Improving Out-of-Distribution Robustness via Selective Augmentation

Huaxiu Yao [*1] **Yu Wang** [*2] **Sai Li** [3] **Linjun Zhang** [4] **Weixin Liang** [1]
James Zou [1] **Chelsea Finn** [1]

Abstract

Machine learning algorithms typically assume that training and test examples are drawn from the same distribution. However, distribution shift is a common problem in real-world applications and can cause models to perform dramatically worse at test time. In this paper, we specifically consider the problems of subpopulation shifts (e.g., imbalanced data) and domain shifts. While prior works often seek to explicitly regularize internal representations or predictors of the model to be domain invariant, we instead aim to learn invariant predictors without restricting the model's internal representations or predictors. This leads to a simple mixup-based technique which learns invariant predictors via selective augmentation called LISA. LISA selectively interpolates samples either with the same labels but different domains or with the same domain but different labels. Empirically, we study the effectiveness of LISA on nine benchmarks ranging from subpopulation shifts to domain shifts, and we find that LISA consistently outperforms other state-of-the-art methods and leads to more invariant predictors. We further analyze a linear setting and theoretically show how LISA leads to a smaller worst-group error. Code is released in https://github.com/huaxiuyao/LISA

1. Introduction

To deploy machine learning algorithms in real-world applications, we must pay attention to distribution shift, i.e. when the test distribution is different from the training distribution, which substantially degrades model performance. In this paper, we refer this problem as out-of-distribution (OOD) generalization and specifically consider performance gaps caused by two kinds of distribution shifts: *subpopulation shifts* and *domain shifts*. In subpopulation shifts, the test domains (or subpopulations) are seen but underrepresented in the training data. When subpopulation shift occurs, models may perform poorly when they falsely rely on spurious correlations between the particular subpopulation and the label. For example, in health risk prediction, a machine learning model trained on the entire population may associate the labels with demographic features (e.g., gender and age), making the model fail on the test set when such an association does not hold in reality. In domain shifts, the test data is from new domains, which requires the trained model to generalize well to test domains without seeing the data from those domains at training time. In the health risk example, we may want to train a model on patients from a few sampled hospitals and then deploy the model to a broader set of hospitals (Koh et al., 2021).

To improve model robustness under these two kinds of distribution shifts, prior works have proposed various regularizers to learn representations or predictors that are invariant to different domains while still containing sufficient information to fulfill the task (Li et al., 2018; Sun & Saenko, 2016; Arjovsky et al., 2019; Krueger et al., 2021; Rosenfeld et al., 2021). However, designing regularizers that are widely suitable to datasets from diverse domains is challenging, and unsuitable regularizers may adversely limit the model's expressive power or yield a difficult optimization problem, leading to inconsistent performance among various real-world datasets. For example, on the WILDS datasets, invariant risk minimization (IRM) (Arjovsky et al., 2019) with reweighting – a representative method for learning invariant predictor – outperforms empirical risk minimization (ERM) on CivilComments, but fails to improve robustness on a variety of other datasets like Camelyon17 and RxRx1 (Koh et al., 2021).

Instead of explicitly imposing regularization, we propose to learn invariant predictors through data interpolation, leading to a simple algorithm called **LISA** (**L**earning **I**nvariant Predictors with **S**elective **A**ugmentation). Concretely, inspired by mixup (Zhang et al., 2018), LISA linearly interpolates the features for a pair of samples and applies the same

[*]Equal contribution . This work was done when Yu Wang was mentored by Huaxiu Yao remotely. [1]Stanford University, CA, USA [2]University of California San Diego, CA, USA [3]Renmin University of China, Beijing, China [4]Rutgers University, NJ, USA. Correspondence to: Huaxiu Yao <huaxiu@cs.stanford.edu>, Sai Li <saili@ruc.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

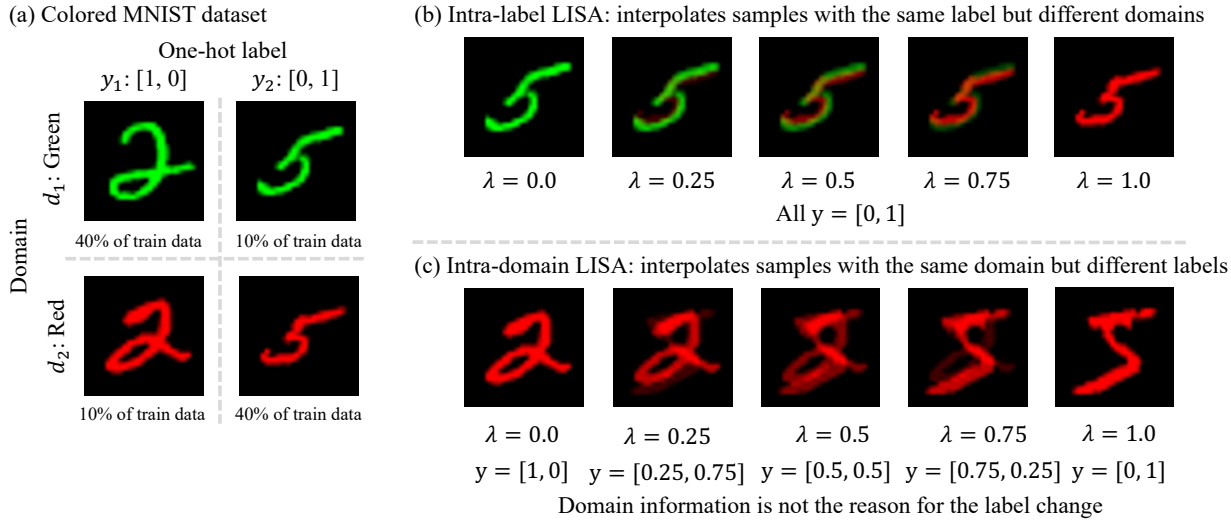

Figure 1. Illustration of the variants of LISA (Intra-label LISA and Intra-domain LISA) on Colored MNIST dataset. λ represents the interpolation ratio, which is sampled from a Beta distribution. (a) Colored MNIST (CMNIST). We classify MNIST digits as two classes, and original digits (0,1,2,3,4) and (5,6,7,8,9) are labeled as class 0 and 1, respectively. Digit color is used as domain information, which is spuriously correlated with labels in training data; (b) Intra-label LISA (LISA-L) cancels out spurious correlation by interpolating samples with the same label; (c) Intra-domain LISA (LISA-D) interpolates samples with the same domain but different labels to encourage the model to learn specific features within a domain.

interpolation strategy on the corresponding labels. Critically, the pairs are selectively chosen according to two selective augmentation strategies – intra-label LISA (LISA-L) and intra-domain LISA (LISA-D), which are described below and illustrated on Colored MNIST dataset in Figure 1. Intra-label LISA (Figure 1(b)) interpolates samples with the same label but from different domains, aiming to eliminate domain-related spurious correlations. Intra-domain LISA (Figure 1(c)) interpolates samples with the same domain but different labels, such that the model should learn to ignore the domain information and generate different predicted values as the interpolation ratio changes. In this way, LISA encourages the model to learn domain-invariant predictors without any explicit constraints or regularizers.

The **primary contributions** of this paper are as follows: (1) We propose a simple yet widely-applicable method for learning domain invariant predictors that is shown to be robust to subpopulation shifts and domain shifts. (2) We conduct broad experiments to evaluate LISA on nine benchmark datasets from diverse domains. In these experiments, we make the following key observations. First, we observe that LISA consistently outperforms seven prior methods to address subpopulation and domain shifts. Second, we find that LISA produces predictors that are consistently more domain invariant than prior approaches. Third, we identify that the performance gains of LISA are from canceling out domain-specific information or spurious correlations and learning invariant predictors, rather than simply involving more data via interpolation. Finally, when the degree of distribution shift increases, LISA achieves more significant performance gains. (3) We provide a theoretical analysis of the phenomena distilled from the empirical studies, where we provably demonstrate that LISA can mitigate spurious correlations and therefore lead to smaller worst-domain error compared with ERM and vanilla mixup. We also note that to the best of our knowledge, our work provides the first theoretical framework of studying how mixup (with or without the selective augmentation strategies) affects misclassification error.

