ERM and DAC with different numbers of augmentations.

Number of Labeled Data	1000	10000	20000
DA-ERM	31.11 ± 0.30	68.89 ± 0.07	$\mathbf{76.79 \pm 0.13}$
DAC ($\lambda = 10$)	$\mathbf{33.59 \pm 0.41}$	$\mathbf{70.71 \pm 0.10}$	$\mathbf{76.86 \pm 0.16}$

Table 3: Testing accuracy of ERM and DAC regularization with different numbers of labeled data.

No Augmentation	DA-ERM	DAC ($\lambda = 0.1$)	DAC ($\lambda = 1$)	DAC ($\lambda = 10$)
62.82 ± 0.21	61.35 ± 0.27	63.73 ± 0.33	$\mathbf{64.30 \pm 0.20}$	64.00 ± 0.26

Table 4: DAC performs well under misspecified augmentations after tuning λ.

Number of Unlabeled Data	5000	10000	20000
FixMatch	67.74	69.23	70.76
FixMatch + DAC ($\lambda = 1$)	**71.24**	**72.7**	**74.04**

Table 5: DAC helps FixMatch when the unlabeled data is scarce.

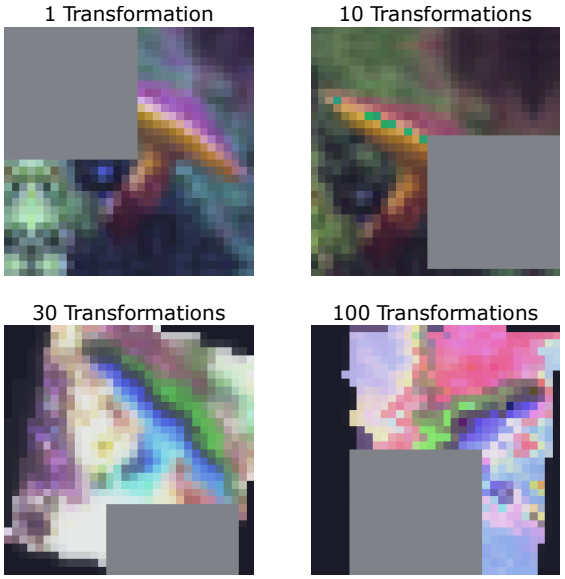

Figure 4: Different numbers of transformations.

mented samples becomes diminishing.

DAC performs well under misspecified augmentations. As suggested by Theorem 2, DAC is more robust to misspecified augmentations with proper λ. We further empirically verify this result with misspecified augmentations - where the augmentations are generated by applying 100 random transformations. When too many transformations are applied (see illustration in Figure 4), the augmentation will alter the label distribution and is thus misspecified. The results are presented in Table 4. Notice that with $\lambda = 1$, the DAC delivers the best accuracy, which supports our theoretical results.

Further, comparing the results of Table 1 and Table 4, we see that the optimal λ is different when the augmentations are misspecified. Because of the flexibility in choosing λ, DAC is able to outperform DA-ERM, which matches the result in Theorem 2.

Combining with a semi-supervised learning algorithm. Here we show that the DAC regularization can be easily extended to the semi-supervised learning setting. We take the previously established semi-supervised learning method FixMatch (Sohn et al., 2020) as the baseline and adapt the FixMatch by combining it with the DAC regularization. Specifically, besides using FixMatch to learn from the unlabeled data, we additionally generate augmentations for the labeled samples and apply DAC. In particular, we focus on the data-scarce regime by only keeping 10,000 labeled samples and at most 20,000 unlabeled samples. Results are listed in Table 5. We see that the DAC regulariza-

tion also improves the performance of FixMatch when the unlabeled samples are scarce. This again demonstrates the efficiency of learning with DAC.

8 CONCLUSION

In this paper, we take a step toward understanding the statistical efficiency of DAC with limited data augmentations. At the core, DAC is statistically more efficient because it reduces problem dimensions by enforcing consistency regularization.

We demonstrate the benefits of DAC compared to DA-ERM (expanding training set with augmented samples) both theoretically and empirically. Theoretically, we show a strictly smaller generalization error under linear regression, and explicitly characterize the generalization upper bound for two-layer neural networks and expansion-based data augmentations. We further show that DAC better handles the label misspecification caused by strong augmentations. Empirically, we provide apples-to-apples comparisons between DAC and DA-ERM. These together demonstrate the superior efficacy of DAC over DA-ERM.

Acknowledgements

SY's research is supported by NSF grants 1564000 and 1934932. YD's research is supported by AFOSR MURI FA9550-19-1-0005, NSF DMS 1952735, NSF HDR-1934932, and NSF 2019844.

References

Bachman, P., Alsharif, O., and Precup, D. (2014). Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373.

Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482.

Benton, G., Finzi, M., Izmailov, P., and Wilson, A. G. (2020). Learning invariances in neural networks. *arXiv preprint arXiv:2010.11882*.

Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., and Kurakin, A. (2021). Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*.

Bietti, A., Venturi, L., and Bruna, J. (2021). On the sample complexity of learning with geometric stability. *arXiv preprint arXiv:2106.07148*.

Cai, T., Gao, R., Lee, J. D., and Lei, Q. (2021). A theory of label propagation for subpopulation shift.

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

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Cohen, T. and Welling, M. (2016). Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Ré, C. (2019). A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pages 1528–1537. PMLR.

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2020). Few-shot learning via learning the representation, provably.

French, G., Mackiewicz, M., and Fisher, M. (2017). Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*.

Garg, S. and Liang, Y. (2020). Functional regularization for representation learning: A unified theoretical perspective. *arXiv preprint arXiv:2008.02447*.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *arXiv preprint arXiv:2106.04156*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, Z., Xie, L., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2019). Data augmentation revisited: Rethinking the distribution gap between clean and augmented data. *arXiv preprint arXiv:1909.09148*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Kuchnik, M. and Smith, V. (2018). Efficient augmentation via data subsampling. *arXiv preprint arXiv:1810.05222*.

Laine, S. and Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Li, Z., Wang, R., Yu, D., Du, S. S., Hu, W., Salakhutdinov, R., and Arora, S. (2019). Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. (2019). Fast autoaugment. *Advances in Neural Information Processing Systems*, 32:6665–6675.

Lyle, C., Kwiatkowksa, M., and Gal, Y. (2019). An analysis of the effect of invariance on generalization in neural networks. In *International conference on machine learning Workshop on Understanding and Improving Generalization in Deep Learning*.

Marcos, D., Volpi, M., Komodakis, N., and Tuia, D. (2017). Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057.

Mei, S., Misiakiewicz, T., and Montanari, A. (2021). Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiopoulos, D. (2019). Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pages 5321–5330. PMLR.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Shao, H., Montasser, O., and Blum, A. (2022). A theory of pac learnability under transformation invariances. *arXiv preprint arXiv:2202.07552*.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. (1998). Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Sosnovik, I., Szmaja, M., and Smeulders, A. (2019). Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*.

Tensmeyer, C. and Martinez, T. (2016). Improving invariance and equivariance properties of convolutional neural networks.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. *arXiv preprint arXiv:2106.04619*.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wei, C., Shen, K., Chen, Y., and Ma, T. (2021). Theoretical analysis of self-training with deep networks on unlabeled data.

Wen, Z. and Li, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. *arXiv preprint arXiv:2105.15134*.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., and Brostow, G. J. (2017). Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037.

Worrall, D. E. and Welling, M. (2019). Deep scale-spaces: Equivariance over scale. *arXiv preprint arXiv:1905.11697*.

Wu, S., Zhang, H., Valiant, G., and Ré, C. (2020). On the generalization effects of linear transformations in data augmentation. In *International Conference on Machine Learning*, pages 10410–10420. PMLR.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhou, Y., Ye, Q., Qiu, Q., and Jiao, J. (2017). Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528.

A Linear Regression Models

In this section, we present formal proofs for the results on linear regression in the fixed design where the training samples $(\mathbf{X}, \mathbf{y})$ and their augmentations $\widetilde{\mathcal{A}}(\mathbf{X})$ are considered to be fixed. We discuss two types of augmentations: the label invariant augmentations in Section 4 and the misspecified augmentations in Section 5.

A.1 Linear Regression with Label Invariant Augmentations

For fixed $\widetilde{\mathcal{A}}(\mathbf{X})$, let $\mathbf{\Delta} \triangleq \widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}$ in this section. We recall that $d_{aug} = \operatorname{rank}(\mathbf{\Delta})$ since there is no randomness in $\widetilde{\mathcal{A}}, \mathbf{X}$ in fix design setting. Assuming that $\widetilde{\mathcal{A}}(\mathbf{X})$ admits full column rank, we have the following theorem on the excess risk of DAC and ERM:

Theorem 5 (Formal restatement of Theorem 1 on linear regression.)**.** *Learning with DAC regularization, we have* $\mathbb{E}\left[L(\widehat{\boldsymbol{\theta}}^{dac}) - L(\boldsymbol{\theta}^*)\right] = \frac{(d-d_{aug})\sigma^2}{N}$*, while learning with ERM directly on the augmented dataset, we have* $\mathbb{E}\left[L(\widehat{\boldsymbol{\theta}}^{da-erm}) - L(\boldsymbol{\theta}^*)\right] = \frac{(d-d_{aug}+d')\sigma^2}{N}$*.* d' *is defined as*

$$d' \triangleq \frac{\operatorname{tr}\left(\widetilde{\mathbf{M}}^\top\left(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} - \mathbf{P}_{\mathcal{S}}\right)\widetilde{\mathbf{M}}\right)}{1+\alpha},$$

where $d' \in [0, d_{aug}]$ *with* $\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} = \widetilde{\mathcal{A}}(\mathbf{X})\left(\widetilde{\mathcal{A}}(\mathbf{X})^\top\widetilde{\mathcal{A}}(\mathbf{X})\right)^{-1}\widetilde{\mathcal{A}}(\mathbf{X})^\top$ *and* $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{(\alpha+1)N\times(\alpha+1)N}$ *is the orthogonal projector onto* $\mathcal{S} \triangleq \left\{\widetilde{\mathbf{M}}\mathbf{X}\boldsymbol{\theta} \mid \forall \boldsymbol{\theta} \in \mathbb{R}^d, s.t.\ \left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}\right)\boldsymbol{\theta} = \boldsymbol{\theta}\right\}$.

Proof. With $L(\boldsymbol{\theta}) \triangleq \frac{1}{(1+\alpha)N}\left\|\widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2$, the excess risk of ERM on the augmented training set satisfies that:

$$\begin{aligned}
\mathbb{E}\left[L(\widehat{\boldsymbol{\theta}}^{da-erm})\right] &= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\widetilde{\mathcal{A}}(\mathbf{X})\widehat{\boldsymbol{\theta}}^{da-erm} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2\right] \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\widetilde{\mathcal{A}}(\mathbf{X})(\widetilde{\mathcal{A}}(\mathbf{X})^\top\widetilde{\mathcal{A}}(\mathbf{X}))^{-1}\widetilde{\mathcal{A}}(\mathbf{X})^\top(\widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* + \widetilde{\mathbf{M}}\boldsymbol{\epsilon}) - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2\right] \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* + \mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathbf{M}}\boldsymbol{\epsilon} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2\right] \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathbf{M}}\boldsymbol{\epsilon}\right\|_2^2\right] \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\operatorname{tr}(\boldsymbol{\epsilon}^\top\widetilde{\mathbf{M}}^\top\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathbf{M}}\boldsymbol{\epsilon})\right] \\
&= \frac{\sigma^2}{(1+\alpha)N}\operatorname{tr}\left(\widetilde{\mathbf{M}}^\top\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathbf{M}}\right).
\end{aligned}$$

Let $\mathcal{C}_{\widetilde{\mathcal{A}}(\mathbf{X})}$ and $\mathcal{C}_{\widetilde{\mathbf{M}}}$ denote the column space of $\widetilde{\mathcal{A}}(\mathbf{X})$ and $\widetilde{\mathbf{M}}$, respectively. Notice that $\mathcal{S}$ is a subspace of both $\mathcal{C}_{\widetilde{\mathcal{A}}(\mathbf{X})}$ and $\mathcal{C}_{\widetilde{\mathbf{M}}}$. Observing that $d_{aug} = \operatorname{rank}(\mathbf{\Delta}) = \operatorname{rank}(\mathbf{P}_{\mathcal{S}})$, we have

$$\begin{aligned}
\mathbb{E}\left[L(\widehat{\boldsymbol{\theta}}^{da-erm})\right] =& \frac{\sigma^2}{(1+\alpha)N}\operatorname{tr}(\widetilde{\mathbf{M}}^\top\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\widetilde{\mathbf{M}}) \\
=& \frac{\sigma^2}{(1+\alpha)N}\operatorname{tr}(\widetilde{\mathbf{M}}^\top\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{M}}) + \frac{\sigma^2}{(1+\alpha)N}\operatorname{tr}(\widetilde{\mathbf{M}}^\top(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} - \mathbf{P}_{\mathcal{S}})\widetilde{\mathbf{M}}) \\
=& \frac{\sigma^2}{(1+\alpha)N}\operatorname{tr}(\widetilde{\mathbf{M}}^\top\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{M}}) + \frac{\sigma^2}{N}\cdot\frac{\operatorname{tr}(\widetilde{\mathbf{M}}^\top(\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})} - \mathbf{P}_{\mathcal{S}})\widetilde{\mathbf{M}})}{1+\alpha}
\end{aligned}$$

By the data augmentation consistency constraint, we are essentially solving the linear regression on the $(d - d_{aug})$-dimensional space $\{\boldsymbol{\theta} \mid \mathbf{\Delta}\boldsymbol{\theta} = 0\}$. The rest of proof is identical to standard regression analysis, with features first projected

to $\mathcal{S}$:

$$\begin{aligned}
\mathbb{E}\left[L(\widehat{\boldsymbol{\theta}}^{dac})\right] &= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\widetilde{\mathcal{A}}(\mathbf{X})\widehat{\boldsymbol{\theta}}^{dac} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2\right] \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\widetilde{\mathcal{A}}(\mathbf{X})(\widetilde{\mathcal{A}}(\mathbf{X})^\top\widetilde{\mathcal{A}}(\mathbf{X}))^{-1}\widetilde{\mathcal{A}}(\mathbf{X})^\top\mathbf{P}_{\mathcal{S}}(\widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* + \widetilde{\mathbf{M}}\boldsymbol{\epsilon}) - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2\right] \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_{\mathcal{S}}\widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* + \mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{M}}\boldsymbol{\epsilon} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*\right\|_2^2\right] \\
&\quad \left(\text{since } \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* \in \mathcal{S}, \text{ and } \mathbf{P}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_{\mathcal{S}} = \mathbf{P}_{\mathcal{S}} \text{ since } \mathcal{S} \subseteq \mathcal{C}_{\widetilde{\mathcal{A}}(\mathbf{X})}\right) \\
&= \frac{1}{(1+\alpha)N}\mathbb{E}\left[\left\|\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{M}}\boldsymbol{\epsilon}\right\|_2^2\right] \\
&= \frac{\sigma^2}{(1+\alpha)N}\operatorname{tr}(\widetilde{\mathbf{M}}^\top\mathbf{P}_{\mathcal{S}}\widetilde{\mathbf{M}}) \\
&= \frac{(d-d_{aug})\sigma^2}{N}.
\end{aligned}$$