2. Preliminaries

In this paper, we consider the setting where one predicts the label $y \in \mathcal{Y}$ based on the input feature $x \in \mathcal{X}$. Given a parameter space Θ and a loss function ℓ, we need to train a model f_θ under the training distribution P_{tr}, where $\theta \in \Theta$. In empirical risk minimization (ERM), the empirical distribution over the training data is $\hat{P}_{tr}$; ERM optimizes the following objective:

$$\theta^* := \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim \hat{P}}[\ell(f_\theta(x), y)]. \tag{1}$$

In a traditional machine learning setting, a test set, sampled from a test distribution P_{ts}, is used to evaluate the generalization of the trained model θ^*, where the test distribution is assumed to be the same as the training distribution, i.e., $P^{tr} = P^{ts}$. In this paper, we are interested in the setting when distribution shift occurs, i.e., $P^{tr} \neq P^{ts}$.

Specifically, following Muandet et al. (2013); Albuquerque et al. (2019); Koh et al. (2021), we regard the overall data distribution containing $\mathcal{D} = \{1, \ldots, D\}$ domains and each

domain $d \in \mathcal{D}$ is associated with a data distribution P_d over a set $(X, Y, d) = \{(x_i, y_i, d)\}_{i=1}^{N^d}$, where N^d is the number of samples in domain d. Then, we formulate the training distribution as the mixture of D domains, i.e., $P^{tr} = \sum_{d\in\mathcal{D}} r_d^{tr} P_d$, where $\{r_d^{tr}\}$ denotes the mixture probabilities in training set. Here, the training domains are defined as $\mathcal{D}^{tr} = \{d \in \mathcal{D} | r_d^{tr} > 0\}$. Similarly, the test distribution could be represented as $P^{ts} = \sum_{d\in\mathcal{D}} r_d^{ts} P_d$, where $\{r_d^{ts}\}$ is the mixture probabilities in test set. The test domains are defined as $\mathcal{D}^{ts} = \{d \in \mathcal{D} | r_d^{ts} > 0\}$.

In subpopulation shifts, the test set has domains that have been seen in the training set, but with a different proportion of subpopulations, i.e., $\mathcal{D}^{ts} \subseteq \mathcal{D}^{tr}$ but $\{r_d^{ts}\} \neq \{r_d^{tr}\}$. Under this setting, following Sagawa et al. (2020a), we consider group-based spurious correlations, where each group $g \in \mathcal{G}$ is defined to be associated with a domain d and a label y, i.e., $g = (d, y)$. We assume that the domain is spuriously correlated with the label. For example, we illustrate the CMNIST dataset in Figure 1, where the digit color d (green or red) is spuriously correlated with the label y ([1, 0] or [0, 1]). Based on the group definition, we evaluate the model via the worst test group error, i.e., $\max_g \mathbb{E}_{(x,y)\sim g}[\ell_{0-1}(f_\theta(x), y)]$, where ℓ_{0-1} represents the 0-1 loss.

In domain shifts, we investigate the problem where the test domains are disjoint from the training domains, i.e., $\mathcal{D}^{tr} \cap \mathcal{D}^{ts} = \emptyset$. In general, we assume the test domains share some common properties with the training domains. For example, in Camelyon17 (Koh et al., 2021), we train the model on some hospitals and test it in a new hospital. We evaluate the worst-domain and/or average performance of the classifier across all test domains.

3. Learning Invariant Predictors with Selective Augmentation

This section presents LISA, a simple way to improve robustness to subpopulation shifts and domain shifts. The key idea behind LISA is to encourage the model to learn invariant predictors by selective data interpolation, which could also alleviates the effects of domain-related spurious correlations. Before detailing how to select interpolated samples, we first provide a general formulation for data interpolation.

In LISA, we perform linear interpolation between training samples. Specifically, given samples (x_i, y_i, d_i) and (x_j, y_j, d_j) drawn from domains d_i and d_j, we apply mixup (Zhang et al., 2018), a simple data interpolation strategy, separately on the input features and corresponding labels as:

$$x_{mix} = \lambda x_i + (1-\lambda) x_j, \; y_{mix} = \lambda y_i + (1-\lambda) y_j, \quad (2)$$

where the interpolation ratio $\lambda \in [0, 1]$ is sampled from a Beta distribution $\mathrm{Beta}(\alpha, \beta)$ and y_i and y_j are one-hot vectors for classification problem. Notice that the mixup approach in (2) can be replaced by CutMix (Yun et al., 2019), which shows stronger empirical performance in vision-based applications. In text-based applications, we can use Manifold Mixup (Verma et al., 2019), interpolating the representations of a pre-trained model, e.g., the output of BERT (Devlin et al., 2019).

After obtaining the interpolated features and labels, we replace the original features and labels in ERM with the interpolated ones. Then, the optimization process in (1) is reformulated as:

$$\theta^* := \arg\min_{\theta\in\Theta} \mathbb{E}_{\{(x_i,y_i,d_i),(x_j,y_j,d_j)\}\sim\hat{P}}[\ell(f_\theta(x_{mix}), y_{mix})]. \quad (3)$$

Without additional selective augmentation strategies, vanilla mixup will regularize the model and reduce overfitting (Zhang et al., 2021b), allowing it to attain good in-distribution generalization. However, vanilla mixup may not be able to cancel out spurious correlations, causing the model to still fail at attaining good OOD generalization (see empirical comparisons in Section 4.3 and theoretical discussion in Section 5). In LISA, we instead adopt a new strategy where mixup is only applied across specific domains or groups, which leans towards learning invariant predictors and thus better OOD performance. Specifically, the two kinds of selective augmentation strategies are presented as:

Intra-label LISA (LISA-L): Interpolating samples with the same label. Intra-label LISA interpolates samples with the same label but different domains (i.e., $d_i \neq d_j$, $y_i = y_j$). As shown in Figure 1(a), this produces datapoints that have both domains partially present, effectively eliminating spurious correlations between domain and label in cases where the pair of domains correlate differently with the label. As a result, intra-label LISA should learn domain-invariant predictors for each class and thus achieve better OOD robustness.

Intra-domain LISA (LISA-D): Interpolating samples with the same domain. Supposing domain information is highly spuriously correlated with the label information, intra-domain LISA (Figure 1(b)) applies the interpolation strategy on samples with the same domain but different labels, i.e., $d_i = d_j, y_i \neq y_j$. Intuitively, even within the same domain, the model is supposed to generate different predicted labels since the interpolation ratio λ is randomly sampled, corresponding to different labels y_{mix}. This causes the model to make predictions that are less dependent on the domain, again improving OOD robustness.

In this paper, we randomly perform intra-label or intra-domain LISA during the training process with probability p_{sel} and $1 - p_{sel}$, where p_{sel} is treated as a hyperparameter and determined via cross-validation. Intuitively, the choice of p_{sel} depends on the number of domains and the strength

of the spurious correlations. Empirically, using intra-label LISA brings more benefits when there are more domains or when the the spurious correlations are not very strong. Intra-domain LISA benefits performance when domain information is highly spuriously correlated with the label. The pseudocode of LISA is in Algorithm 1.

Algorithm 1 Training Procedure of LISA

Require: Training data $\mathcal{D}$, step size η, learning rate γ, shape parameters α, β of Beta distribution

while not converge **do**
  Sample $\lambda \sim \text{Beta}(\alpha, \beta)$
  Sample minibatch $B_1 \sim \mathcal{D}$
  Initialize $B_2 \leftarrow \{\}$
  Select strategy $s \sim Bernoulli(p_{sel})$
  if s is True **then**
    for $(x_i, y_i, d_i) \in B_1$ **do**
      Randomly sample $(x_j, y_j, d_j) \sim \{(x, y, d) \in \mathcal{D}\}$ which satisfies $(y_i = y_j)$ and $(d_i \neq d_j)$.
      Put (x_j, y_j, d_j) into B_2.
  else
    for $(x_i, y_i, d_i) \in B_1$ **do**
      Randomly sample $(x_j, y_j, d_j) \sim \{(x, y, d) \in \mathcal{D}\}$ which satisfies $(y_i \neq y_j)$ and $(d_i = d_j)$.
      Put (x_j, y_j, d_j) into B_2.
  Update θ with data $\lambda B_1 + (1 - \lambda) B_2$ with learning rate γ.

4. Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of LISA. Specifically, we aim to answer the following questions: **Q1**: Compared to prior methods, can LISA improve robustness to subpopulation shifts and domain shifts (Section 4.1 and Section 4.2)? **Q2**: Which aspects of LISA are the most important for improving robustness (Section 4.3)? **Q3**: Does LISA successfully produce more invariant predictors (Section 4.4)? **Q4**: How does LISA perform with varying degrees of distribution shifts (Section 4.5)?

To answer Q1, we compare to ERM, IRM (Arjovsky et al., 2019), IB-IRM (Ahuja et al., 2021), V-REx (Krueger et al., 2021), CORAL (Li et al., 2018), DRNN (Ganin & Lempitsky, 2015), GroupDRO (Sagawa et al., 2020a), DomainMix (Xu et al., 2020), and Fish (Shi et al., 2021). Upweighting (UW) is particularly suitable for subpopulation shifts, so we also use it for comparison. We adopt the same model architectures for all approaches. The strategy selection probability p_{sel} is selected via cross-validation.

4.1. Evaluating Robustness to Subpopulation Shifts

Evaluation Protocol. In subpopulation shifts, we evaluate the performance on four binary classification datasets, including Colored MNIST (CMNIST), Waterbirds (Sagawa et al., 2020a), CelebA (Liu et al., 2015), and Civilcomments (Borkan et al., 2019). We detail the data descriptions of subpopulation shifts in Appendix A.1.1 and report the detailed data statistics in Table 1, covering domain information, model architecture, and class information. Following Sagawa et al. (2020a), in subpopulation shifts, we use the worst-group accuracy to evaluate the performance of all approaches. In these datasets, the domain information is highly spurious correlated with the label information. For example, as suggested in Figure 1, 80% images in the CMNIST dataset have the same color in each specific class, i.e., green color for label [1, 0] and red color for label [0, 1].

In CMNIST, Waterbirds, and CelebA, we find that $p_{sel} = 0.5$ works well for choosing selective augmentation strategies, while in CivilComments, we set p_{sel} as 1.0 . This is not surprising because it might be more beneficial to use intra-label LISA more often to eliminate domain effects when there are more domains, i.e., eight domains in CivilComments v.s. two domains in others. The rest hyperparameter settings and training details are listed in Appendix A.1.2.

Results. In Table 2, we report the overall performance of LISA and other methods. According to Table 2, we observe that the performance of approaches that learn invariant predictors with explicit regularizers (e.g., IRM, IB-IRM, V-REx) is not consistent across datasets. For example, IRM and V-REx outperform UW on CMNIST, but they fail to achieve better performance than UW on Waterbirds. The results corroborate our hypothesis that designing widely effective regularizers is challenging, and that inappropriate regularizers may even hurt the performance. LISA instead consistently outperforms other invariant learning methods (e.g., IRM, IB-IRM, V-REx, CORAL, DomainMix, Fish) in all datasets. LISA further shows the best performance on CMNIST, CelebA, and CivilComments. In Waterbirds, it is slightly worse than GroupDRO, but the performance is comparable. These results demonstrate the effectiveness of LISA in improving robustness to subpopulation shifts.

Effects of Intra-label and Intra-domain LISA. For CMNIST, Waterbirds and CelebA, both intra-label and intra-domain LISA are used (i.e., $p_{sel} = 0.5$), we illustrate the separate results in Figure 2 and observe that both variants contribute to the final performance. In addition, intra-domain LISA performs slightly better than intra-label LISA, corroborating our assumption that intra-domain LISA benefits more when domain information is highly spuriously correlated with the label (see the discussion of the strength of spurious correlation in Appendix A.3).

4.2. Evaluating Robustness to Domain Shifts

Experimental Setup. In domain shifts, we study five datasets. Four of them (Camelyon17, FMoW, RxRx1, and

Table 1. Dataset Statistics for Subpopulation Shifts. All datasets are binary classification tasks and we use the worst group accuracy as the evaluation metric.

Datasets	Domains	Model Architecture	Class Information
CMNIST	2 digit colors	ResNet-50	digit (0,1,2,3,4) v.s. (5,6,7,8,9)
Waterbirds	2 backgrounds	ResNet-50	waterbirds v.s. landbirds
CelebA	2 hair colors	ResNet-50	man v.s. women
CivilComments	8 demographic identities	DistilBERT-uncased	toxic v.s. non-toxic

Table 2. Results of subpopulation shifts. Here, we show the average and worst group accuracy. We repeat the experiments three times and put full results with standard deviation in Table 10.

	CMNIST		Waterbirds		CelebA		CivilComments	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
ERM	27.8%	0.0%	97.0%	63.7%	94.9%	47.8%	92.2%	56.0%
UW	72.2%	66.0%	95.1%	88.0%	92.9%	83.3%	89.8%	69.2%
IRM	72.1%	70.3%	87.5%	75.6%	94.0%	77.8%	88.8%	66.3%
IB-IRM	72.2%	70.7%	88.5%	76.5%	93.6%	85.0%	89.1%	65.3%
V-REx	71.7%	70.2%	88.0%	73.6%	92.2%	86.7%	90.2%	64.9%
CORAL	71.8%	69.5%	90.3%	79.8%	93.8%	76.9%	88.7%	65.6%
GroupDRO	72.3%	68.6%	91.8%	**90.6%**	92.1%	87.2%	89.9%	70.0%
DomainMix	51.4%	48.0%	76.4%	53.0%	93.4%	65.6%	90.9%	63.6%
Fish	46.9%	35.6%	85.6%	64.0%	93.1%	61.2%	89.8%	71.1%
LISA (ours)	74.0%	**73.3%**	91.8%	89.2%	92.4%	**89.3%**	89.2%	**72.6%**

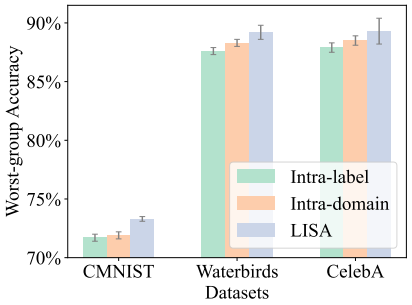

Figure 2. Effects of intra-label and intra-domain LISA in CMNIST, Waterbirds and CelebA. The experiments are repeated three times with different seeds.

Amazon) are selected from WILDS (Koh et al., 2021), covering natural distribution shifts across diverse domains (e.g., health, language, and vision). Besides the WILDS data, we also apply LISA on the MetaShift datasets (Liang & Zou, 2021), constructed using the real-world images and natural heterogeneity of Visual Genome (Krishna et al., 2016). We summarize these datasets in Table 4, including domain information, evaluation metric, model architecture, and the number of classes. Detailed dataset descriptions and other training details are discussed in Appendix A.2.1 and A.2.2, respectively.

The strategy selection probability p_{sel} is set as 1.0 for these domain shifts datasets, i.e., only intra-label LISA is used. Additionally, we only interpolate samples with the same labels without considering the domain information in Camelyon17, FMoW, and RxRx1, which empirically leads to the best performance. One potential reason is that the spurious correlations between labels and domains are not very strong in datasets with natural domain shifts under the existing domain partitions. Here, to evaluate the strength of spurious correlation, we adopt Cramér's V (Cramér, 2016) (see the detailed definition in Appendix A.3) to measure the association between the domain set $\mathcal{D}$ and the label set $\mathcal{Y}$, where the results are reported in Table 13 of Appendix A.3. The Cramér's V values in Camelyon17, FMoW, and RxRx1 are significantly smaller than other datasets, indicating relatively weak spurious correlations. Under this setting, enlarging the interpolation scope by directly interpolating samples within the same class regardless of existing domain information may bring more benefits.

Results. We report the results of domain shifts in Table 3, where full results that include validation performance and other metrics are listed in Appendix A.6. Aligning with the observation in subpopulation shifts, the performance of prior regularization-based invariant predictor learning methods (e.g., IRM, IB-IRM, V-REx) is still unstable across different datasets. For example, V-REx outperforms ERM on Camelyon17, while it fails in RxRx1. However, LISA consistently outperforms all these methods on five datasets regardless of the model architecture and data types (i.e., image or text), indicating its effectiveness in improving robustness to domain shifts with selective augmentation.

4.3. Are the Performance Gains of LISA from Data Augmentation?

In LISA, we apply selective augmentation strategies on samples either with the same label but different domains or

Table 3. Main domain shifts results. LISA outperforms prior methods on all five datasets. Following the instructions of Koh et al. (2021), we report the performance of Camelyon17 over 10 different seeds and the results of other datasets are obtained over 3 different seeds.