■

A.2 Linear Regression Beyond Label Invariant Augmentations

Proof of Theorem 2. With $L(\boldsymbol{\theta}) \triangleq \frac{1}{N}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2$, we start by partitioning the excess risk into two parts – the variance from label noise and the bias from feature-label mismatch due to augmentations (*i.e.*, $\widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^* \neq \widetilde{\mathbf{M}}\mathbf{X}\boldsymbol{\theta}^*$):

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)\right] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2\right] = \underbrace{\mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\theta} - \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\theta}]\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2\right]}_{\text{Variance}} + \underbrace{\|\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\theta}] - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2}_{\text{Bias}}.$$

First, we consider learning with DAC regularization with some finite $0 < \lambda < \infty$,

$$\widehat{\boldsymbol{\theta}}^{dac} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{\lambda}{(1+\alpha)N}\left\|\left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}\right)\boldsymbol{\theta}\right\|_2^2.$$

By setting the gradient of Equation (4) with respect to $\boldsymbol{\theta}$ to $\mathbf{0}$, with $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$, we have

$$\widehat{\boldsymbol{\theta}}^{dac} = \frac{1}{N}\left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}\right),$$

Then with $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\boldsymbol{\theta}}^{dac}\right] = \left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\boldsymbol{\Sigma}_\mathbf{X}\boldsymbol{\theta}^*$,

$$\text{Var} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\frac{1}{N}\left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\mathbf{X}^\top\boldsymbol{\epsilon}\right\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2\right], \quad \text{Bias} = \left\|\left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\boldsymbol{\Sigma}_\mathbf{X}\boldsymbol{\theta}^* - \boldsymbol{\theta}^*\right\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2.$$

For the variance term, we have

$$\begin{aligned}
\text{Var} =& \frac{\sigma^2}{N}\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\boldsymbol{\Sigma}_\mathbf{X}\left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\boldsymbol{\Sigma}_\mathbf{X}\right) \\
=& \frac{\sigma^2}{N}\operatorname{tr}\left(\left[\boldsymbol{\Sigma}_\mathbf{X}^{1/2}\left(\boldsymbol{\Sigma}_\mathbf{X} + \lambda\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\right)^\dagger\boldsymbol{\Sigma}_\mathbf{X}^{1/2}\right]^2\right) \\
=& \frac{\sigma^2}{N}\operatorname{tr}\left(\left(\mathbf{I}_d + \lambda\boldsymbol{\Sigma}_\mathbf{X}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_\mathbf{X}^{-1/2}\right)^{-2}\right)
\end{aligned}$$

For the semi-positive definite matrix $\boldsymbol{\Sigma}_\mathbf{X}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_\mathbf{X}^{-1/2}$, we introduce the spectral decomposition:

$$\boldsymbol{\Sigma}_\mathbf{X}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_\mathbf{X}^{-1/2} = \underset{d\times d_{aug}}{\mathbf{Q}}\ \underset{d_{aug}\times d_{aug}}{\boldsymbol{\Gamma}}\ \mathbf{Q}^\top, \quad \boldsymbol{\Gamma} = \operatorname{diag}\left(\gamma_1, \ldots, \gamma_{d_{aug}}\right),$$

where $\mathbf{Q}$ consists of orthonormal columns and $\gamma_1 \geq \cdots \geq \gamma_{d_{aug}} > 0$. Then

$$\text{Var} = \frac{\sigma^2}{N} \operatorname{tr}\left(\left(\mathbf{I}_d - \mathbf{Q}\mathbf{Q}^\top\right) + \mathbf{Q}\left(\mathbf{I}_{d_{aug}} + \lambda\mathbf{\Gamma}\right)^{-2}\mathbf{Q}^\top\right) = \frac{\sigma^2\left(d - d_{aug}\right)}{N} + \frac{\sigma^2}{N}\sum_{i=1}^{d_{aug}} \frac{1}{\left(1+\lambda\gamma_i\right)^2}.$$

For the bias term, we observe that

$$\begin{aligned}
\text{Bias} &= \left\|\left(\mathbf{\Sigma_X} + \lambda\mathbf{\Sigma_\Delta}\right)^\dagger \mathbf{\Sigma_X}\boldsymbol{\theta}^* - \boldsymbol{\theta}^*\right\|_{\mathbf{\Sigma_X}}^2 \\
&= \left\|\left(\mathbf{\Sigma_X} + \lambda\mathbf{\Sigma_\Delta}\right)^\dagger \left(-\lambda\mathbf{\Sigma_\Delta}\right)\boldsymbol{\theta}^*\right\|_{\mathbf{\Sigma_X}}^2 \\
&= \left\|\left(\mathbf{I}_d + \lambda\mathbf{\Sigma_X}^{-\frac{1}{2}}\mathbf{\Sigma_\Delta}\mathbf{\Sigma_X}^{-\frac{1}{2}}\right)^{-1}\left(\lambda\mathbf{\Sigma_X}^{-\frac{1}{2}}\mathbf{\Sigma_\Delta}\mathbf{\Sigma_X}^{-\frac{1}{2}}\right)\left(\mathbf{\Sigma_X}^{1/2}\mathbf{P_\Delta}\boldsymbol{\theta}^*\right)\right\|_2^2.
\end{aligned}$$

Then with $\boldsymbol{\vartheta} \triangleq \mathbf{\Sigma_X}^{1/2}\mathbf{P_\Delta}\boldsymbol{\theta}^*$, we have

$$\text{Bias} = \sum_{i=1}^{d_{aug}} \vartheta_i^2 \left(\frac{\lambda\gamma_i}{1+\lambda\gamma_i}\right)^2$$

To simply the optimization of regularization parameter λ, we leverage upper bounds of the variance and bias terms:

$$\begin{aligned}
\text{Var} - \frac{\sigma^2\left(d - d_{aug}\right)}{N} &\leq \frac{\sigma^2}{N}\sum_{i=1}^{d_{aug}} \frac{1}{\left(1+\lambda\gamma_i\right)^2} \leq \frac{\sigma^2}{2N\lambda}\sum_{i=1}^{d_{aug}} \frac{1}{\gamma_i} \leq \frac{\sigma^2}{2N\lambda}\operatorname{tr}\left(\mathbf{\Sigma_X}\mathbf{\Sigma_\Delta}^\dagger\right), \\
\text{Bias} = \sum_{i=1}^{d_{aug}} \vartheta_i^2 \left(\frac{\lambda\gamma_i}{1+\lambda\gamma_i}\right)^2 &\leq \frac{\lambda}{2}\sum_{i=1}^{d_{aug}} \vartheta_i^2\gamma_i = \frac{\lambda}{2}\left\|\mathbf{P_\Delta}\boldsymbol{\theta}^*\right\|_{\mathbf{\Sigma_\Delta}}^2.
\end{aligned}$$

Then with $\lambda = \sqrt{\frac{\sigma^2 \operatorname{tr}\left(\mathbf{\Sigma_X}\mathbf{\Sigma_\Delta}^\dagger\right)}{N\left\|\mathbf{P_\Delta}\boldsymbol{\theta}^*\right\|_{\mathbf{\Sigma_\Delta}}^2}}$, we have the generalization bound for $\widehat{\boldsymbol{\theta}}^{dac}$ in Theorem 2.

Second, we consider learning with DA-ERM:

$$\widehat{\boldsymbol{\theta}}^{da-erm} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{\left(1+\alpha\right)N}\left\|\widetilde{\mathcal{A}}\left(\mathbf{X}\right)\boldsymbol{\theta} - \widetilde{\mathbf{M}}\mathbf{y}\right\|_2^2.$$

With

$$\widehat{\boldsymbol{\theta}}^{da-erm} = \frac{1}{(1+\alpha)N}\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{-1}\widetilde{\mathcal{A}}\left(\mathbf{X}\right)^\top\widetilde{\mathbf{M}}\left(\mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}\right),$$

we again partition the excess risk into the variance and bias terms. For the variance term, with the assumptions $\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})} \preccurlyeq c_X\mathbf{\Sigma_X}$ and $\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})} \preccurlyeq c_S\mathbf{\Sigma_S}$, we have

$$\begin{aligned}
\text{Var} =& \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\frac{1}{(1+\alpha)N}\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{-1}\widetilde{\mathcal{A}}\left(\mathbf{X}\right)^\top\widetilde{\mathbf{M}}\boldsymbol{\epsilon}\right\|_{\mathbf{\Sigma_X}}^2\right] \\
=& \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\frac{1}{N}\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{-1}\mathbf{S}^\top\boldsymbol{\epsilon}\right\|_{\mathbf{\Sigma_X}}^2\right] \\
=& \frac{\sigma^2}{N}\operatorname{tr}\left(\mathbf{\Sigma_X}\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{-1}\mathbf{\Sigma_S}\mathbf{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{-1}\right) \\
\geq& \frac{\sigma^2}{N}\operatorname{tr}\left(\frac{1}{c_X c_S}\mathbf{I}_d\right) = \frac{\sigma^2 d}{N c_X c_S}.
\end{aligned}$$

Additionally, for the bias term, we have

$$\begin{aligned}\text{Bias} &= \left\| \frac{1}{(1+\alpha)N} \boldsymbol{\Sigma}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{-1} \widetilde{\mathcal{A}}(\mathbf{X})^\top \widetilde{\mathbf{M}}\mathbf{X}\boldsymbol{\theta}^* - \boldsymbol{\theta}^* \right\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2 \\ &= \left\| \left(\widetilde{\mathcal{A}}(\mathbf{X})^\top \widetilde{\mathcal{A}}(\mathbf{X}) \right)^{-1} \widetilde{\mathcal{A}}(\mathbf{X})^\top \boldsymbol{\Delta} \left(\mathbf{P}_{\boldsymbol{\Delta}} \boldsymbol{\theta}^* \right) \right\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2 \\ &= \left\| \widetilde{\mathcal{A}}(\mathbf{X})^\dagger \boldsymbol{\Delta} \left(\mathbf{P}_{\boldsymbol{\Delta}} \boldsymbol{\theta}^* \right) \right\|_{\boldsymbol{\Sigma}_\mathbf{X}}^2 = \| \mathbf{P}_{\boldsymbol{\Delta}} \boldsymbol{\theta}^* \|_{\boldsymbol{\Sigma}_{\widetilde{\mathbf{A}}}}^2 .\end{aligned}$$

Combining the variance and bias leads to the generalization bound for $\widehat{\boldsymbol{\theta}}^{da-erm}$ in Theorem 2. ■

B Two-layer Neural Network Regression

In the two-layer neural network regression setting with $\mathcal{X} = \mathbb{R}^d$ described in Section 6.1, let $\mathbf{X} \sim P^N(\mathbf{x})$ be a set of N *i.i.d.* samples drawn from the marginal distribution $P(\mathbf{x})$ that satifies the following.

Assumption 1 (Regularity of marginal distribution)**.** *Let* $\mathbf{x} \sim P(\mathbf{x})$ *be zero-mean* $\mathbb{E}[\mathbf{x}] = \boldsymbol{0}$*, with covairance matrix* $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma}_\mathbf{x} \succ 0$ *whose eigenvalues are bounded by constant factors* $\Omega(1) = \sigma_{\min}(\boldsymbol{\Sigma}_\mathbf{x}) \leq \sigma_{\max}(\boldsymbol{\Sigma}_\mathbf{x}) = O(1)$*, such that* $(\boldsymbol{\Sigma}_\mathbf{x}^{-1/2}\mathbf{x})$ *is* ρ^2*-subgaussian* [3].

For the sake of analysis, we isolate the augmented part in $\widetilde{\mathcal{A}}(\mathbf{X})$ and denote the set of these augmentations as

$$\mathcal{A}(\mathbf{X}) = [\mathbf{x}_{1,1}; \cdots; \mathbf{x}_{N,1}; \cdots; \mathbf{x}_{1,\alpha}; \cdots; \mathbf{x}_{N,\alpha}] \in \mathcal{X}^{\alpha N},$$

where for each sample $i \in [N]$, $\{\mathbf{x}_{i,j}\}_{j\in[\alpha]}$ is a set of α augmentations generated from $\mathbf{x}_i$, and $\mathbf{M} \in \mathbb{R}^{\alpha N \times N}$ is the vertical stack of α $N \times N$ identity matrices. Analogous to the notions with respect to $\widetilde{\mathcal{A}}(\mathbf{X})$ in the linear regression cases in Appendix A, in this section, we denote $\boldsymbol{\Delta} \triangleq \mathcal{A}(\mathbf{X}) - \mathbf{M}\mathbf{X}$ and quantify the augmentation strength as

$$d_{aug} \triangleq \operatorname{rank}(\boldsymbol{\Delta}) = \operatorname{rank}\left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X} \right)$$

such that $0 \leq d_{aug} \leq \min(d, \alpha N)$ can be intuitively interpreted as the number of dimensions in the span of the unlabeled samples, $\operatorname{Row}(\mathbf{X})$, perturbed by the augmentations.

Then, to learn the ground truth distribution $\mathbf{y} = h^*(\mathbf{X}) + \boldsymbol{\epsilon} = (\mathbf{X}\mathbf{B}^*)_+ \mathbf{w}^* + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$, training with the DAC regularization can be formulated explicitly as

$$\begin{aligned}\widehat{\mathbf{B}}^{dac}, \widehat{\mathbf{w}}^{dac} = \underset{\mathbf{B}\in\mathbb{R}^{d\times q}, \mathbf{w}\in\mathbb{R}^q}{\operatorname{argmin}} & \frac{1}{N} \left\| \mathbf{y} - (\mathbf{X}\mathbf{B})_+ \mathbf{w} \right\|_2^2 \\ \text{s.t.} \quad & \mathbf{B} = \left[\mathbf{b}_1 \dots \mathbf{b}_k \dots \mathbf{b}_q \right], \ \mathbf{b}_k \in \mathbb{S}^{d-1} \ \forall\, k \in [q], \quad \|\mathbf{w}\|_1 \leq C_w \\ & (\mathcal{A}(\mathbf{X})\mathbf{B})_+ = (\mathbf{M}\mathbf{X}\mathbf{B})_+ .\end{aligned}$$

For the resulted minimizer $\widehat{h}^{dac}(\mathbf{x}) \triangleq (\mathbf{x}^\top \widehat{\mathbf{B}}^{dac})_+ \widehat{\mathbf{w}}^{dac}$, we have the following.