	Camelyon17	FMoW	RxRx1	Amazon	MetaShift
	Avg. Acc.	Worst Acc.	Avg. Acc.	10-th Per. Acc.	Worst Acc.
ERM	70.3 ± 6.4%	32.3 ± 1.25%	29.9 ± 0.4%	53.8 ± 0.8%	52.1 ± 0.4%
IRM	64.2 ± 8.1%	30.0 ± 1.37%	8.2 ± 1.1%	52.4 ± 0.8%	51.8 ± 0.8%
IB-IRM	68.9 ± 6.1%	28.4 ± 0.90%	6.4 ± 0.6%	53.8 ± 0.7%	52.3 ± 1.0%
V-REx	71.5 ± 8.3%	27.2 ± 0.78%	7.5 ± 0.8%	53.3 ± 0.0%	51.6 ± 1.8%
CORAL	59.5 ± 7.7%	31.7 ± 1.24%	28.4 ± 0.3%	52.9 ± 0.8%	47.6 ± 1.9%
GroupDRO	68.4 ± 7.3%	30.8 ± 0.81%	23.0 ± 0.3%	53.3 ± 0.0%	51.9 ± 0.7%
DomainMix	69.7 ± 5.5%	34.2 ± 0.76%	30.8 ± 0.4%	53.3 ± 0.0%	51.3 ± 0.5%
Fish	74.7 ± 7.1%	34.6 ± 0.18%	10.1 ± 1.5%	53.3 ± 0.0%	49.2 ± 2.1%
LISA (ours)	**77.1 ± 6.5%**	**35.5 ± 0.65%**	**31.9 ± 0.8%**	**54.7 ± 0.0%**	**54.2 ± 0.7%**

Table 4. Dataset Statistics for Domain Shifts.

Datasets	Domains	Metric	Base Model	Num. of classes
Camelyon17	5 hospitals	Avg. Acc.	DenseNet-121	2
FMoW	16 years x 5 regions	Worst-group Acc.	DenseNet-121	62
RxRx1	51 experimental batches	Avg. Acc.	ResNet-50	1,139
Amazon	7,676 reviewers	10th Percentile Acc.	DistilBERT-uncased	5
MetaShift	4 backgrounds	Worst-group Acc.	ResNet-50	2

with the same domain but different labels. Here, we explore two substitute interpolation strategies:

- *Vanilla mixup*: in Vanilla mixup, we do not add any constraints on the sample selection, i.e., the mixup is performed on any pairs of samples.
- *In-group mixup*: this strategy applies data interpolation on samples with the same labels and from the same domains.

Notice that all substitute interpolation strategies use the same variant of mixup as LISA (e.g., mixup/CutMix). Finally, as upweighting (UW) small groups significantly improves performance in subpopulation shifts, we evaluate UW combined with Vanilla/In-group mixup.

The results of substitute interpolation strategies on domain shifts and subpopulation shifts are in Table 5 and Table 6, respectively. Furthermore, we also conduct experiments on datasets without spurious correlation in Table 14 of Appendix A.4. From the results, we make the following three key observations. *First*, compared with Vanilla mixup, the performance of LISA verifies that selective data interpolation indeed improve the out-of-distribution robustness by canceling out the spurious correlations and encouraging learning invariant predictors rather than simple data augmentation. These findings are further strengthened by the results in Table 14 of Appendix A.4, where Vanilla mixup outperforms LISA and ERM without spurious correlations but LISA achieves the best performance with spurious correlations. *Second*, the superiority of LISA over In-group mixup verifies that only interpolating samples within each group is incapable of eliminating out the spurious information, where In-group mixup still performs the role of data augmentation. *Third*, though incorporating UW significantly improves the performance of Vanilla mixup and In-group mixup in subpopulation shifts, LISA still achieves larger benefits than these enhanced substitute strategies, demonstrating its stronger power in improving OOD robustness.

4.4. Does LISA Lead to More Invariant Predictors?

We further analyze the model invariance learned by LISA. Specifically, for each sample (x_i, y_i, d) in domain d, we denote the unscaled output (i.e., logits) of the model as $g_{i,d}$. We use two metrics to measure the invariance (see Appendix A.5.1 for additional metrics and the corresponding results):

- **Accuracy of domain prediction** (IP_{adp}). In the first metric, we use the unscaled output to predict the domain. Concretely, the entire dataset is re-split into training, validation, and test sets, where logits are used as features and labels represent the corresponding domain ID. A logistic regression model is trained to predict the domain.
- **Pairwise divergence of prediction** (IP_{kl}). We calculate the KL divergence of the distribution of logits among all domains, where kernel density estimation is used to estimate the probability density function $P(g_d^y)$ of logits from domain d with label y. The pairwise divergence of the predictions is defined as $\frac{1}{|\mathcal{Y}||\mathcal{D}|^2}\sum_{y\in\mathcal{Y}}\sum_{d',d\in\mathcal{D}}\text{KL}(P(g_D^y \mid D=d)|P(g_D^y \mid D=d'))$.

Small values of IP_{adp} and IP_{kl} represent strong function-level invariance. In Table 7, we report the results of LISA and other approaches on CMNIST, Waterbirds, Camelyon17

Table 5. Compared LISA with substitute mixup strategies in domain shifts.

	Camelyon17	FMoW	RxRx1	Amazon	MetaShift
	Avg. Acc.	Worst Acc.	Avg. Acc.	10-th Per. Acc.	Worst Acc.
ERM	$70.3 \pm 6.4\%$	$32.8 \pm 0.45\%$	$29.9 \pm 0.4\%$	$53.8 \pm 0.8\%$	$52.1 \pm 0.4\%$
Vanilla mixup	$71.2 \pm 5.3\%$	$34.2 \pm 0.45\%$	$26.5 \pm 0.5\%$	$53.3 \pm 0.0\%$	$51.3 \pm 0.7\%$
In-group mixup	$75.5 \pm 6.7\%$	$32.2 \pm 1.18\%$	$24.4 \pm 0.2\%$	$53.8 \pm 0.6\%$	$52.7 \pm 0.5\%$
LISA (ours)	$\mathbf{77.1 \pm 6.5\%}$	$\mathbf{35.5 \pm 0.65\%}$	$\mathbf{31.9 \pm 0.8\%}$	$\mathbf{54.7 \pm 0.0\%}$	$\mathbf{54.2 \pm 0.7\%}$

Table 6. Compared LISA with substitute mixup strategies in subpopulation shifts. UW represents upweighting. Full results with standard deviation is listed in Table 11.

	CMNIST		Waterbirds		CelebA		CivilComments	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
ERM	27.8%	0.0%	97.0%	63.7%	94.9%	47.8%	92.2%	56.0%
Vanilla mixup	32.6%	3.1%	81.0%	56.2%	95.8%	46.4%	90.8%	67.2%
Vanilla mixup + UW	72.2%	71.8%	92.1%	85.6%	91.5%	88.0%	87.8%	66.1%
In-group mixup	33.6%	24.0%	88.7%	68.0%	95.2%	58.3%	90.8%	69.2%
In-group mixup + UW	72.6%	71.6%	91.4%	87.1%	92.4%	87.8%	84.8%	69.3%
LISA (ours)	74.0%	**73.3%**	91.8%	**89.2%**	92.4%	**89.3%**	89.2%	**72.6%**

and MetaShift. The results verify that LISA learns predictors with greater domain invariance. Besides having more invariant predictors, we observe that LISA also leads to more invariant representations, as detailed in Appendix A.5.2.

4.5. Effect of the Degree of Distribution Shifts

We investigate the performance of LISA with respect to the degree of distribution shifts. Here, we use MetaShift to evaluate performance, where the distance between training and test domains is measured as the node similarity on a meta-graph (Liang & Zou, 2021). To vary the distance between training and test domains, we change the backgrounds of training objects (see full experimental details in Appendix A.2.1). The performance with varied distances is illustrated in Table 8, where the top four best methods (i.e., ERM, IRM, IB-IRM, GroupDRO) are reported for comparison. We observe that LISA consistently outperforms other methods under all scenarios. Another interesting finding is that LISA achieves more substantial improvements with the increases of distance. A potential reason is that the effects of eliminating domain information is more obvious when there is a larger distance between training and test domains.

5. Theoretical Analysis

In this section, we provide some theoretical understandings that explain several of the empirical phenomena from the previous experiments and theoretically compare the worst-group errors of three methods: the proposed LISA, ERM, and vanilla mixup. Specifically, we consider a Gaussian mixture model with subpopulation and domain shifts, which has been widely adopted in theory to shed light upon complex machine learning phenomenon such as in (Montanari et al., 2019; Zhang et al., 2021c; Liu et al., 2021b). We note here that despite the popularity of mixup in practice, the theoretical analysis of how mixup (w/ or w/o the selective augmentation strategies) affects the misclassification error is still largely unexplored in the literature even in the simple models. As discussed in Section 2, here, we define $y \in \{0, 1\}$ as the label, and $d \in \{R, G\}$ as the domain information. For $y \in \{0, 1\}$ and $d \in \{R, G\}$, we consider the following model:

$$x_i | y_i = y, d_i = d \sim N(\mu^{(y,d)}, \Sigma^{(d)}), i = 1, \ldots, n^{(y,d)}, \quad (4)$$

where $\mu^{(y,d)} \in \mathbb{R}^p$ is the conditional mean vector and $\Sigma^{(d)} \in \mathbb{R}^{p \times p}$ is the covariance matrix. Let $n = \sum_{y \in \{0,1\}, d \in \{R,G\}} n^{(y,d)}$. Let $\pi^{(y,d)} = \mathbb{P}(y_i = y, d_i = d)$, $\pi^{(y)} = \mathbb{P}(y_i = y)$, and $\pi^{(d)} = \mathbb{P}(d_i = d)$.

To account for the spurious correlation brought by domains, we consider $\mu^{(y,R)} \neq \mu^{(y,G)}$ in general for $y \in \{0, 1\}$ and the imbalanced case where $\pi^{(0,R)}, \pi^{(1,G)} < 1/4$. Moreover, we assume there exists some invariance across different domains. Specifically, we assume

$$\mu^{(1,R)} - \mu^{(0,R)} = \mu^{(1,G)} - \mu^{(0,G)} := \Delta \text{ and } \Sigma^{(G)} = \Sigma^{(R)} := \Sigma.$$

According to Fisher's linear discriminant analysis (Anderson, 1962; Tony Cai & Zhang, 2019; Cai & Zhang, 2021), the optimal classification rule is linear with slope $\Sigma^{-1}\Delta$. The assumption above implies that $(\Sigma^{-1}\Delta)^\top x$ is the (unknown) invariant prediction rule for model (4).

Suppose we use some method A and obtain a linear classifier $x^T b + b_0 > 0$ from a training data, we will apply it to a test data and compute the worst-group misclassification error,

Table 7. Results of the analysis of learned invariant predictors. Accuracy of domain prediction (IP_{adp}) and pairwise divergence of prediction among all domains (IP_{kl}) are used to measure the invariance. Smaller values denote stronger invariance.

	$\text{IP}_{adp} \downarrow$				$\text{IP}_{kl} \downarrow$			
	CMNIST	Waterbirds	Camelyon17	MetaShift	CMNIST	Waterbirds	Camelyon17	MetaShift
ERM	82.85%	94.99%	49.43%	67.98%	6.286	1.888	1.536	1.205
Vanilla mixup	92.34%	94.49%	52.79%	69.36%	4.737	2.912	0.790	1.171
IRM	69.42%	95.12%	47.96%	67.59%	7.755	1.122	0.875	1.148
IB-IRM	74.72%	94.78%	48.37%	67.39%	1.004	3.563	0.756	1.115
V-REx	63.58%	93.32%	61.38%	68.38%	3.190	3.791	1.281	1.094
LISA (ours)	**58.42%**	**90.28%**	**45.15%**	**66.01%**	**0.567**	**0.134**	**0.723**	**1.001**

Table 8. Effects of the degree of distribution shifts w.r.t. the performance on the MetaShift benchmark. Distance represents the distribution distance between training and test domains. Best B/L represents best baseline.

Distance	0.44	0.71	1.12	1.43
ERM	80.1%	68.4%	52.1%	33.2%
IRM	79.5%	67.4%	51.8%	32.0%
IB-IRM	79.7%	66.9%	52.3%	33.6%
GroupDRO	77.0%	68.9%	51.9%	34.2%
LISA (ours)	**81.3%**	**69.7%**	**54.2%**	**37.5%**
LISA v.s. Best B/L	+1.5%	+1.2%	+3.6%	+9.6%

where the mis-classification error for domain d and class y is $E^{(y,d)}(b, b_0) := \mathbb{P}(\mathbb{1}(x_i^T b + b_0 > \frac{1}{2}) \neq y | d_i = d, y_i = y)$, and we denote the worst-group error with the method A as

$$E_A^{(wst)} = \max_{d \in \{R,G\}, y \in \{0,1\}} E^{(y,d)}(b_A, b_{0,A}),$$

where b_A and $b_{0,A}$ are the slope and intercept based on the method A. Specifically, $A = \text{ERM}$ denotes the ERM method (by minimizing the sum of squares loss on the training data altogether), $A = \text{mix}$ denotes the vanilla mixup method (without any selective augmentation strategy), and $A = \text{LISA}$ denotes the mixup strategy for LISA. We also denote its finite sample version by $\hat{E}_A^{(wst)}$.

Let $\widetilde{\Delta} = \mathbb{E}[x_i | y_i = 1] - \mathbb{E}[x_i | y_i = 0]$ denote the marginal difference and $\xi = \frac{\Delta^T \Sigma^{-1} \widetilde{\Delta}}{\|\Delta\|_\Sigma \|\widetilde{\Delta}\|_\Sigma}$ denote the correlation operator between the domain-specific difference Δ and the marginal difference $\widetilde{\Delta}$ with respect to Σ. We see that smaller ξ indicates larger discrepancy between the marginal difference and the domain-specific difference and therefore implies stronger spurious correlation between the domains and labels. We present the following theorem showing that our proposed LISA algorithm outperforms the ERM and vanilla mixup in the subpopulation shifts setting.

Theorem 1 (Error comparison with subpopulation shifts) *Consider n independent samples generated from model (4), $\pi^{(R)} = \pi^{(1)} = 1/2$, $\pi^{(0,R)} = \pi^{(1,G)} = \alpha < 1/4$, $\max_{y,d} \|\mu^{(y,d)}\|_2 \leq C$, and Σ is positive definite. Suppose (ξ, α) satisfies that $\xi < \min\{\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, \frac{\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma}\} - C\alpha$ for some large enough constant C and $\mathbb{E}[\lambda_i^2] / \max\{var(\lambda_i), 1/4\} \geq \|\widetilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma \|\Delta\|_\Sigma$. Then for any $p_{sel} \in [0, 1]$,*

$$\widehat{E}_{\text{LISA}}^{(wst)} < \min\{\widehat{E}_{\text{ERM}}^{(wst)}, \widehat{E}_{\text{mix}}^{(wst)}\} + O_P\left(\frac{p \log n}{n} + \frac{p}{\alpha n}\right).$$

In Theorem 1, λ_i is the random mixup coefficient for the i-th sample. If $\lambda_i = \lambda$ are the same for all the samples in a mini-batch, the results still hold. Theorem 1 implies that when ξ is small (indicating that the domain has strong spurious correlation with the label) and $p = o(\alpha n)$, the worst-group classification errors of LISA are asymptotically smaller than that of ERM and vanilla mixup. In fact, our analysis shows that LISA yields a classification rule closer to the invariant classification rules by leveraging the domain information.

In the next theorem, we present the mis-classification error comparisons with domain shifts. That is, consider samples from a new unseen domain:

$$x_i^{(0,*)} \sim N(\mu^{(0,*)}, \Sigma), \;\; x_i^{(1,*)} \sim N(\mu^{(1,*)}, \Sigma). \tag{5}$$

Let $\widetilde{\Delta}^* = 2(\mu^{(0,*)} - \mathbb{E}[x_i])$, where $\mathbb{E}[x_i]$ is the mean of the training distribution, and assume $\mu^{(1,*)} - \mu^{(0,*)} = \Delta$. Let $\xi^* = \frac{\widetilde{\Delta}^T \Sigma^{-1} \widetilde{\Delta}^*}{\|\widetilde{\Delta}\|_\Sigma \|\Delta\|_\Sigma}$ and $\gamma = \frac{\Delta^T \Sigma^{-1} \widetilde{\Delta}^*}{\|\widetilde{\Delta}\|_\Sigma \|\Delta\|_\Sigma}$ denote the correlation for $(\widetilde{\Delta}^*, \widetilde{\Delta})$ and for $(\widetilde{\Delta}^*, \Delta)$, respectively, with respect to Σ^{-1}. Let $E_A^{(wst*)} = \max_{y \in \{0,1\}} E^{(y,*)}(b_A, b_{0,A})$ and its sample version be $\hat{E}_A^{(wst*)}$.