Theorem 6 (Formal restatement of Theorem 3 on two-layer neural network with DAC)**.** *Under Assumption 1, we suppose* $\mathbf{X}$ *and* $\boldsymbol{\Delta}$ *satisfy that (a)* $\alpha N \geq 4 d_{aug}$*; and (b)* $\boldsymbol{\Delta}$ *admits an absolutely continuous distribution. Then conditioned on* $\mathbf{X}$ *and* $\boldsymbol{\Delta}$*, with* $L(h) = \frac{1}{N}\|h(\mathbf{X}) - h^*(\mathbf{X})\|_2^2$ *and* $\frac{1}{N}\sum_{i=1}^N \left\| \mathbf{P}_{\boldsymbol{\Delta}}^\perp \mathbf{x}_i \right\|_2^2 \leq C_\mathcal{N}^2$ *for some* $C_\mathcal{N} > 0$*, for all* $\delta \in (0,1)$*, with probability at least* $1-\delta$ *(over* $\boldsymbol{\epsilon}$*),*

$$L\left(\widehat{h}^{dac}\right) - L(h^*) \lesssim \sigma C_w C_\mathcal{N} \left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{N}} \right).$$

Moreover, to account for randomness in $\mathbf{X}$ and $\boldsymbol{\Delta}$, we introduce the following notion of augmentation strength.

[3] A random vector $\mathbf{v} \in \mathbb{R}^d$ is ρ^2-subgaussian if for any unit vector $\mathbf{u} \in \mathbb{S}^{d-1}$, $\mathbf{u}^\top \mathbf{v}$ is ρ^2-subgaussian, $\mathbb{E}\left[\exp(s \cdot \mathbf{u}^\top \mathbf{v})\right] \leq \exp\left(s^2\rho^2/2\right)$.

Definition 3 (Augmentation strength)**.** *For any $\delta \in [0, 1)$, let*

$$d_{aug}(\delta) \triangleq \underset{d'}{\operatorname{argmax}}\ \mathbb{P}_{\mathbf{\Delta}}\left[\operatorname{rank}\left(\mathbf{\Delta}\right) < d'\right] \leq \delta.$$

Intuitively, the *augmentation strength* d_{aug} ensures that the feature subspace perturbed by the augmentations in $\mathcal{A}(\mathbf{X})$ has a minimum dimension $d_{aug}(\delta)$ with probability at least $1-\delta$. A larger $d_{aug}(\delta)$ corresponds to stronger augmentations. For instance, when $\mathcal{A}(\mathbf{X}) = \mathbf{MX}$ almost surely (*e.g.*, when the augmentations are identical copies of the original samples, corresponding to the weakest augmentation – no augmentations at all), we have $d_{aug}(\delta) = d_{aug} = 0$ for all $\delta < 1$. Whereas for randomly generated augmentations, d_{aug} is likely to be larger (i.e., with more dimensions being perturbed). For example in Example 2, for a given d_{aug}, with random augmentations $\mathcal{A}\left(\mathbf{X}\right) = \mathbf{X}'$ where $\mathbf{X}'_{ij} = \mathbf{X}_{ij} + \mathcal{N}\left(0, 0.1\right)$ for all $i \in [N]$, $d - d_{aug} + 1 \leq j \leq d$, we have $\operatorname{rank}\left(\mathbf{\Delta}\right) = d_{aug}$ with probability 1. That is $d_{aug}(\delta) = d_{aug}$ for all $\delta \geq 0$.

Leveraging the notion of augmentation strength in Definition 3, we show that the stronger augmentations lead to the better generalization by reducing $C_{\mathcal{N}}$ in Theorem 6.

Corollary 1. *When $N \gg \rho^4 d$ and $\alpha N \geq d$, for any $\delta \in (0, 1)$, with probability at least $1-\delta$ (over $\mathbf{X}$ and $\mathbf{\Delta}$), we have $C_{\mathcal{N}} \lesssim \sqrt{d - d_{aug}(\delta)}$.*

To prove Theorem 6, we start by showing that, with sufficient samples ($\alpha N \geq 4 d_{aug}$), consistency of the first layer outputs over the samples implies consistency of those over the population.

Lemma 1. *Under the assumptions in Theorem 6, every size-d_{aug} subset of rows in $\mathbf{\Delta} = \mathcal{A}(\mathbf{X}) - \mathbf{MX}$ is linearly independent almost surely.*

Proof of Lemma 1. Since $\alpha N > d_{aug}$, it is sufficient to show that a random matrix with an absolutely continuous distribution is totally invertible [4] almost surely.

It is known that for any dimension $m \in \mathbb{N}$, an $m \times m$ square matrix $\mathbf{S}$ is singular if $\det(\mathbf{S}) = 0$ where entries of $\mathbf{S}$ lie within the roots of the polynomial equation specified by the determinant. Therefore, the set of all singular matrices in $\mathbb{R}^{m \times m}$ has Lebesgue measure zero,

$$\lambda\left(\left\{\mathbf{S} \in \mathbb{R}^{m \times m} \;\middle|\; \det(\mathbf{S}) = 0\right\}\right) = 0.$$

Then, for an absolutely continuous probability measure μ with respect to λ, we also have

$$\mathbb{P}_{\mu}\left[\mathbf{S} \in \mathbb{R}^{m \times m} \text{ is singular}\right] = \mu\left(\left\{\mathbf{S} \in \mathbb{R}^{m \times m} \;\middle|\; \det(\mathbf{S}) = 0\right\}\right) = 0.$$

Since a general matrix $\mathbf{R}$ contains only finite number of submatrices, when $\mathbf{R}$ is drawn from an absolutely continuous distribution, by the union bound, $\mathbb{P}\left[\mathbf{R} \text{ cotains a singular submatrix}\right] = 0$. That is, $\mathbf{R}$ is totally invertible almost surely. ■

Lemma 2. *Under the assumptions in Theorem 6, the hidden layer in the two-layer ReLU network learns $\operatorname{Null}\left(\mathbf{\Delta}\right)$, the invariant subspace under data augmentations : with high probability,*

$$\left(\mathbf{x}^{\top}\widehat{\mathbf{B}}^{dac}\right)_{+} = \left(\mathbf{x}^{\top}\mathbf{P}_{\mathbf{\Delta}}^{\perp}\widehat{\mathbf{B}}^{dac}\right)_{+} \quad \forall\, \mathbf{x} \in \mathcal{X}.$$

Proof of Lemma 2. We will show that for all $\mathbf{b}_k = \mathbf{P}_{\mathbf{\Delta}}^{\perp}\mathbf{b}_k + \mathbf{P}_{\mathbf{\Delta}}\mathbf{b}_k$, $k \in [q]$, $\mathbf{P}_{\mathbf{\Delta}}\mathbf{b}_k = \mathbf{0}$ with high probability, which then implies that given any $\mathbf{x} \in \mathcal{X}$, $(\mathbf{x}^{\top}\mathbf{b}_k)_+ = (\mathbf{x}^{\top}\mathbf{P}_{\mathbf{\Delta}}^{\perp}\mathbf{b}_k)_+$ for all $k \in [q]$.

For any $k \in [q]$ associated with an arbitrary fixed $\mathbf{b}_k \in \mathbb{S}^{d-1}$, let $\mathbf{X}_k \triangleq \mathbf{X}_k\mathbf{P}_{\mathbf{\Delta}}^{\perp} + \mathbf{X}_k\mathbf{P}_{\mathbf{\Delta}} \in \mathcal{X}^{N_k}$ be the inclusion-wisely maximum row subset of $\mathbf{X}$ such that $\mathbf{X}_k\mathbf{b}_k > \mathbf{0}$ element-wisely. Meanwhile, we denote $\mathcal{A}(\mathbf{X}_k) = \mathbf{M}_k\mathbf{X}_k\mathbf{P}_{\mathbf{\Delta}}^{\perp} + \mathcal{A}(\mathbf{X}_k)\mathbf{P}_{\mathbf{\Delta}} \in \mathcal{X}^{\alpha N_k}$ as the augmentation of $\mathbf{X}_k$ where $\mathbf{M}_k \in \mathbb{R}^{\alpha N_k \times N_k}$ is the vertical stack of α identity matrices with size $N_k \times N_k$. Then the DAC constraint implies that $(\mathcal{A}(\mathbf{X}_k) - \mathbf{M}_k\mathbf{X}_k)\mathbf{P}_{\mathbf{\Delta}}\mathbf{b}_k = \mathbf{0}$.

With Assumption 1, for a fixed $\mathbf{b}_k \in \mathbb{S}^{d-1}$, $\mathbb{P}[\mathbf{x}^{\top}\mathbf{b}_k > 0] = \frac{1}{2}$. Then, with the Chernoff bound,

$$\mathbb{P}\left[N_k < \frac{N}{2} - t\right] \leq e^{-\frac{2t^2}{N}},$$

[4] A matrix is totally invertible if all its square submatrices are invertible.

which implies that, $N_k \geq \frac{N}{4}$ with high probability.

Leveraging the assumptions in Theorem 6, $\alpha N \geq 4d_{aug}$ implies that $\alpha N_k \geq d_{aug}$. Therefore by Lemma 1, $\text{Row}\left(\mathcal{A}(\mathbf{X}_k) - \mathbf{M}_k\mathbf{X}_k\right) = \text{Row}\left(\mathbf{\Delta}\right)$ with probability 1. Thus, $(\mathcal{A}(\mathbf{X}_k) - \mathbf{M}_k\mathbf{X}_k)\mathbf{P}_{\mathbf{\Delta}}\mathbf{b}_k = \mathbf{0}$ enforces that $\mathbf{P}_{\mathbf{\Delta}}\mathbf{b}_k = \mathbf{0}$. ■

Proof of Theorem 6. Conditioned on $\mathbf{X}$ and $\mathbf{\Delta}$, we are interested in the excess risk $L(\widehat{h}^{dac}) - L(h^*) = \frac{1}{N}\left\|(\mathbf{X}\widehat{\mathbf{B}}^{dac})_+\widehat{\mathbf{w}}^{dac} - (\mathbf{X}\mathbf{B}^*)_+\mathbf{w}^*\right\|_2^2$ with randomness on $\boldsymbol{\epsilon}$.

We first recall that Lemma 2 implies $\widehat{h}^{dac} \in \mathcal{H}_{dac} = \left\{h(\mathbf{x}) = \left(\mathbf{x}^\top\mathbf{B}\right)_+\mathbf{w} \;\middle|\; \mathbf{B} \in \mathcal{B},\ \|\mathbf{w}\|_1 \leq C_w\right\}$ where

$$\mathcal{B} \triangleq \left\{\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_q] \mid \|\mathbf{b}_k\| = 1\ \forall\, k \in [q], (\mathbf{X}\mathbf{B})_+ = (\mathbf{X}\mathbf{P}_{\mathbf{\Delta}}^{\perp}\mathbf{B})_+\right\}.$$

Leveraging Equation (21) and (22) in Du et al. (2020), since $(\mathbf{B}^*, \mathbf{w}^*)$ is feasible under the constraint, by the basic inequality,

$$\left\|\mathbf{y} - (\mathbf{X}\widehat{\mathbf{B}}^{dac})_+\widehat{\mathbf{w}}^{dac}\right\|_2^2 \leq \|\mathbf{y} - (\mathbf{X}\mathbf{B}^*)_+\mathbf{w}^*\|_2^2. \tag{5}$$

Knowing that $\mathbf{y} = (\mathbf{X}\mathbf{B}^*)_+\mathbf{w}^* + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_N\right)$, we can rewrite Equation (5) as

$$\begin{aligned}\frac{1}{N}\left\|(\mathbf{X}\widehat{\mathbf{B}}^{dac})_+\widehat{\mathbf{w}}^{dac} - (\mathbf{X}\mathbf{B}^*)_+\mathbf{w}^*\right\|_2^2 \leq& \frac{2}{N}\boldsymbol{\epsilon}^\top\left((\mathbf{X}\widehat{\mathbf{B}}^{dac})_+\widehat{\mathbf{w}}^{dac} - (\mathbf{X}\mathbf{B}^*)_+\mathbf{w}^*\right)\\ \leq& 4\sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}\boldsymbol{\epsilon}^\top h(\mathbf{X})\end{aligned}$$

First, we observe that $\sigma^{-1}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}\boldsymbol{\epsilon}^\top h(\mathbf{X})\right] = \widehat{\mathfrak{G}}_{\mathbf{X}}\left(\mathcal{H}_{dac}\right)$ measures the empirical Gaussian width of $\mathcal{H}_{dac}$ over $\mathbf{X}$. Moreover, by observing that for any $h \in \mathcal{H}_{dac}$ and $\mathbf{x}_i \in \mathbf{X}$,

$$\begin{aligned}&|h(\mathbf{x}_i)| \leq \left\|\left(\mathbf{B}^\top\mathbf{x}_i\right)_+\right\|_\infty\|\mathbf{w}\|_1 \leq \max_{k\in[q]}\left|\mathbf{b}_k^\top\mathbf{P}_{\mathbf{\Delta}}^{\perp}\mathbf{x}_i\right|\|\mathbf{w}\|_1 \leq \left\|\mathbf{P}_{\mathbf{\Delta}}^{\perp}\mathbf{x}_i\right\|_2\|\mathbf{w}\|_1,\\ &\frac{1}{N}\|h(\mathbf{X})\|_2^2 = \frac{1}{N}\sum_{i=1}^N|h(\mathbf{x}_i)|^2 \leq \|\mathbf{w}\|_1^2\cdot\frac{1}{N}\sum_{i=1}^N\left\|\mathbf{P}_{\mathbf{\Delta}}^{\perp}\mathbf{x}_i\right\|_2^2 \leq C_w^2C_{\mathcal{N}}^2\end{aligned}$$

and

$$\begin{aligned}&\left|\sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}\boldsymbol{\epsilon}_1^\top h(\mathbf{X}) - \sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}\boldsymbol{\epsilon}_2^\top h(\mathbf{X})\right|\\ \leq&\left|\sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}h(\mathbf{X})^\top(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_2)\right|\\ \leq&\frac{1}{\sqrt{N}}\left\|\frac{1}{\sqrt{N}}h(\mathbf{X})\right\|_2\|\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_2\|_2\\ \leq&\frac{C_wC_{\mathcal{N}}}{\sqrt{N}}\|\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_2\|_2,\end{aligned}$$

we know that the function $\boldsymbol{\epsilon} \to \sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}\boldsymbol{\epsilon}^\top h(\mathbf{X})$ is $\frac{C_{\mathcal{N}}C_w}{\sqrt{N}}$-Lipschitz in ℓ_2 norm. Therefore, by Wainwright (2019) Theorem 2.26, we have that with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}_{dac}}\frac{1}{N}\boldsymbol{\epsilon}^\top h(\mathbf{X}) \leq \sigma\cdot\left(\widehat{\mathfrak{G}}_{\mathbf{X}}\left(\mathcal{H}_{dac}\right) + C_wC_{\mathcal{N}}\sqrt{\frac{2\log(1/\delta)}{N}}\right),$$

where the empirical Gaussian complexity is upper bounded by

$$\begin{aligned}
\widehat{\mathfrak{G}}_{\mathbf{X}}\left(\mathcal{H}_{dac}\right) =& \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_N)}\left[\sup_{\mathbf{B}\in\mathcal{B},\|\mathbf{w}\|_1\leq R}\frac{1}{N}\mathbf{g}^\top(\mathbf{X}\mathbf{B})_+\mathbf{w}\right] \\
\leq& \frac{C_w}{N}\mathop{\mathbb{E}}_{\mathbf{g}}\left[\sup_{\mathbf{B}\in\mathcal{B}}\left\|(\mathbf{X}\mathbf{B})_+^\top\mathbf{g}\right\|_\infty\right] \\
=& \frac{C_w}{N}\mathop{\mathbb{E}}_{\mathbf{g}}\left[\sup_{\mathbf{b}\in\mathbb{S}^{d-1}}\mathbf{g}^\top\left(\mathbf{X}\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{b}\right)_+\right] \quad \text{(Lemma 6, } (\cdot)_+ \text{ is 1-Lipschitz)} \\
\leq& \frac{C_w}{N}\mathop{\mathbb{E}}_{\mathbf{g}}\left[\sup_{\mathbf{b}\in\mathbb{S}^{d-1}}\mathbf{g}^\top\mathbf{X}\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{b}\right] \\
=& \frac{C_w}{N}\mathop{\mathbb{E}}_{\mathbf{g}}\left[\left\|\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{X}^\top\mathbf{g}\right\|_2\right] \\
\leq& \frac{C_w}{N}\left(\mathop{\mathbb{E}}_{\mathbf{g}}\left[\left\|\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{X}^\top\mathbf{g}\right\|_2^2\right]\right)^{1/2} \\
=& \frac{C_w}{N}\sqrt{\operatorname{tr}(\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{X}^\top\mathbf{X}\mathbf{P}_{\boldsymbol{\Delta}}^\perp)} \\
=& \frac{C_w C_{\mathcal{N}}}{\sqrt{N}}.
\end{aligned}$$

■

Proof of Corollary 1. By Definition 3, we have with probability at least $1-\delta$ that $d_{aug} = \operatorname{rank}(\mathbf{P}_{\boldsymbol{\Delta}}) \geq d_{aug}(\delta)$ and $\operatorname{rank}(\mathbf{P}_{\boldsymbol{\Delta}}^\perp) \leq d - d_{aug}(\delta)$. Meanwhile, leveraging Lemma 5, we have that under Assumption 1 and with $N \gg \rho^4 d$, with high probability,

$$\left\|\frac{1}{N}\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{X}^\top\mathbf{X}\mathbf{P}_{\boldsymbol{\Delta}}^\perp\right\|_2 \leq \left\|\frac{1}{N}\mathbf{X}^\top\mathbf{X}\right\|_2 \leq 1.1C \lesssim 1.$$

Therefore, there exists $C_{\mathcal{N}} > 0$ with $\frac{1}{N}\sum_{i=1}^n \left\|\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{x}_i\right\|_2^2 \leq C_{\mathcal{N}}^2$ such that, with probability at least $1-\delta$,

$$C_{\mathcal{N}}^2 \leq (d - d_{aug}) \cdot \left\|\frac{1}{N}\mathbf{P}_{\boldsymbol{\Delta}}^\perp\mathbf{X}^\top\mathbf{X}\mathbf{P}_{\boldsymbol{\Delta}}^\perp\right\|_2 \lesssim d - d_{aug}(\delta).$$

■

C Classification with Expansion-based Augmentations

We first recall the multi-class classification problem setup in Section 6.2, while introducing some helpful notions. For an arbitrary set $\mathcal{X}$, let $\mathcal{Y} = [K]$, and $h^* : \mathcal{X} \to [K]$ be the ground truth classifier that partitions $\mathcal{X}$: for each $k \in [K]$, let $\mathcal{X}_k \triangleq \{\mathbf{x} \in \mathcal{X} \mid h^*(\mathbf{x}) = k\}$, with $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset, \forall i \neq j$. In addition, for an arbitrary classifier $h : \mathcal{X} \to [K]$, we denote the majority label with respect to h for each class,

$$\widehat{y}_k \triangleq \mathop{\operatorname{argmax}}_{y\in[K]} \mathbb{P}_P\left[h(\mathbf{x}) = y \,\middle|\, \mathbf{x} \in \mathcal{X}_k\right] \quad \forall\, k \in [K],$$

along with the respective class-wise local and global minority sets,

$$M_k \triangleq \left\{\mathbf{x} \in \mathcal{X}_k \,\middle|\, h(\mathbf{x}) \neq \widehat{y}_k\right\} \subsetneq \mathcal{X}_k \quad \forall\, k \in [K], \quad M \triangleq \bigcup_{k=1}^K M_k.$$

Given the marginal distribution $P(\mathbf{x})$, we introduce the *expansion-based data augmentations* that concretizes Definition 1 in the classification setting:

Definition 4 (Expansion-based data augmentations, Cai et al. (2021)). *We call $\mathcal{A}: \mathcal{X} \to 2^{\mathcal{X}}$ an augmentation function that induces expansion-based data augmentations if $\mathcal{A}$ is class invariant: $\{\mathbf{x}\} \subsetneq \mathcal{A}(\mathbf{x}) \subseteq \{\mathbf{x}' \in \mathcal{X} \mid h^*(\mathbf{x}) = h^*(\mathbf{x}')\}$ for all $\mathbf{x} \in \mathcal{X}$. Let*

$$NB(\mathbf{x}) \triangleq \left\{\mathbf{x}' \in \mathcal{X} \;\middle|\; \mathcal{A}(\mathbf{x}) \cap \mathcal{A}(\mathbf{x}') \neq \emptyset\right\}, \quad NB(S) \triangleq \cup_{\mathbf{x} \in S} NB(\mathbf{x})$$

be the neighborhoods of $\mathbf{x} \in \mathcal{X}$ and $S \subseteq \mathcal{X}$ with respect to $\mathcal{A}$. Then, $\mathcal{A}$ satisfies

(a) (q, ξ)-constant expansion if given any $S \subseteq \mathcal{X}$ with $P(S) \geq q$ and $P(S \cap \mathcal{X}_k) \leq \frac{1}{2}$ for all $k \in [K]$, $P(NB(S)) \geq \min\{P(S), \xi\} + P(S)$;

(b) (a, c)-multiplicative expansion if for all $k \in [K]$, given any $S \subseteq \mathcal{X}$ with $P(S \cap \mathcal{X}_k) \leq a$, $P(NB(S) \cap \mathcal{X}_k) \geq \min\{c \cdot P(S \cap \mathcal{X}_k), 1\}$.

On Definition 4, we first point out that the ground truth classifier is invariant throughout the neighborhood: given any $\mathbf{x} \in \mathcal{X}$, $h^*(\mathbf{x}) = h^*(\mathbf{x}')$ for all $\mathbf{x}' \in NB(\mathbf{x})$. Second, in contrast to the linear regression and two-layer neural network cases where we assume $\mathcal{X} \subseteq R^d$, with the expansion-based data augmentation over a general $\mathcal{X}$, the notion of d_{aug} in Definition 3 is not well-established. Alternatively, we leverage the concept of constant / multiplicative expansion from Cai et al. (2021), and quantify the augmentation strength with parameters (q, ξ) or (a, c). Intuitively, the strength of expansion-based data augmentations is characterized by expansion capability of $\mathcal{A}$: for a neighborhood $S \subseteq \mathcal{X}$ of proper size (characterized by q or a under measure P), the stronger augmentation $\mathcal{A}$ leads to more expansion in $NB(S)$, and therefore larger ξ or c. For example in Definition 2, we use an expansion-based augmentation function $\mathcal{A}$ that satisfies $\left(\frac{1}{2}, c\right)$-multiplicative expansion.

Adapting the existing setting in Wei et al. (2021); Cai et al. (2021), we concretize the classifier class $\mathcal{H}$ with a function class $\mathcal{F} \subseteq \left\{f: \mathcal{X} \to \mathbb{R}^K\right\}$ of fully connected neural networks such that $\mathcal{H} = \left\{h(\mathbf{x}) \triangleq \operatorname{argmax}_{k \in [K]} \; f(\mathbf{x})_k \;\middle|\; f \in \mathcal{F}\right\}$. To constrain the feasible hypothesis class through the DAC regularization with finite unlabeled samples, we recall the notion of all-layer-margin, $m: \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ (from Wei et al. (2021)) that measures the maximum possible perturbation in all layers of f while maintaining the prediction y. Precisely, given any $f \in \mathcal{F}$ such that $f(\mathbf{x}) = \mathbf{W}_p \varphi(\ldots \varphi(\mathbf{W}_1 \mathbf{x}) \ldots)$ for some activation function $\varphi: \mathbb{R} \to \mathbb{R}$ and parameters $\left\{\mathbf{W}_\iota \in \mathbb{R}^{d_\iota \times d_{\iota-1}}\right\}_{\iota=1}^p$, we can write $f = f_{2p-1} \circ \cdots \circ f_1$ where $f_{2\iota-1}(\mathbf{x}) = \mathbf{W}_\iota \mathbf{x}$ for all $\iota \in [p]$ and $f_{2\iota}(\mathbf{z}) = \varphi(\mathbf{z})$ for $\iota \in [p-1]$. For an arbitrary set of perturbation vectors $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{2p-1})$ such that $\boldsymbol{\delta}_{2\iota-1}, \boldsymbol{\delta}_{2\iota} \in \mathbb{R}^{d_\iota}$ for all ι, let $f(\mathbf{x}, \boldsymbol{\delta})$ be the perturbed neural network defined recursively such that

$$\begin{aligned}
&\widetilde{\mathbf{z}}_1 = f_1(\mathbf{x}) + \|\mathbf{x}\|_2 \boldsymbol{\delta}_1, \\
&\widetilde{\mathbf{z}}_\iota = f_\iota(\widetilde{\mathbf{z}}_{\iota-1}) + \|\widetilde{\mathbf{z}}_{\iota-1}\|_2 \boldsymbol{\delta}_\iota \quad \forall\, \iota = 2, \ldots, 2p-1, \\
&f(\mathbf{x}, \boldsymbol{\delta}) = \widetilde{\mathbf{z}}_{2p-1}.
\end{aligned}$$

The all-layer-margin $m(f, \mathbf{x}, y)$ measures the minimum norm of the perturbation $\boldsymbol{\delta}$ such that $f(\mathbf{x}, \boldsymbol{\delta})$ fails to provide the classification y,

$$m(f, \mathbf{x}, y) \triangleq \min_{\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{2p-1})} \sqrt{\sum_{\iota=1}^{2p-1} \|\boldsymbol{\delta}_\iota\|_2^2} \quad \text{s.t.} \quad \operatorname*{argmax}_{k \in [K]} f(\mathbf{x}, \boldsymbol{\delta})_k \neq y. \tag{6}$$

With the notion of all-layer-margin established, for any $\mathcal{A}: \mathcal{X} \to 2^{\mathcal{X}}$ that satisfies conditions in Definition 4, the robust margin is defined as

$$m_{\mathcal{A}}(f, \mathbf{x}) \triangleq \sup_{\mathbf{x}' \in \mathcal{A}(\mathbf{x})} m\left(f, \mathbf{x}', \operatorname*{argmax}_{k \in [K]} f(\mathbf{x})_k\right).$$

Intuitively, the robust margin measures the maximum possible perturbation in all-layer weights of f such that predictions on all data augmentations of $\mathbf{x}$ remain consistent. For instance, $m_{\mathcal{A}}(f, \mathbf{x}) > 0$ is equivalent to enforcing $h(\mathbf{x}) = h(\mathbf{x}')$ for all $\mathbf{x}' \in \mathcal{A}(\mathbf{x})$.

To achieve finite sample guarantees, DAC regularization requires stronger consistency conditions than merely consistent classification outputs (*i.e.*, $m_{\mathcal{A}}(f, \mathbf{x}) > 0$). Instead, we enforce $m_{\mathcal{A}}(f, \mathbf{x}) > \tau$ for any $0 < \tau < \max_{f \in \mathcal{F}} \inf_{\mathbf{x} \in \mathcal{X}} m_{\mathcal{A}}(f, \mathbf{x})$[5] over an finite set of unlabeled samples $\mathbf{X}^u$ that is independent of $\mathbf{X}$ and drawn *i.i.d.* from

[5]The upper bound on τ ensures the proper learning setting, *i.e.*, there exists $f \in \mathcal{F}$ such that $m_{\mathcal{A}}(f, \mathbf{x}) > \tau$ for all $\mathbf{x} \in \mathcal{X}$.

$P(\mathbf{x})$. Then, learning the classifier with zero-one loss $l_{01}\left(h(\mathbf{x}), y\right) = \mathbf{1}\left\{h(\mathbf{x}) \neq y\right\}$ from a class of p-layer fully connected neural networks with maximum width q,

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R}^K \;\middle|\; f = f_{2p-1} \circ \cdots \circ f_1,\ f_{2\iota-1}(\mathbf{x}) = \mathbf{W}_\iota \mathbf{x},\ f_{2\iota}(\mathbf{z}) = \varphi(\mathbf{z}) \right\},$$

where $\mathbf{W}_\iota \in \mathbb{R}^{d_\iota \times d_{\iota-1}}$ for all $\iota \in [p]$, and $q \triangleq \max_{\iota=0,\ldots,p} d_\iota$, we solve

$$\begin{aligned} \widehat{h}^{dac} \;\triangleq \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L}_{01}^{dac}(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{h\left(\mathbf{x}_i\right) \neq h^*\left(\mathbf{x}_i\right)\right\} \\ \text{s.t.} \quad m_{\mathcal{A}}(f, \mathbf{x}_i^u) > \tau \quad \forall\, i \in [|\mathbf{X}^u|] \end{aligned} \tag{7}$$

for any $0 < \tau < \max_{f \in \mathcal{F}} \inf_{\mathbf{x} \in \mathcal{X}} m_{\mathcal{A}}(f, \mathbf{x})$. The corresponding reduced function class is given by

$$\mathcal{H}_{dac} \triangleq \left\{ h \in \mathcal{H} \mid m_{\mathcal{A}}(f, \mathbf{x}_i^u) > \tau \quad \forall\, i \in [|\mathbf{X}^u|] \right\}.$$

Specifically, with $\mu \triangleq \sup_{h \in \mathcal{H}_{dac}} \mathbb{P}_P\left[\exists\, \mathbf{x}' \in \mathcal{A}(\mathbf{x}) : h(\mathbf{x}) \neq h(\mathbf{x}')\right]$, Wei et al. (2021); Cai et al. (2021) demonstrate the following for $\mathcal{H}_{dac}$:

Proposition 7 (Wei et al. (2021) Theorem 3.7, Cai et al. (2021) Proposition 2.2)**.** *For any $\delta \in (0,1)$, with probability at least $1 - \delta/2$ (over $\mathbf{X}^u$),*

$$\mu \leq \widetilde{O}\left(\frac{\sum_{\iota=1}^{p} \sqrt{q}\, \|\mathbf{W}_\iota\|_F}{\tau \sqrt{|\mathbf{X}^u|}} + \sqrt{\frac{\log\left(1/\delta\right) + p \log |\mathbf{X}^u|}{|\mathbf{X}^u|}} \right),$$

where $\widetilde{O}\left(\cdot\right)$ hides polylogarithmic factors in $|\mathbf{X}^u|$ and d.