Theorem 2 (Error comparison with domain shifts) *Suppose n samples are independently generated from model (4), $\pi^{(R)} = \pi^{(1)} = 1/2, \pi^{(0,R)} = \pi^{(1,G)} = \alpha < 1/4$, $\max_{y,d} \|\mu^{(y,d)}\|_2 \leq C$ and Σ is positive definite. Suppose that (ξ, ξ^*, γ) satisfy that $0 \leq \xi^* \leq \gamma\xi$ and $\xi < \min\{\frac{\gamma}{2} \frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, \frac{\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma}\} - C\alpha$ for some large enough constant C and $\mathbb{E}[\lambda_i^2] / \max\{var(\lambda_i), 1/4\} \geq \|\widetilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma \|\Delta\|_\Sigma$. Then for any $p_{sel} \in [0, 1]$,*

$$\widehat{E}_{\text{LISA}}^{(wst*)} < \min\{\widehat{E}_{\text{ERM}}^{(wst*)}, \widehat{E}_{\text{mix}}^{(wst*)}\} + O_P\left(\frac{p \log n}{n} + \frac{p}{\alpha n}\right).$$

Similar to Theorem 1, this result shows that when domain has strong spurious correlation with the label (corresponding to small ξ), such a spurious correlation leads to the downgraded performance of ERM and vanilla mixup, while

our proposed LISA method is able to mitigate such an issue by selective data interpolation. Proofs of Theorem 1 and Theorem 2 are provided in Appendix B.

6. Related Work and Discussion

In this paper, we focus on improving the robustness of machine learning models to subpopulation shifts and domain shifts. Here, we discuss related approaches from the following three categories:

Learning Invariant Representations. Motivated by unsupervised domain adaptation (Ben-David et al., 2010; Ganin et al., 2016), the first category of works learns invariant representations by aligning representations across domains. The major research line of this category aims to eliminate the domain dependency by minimizing the divergence of feature distributions with different distance metrics, e.g., maximum mean discrepancy (Tzeng et al., 2014; Long et al., 2015), an adversarial loss (Ganin et al., 2016; Li et al., 2018), Wassertein distance (Zhou et al., 2020a). Follow-up works applied data augmentation to (1) generate more domains and enhance the consistency of representations during training (Yue et al., 2019; Zhou et al., 2020b; Xu et al., 2020; Yan et al., 2020; Shu et al., 2021; Wang et al., 2020; Yao et al., 2021) or (2) generate new domains in an adversarial way to imitate the challenging domains without using training domain information (Zhao et al., 2020; Qiao et al., 2020; Volpi et al., 2018). Unlike these latter methods, LISA instead focuses on learning invariant predictors without restricting the internal representations, leading to stronger empirical performance.

Learning Invariant Predictors. Beyond using domain alignment to learning invariant representations, recent work aims to further enhance the correlations between the invariant representations and the labels (Koyama & Yamaguchi, 2020), leading to invariant predictors. Representatively, motivated by casual inference, invariant risk minimization (IRM) (Arjovsky et al., 2019) and its variants (Guo et al., 2021; Khezeli et al., 2021; Ahuja et al., 2021) aim to find a predictor that performs well across all domains through regularizations. Other follow-up works leverage regularizers to penalize the variance of risks across all domains (Krueger et al., 2021), to align the gradient across domains (Koyama & Yamaguchi, 2020), to smooth the cross-domain interpolation paths (Chuang & Mroueh, 2021), or to involve game-theoretic invariant rationalization criterion (Chang et al., 2020). Instead of using regularizers, LISA instead learns domain-invariant predictors via data interpolation.

Group Robustness. The last category of methods combating spurious correlations and are particularly suitable for subpopulation shifts. These approaches include directly optimizing the worst-group performance with Distributionally Robust Optimization (Sagawa et al., 2020a; Zhang et al., 2021a; Zhou et al., 2021), generating samples around the minority groups (Goel et al., 2021), and balancing the majority and minority groups via reweighting (Sagawa et al., 2020b) or regularizing (Cao et al., 2019; 2020). A few recent approaches in this category target on subpopulation shifts without annotated group labels (Nam et al., 2020; Liu et al., 2021a; Zhang et al., 2021d; Creager et al., 2021; Lee et al., 2022). LISA proposes a more general strategy that is suitable for both domain shifts and subpopulation shifts.

7. Conclusion

To tackle distribution shifts, we propose LISA, a simple and efficient algorithm, to improve the out-of-distribution robustness. LISA aims to eliminate the domain-related spurious correlations among the training set with selective interpolation. We evaluate the effectiveness of LISA on nine datasets under subpopulation shifts and domain shifts settings, demonstrating its promise. Besides, detailed analyses verify that the performance gains caused by LISA result from encouraging learning invariant predictors and representations. Theoretical results further strengthen the superiority of LISA by showing smaller worst-group mis-classification error compared with ERM and vanilla data interpolation.

While we have made progress in learning invariant predictors with selective augmentation, a limitation of LISA is how to make it compatible with problems in which it is difficult to obtain examples with the same label (e.g., object detection, generative modeling). It would be interesting to explore more general selective augmentation strategies in the future. Additionally, we empirically find that intra-label LISA works without domain information in some domain shift situations. Systematically exploring domain-free intra-label LISA with a theoretical guarantee would be another interesting future direction.

Acknowledgement

We thank Pang Wei Koh for the many insightful discussions. This research was funded in part by JPMorgan Chase & Co. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material should not be construed as an individual recommendation for any particular client and is not intended as a recommendation of particular securities, financial instruments or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction. The research was also supported by Apple and Juniper Networks. The research of Linjun Zhang is partially supported by NSF DMS-2015378.

References

Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. 2021.

Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., and Mitliagkas, I. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

Anderson, T. W. An introduction to multivariate statistical analysis. Technical report, Wiley New York, 1962.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Cai, T. T. and Zhang, L. A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *The Annals of Statistics*, 49(3):1537–1568, 2021.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 2019.

Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., and Ma, T. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Chuang, C.-Y. and Mroueh, Y. Fair mixup: Fairness via interpolation. *ICLR*, 2021.

Cramér, H. *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton university press, 2016.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Goel, K., Gu, A., Li, Y., and Ré, C. Model patching: Closing the subgroup performance gap with data augmentation. In *ICLR*, 2021.

Guo, R., Zhang, P., Liu, H., and Kiciman, E. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

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

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Khezeli, K., Blaas, A., Soboczenski, F., Chia, N., and Kalantari, J. On invariance penalties for risk minimization. *arXiv preprint arXiv:2106.09777*, 2021.

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Koyama, M. and Yamaguchi, S. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL https://arxiv.org/abs/1602.07332.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Lee, H. B., Nam, T., Yang, E., and Hwang, S. J. Meta dropout: Learning to perturb latent features for generalization. In *International Conference on Learning Representations*, 2019.

Lee, Y., Yao, H., and Finn, C. Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*, 2022.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018.

Liang, W. and Zou, J. Metadataset: A dataset of datasets for evaluating distribution shifts and training conflicts. In *ICML2021 ML4data Workshop*, 2021.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *ICML*, pp. 6781–6792. PMLR, 2021a.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021b.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *ICCV*, 2015.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Nam, J., Cha, H., Ahn, S.-S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33, 2020.

Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Qiao, F., Zhao, L., and Peng, X. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12556–12565, 2020.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. In *ICLR*, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020a.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, pp. 8346–8356. PMLR, 2020b.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Shu, Y., Cao, Z., Wang, C., Wang, J., and Long, M. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9624–9633, 2021.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Taylor, J., Earnshaw, B., Mabey, B., Victors, M., and Yosinski, J. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, 2019.

Tony Cai, T. and Zhang, L. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wang, Y., Li, H., and Kot, A. C. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3622–3626. IEEE, 2020.

Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., and Zhang, W. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6502–6509, 2020.

Yan, S., Song, H., Li, N., Zou, L., and Ren, L. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

Yao, H., Zhang, L., and Finn, C. Meta-learning with fewer tasks through task interpolation. In *International Conference on Learning Representations*, 2021.

Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., and Gong, B. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2100–2110, 2019.

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

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. 2018.