Leveraging the existing theory above on finite sample guarantee of the maximum possible inconsistency, we have the following.

Theorem 8 (Formal restatement of Theorem 4 on classification with DAC)**.** *Learning the classifier with DAC regularization in Equation (7) provides that, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$L_{01}\left(\widehat{h}^{dac}\right) - L_{01}\left(h^*\right) \leq 4\mathfrak{R} + \sqrt{\frac{2\log(4/\delta)}{N}}, \tag{8}$$

where with $0 < \mu < 1$ defined in Proposition 7, for any $0 \leq q < \frac{1}{2}$ and $c > 1 + 4\mu$,

(a) when $\mathcal{A}$ satisfies $(q, 2\mu)$-constant expansion, $\mathfrak{R} \leq \sqrt{\frac{2K\log(2N)}{N} + 2\max\left\{q, 2\mu\right\}}$;

(b) when $\mathcal{A}$ satisfies $(\frac{1}{2}, c)$-multiplicative expansion, $\mathfrak{R} \leq \sqrt{\frac{2K\log(2N)}{N} + \frac{4\mu}{\min\{c-1,1\}}}$.

First, to quantify the function class complexity and relate it to the generalization error, we leverage the notion of Rademacher complexity and the associated standard generalization bound.

Lemma 3. *Given a fixed function class $\mathcal{H}_{dac}$ (i.e., conditioned on $\mathbf{X}^u$) and a B-bounded and C_l-Lipschitz loss function l, let $\widehat{L}(h) = \frac{1}{N}\sum_{i=1}^{N} l(h(\mathbf{x}_i), y_i)$, $L(h) = \mathbb{E}\left[l(h(\mathbf{x}_i), y_i)\right]$, and $\widehat{h}^{dac} = \operatorname{argmin}_{h \in \mathcal{H}_{dac}} \widehat{L}(h)$. Then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over $\mathbf{X}$,*

$$L(\widehat{h}^{dac}) - L(h^*) \leq 4C_l \cdot \mathfrak{R}_N\left(\mathcal{H}_{dac}\right) + \sqrt{\frac{2B^2\log(4/\delta)}{N}}.$$

Proof of Lemma 3. We first decompose the expected excess risk as

$$L(\widehat{h}^{dac}) - L(h^*) = \left(L(\widehat{h}^{dac}) - \widehat{L}(\widehat{h}^{dac})\right) + \left(\widehat{L}(\widehat{h}^{dac}) - \widehat{L}(h^*)\right) + \left(\widehat{L}(h^*) - L(h^*)\right),$$

where $\widehat{L}(\widehat{h}^{dac}) - \widehat{L}(h^*) \leq 0$ by the basic inequality. Since both $\widehat{h}^{dac}, h^* \in \mathcal{H}_{dac}$, we then have

$$L(\widehat{h}^{dac}) - L(h^*) \leq 2 \sup_{h \in \mathcal{H}_{dac}} \left| L(h) - \widehat{L}(h) \right|.$$

Let $g^+(\mathbf{X},\mathbf{y}) = \sup_{h\in\mathcal{H}_{dac}} : L(h) - \widehat{L}(h)$ and $g^-(\mathbf{X},\mathbf{y}) = \sup_{h\in\mathcal{H}_{dac}} : -L(h) + \widehat{L}(h)$. Then,

$$\mathbb{P}\left[L(\widehat{h}^{dac}) - L(h^*) \geq \epsilon\right] \leq \mathbb{P}\left[g^+(\mathbf{X},\mathbf{y}) \geq \frac{\epsilon}{2}\right] + \mathbb{P}\left[g^-(\mathbf{X},\mathbf{y}) \geq \frac{\epsilon}{2}\right].$$

We will derive a tail bound for $g^+(\mathbf{X},\mathbf{y})$ with the standard inequalities and symmetrization argument Wainwright (2019); Bartlett and Mendelson (2003), while the analogous statement holds for $g^-(\mathbf{X},\mathbf{y})$.

Let $(\mathbf{X}^{(1)},\mathbf{y}^{(1)})$ be a sample set generated by replacing an arbitrary sample in $(\mathbf{X},\mathbf{y})$ with an independent sample $(\mathbf{x},y) \sim P(\mathbf{x},y)$. Since l is B-bounded, we have $\left|g^+(\mathbf{X},\mathbf{y}) - g^+(\mathbf{X}^{(1)},\mathbf{y}^{(1)})\right| \leq B/N$. Then, via McDiarmid's inequality Bartlett and Mendelson (2003),

$$\mathbb{P}\left[g^+(\mathbf{X},\mathbf{y}) \geq \mathbb{E}[g^+(\mathbf{X},\mathbf{y})] + t\right] \leq \exp\left(-\frac{2Nt^2}{B^2}\right).$$

For an arbitrary sample set $(\mathbf{X},\mathbf{y})$, let $\widehat{L}_{(\mathbf{X},\mathbf{y})}(h) = \frac{1}{N}\sum_{i=1}^N l\left(h(\mathbf{x}_i),y_i\right)$ be the empirical risk of h with respect to $(\mathbf{X},\mathbf{y})$. Then, by a classical symmetrization argument (e.g., proof of Wainwright (2019) Theorem 4.10), we can bound the expectation: for an independent sample set $(\mathbf{X}',\mathbf{y}') \in \mathcal{X}^N \times \mathcal{Y}^N$ drawn *i.i.d.* from P,

$$\begin{aligned}
\mathbb{E}\left[g^+(\mathbf{X},\mathbf{y})\right] =& \mathbb{E}_{(\mathbf{X},\mathbf{y})}\left[\sup_{h\in\mathcal{H}_{dac}} L(h) - \widehat{L}_{(\mathbf{X},\mathbf{y})}(h)\right] \\
=& \mathbb{E}_{(\mathbf{X},\mathbf{y})}\left[\sup_{h\in\mathcal{H}_{dac}} \mathbb{E}_{(\mathbf{X}',\mathbf{y}')}\left[\widehat{L}_{(\mathbf{X}',\mathbf{y}')}(h)\right] - \widehat{L}_{(\mathbf{X},\mathbf{y})}(h)\right] \\
=& \mathbb{E}_{(\mathbf{X},\mathbf{y})}\left[\sup_{h\in\mathcal{H}_{dac}} \mathbb{E}_{(\mathbf{X}',\mathbf{y}')}\left[\widehat{L}_{(\mathbf{X}',\mathbf{y}')}(h) - \widehat{L}_{(\mathbf{X},\mathbf{y})}(h) \,\middle|\, (\mathbf{X},\mathbf{y})\right]\right] \\
\leq& \mathbb{E}_{(\mathbf{X},\mathbf{y})}\left[\mathbb{E}_{(\mathbf{X}',\mathbf{y}')}\left[\sup_{h\in\mathcal{H}_{dac}} \widehat{L}_{(\mathbf{X}',\mathbf{y}')}(h) - \widehat{L}_{(\mathbf{X},\mathbf{y})}(h) \,\middle|\, (\mathbf{X},\mathbf{y})\right]\right] \\
& \text{(Law of iterated conditional expectation)} \\
=& \mathbb{E}_{(\mathbf{X},\mathbf{y},\mathbf{X}',\mathbf{y}')}\left[\sup_{h\in\mathcal{H}_{dac}} \widehat{L}_{(\mathbf{X}',\mathbf{y}')}(h) - \widehat{L}_{(\mathbf{X},\mathbf{y})}(h)\right]
\end{aligned}$$

Since $(\mathbf{X},\mathbf{y}),(\mathbf{X}',\mathbf{y}')$ are drawn *i.i.d.* from P, we can introduce *i.i.d.* Rademacher random variables $\mathbf{r} = \{r_i \in \{-1,+1\} \mid i \in [N]\}$ (independent of both $(\mathbf{X},\mathbf{y})$ and $(\mathbf{X}',\mathbf{y}')$) such that

$$\begin{aligned}
\mathbb{E}\left[g^+(\mathbf{X},\mathbf{y})\right] \leq& \mathbb{E}_{(\mathbf{X},\mathbf{y},\mathbf{X}',\mathbf{y}',\mathbf{r})}\left[\sup_{h\in\mathcal{H}_{dac}} \frac{1}{N}\sum_{i=1}^N r_i \cdot \left(l\left(h\left(\mathbf{x}_i'\right),y_i'\right) - l\left(h\left(\mathbf{x}_i\right),y_i\right)\right)\right] \\
\leq& 2\, E_{(\mathbf{X},\mathbf{y},\mathbf{r})}\left[\sup_{h\in\mathcal{H}_{dac}} \frac{1}{N}\sum_{i=1}^N r_i \cdot l\left(h\left(\mathbf{x}_i\right),y_i\right)\right] \\
\leq& 2\,\mathfrak{R}_N\left(l\circ\mathcal{H}_{dac}\right)
\end{aligned}$$

where $l\circ\mathcal{H}_{dac} = \{l(h(\cdot),\cdot) : \mathcal{X}\times\mathcal{Y}\to\mathbb{R} : h\in\mathcal{H}_{dac}\}$ is the loss function class, and

$$\mathfrak{R}_N\left(\mathcal{F}\right) \triangleq E_{(\mathbf{X},\mathbf{y},\mathbf{r})}\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N r_i \cdot f\left(\mathbf{x}_i,y_i\right)\right]$$

denotes the Rademacher complexity. Analogously, $\mathbb{E}[g^-(\mathbf{X},\mathbf{y})] \leq 2\mathfrak{R}_N\left(l\circ\mathcal{H}_{dac}\right)$. Therefore, assuming that $T^{dac}_{\tilde{A},\mathbf{X}}(\mathcal{H}) \subseteq \mathcal{H}_{dac}(\mathcal{H})$ holds, with probability at least $1-\delta/2$,

$$L(\widehat{h}^{dac}) - L(h^*) \leq 4\mathfrak{R}_N\left(l\circ\mathcal{H}_{dac}\right) + \sqrt{\frac{2B^2\log(4/\delta)}{N}}$$

Finally, since $l(\cdot,y)$ is C_l-Lipschitz for all $y \in \mathcal{Y}$, by Ledoux and Talagrand (2013) Theorem 4.12, we have $\mathfrak{R}_N\left(l\circ\mathcal{H}_{dac}\right) \leq C_l\cdot\mathfrak{R}_N\left(\mathcal{H}_{dac}\right)$. ■

Lemma 4 (Cai et al. (2021), Lemma A.1)**.** *For any $h\in\mathcal{H}_{dac}$, when P satisfies*

(a) $(q, 2\mu)$*-constant expansion with* $q < \frac{1}{2}$, $P(M) \leq \max\{q, 2\mu\}$;
(b) $\left(\frac{1}{2}, c\right)$*-multiplicative expansion with* $c > 1 + 4\mu$, $P(M) \leq \max\left\{\frac{2\mu}{c-1}, 2\mu\right\}$.

Proof of Lemma 4. We start with the proof for Lemma 4 (a). By definition of M_k and $\widehat{y}_k$, we know that $M_k = M \cap \mathcal{X}_k \leq \frac{1}{2}$. Therefore, for any $0 < q < \frac{1}{2}$, one of the following two cases holds:

(i) $P(M) < q$;
(ii) $P(M) \geq q$. Since $P(M \cap \mathcal{X}_k) < \frac{1}{2}$ for all $k \in [K]$ holds by construction, with the $(q, 2\mu)$-constant expansion, $P(NB(M)) \geq \min\{P(M), 2\mu\} + P(M)$.
Meanwhile, since the ground truth classifier h^* is invariant throughout the neighborhoods, $NB(M_k) \cap NB(M_{k'}) = \emptyset$ for $k \neq k'$, and therefore $NB(M) \setminus M = \bigcup_{k=1}^{K} NB(M_k) \setminus M_k$ with each $NB(M_k) \setminus M_k$ disjoint. Then, we observe that for each $\mathbf{x} \in NB(M) \setminus M$, here exists some $k = h^*(\mathbf{x})$ such that $\mathbf{x} \in NB(M_k) \setminus M_k$. $\mathbf{x} \in \mathcal{X}_k \setminus M_k$ implies that $h(\mathbf{x}) = \widehat{y}_k$, while $\mathbf{x} \in NB(M_k)$ suggests that there exists some $\mathbf{x}' \in \mathcal{A}(\mathbf{x}) \cap \mathcal{A}(\mathbf{x}'')$ where $\mathbf{x}'' \in M_k$ such that either $h(\mathbf{x}') = \widehat{y}_k$ and $h(\mathbf{x}') \neq h(\mathbf{x}'')$ for $\mathbf{x}' \in \mathcal{A}(\mathbf{x}'')$, or $h(\mathbf{x}') \neq \widehat{y}_k$ and $h(\mathbf{x}') \neq h(\mathbf{x})$ for $\mathbf{x}' \in \mathcal{A}(\mathbf{x})$. Therefore, we have

$$P(NB(M) \setminus M) \leq 2\mathbb{P}_P[\exists\, \mathbf{x}' \in \mathcal{A}(\mathbf{x}) \text{ s.t. } h(\mathbf{x}) \neq h(\mathbf{x}')] \leq 2\mu.$$

Moreover, since $P(NB(M)) - P(M) \leq P(NB(M) \setminus M) \leq 2\mu$, we know that

$$\min\{P(M), 2\mu\} + P(M) \leq P(NB(M)) \leq P(M) + 2\mu.$$

That is, $P(M) \leq 2\mu$.

Overall, we have $P(M) \leq \max\{q, 2\mu\}$.

To show Lemma 4 (b), we recall from Wei et al. (2021) Lemma B.6 that for any $c > 1 + 4\mu$, $\left(\frac{1}{2}, c\right)$-multiplicative expansion implies $\left(\frac{2\mu}{c-1}, 2\mu\right)$-constant expansion. Then leveraging the proof for Lemma 4 (a), with $q = \frac{2\mu}{c-1}$, we have $P(M) \leq \max\left\{\frac{2\mu}{c-1}, 2\mu\right\}$. ■

Proof of Theorem 8. To show Equation (8), we leverage Lemma 3 and observe that $B = 1$ with the zero-one loss. Therefore, conditioned on $\mathcal{H}_{dac}$ (which depends only on $\mathbf{X}^u$ but not on $\mathbf{X}$), for any $\delta \in (0, 1)$, with probability at least $1 - \delta/2$,

$$L_{01}\left(\widehat{h}^{dac}\right) - L_{01}(h^*) \leq 4\mathfrak{R}_N(l_{01} \circ \mathcal{H}_{dac}) + \sqrt{\frac{2\log(4/\delta)}{N}}.$$

For the upper bounds of the Rademacher complexity, let $\widetilde{\mu} \triangleq \sup_{h \in \mathcal{H}_{dac}} P(M)$ where M denotes the global minority set with respect to $h \in \mathcal{H}_{dac}$. Lemma 4 suggests that

(a) when P satisfies $(q, 2\mu)$-constant expansion for some $q < \frac{1}{2}$, $\widetilde{\mu} \leq \max\{q, 2\mu\}$; while
(b) when P satisfies $(\frac{1}{2}, c)$-multiplicative expansion for some $c > 1 + 4\mu$, $\widetilde{\mu} \leq \frac{2\mu}{\min\{c-1, 1\}}$.

Then, it is sufficient to show that, conditioned on $\mathcal{H}_{dac}$,

$$\mathfrak{R}_N(l_{01} \circ \mathcal{H}_{dac}) \leq \sqrt{\frac{2K\log(2N)}{N} + 2\widetilde{\mu}}. \tag{9}$$

To show this, we first consider a fixed set of n observations in $\mathcal{X}$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathcal{X}^N$. Let the number of distinct behaviors over $\mathbf{X}$ in $\mathcal{H}_{dac}$ be

$$\mathfrak{s}(l_{01} \circ \mathcal{H}_{dac}, \mathbf{X}) \triangleq \left|\left\{[l_{01} \circ h(\mathbf{x}_1), \ldots, l_{01} \circ h(\mathbf{x}_N)] \;\middle|\; h \in \mathcal{H}_{dac}\right\}\right|.$$

Then, by the Massart's finite lemma, the empirical rademacher complexity with respect to $\mathbf{X}$ is upper bounded by

$$\widehat{\mathfrak{R}}_{\mathbf{X}}(l_{01} \circ \mathcal{H}_{dac}) \leq \sqrt{\frac{2\log \mathfrak{s}(l_{01} \circ \mathcal{H}_{dac}, \mathbf{X})}{N}}.$$

By the concavity of $\sqrt{\log(\cdot)}$, we know that,

$$\begin{aligned}\mathfrak{R}_N\left(l_{01}\circ\mathcal{H}_{dac}\right) =& \mathbb{E}_{\mathbf{X}}\left[\widehat{\mathfrak{R}}_{\mathbf{X}}\left(l_{01}\circ\mathcal{H}_{dac}\right)\right] \leq \mathbb{E}_{\mathbf{X}}\left[\sqrt{\frac{2\log\mathfrak{s}\left(l_{01}\circ\mathcal{H}_{dac},\mathbf{X}\right)}{N}}\right]\\ \leq& \sqrt{\frac{2\log\mathbb{E}_{\mathbf{X}}\left[\mathfrak{s}\left(l_{01}\circ\mathcal{H}_{dac},\mathbf{X}\right)\right]}{N}}.\end{aligned} \tag{10}$$

Since $P(M)\leq\widetilde{\mu}\leq\frac{1}{2}$ for all $h\in\mathcal{H}_{dac}$, we have that, conditioned on $\mathcal{H}_{dac}$,

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\mathfrak{s}\left(l_{01}\circ\mathcal{H}_{dac},\mathbf{X}\right)\right] \leq& \sum_{r=0}^{N}\binom{N}{r}\widetilde{\mu}^r\left(1-\widetilde{\mu}\right)^{N-r}\cdot\binom{N-r-1}{\min(K,N-r)-1}2^{K+r}\\ \leq& (2N)^K\sum_{r=0}^{N}\binom{N}{r}\left(2\widetilde{\mu}\right)^r\left(1-\widetilde{\mu}\right)^{N-r}\\ =& (2N)^K\left(1-\widetilde{\mu}+2\widetilde{\mu}\right)^N\\ \leq& (2N)^K\cdot e^{N\widetilde{\mu}}.\end{aligned}$$

Plugging this into Equation (10) yields Equation (9). Finally, the randomness in $\mathcal{H}_{dac}$ is quantified by $\widetilde{\mu},\mu$, and upper bounded by Proposition 7. ■

D Supplementary Application: Domain Adaptation

As a supplementary example, we demonstrate the possible failure of DA-ERM, and alternatively how DAC regularization can serve as a remedy. Concretely, we consider an illustrative linear regression problem in the domain adaptation setting: with training samples drawn from a source distribution P^s and generalization (in terms of excess risk) evaluated over a related but different target distribution P^t. With distinct $\mathbb{E}_{P^s}[y|\mathbf{x}]$ and $\mathbb{E}_{P^t}[y|\mathbf{x}]$, we assume the existence of an unknown but unique inclusionwisely maximal invariant feature subspace $\mathcal{X}_r\subset\mathcal{X}$ such that $P^s[y|\mathbf{x}\in\mathcal{X}_r]=P^t[y|\mathbf{x}\in\mathcal{X}_r]$, we aim to demonstrate the advantage of the DAC regularization over the ERM on augmented training set, with a provable separation in the respective excess risks.

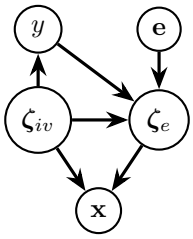

Figure 5: Causal graph shared by P^s and P^t.

Source and target distributions. Formally, the source and target distributions are concretized with the causal graph in Figure 5. For both P^s and P^t, the observable feature $\mathbf{x}$ is described via a linear generative model in terms of two latent features, the 'invariant' feature $\boldsymbol{\zeta}_{iv}\in\mathbb{R}^{d_{iv}}$ and the 'environmental' feature $\boldsymbol{\zeta}_e\in\mathbb{R}^{d_e}$:

$$\mathbf{x}=g(\boldsymbol{\zeta}_{iv},\boldsymbol{\zeta}_e)\triangleq\mathbf{S}\left[\boldsymbol{\zeta}_{iv};\boldsymbol{\zeta}_e\right]=\mathbf{S}_{iv}\boldsymbol{\zeta}_{iv}+\mathbf{S}_e\boldsymbol{\zeta}_e,$$

where $\mathbf{S}=\left[\mathbf{S}_{iv},\mathbf{S}_e\right]\in\mathbb{R}^{d\times(d_{iv}+d_e)}$ $(d_{iv}+d_e\leq d)$ consists of orthonormal columns. Let the label y depends only on the invariant feature $\boldsymbol{\zeta}_{iv}$ for both domains,

$$y=\left(\boldsymbol{\theta}^*\right)^\top\mathbf{x}+z=\left(\boldsymbol{\theta}^*\right)^\top\mathbf{S}_{iv}\boldsymbol{\zeta}_{iv}+z,\quad z\sim\mathcal{N}\left(0,\sigma^2\right),\quad z\perp\boldsymbol{\zeta}_{iv},$$

for some $\boldsymbol{\theta}^*\in\mathrm{Range}\left(\mathbf{S}_{iv}\right)$ such that $P^s[y|\boldsymbol{\zeta}_{iv}]=P^t[y|\boldsymbol{\zeta}_{iv}]$, while the environmental feature $\boldsymbol{\zeta}_e$ is conditioned on y, $\boldsymbol{\zeta}_{iv}$, (along with the Gaussian noise z), and varies across different domains $\mathbf{e}$ with $\mathbb{E}_{P^s}[y|\mathbf{x}]\neq\mathbb{E}_{P^t}[y|\mathbf{x}]$. In other words, with

the square loss $l(h(\mathbf{x}), y) = \frac{1}{2}(h(\mathbf{x}) - y)^2$, the optimal hypotheses that minimize the expected excess risk over the source and target distributions are distinct. Therefore, learning via the ERM with training samples from P^s can overfit the source distribution, in which scenario identifying the invariant feature subspace $\mathrm{Range}\,(\mathbf{S}_{iv})$ becomes indispensable for achieving good generalization in the target domain.

For P^s and P^t, we assume the following regularity conditions:

Assumption 2 (Regularity conditions for P^s and P^t)**.** *Let P^s satisfy Assumption 1. While P^t satisfies that $\mathbb{E}_{P^t}[\mathbf{x}\mathbf{x}^\top] \succ 0$, and*

(a) for the invariant feature, $c_{t,iv}\mathbf{I}_{d_{iv}} \preccurlyeq \mathbb{E}_{P^t}[\boldsymbol{\zeta}_{iv}\boldsymbol{\zeta}_{iv}^\top] \preccurlyeq C_{t,iv}\mathbf{I}_{d_{iv}}$ for some $C_{t,iv} \geq c_{t,iv} = \Theta(1)$;
(b) for the environmental feature, $\mathbb{E}_{P^t}[\boldsymbol{\zeta}_e\boldsymbol{\zeta}_e^\top] \succcurlyeq c_{t,e}\mathbf{I}_{d_e}$ for some $c_{t,e} > 0$, and $\mathbb{E}_{P^t}[z \cdot \boldsymbol{\zeta}_e] = \boldsymbol{0}$.

Training samples and data augmentations. Let $\mathbf{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_N]$ be a set of N samples drawn *i.i.d.* from $P^s(\mathbf{x})$ such that $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Recall that we denote the augmented training sets, including/excluding the original samples, respectively, with

$$\widetilde{\mathcal{A}}(\mathbf{X}) = [\mathbf{x}_1; \cdots; \mathbf{x}_N; \mathbf{x}_{1,1}; \cdots; \mathbf{x}_{N,1}; \cdots; \mathbf{x}_{1,\alpha}; \cdots; \mathbf{x}_{N,\alpha}] \in \mathcal{X}^{(1+\alpha)N},$$
$$\mathcal{A}(\mathbf{X}) = [\mathbf{x}_{1,1}; \cdots; \mathbf{x}_{N,1}; \cdots; \mathbf{x}_{1,\alpha}; \cdots; \mathbf{x}_{N,\alpha}] \in \mathcal{X}^{\alpha N}.$$

In particular, we consider a set of augmentations that only perturb the environmental feature $\boldsymbol{\zeta}_e$, while keep the invariant feature $\boldsymbol{\zeta}_{iv}$ intact:

$$\mathbf{S}_{iv}^\top\mathbf{x}_i = \mathbf{S}_{iv}^\top\mathbf{x}_{i,j}, \quad \mathbf{S}_e^\top\mathbf{x}_i \neq \mathbf{S}_e^\top\mathbf{x}_{i,j} \quad \forall\, i \in [n],\ j \in [\alpha]. \tag{11}$$

We recall the notion $\boldsymbol{\Delta} \triangleq \mathcal{A}(\mathbf{X}) - \mathbf{M}\mathbf{X}$ such that $d_{aug} \triangleq \mathrm{rank}\,(\boldsymbol{\Delta}) = \mathrm{rank}\left(\widetilde{\mathcal{A}}(\mathbf{X}) - \widetilde{\mathbf{M}}\mathbf{X}\right)$ $(0 \leq d_{aug} \leq d_e)$, and assume that $\mathbf{X}$ and $\mathcal{A}(\mathbf{X})$ are representative enough:

Assumption 3 (Diversity of $\mathbf{X}$ and $\mathcal{A}(\mathbf{X})$)**.** $(\mathbf{X}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ *is sufficiently large with* $n \gg \rho^4 d$, $\boldsymbol{\theta}^* \in \mathrm{Row}(\mathbf{X})$, *and* $d_{aug} = d_e$.

Excess risks in target domain. Learning from the linear hypothesis class $\mathcal{H} = \left\{h(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta} \;\middle|\; \boldsymbol{\theta} \in \mathbb{R}^d\right\}$, with the DAC regularization on $h\,(\mathbf{x}_i) = h\,(\mathbf{x}_{i,j})$, we have

$$\widehat{\boldsymbol{\theta}}^{dac} = \underset{\boldsymbol{\theta} \in \mathcal{H}_{dac}}{\mathrm{argmin}} \frac{1}{2N}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|_2^2, \quad \mathcal{H}_{dac} = \left\{h\,(\mathbf{x}) = \boldsymbol{\theta}^\top\mathbf{x} \;\middle|\; \boldsymbol{\Delta}\boldsymbol{\theta} = \mathbf{0}\right\},$$

while with the ERM on augmented training set,

$$\widehat{\boldsymbol{\theta}}^{da-erm} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\mathrm{argmin}} \frac{1}{2(1+\alpha)N}\left\|\widetilde{\mathbf{M}}\mathbf{y} - \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}\right\|_2^2,$$

where $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ denote the vertical stacks of α and $1 + \alpha$ identity matrices of size $n \times n$, respectively as denoted earlier.

We are interested in the excess risk on P^t: $L_t\,(\boldsymbol{\theta}) - L_t\,(\boldsymbol{\theta}^*)$ where $L_t\,(\boldsymbol{\theta}) \triangleq \mathbb{E}_{P^t(\mathbf{x},y)}\left[\frac{1}{2}(y - \mathbf{x}^\top\boldsymbol{\theta})^2\right]$.

Theorem 9 (Domain adaptation with DAC)**.** *Under Assumption 2(a) and Assumption 3, $\widehat{\boldsymbol{\theta}}^{dac}$ satisfies that, with constant probability,*

$$\mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{dac}) - L_t(\boldsymbol{\theta}^*)\right] \lesssim \frac{\sigma^2 d_{iv}}{N}. \tag{12}$$

Theorem 10 (Domain adaptation with ERM on augmented samples)**.** *Under Assumption 2 and Assumption 3, $\widehat{\boldsymbol{\theta}}^{dac}$ and $\widehat{\boldsymbol{\theta}}^{da-erm}$ satisfies that,*

$$\mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{da-erm}) - L_t(\boldsymbol{\theta}^*)\right] \geq \mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{dac}) - L_t(\boldsymbol{\theta}^*)\right] + c_{t,e} \cdot EER_e, \tag{13}$$

for some $EER_e > 0$.

In contrast to $\widehat{\boldsymbol{\theta}}^{dac}$ where the DAC constraints enforce $\mathbf{S}_e^\top\widehat{\boldsymbol{\theta}}^{dac} = \mathbf{0}$ with a sufficiently diverse $\mathcal{A}\,(\mathbf{X})$ (Assumption 3), the ERM on augmented training set fails to filter out the environmental feature in $\widehat{\boldsymbol{\theta}}^{da-erm}$: $\mathbf{S}_e^\top\widehat{\boldsymbol{\theta}}^{da-erm} \neq \mathbf{0}$. As a consequence, the expected excess risk of $\widehat{\boldsymbol{\theta}}^{da-erm}$ in the target domain can be catastrophic when $c_{t,e} \to \infty$, as instantiated by Example 3.

Proofs and instantiation. Recall that for $\mathbf{\Delta} \triangleq \mathcal{A}(\mathbf{X}) - \mathbf{M}\mathbf{X}$, $\mathbf{P}_{\mathbf{\Delta}}^{\perp} \triangleq \mathbf{I}_d - \mathbf{\Delta}^{\dagger}\mathbf{\Delta}$ denotes the orthogonal projector onto the dimension-$(d - d_{aug})$ null space of $\mathbf{\Delta}$. Furthermore, let $\mathbf{P}_{iv} \triangleq \mathbf{S}_{iv}\mathbf{S}_{iv}^{\top}$ and $\mathbf{P}_e \triangleq \mathbf{S}_e\mathbf{S}_e^{\top}$ be the orthogonal projectors onto the invariant and environmental feature subspaces, respectively, such that $\mathbf{x} = \mathbf{S}_{iv}\boldsymbol{\zeta}_{iv} + \mathbf{S}_e\boldsymbol{\zeta}_e = (\mathbf{P}_{iv} + \mathbf{P}_e)\,\mathbf{x}$ for all $\mathbf{x}$.