Zhang, J., Menon, A., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. Coping with label shift via distributionally robust optimisation. In *ICLR*, 2021a.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? In *ICLR*, 2021b.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. When and how mixup improves calibration. *arXiv preprint arXiv:2102.06289*, 2021c.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021d.

Zhao, L., Liu, T., Peng, X., and Metaxas, D. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *arXiv preprint arXiv:2010.08001*, 2020.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhou, C., Ma, X., Michel, P., and Neubig, G. Examining and combating spurious features under distribution shift. In *ICML*, 2021.

Zhou, F., Jiang, Z., Shui, C., Wang, B., and Chaib-draa, B. Domain generalization with optimal transport and metric learning. *arXiv preprint arXiv:2007.10573*, 2020a.

Zhou, K., Yang, Y., Hospedales, T., and Xiang, T. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13025–13032, 2020b.

A. Additional Experiments

A.1. Additional Experiments on Subpopulation Shifts

A.1.1. Dataset Details

Colored MNIST (CMNIST): We classify MNIST digits from 2 classes, where classes 0 and 1 indicate original digits (0,1,2,3,4) and (5,6,7,8,9). The color is treated as a spurious attribute. Concretely, in the training set, the proportion between red samples and green samples is 8:2 in class 0, while the proportion is set as 2:8 in class 1. In the validation set, the proportion between green and red samples is 1:1 for all classes. In the test set, the proportion between green and red samples is 1:9 in class 0, while the ratio is 9:1 in class 1. The data sizes of train, validation, and test sets are 30000, 10000, and 20000, respectively. Follow (Arjovsky et al., 2019), we flip labels with probability 0.25.

Waterbirds (Sagawa et al., 2020a): The Waterbirds dataset aims to classify birds as "waterbird" or "landbird", where each bird image is spuriously associated with the background "water" or "land". Waterbirds is a synthetic dataset where each image is composed by pasting a bird image sampled from CUB dataset (Wah et al., 2011) to a background drawn from the Places dataset (Zhou et al., 2017). The bird categories in CUB are stratified as land birds or water birds. Specifically, the following bird species are selected to construct the waterbird class: albatross, auklet, cormorant, frigatebird, fulmar, gull, jaeger, kittiwake, pelican, puffin, tern, gadwall, grebe, mallard, merganser, guillemot, or Pacific loon. All other bird species are combined as the landbird class. We define (land background, waterbird) and (water background, landbird) are minority groups. There are 4,795 training samples while only 56 samples are "waterbirds on land" and 184 samples are "landbirds on water". The remaining training data include 3,498 samples from "landbirds on land", and 1,057 samples from "waterbirds on water".

CelebA (Liu et al., 2015; Sagawa et al., 2020a): For the CelebA data (Liu et al., 2015), we follow the data preprocess procedure from (Sagawa et al., 2020a). CelebA defines a image classification task where the input is a face image of celebrities and the classification label is its corresponding hair color – "blond" or "not blond." The label is spuriously correlated with gender, i.e., male or female. In CelebA, the minority groups are (blond, male) and (not blond, female). The number of samples for each group are 71,629 "dark hair, female", 66,874 "dark hair, male", 22,880 "blond hair, female", 1,387 "blond hair, male".

CivilComments (Borkan et al., 2019; Koh et al., 2021): We use CivilComments from the WILDS benchmark (Koh et al., 2021). CivilComments is a text classification task, aiming to predict whether an online comment is toxic or non-toxic. The spurious domain identifications are defined as the demographic features, including male, female, LGBTQ, Christian, Muslim, other religion, Black, and White. CivilComments contains 450,000 comments collected from online articles. The number of samples for training, validation, and test are 269,038, 45,180, and 133,782, respectively. The readers may kindly refer to Table 17 in (Koh et al., 2021) for the detailed group information.

A.1.2. Training Details

We adopt pre-trained ResNet-50 (He et al., 2016) and BERT (Sanh et al., 2019) as the model for image data (i.e., CMNIST, Waterbirds, CelebA) and text data (i.e., CivilComments), respectively. In each training iteration, we sample a batch of data per group. For intra-label LISA, we randomly apply mixup on sample batches with the same labels but different domains. For intra-domain LISA, we instead apply mixup on sample batches with the same domain but different labels. The interpolation ratio λ is sampled from the distribution $\mathrm{Beta}(2, 2)$. All hyperparameters are selected via cross-validation and are listed in Table 9.

A.1.3. Additional Results

In this section, we have added the full results of subpopulation shifts in Table 10 and Table 11.

A.2. Additional Experimental Settings on Domain Shifts

A.2.1. Dataset Details

In this section, we provide detailed descriptions of datasets used in the experiments of domain shifts and report the data statistics in Table 4.

Table 9. Hyperparameter settings for the subpopulation shifts.

Dataset	CMNIST	Waterbirds	CelebA	CivilComments
Learning rate	1e-3	1e-3	1e-4	1e-5
Weight decay	1e-4	1e-4	1e-4	0
Scheduler	n/a	n/a	n/a	n/a
Batch size	16	16	16	8
Type of mixup	mixup	mixup	CutMix	ManifoldMix
Architecture	ResNet50	ResNet50	ResNet50	DistilBert
Optimizer	SGD	SGD	SGD	Adam
Maximum Epoch	300	300	50	3
Strategy sel. prob. p_{sel}	0.5	0.5	0.5	1.0

Table 10. Full results of subpopulation shifts with standard deviation. All the results are performed with three random seed.

	CMNIST		Waterbirds	
	Avg.	Worst	Avg.	Worst
ERM	27.8 ± 1.9%	0.0 ± 0.0%	97.0 ± 0.2%	63.7 ± 1.9%
UW	72.2 ± 1.1%	66.0 ± 0.7%	95.1 ± 0.3%	88.0 ± 1.3%
IRM	72.1 ± 1.2%	70.3 ± 0.8%	87.5 ± 0.7%	75.6 ± 3.1%
IB-IRM	72.2 ± 1.3%	70.7 ± 1.2%	88.5 ± 0.6%	76.5 ± 1.2 %
V-REx	71.7 ± 1.2%	70.2 ± 0.9%	88.0 ± 1.0%	73.6 ± 0.2%
Coral	71.8 ± 1.7%	69.5 ± 0.9%	90.3 ± 1.1%	79.8 ± 1.8%
GroupDRO	72.3 ± 1.2%	68.6 ± 0.8%	91.8 ± 0.3%	**90.6 ± 1.1%**
DomainMix	51.4 ± 1.3%	48.0 ± 1.3%	76.4 ± 0.3%	53.0 ± 1.3%
Fish	46.9 ± 1.4%	35.6 ± 1.7%	85.6 ± 0.4%	64.0 ± 0.3%
LISA	74.0 ± 0.1%	**73.3 ± 0.2%**	91.8 ± 0.3%	89.2 ± 0.6%

	CelebA		CivilComments	
	Avg.	Worst	Avg.	Worst
ERM	94.9 ± 0.2%	47.8 ± 3.7%	92.2 ± 0.1%	56.0 ± 3.6%
UW	92.9 ± 0.2%	83.3 ± 2.8%	89.8 ± 0.5%	69.2 ± 0.9%
IRM	94.0 ± 0.4%	77.8 ± 3.9%	88.8 ± 0.7%	66.3 ± 2.1%
IB-IRM	93.6 ± 0.3%	85.0 ± 1.8%	89.1 ± 0.3%	65.3 ± 1.5%
V-REx	92.2 ± 0.1%	86.7 ± 1.0%	90.2 ± 0.3%	64.9 ± 1.2%
Coral	93.8 ± 0.3%	76.9 ± 3.6%	88.7 ± 0.5%	65.6 ± 1.3%
GroupDRO	92.1 ± 0.4%	87.2 ± 1.6%	89.9 ± 0.5%	70.0 ± 2.0%
DomainMix	93.4 ± 0.1%	65.6 ± 1.7%	90.9 ± 0.4%	63.6 ± 2.5%
Fish	93.1 ± 0.3%	61.2 ± 2.5%	89.8 ± 0.4%	71.1 ± 0.4%
LISA (ours)	92.4 ± 0.4%	**89.3 ± 1.1%**	89.2 ± 0.9%	**72.6 ± 0.1%**

Camelyon17 We use Camelyon17 from the WILDS benchmark (Koh et al., 2021; Bandi et al., 2018), which provides $450,000$ lymph-node scans sampled from 5 hospitals. Camelyon17 is a medical image classification task where the input x is a 96×96 image and the label y is whether there exists tumor tissue in the image. The domain d denotes the hospital that the patch was taken from. The training dataset is drawn from the first 3 hospitals, while out-of-distribution validation and out-of-distribution test datasets are sampled from the 4-th hospital and 5-th hospital respectively.