Proof of Theorem 9. By construction Equation (11), $\mathbf{\Delta}\mathbf{P}_{iv} = \mathbf{0}$, and it follows that $\mathbf{P}_{iv} \preccurlyeq \mathbf{P}_{\mathbf{\Delta}}^{\perp}$. Meanwhile from Assumption 3, $d_{aug} = d_e$ implies that $\dim\left(\mathbf{P}_{\mathbf{\Delta}}^{\perp}\right) = d_{iv}$. Therefore, $\mathbf{P}_{iv} = \mathbf{P}_{\mathbf{\Delta}}^{\perp}$, and the data augmentation consistency constraints can be restated as

$$\mathcal{H}_{dac} = \left\{h\left(\mathbf{x}\right) = \boldsymbol{\theta}^{\top}\mathbf{x} \;\middle|\; \mathbf{P}_{\mathbf{\Delta}}^{\perp}\boldsymbol{\theta} = \boldsymbol{\theta}\right\} = \left\{h\left(\mathbf{x}\right) = \boldsymbol{\theta}^{\top}\mathbf{x} \;\middle|\; \mathbf{P}_{iv}\boldsymbol{\theta} = \boldsymbol{\theta}\right\}$$

Then with $\boldsymbol{\theta}^* \in \mathrm{Row}(\mathbf{X})$ from Assumption 3,

$$\widehat{\boldsymbol{\theta}}^{dac} - \boldsymbol{\theta}^* = \frac{1}{N}\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{iv}}^{\dagger}\mathbf{P}_{iv}\mathbf{X}^{\top}(\mathbf{X}\mathbf{P}_{iv}\boldsymbol{\theta}^* + \mathbf{z}) - \boldsymbol{\theta}^* = \frac{1}{N}\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{iv}}^{\dagger}\mathbf{P}_{iv}\mathbf{X}^{\top}\mathbf{z},$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{iv}} \triangleq \frac{1}{N}\mathbf{P}_{iv}\mathbf{X}^{\top}\mathbf{X}\mathbf{P}_{iv}$. Since $\widehat{\boldsymbol{\theta}}^{dac} - \boldsymbol{\theta}^* \in \mathrm{Col}\left(\mathbf{S}_{iv}\right)$, we have $\mathbb{E}_{P^t}\left[z \cdot \mathbf{x}^{\top}\mathbf{P}_e(\widehat{\boldsymbol{\theta}}^{dac} - \boldsymbol{\theta}^*)\right] = 0$. Therefore, let $\mathbf{\Sigma}_{\mathbf{x},t} \triangleq \mathbb{E}_{P^t}[\mathbf{x}\mathbf{x}^{\top}]$, with high probability,

$$\begin{aligned}
E_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{dac}) - L_t(\boldsymbol{\theta}^*)\right] &= E_{P^s}\left[\frac{1}{2}\left\|\widehat{\boldsymbol{\theta}}^{dac} - \boldsymbol{\theta}^*\right\|_{\mathbf{\Sigma}_{\mathbf{x},t}}^2\right] \\
&= \mathrm{tr}\left(\frac{1}{2N}\mathbb{E}_{P^s}\left[\mathbf{z}\mathbf{z}^{\top}\right]\,\mathbb{E}_{P^s}\left[\left(\frac{1}{N}\mathbf{P}_{iv}\mathbf{X}^{\top}\mathbf{X}\mathbf{P}_{iv}\right)^{\dagger}\right]\,\mathbf{\Sigma}_{\mathbf{x},t}\right) \\
&= \mathrm{tr}\left(\frac{\sigma^2}{2N}\,\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{iv}}^{\dagger}\right]\,\mathbf{\Sigma}_{\mathbf{x},t}\right) \\
&\leq C_{t,iv}\,\frac{\sigma^2}{2N}\,tr\left(\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}_{\mathbf{X}_{iv}}^{\dagger}\right]\right) \quad \text{(Lemma 5, } w.h.p.\text{)} \\
&\lesssim \frac{\sigma^2}{2N}\,\mathrm{tr}\left(\left(\mathbb{E}_{P^s}\left[\mathbf{P}_{iv}\mathbf{x}\mathbf{x}^{\top}\mathbf{P}_{iv}\right]\right)^{\dagger}\right) \\
&\leq \frac{\sigma^2 d_{iv}}{2Nc} \lesssim \frac{\sigma^2 d_{iv}}{2N}.
\end{aligned}$$

■

Proof of Theorem 10. Let $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})} \triangleq \frac{1}{(1+\alpha)N}\widetilde{\mathcal{A}}\left(\mathbf{X}\right)^{\top}\widetilde{\mathcal{A}}\left(\mathbf{X}\right)$. Then with $\boldsymbol{\theta}^* \in \mathrm{Row}(\mathbf{X})$ from Assumption 3, we have $\boldsymbol{\theta}^* = \widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{\dagger}\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}\boldsymbol{\theta}^*$. Since $\boldsymbol{\theta}^* \in \mathrm{Col}\left(\mathbf{S}_{iv}\right)$, $\widetilde{\mathbf{M}}\mathbf{X}\boldsymbol{\theta}^* = \widetilde{\mathbf{M}}\mathbf{X}\mathbf{P}_{iv}\boldsymbol{\theta}^* = \widetilde{\mathcal{A}}(\mathbf{X})\boldsymbol{\theta}^*$. Then, the ERM on the augmented training set yields

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{da-erm} - \boldsymbol{\theta}^* &= \frac{1}{(1+\alpha)N}\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{\dagger}\widetilde{\mathcal{A}}(\mathbf{X})^{\top}\widetilde{\mathbf{M}}(\mathbf{X}\boldsymbol{\theta}^* + \mathbf{z}) - \widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{\dagger}\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}\boldsymbol{\theta}^* \\
&= \frac{1}{(1+\alpha)N}\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{\dagger}\widetilde{\mathcal{A}}(\mathbf{X})^{\top}\widetilde{\mathbf{M}}\mathbf{z}.
\end{aligned}$$

Meanwhile with $\mathbb{E}_{P^t}\left[z \cdot \boldsymbol{\zeta}_e\right] = \mathbf{0}$ from Assumption 2, we have $\mathbb{E}_{P^t}\left[z \cdot \mathbf{P}_e\mathbf{x}\right] = \mathbf{0}$. Therefore, by recalling that $\mathbf{\Sigma}_{\mathbf{x},t} \triangleq \mathbb{E}_{P^t}[\mathbf{x}\mathbf{x}^{\top}]$,

$$L_t(\boldsymbol{\theta}) - L_t(\boldsymbol{\theta}^*) \;=\; \mathbb{E}_{P^t(\mathbf{x})}\left[\frac{1}{2}\left(\mathbf{x}^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right)^2 + z \cdot \mathbf{x}^{\top}\mathbf{P}_e(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right] \;=\; \frac{1}{2}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_{\mathbf{\Sigma}_{\mathbf{x},t}}^2,$$

such that the expected excess risk can be expressed as

$$\mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{da-erm}) - L_t(\boldsymbol{\theta}^*)\right] = \frac{1}{2(1+\alpha)^2N^2}\,\mathrm{tr}\left(\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{\dagger}\left(\widetilde{\mathcal{A}}(\mathbf{X})^{\top}\widetilde{\mathbf{M}}\mathbf{z}\mathbf{z}^{\top}\widetilde{\mathbf{M}}^{\top}\widetilde{\mathcal{A}}(\mathbf{X})\right)\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})}^{\dagger}\right]\mathbf{\Sigma}_{\mathbf{x},t}\right),$$

where let $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)} \triangleq \mathbf{P}_e \widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X})} \mathbf{P}_e$,

$$
\begin{aligned}
&\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X})}\left(\widetilde{\mathcal{A}}(\mathbf{X})^{\top}\widetilde{\mathbf{M}}\mathbf{z}\mathbf{z}^{\top}\widetilde{\mathbf{M}}^{\top}\widetilde{\mathcal{A}}(\mathbf{X})\right)\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X})}\right] \\
\succcurlyeq\ &\mathbb{E}_{P^s}\left[\left(\mathbf{P}_{iv}\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_{iv}+\mathbf{P}_e\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_e\right)\widetilde{\mathcal{A}}(\mathbf{X})^{\top}\widetilde{\mathbf{M}}\mathbf{z}\mathbf{z}^{\top}\widetilde{\mathbf{M}}^{\top}\widetilde{\mathcal{A}}(\mathbf{X})\left(\mathbf{P}_{iv}\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_{iv}+\mathbf{P}_e\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X})}\mathbf{P}_e\right)\right] \\
\succcurlyeq\ &\sigma^2(1+\alpha)^2N\cdot\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}^{\dagger}_{\mathbf{X}_{iv}}\right]+\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\widetilde{\mathcal{A}}(\mathbf{X}_e)^{\top}\widetilde{\mathbf{M}}\mathbf{z}\mathbf{z}^{\top}\widetilde{\mathbf{M}}^{\top}\widetilde{\mathcal{A}}(\mathbf{X}_e)\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\right].
\end{aligned}
$$

We denote

$$
\mathrm{EER}_e \triangleq \operatorname{tr}\left(\mathbb{E}_{P^s}\left[\frac{1}{2(1+\alpha)^2N^2}\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\widetilde{\mathcal{A}}(\mathbf{X}_e)^{\top}\widetilde{\mathbf{M}}\mathbf{z}\mathbf{z}^{\top}\widetilde{\mathbf{M}}^{\top}\widetilde{\mathcal{A}}(\mathbf{X}_e)\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\right]\right),
$$

and observe that

$$
\mathrm{EER}_e = \mathbb{E}_{P^s}\left[\frac{1}{2}\left\|\frac{1}{(1+\alpha)N}\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\widetilde{\mathcal{A}}(\mathbf{X}_e)^{\top}\widetilde{\mathbf{M}}\mathbf{z}\right\|_2^2\right] > 0.
$$

Finally, we complete the proof by partitioning the lower bound for the target expected excess risk of $\widehat{\boldsymbol{\theta}}^{da-erm}$ into the invariantand environmental parts such that

$$
\begin{aligned}
&\mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{da-erm})-L_t(\boldsymbol{\theta}^*)\right] \\
\geq\ &\underbrace{\operatorname{tr}\left(\frac{\sigma^2}{2N}\,\mathbb{E}_{P^s}\left[\widehat{\mathbf{\Sigma}}^{\dagger}_{\mathbf{X}_{iv}}\right]\mathbf{\Sigma}_{\mathbf{x},t}\right)}_{=\mathbb{E}\left[L_t(\widehat{\boldsymbol{\theta}}^{dac})-L_t(\boldsymbol{\theta}^*)\right]} \\
&+\ \underbrace{\operatorname{tr}\left(\mathbb{E}_{P^s}\left[\frac{1}{2(1+\alpha)^2N^2}\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\widetilde{\mathcal{A}}(\mathbf{X}_e)^{\top}\widetilde{\mathbf{M}}\mathbf{z}\mathbf{z}^{\top}\widetilde{\mathbf{M}}^{\top}\widetilde{\mathcal{A}}(\mathbf{X}_e)\widehat{\mathbf{\Sigma}}^{\dagger}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)}\right]\mathbf{\Sigma}_{\mathbf{x},t}\right)}_{\text{expected excess risk from environmental feature subspace}\geq c_{t,e}\cdot\mathrm{EER}_e} \\
\geq\ &\mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{dac})-L_t(\boldsymbol{\theta}^*)\right]+c_{t,e}\cdot\mathrm{EER}_e.
\end{aligned}
$$

■

Now we construct a specific domain adaptation example with a large separation (*i.e.*, proportional to d_e) in the target excess risk between learning with the DAC regularization (*i.e.*, $\widehat{\boldsymbol{\theta}}^{dac}$) and with the ERM on augmented training set (*i.e.*, $\widehat{\boldsymbol{\theta}}^{da-erm}$).

Example 3. *We consider P^s and P^t that follow the same set of relations in Figure 5, except for the distributions over $\mathbf{e}$ where $P^s(\mathbf{e}) \neq P^t(\mathbf{e})$. Precisely, let the environmental feature $\boldsymbol{\zeta}_e$ depend on $(\boldsymbol{\zeta}_{iv}, y, \mathbf{e})$:*

$$
\boldsymbol{\zeta}_e = \operatorname{sign}\left(y-(\boldsymbol{\theta}^*)^{\top}\mathbf{S}_{iv}\boldsymbol{\zeta}_{iv}\right)\mathbf{e} = \operatorname{sign}(z)\mathbf{e}, \quad z\sim\mathcal{N}(0,\sigma^2), \quad z\perp\mathbf{e},
$$

where $\mathbf{e}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_{d_e})$ for $P^s(\mathbf{e})$ and $\mathbf{e}\sim\mathcal{N}(\mathbf{0},\sigma_t^2\mathbf{I}_{d_e})$ for $P^t(\mathbf{e})$, $\sigma_t \geq c_{t,e}$ (recall $c_{t,e}$ from Assumption 2). Assume that the training set $\mathbf{X}$ is sufficiently large, $n \gg d_e + \log(1/\delta)$ for some given $\delta\in(0,1)$. Augmenting $\mathbf{X}$ with a simple by common type of data augmentations – the linear transforms, we let

$$
\widetilde{\mathcal{A}}(\mathbf{X}) = [\mathbf{X};(\mathbf{X}\mathbf{A}_1);\ldots;(\mathbf{X}\mathbf{A}_\alpha)], \quad \mathbf{A}_j = \mathbf{P}_{iv}+\mathbf{u}_j\mathbf{v}_j^{\top}, \quad \mathbf{u}_j,\mathbf{v}_j\in\mathrm{Col}(\mathbf{S}_e) \quad \forall\, j\in[\alpha],
$$

and define

$$
\nu_1 \triangleq \max\{1\}\cup\{\sigma_{\max}(\mathbf{A}_j)\mid j\in[\alpha]\} \quad \textit{and} \quad \nu_2 \triangleq \sigma_{\min}\left(\frac{1}{1+\alpha}\left(\mathbf{I}_d+\sum_{j=1}^{\alpha}\mathbf{A}_k\right)\right),
$$

where $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ refer to the minimum and maximum singular values, respectively. Then under Assumption 2 and Assumption 3, with constant probability,

$$
\mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{da-erm})-L_t(\boldsymbol{\theta}^*)\right] \gtrsim \mathbb{E}_{P^s}\left[L_t(\widehat{\boldsymbol{\theta}}^{dac})-L_t(\boldsymbol{\theta}^*)\right]+c_{t,e}\cdot\frac{\sigma^2 d_e}{2N}.
$$

Proof of Example 3. With the specified distribution, for $\mathbf{E} = [\mathbf{e}_1; \dots; \mathbf{e}_N] \in \mathbb{R}^{N \times d_e}$,

$$\widehat{\mathbf{\Sigma}}_{\widetilde{\mathcal{A}}(\mathbf{X}_e)} = \frac{1}{(1+\alpha)N} \mathbf{S}_e \left(\mathbf{E}^\top \mathbf{E} + \sum_{j=1}^{\alpha} \mathbf{A}_j^\top \mathbf{E}^\top \mathbf{E} \mathbf{A}_j \right) \mathbf{S}_e^\top \preccurlyeq \frac{\nu_1^2}{N} \mathbf{S}_e \mathbf{E}^\top \mathbf{E} \mathbf{S}_e^\top,$$

$$\frac{1}{(1+\alpha)N} \widetilde{\mathcal{A}}(\mathbf{X}_e)^\top \widetilde{\mathbf{M}} \mathbf{z} = \left(\frac{1}{1+\alpha} \left(\mathbf{I}_d + \sum_{j=1}^{\alpha} \mathbf{A}_j \right) \right)^\top \frac{1}{N} \mathbf{S}_e \mathbf{E}^\top |\mathbf{z}|.$$

By Lemma 5, under Assumption 2 and Assumption 3, we have that with high probability, $0.9\mathbf{I}_{d_e} \preccurlyeq \frac{1}{N}\mathbf{E}^\top \mathbf{E} \preccurlyeq 1.1\mathbf{I}_{d_e}$. Therefore with $\mathbf{E}$ and $\mathbf{z}$ being independent,

$$\begin{aligned}
\mathrm{EER}_e &= \mathbb{E}_{P^s} \left[\frac{1}{2} \left\| \frac{1}{(1+\alpha)N} \widehat{\mathbf{\Sigma}}^\dagger_{\widetilde{\mathcal{A}}(\mathbf{X}_e)} \widetilde{\mathcal{A}}(\mathbf{X}_e)^\top \widetilde{\mathbf{M}} \mathbf{z} \right\|_2^2 \right] \\
&\geq \frac{\sigma^2}{2N} \frac{\nu_2^2}{\nu_1^4} \operatorname{tr} \left(\mathbb{E}_{P^s} \left[\left(\frac{1}{N} \mathbf{S}_e \mathbf{E}^\top \mathbf{E} \mathbf{S}_e^\top \right)^\dagger \right] \right) \\
&\gtrsim \frac{\sigma^2}{2N} \frac{\nu_2^2}{\nu_1^4} d_e \\
&\gtrsim \frac{\sigma^2 d_e}{2N},
\end{aligned}$$

and the rest follows from Theorem 10. ■

E Technical Lemmas

Lemma 5. *We consider a random vector $\mathbf{x} \in \mathbb{R}^d$ with $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}$, and $\overline{\mathbf{x}} = \mathbf{\Sigma}^{-1/2}\mathbf{x}$ [6] being ρ^2-subgaussian. Given an* i.i.d. *sample of $\mathbf{x}$, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top$, for any $\delta \in (0,1)$, if $n \gg \rho^4 d$, then $0.9\mathbf{\Sigma} \preccurlyeq \frac{1}{n}\mathbf{X}^\top \mathbf{X} \preccurlyeq 1.1\mathbf{\Sigma}$ with high probability.*

Proof. We first denote $\mathbf{P}_{\mathcal{X}} \triangleq \mathbf{\Sigma}\mathbf{\Sigma}^\dagger$ as the orthogonal projector onto the subspace $\mathcal{X} \subseteq \mathbb{R}^d$ supported by the distribution of $\mathbf{x}$. With the assumptions $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}$, we observe that $\mathbb{E}[\overline{\mathbf{x}}] = \mathbf{0}$ and $\mathbb{E}\left[\overline{\mathbf{x}}\overline{\mathbf{x}}^\top\right] = \mathbb{E}\left[\mathbf{x}\mathbf{\Sigma}^{-1}\mathbf{x}^\top\right] = \mathbf{P}_{\mathcal{X}}$. Given the sample set $\mathbf{X}$ of size $n \gg \rho^4 (d + \log(1/\delta))$ for some $\delta \in (0,1)$, we let $\mathbf{U} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{\Sigma}^{-1} \mathbf{x}_i^\top - \mathbf{P}_{\mathcal{X}}$. Then the problem can be reduced to showing that, with probability at least $1-\delta$, $\|\mathbf{U}\|_2 \leq 0.1$. For this, we leverage the ϵ-net argument as following.

For an arbitrary $\mathbf{v} \in \mathcal{X} \cap \mathbb{S}^{d-1}$, we have

$$\mathbf{v}^\top \mathbf{U} \mathbf{v} = \frac{1}{n} \sum_{i=1}^n \left(\mathbf{v}^\top \mathbf{x}_i \mathbf{\Sigma}^{-1} \mathbf{x}_i^\top \mathbf{v} - 1 \right) = \frac{1}{n} \sum_{i=1}^n \left(\left(\mathbf{v}^\top \overline{\mathbf{x}}_i \right)^2 - 1 \right),$$

where, given $\overline{\mathbf{x}}_i$ being ρ^2-subgaussian, $\mathbf{v}^\top \overline{\mathbf{x}}_i$ is ρ^2-subgaussian. Since

$$\mathbb{E}\left[\left(\mathbf{v}^\top \overline{\mathbf{x}}_i \right)^2 \right] = \mathbf{v}^\top \mathbb{E}\left[\overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^\top \right] \mathbf{v} = 1,$$

we know that $\left(\mathbf{v}^\top \overline{\mathbf{x}}_i \right)^2 - 1$ is $16\rho^2$-subexponential. Then, we recall the Bernstein's inequality,

$$\mathbb{P}\left[\left| \mathbf{v}^\top \mathbf{U} \mathbf{v} \right| > \epsilon \right] \leq 2 \exp\left(-\frac{n}{2} \min\left(\frac{\epsilon^2}{(16\rho^2)^2}, \frac{\epsilon}{16\rho^2} \right) \right).$$

[6] In the case where $\mathbf{\Sigma}$ is rank-deficient, we slightly abuse the notation such that $\mathbf{\Sigma}^{-1/2}$ and $\mathbf{\Sigma}^{-1}$ refer to the respective pseudo-inverses.

Let $N \subset \mathcal{X} \cap \mathbb{S}^{d-1}$ be an ϵ_1-net such that $|N| = e^{O(d)}$. Then for some $0 < \epsilon_2 \leq 16\rho^2$, by the union bound,

$$\begin{aligned}\mathbb{P}\left[\max_{\mathbf{v}\in N} : \left|\mathbf{v}^\top \mathbf{U}\mathbf{v}\right| > \epsilon_2\right] &\leq 2\,|N| \exp\left(-\frac{n}{2}\min\left(\frac{\epsilon_2^2}{\left(16\rho^2\right)^2}, \frac{\epsilon_2}{16\rho^2}\right)\right) \\ &\leq \exp\left(O\left(d\right) - \frac{n}{2}\cdot\frac{\epsilon_2^2}{\left(16\rho^2\right)^2}\right) \leq \delta\end{aligned}$$

whenever $n > \frac{2\left(16\rho^2\right)^2}{\epsilon_2^2}\left(\Theta\left(d\right) + \log\frac{1}{\delta}\right)$. By taking $\delta = \exp\left(-\frac{1}{4}\left(\frac{\epsilon_2}{16\rho^2}\right)^2 n\right)$, we have that $\max_{\mathbf{v}\in N}\left|\mathbf{v}^\top\mathbf{U}\mathbf{v}\right| \leq \epsilon_2$ with high probability when $n > 4\left(\frac{16\rho^2}{\epsilon_2}\right)^2\Theta\left(d\right)$, and taking $n \gg \rho^4 d$ is sufficient.

Now for any $\mathbf{v} \in \mathcal{X} \cap \mathbb{S}^{d-1}$, there exists some $\mathbf{v}' \in N$ such that $\left\|\mathbf{v} - \mathbf{v}'\right\|_2 \leq \epsilon_1$. Therefore,

$$\begin{aligned}\left|\mathbf{v}^\top\mathbf{U}\mathbf{v}\right| &= \left|\mathbf{v}'^\top\mathbf{U}\mathbf{v}' + 2\mathbf{v}'^\top\mathbf{U}\left(\mathbf{v}-\mathbf{v}'\right) + \left(\mathbf{v}-\mathbf{v}'\right)^\top\mathbf{U}\left(\mathbf{v}-\mathbf{v}'\right)\right| \\ &\leq \left(\max_{\mathbf{v}\in N} : \left|\mathbf{v}^\top\mathbf{U}\mathbf{v}\right|\right) + 2\left\|\mathbf{U}\right\|_2\left\|\mathbf{v}'\right\|_2\left\|\mathbf{v}-\mathbf{v}'\right\|_2 + \left\|\mathbf{U}\right\|_2\left\|\mathbf{v}-\mathbf{v}'\right\|_2^2 \\ &\leq \left(\max_{\mathbf{v}\in N} : \left|\mathbf{v}^\top\mathbf{U}\mathbf{v}\right|\right) + \left\|\mathbf{U}\right\|_2\left(2\epsilon_1 + \epsilon_1^2\right).\end{aligned}$$

Taking the supremum over $\mathbf{v} \in \mathbb{S}^{d-1}$, with probability at least $1-\delta$,

$$\max_{\mathbf{v}\in\mathcal{X}\cap\mathbb{S}^{d-1}} : \left|\mathbf{v}^\top\mathbf{U}\mathbf{v}\right| = \left\|\mathbf{U}\right\|_2 \leq \epsilon_2 + \left\|\mathbf{U}\right\|_2\left(2\epsilon_1+\epsilon_1^2\right), \qquad \left\|\mathbf{U}\right\|_2 \leq \frac{\epsilon_2}{2-\left(1+\epsilon_1\right)^2}.$$

With $\epsilon_1 = \frac{1}{3}$ and $\epsilon_2 = \frac{1}{45}$, we have $\frac{\epsilon_2}{2-(1+\epsilon_1)^2} = \frac{1}{10}$.

Overall, if $n \gg \rho^4 d$, then with high probability, we have $\left\|\mathbf{U}\right\|_2 \leq 0.1$. ■

Lemma 6. *Let $U \subseteq \mathbb{R}^d$ be an arbitrary subspace in $\mathbb{R}^d$, and $\mathbf{g} \sim \mathcal{N}\left(\boldsymbol{0}, \mathbf{I}_d\right)$ be a Gaussian random vector. Then for any continuous and C_l-Lipschitz function $\varphi : \mathbb{R} \to \mathbb{R}$ (*i.e.*, $|\varphi(u) - \varphi(u')| \leq C_l \cdot |u - u'|$ for all $u, u' \in \mathbb{R}$),*

$$\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in U}\mathbf{g}^\top\varphi(\mathbf{u})\right] \leq C_l \cdot \mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in U}\mathbf{g}^\top\mathbf{u}\right],$$

where φ acts on $\mathbf{u}$ entry-wisely, $\left(\varphi(\mathbf{u})\right)_j = \varphi(u_j)$. In other words, the Gaussian width of the image set $\varphi(U) \triangleq \left\{\varphi(\mathbf{u}) \in \mathbb{R}^d \mid \mathbf{u} \in U\right\}$ is upper bounded by that of U scaled by the Lipschitz constant.

Proof.

$$\begin{aligned}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in U}\mathbf{g}^\top\varphi(\mathbf{u})\right] =& \frac{1}{2}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in U}\mathbf{g}^\top\varphi(\mathbf{u}) + \sup_{\mathbf{u}'\in U}\mathbf{g}^\top\varphi(\mathbf{u})\right] \\ =& \frac{1}{2}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u},\mathbf{u}'\in U}\mathbf{g}^\top\left(\varphi(\mathbf{u}) - \varphi(\mathbf{u}')\right)\right] \\ \leq& \frac{1}{2}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u},\mathbf{u}'\in U}\sum_{j=1}^d |g_j|\left|\varphi(u_j) - \varphi(u_j')\right|\right] \quad \text{(since } \varphi \text{ is } C_l\text{-Lipschitz)} \\ \leq& \frac{C_l}{2}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u},\mathbf{u}'\in U}\sum_{j=1}^d |g_j|\left|u_j - u_j'\right|\right] \\ =& \frac{C_l}{2}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u},\mathbf{u}'\in U}\mathbf{g}^\top\left(\mathbf{u}-\mathbf{u}'\right)\right] \\ =& \frac{C_l}{2}\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in U}\mathbf{g}^\top\mathbf{u} + \sup_{\mathbf{u}'\in U}\mathbf{g}^\top\left(-\mathbf{u}'\right)\right] \\ =& C_l \cdot \mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in U}\mathbf{g}^\top\mathbf{u}\right]\end{aligned}$$

■

F Experiment Details

In this section, we provide the details of our experiments. Our code is adapted from the publicly released repo: https://github.com/kekmodel/FixMatch-pytorch.

Dataset: Our training dataset is derived from CIFAR-100, where the original dataset contains 50,000 training samples of 100 different classes. Out of the original 50,000 samples, we randomly select 10,000 labeled data as training set (i.e., 100 labeled samples per class). To see the impact of different training samples, we also trained our model with dataset that contains 1,000 and 20,000 samples. Evaluations are done on standard test set of CIFAR-100, which contains 10,000 testing samples.

Data Augmentation: During the training time, given a training batch, we generate corresponding augmented samples by RandAugment (Cubuk et al., 2020). We set the number of augmentations per sample to 7, unless otherwise mentioned.

To generate an augmented image, the RandAugment draws n transformations uniformaly at random from 14 different augmentations, namely {identity, autoContrast, equalize, rotate, solarize, color, posterize, contrast, brightness, sharpness, shear-x, shear-y, translate-x, translate-y}. The RandAugment provides each transformation with a single scalar (1 to 10) to control the strength of each of them, which we always set to 10 for all transformations. By default, we set $n = 2$ (i.e., using 2 random transformations to generate an augmented sample). To see the impact of different augmentation strength, we choose $n \in \{1, 2, 5, 10\}$. Examples of augmented samples are shown in Figure 4.

Parameter Setting: The batch size is set to 64 and the entire training process takes 2^{15} steps. During the training, we adopt the SGD optimizer with momentum set to 0.9, with learning rate for step i being $0.03 \times \cos\left(\frac{i \times 7\pi}{2^{15} \times 16}\right)$.

Additional Settings for the semi-supervised learning results: For the results on semi-supervised learning, besides the 10,000 labeled samples, we also draw additionally samples (ranging from 5,000 to 20,000) from the training set of the original CIFAR-100. We remove the labels of those additionally sampled images, as they serve as “unlabeled" samples in the semi-supervised learning setting. The FixMatch implementation follows the publicly available on in https://github.com/kekmodel/FixMatch-pytorch.