FMoW The FMoW dataset is from the WILDS benchmark (Koh et al., 2021; Christie et al., 2018) — a satellite image classification task which includes 62 classes and 80 domains (16 years x 5 regions). Concretely, the input x is a 224×224 RGB satellite image, the label y is one of the 62 building or land use categories, and the domain d represents the year that the image was taken as well as its corresponding geographical region – Africa, the Americas, Oceania, Asia, or Europe. The train/test/validation splits are based on the time when the images are taken. Specifically, images taken before 2013 are used as the training set. Images taken between 2013 and 2015 are used as the validation set. Images taken after 2015 are used for testing.

RxRx1 RxRx1 (Koh et al., 2021; Taylor et al., 2019) from the WILDS benchmark is a cell image classification task. In the dataset, some cells have been genetically perturbed by siRNA. The goal of RxRx1 is to predict which siRNA that

Table 11. Full table of the comparison between LISA and other substitute mixup strategies in subpopulation shifts. UW represents upweighting.

	CMNIST		Waterbirds	
	Avg.	Worst	Avg.	Worst
ERM	$27.8 \pm 1.9\%$	$0.0 \pm 0.0\%$	$97.0 \pm 0.2\%$	$63.7 \pm 1.9\%$
Vanilla mixup	$32.6 \pm 3.1\%$	$3.1 \pm 2.4\%$	$81.0 \pm 0.2\%$	$56.2 \pm 0.2\%$
Vanilla mixup + UW	$72.2 \pm 0.7\%$	$71.8 \pm 0.1\%$	$92.1 \pm 0.1\%$	$85.6 \pm 1.0\%$
In-group Group	$33.6 \pm 1.9\%$	$24.0 \pm 1.1\%$	$88.7 \pm 0.3\%$	$68.0 \pm 0.4\%$
In-group + UW	$72.6 \pm 0.1\%$	$71.6 \pm 0.2\%$	$91.4 \pm 0.6\%$	$87.1 \pm 0.6\%$
LISA (ours)	$74.0 \pm 0.1\%$	$\mathbf{73.3 \pm 0.2\%}$	$91.8 \pm 0.3\%$	$\mathbf{89.2 \pm 0.6\%}$
	CelebA		CivilComments	
	Avg.	Worst	Avg.	Worst
ERM	$94.9 \pm 0.2\%$	$47.8 \pm 3.7\%$	$92.2 \pm 0.1\%$	$56.0 \pm 3.6\%$
Vanilla mixup	$95.8 \pm 0.0\%$	$46.4 \pm 0.5\%$	$90.8 \pm 0.8\%$	$67.2 \pm 1.2\%$
Vanilla mixup + UW	$91.5 \pm 0.2\%$	$88.0 \pm 0.3\%$	$87.8 \pm 1.2\%$	$66.1 \pm 1.4\%$
Within Group	$95.2 \pm 0.3\%$	$58.3 \pm 0.9\%$	$90.8 \pm 0.6\%$	$69.2 \pm 0.8\%$
Within Group + UW	$92.4 \pm 0.4\%$	$87.8 \pm 0.6\%$	$84.8 \pm 0.7\%$	$69.3 \pm 1.1\%$
LISA (ours)	$92.4 \pm 0.4\%$	$\mathbf{89.3 \pm 1.1\%}$	$89.2 \pm 0.9\%$	$\mathbf{72.6 \pm 0.1\%}$

the cells have been treated with. Concretely, the input x is an image of cells obtained by fluorescent microscopy, the label y indicates which of the $1,139$ genetic treatments the cells received, and the domain d denotes the experimental batches. Here, 33 different batches of images are used for training, where each batch contains one sample for each class. The out-of-distribution validation set has images from 4 experimental batches. The out-of-distribution test set has 14 experimental batches. The average accuracy on out-of-distribution test set is reported.

Amazon Each task in the Amazon benchmark (Koh et al., 2021; Ni et al., 2019) is a multi-class sentiment classification task. The input x is the text of a review, the label y is the corresponding star rating ranging from 1 to 5, and the domain d is the corresponding reviewer. The training set has $245,502$ reviews from $1,252$ reviewers, while the out-of-distribution validation set has $100,050$ reviews from another $1,334$ reviewers. The out-of-distribution test set also has $100,050$ reviews from the rest $1,252$ reviewers. We evaluate the models by the 10th percentile of per-user accuracies in the test set.

MetaShift We use the MetaShift (Liang & Zou, 2021), which is derived from Visual Genome (Krishna et al., 2016). MetaShift leverages the natural heterogeneity of Visual Genome to provide many distinct data distributions for a given class (e.g. "cats with cars" or "cats in bathroom" for the "cat" class). A key feature of MetaShift is that it provides explicit explanations of the dataset correlation and a distance score to measure the degree of distribution shift between any pair of sets.

We adopt the "Cat vs. Dog" task in MetaShift, where we evaluate the model on the "dog(*shelf*)" domain with 306 images, and the "cat(*shelf*)" domain with 235 images. The training data for the "Cat" class is the cat(*sofa* + *bed*), including cat(*sofa*) domain and cat(*bed*) domain. MetaShift provides 4 different sets of training data for the "Dog" class in an increasingly challenging order, i.e., increasing the amount of distribution shift. Specifically, dog(*cabinet* + *bed*), dog(*bag* + *box*), dog(*bench* + *bike*), dog(*boat* + *surfboard*) are selected for training, where their corresponding distances to dog(*shelf*) are 0.44, 0.71, 1.12, 1.43.

A.2.2. Training Details

Follow WILDS Koh et al. (2021), we adopt pre-trained DenseNet121 (Huang et al., 2017) for Camelyon17 and FMoW datasets, ResNet-50 (He et al., 2016) for RxRx1 and MetaShift datasets, and DistilBert (Sanh et al., 2019) for Amazon datasets.

In each training iteration, we first draw a batch of samples B_1 from the training set. With B_1, we then select another sample batch B_2 with same labels as B_1 for data interpolation. The interpolation ratio λ is drawn from the distribution $\mathrm{Beta}(2,2)$. We use the same image transformers as Koh et al. (2021), and all other hyperparameters are selected via cross-validation and are listed in Table 12.

Table 12. Hyperparameter settings for the domain shifts.

Dataset	Camelyon17	FMoW	RxRx1	Amazon	MetaShift
Learning rate	1e-4	1e-4	1e-3	2e-6	1e-3
Weight decay	0	0	1e-5	0	1e-4
Scheduler	n/a	n/a	Cosine Warmup	n/a	n/a
Batch size	32	32	72	8	16
Type of mixup	CutMix	CutMix	CutMix	ManifoldMix	CutMix
Architecture	DenseNet121	DenseNet121	ResNet50	DistilBert	ResNet50
Optimizer	SGD	Adam	Adam	Adam	SGD
Maximum Epoch	2	5	90	3	100
Strategy sel. prob. p_{sel}	1.0	1.0	1.0	1.0	1.0

A.3. Strength of Spurious Correlation

In Section 2, the spurious correlation is defined as the association between the domain d and label y, measured by Cramér's V (Cramér, 2016). Specifically, let $k_{y,d}$ be the number of samples from domain d with label y. The Cramér's V is formulated as

$$V = \sqrt{\frac{\chi^2}{N \min(|Y-1|, |D-1|)}} = \sqrt{\frac{\sum_{y\in\mathcal{Y}, d\in\mathcal{D}} \frac{(k_{y,d}-\tilde{k}_{y,d})^2}{\tilde{k}_{y,d}}}{N \min(|\mathcal{Y}|-1, |\mathcal{D}|-1|)}}, \tag{6}$$

where N represents the number of samples in the entire dataset and $\tilde{k}_{y,d} = \frac{\sum_{y\in\mathcal{Y}} k_{y,d} \sum_{d\in\mathcal{D}} k_{y,d}}{\sum_{y\in\mathcal{Y}, d\in\mathcal{D}} k_{y,d}}$. Cramér's V varies from 0 to 1 and higher Cramér's V represents stronger correlation.

According to Eq. (6), we calculate the strength of spurious correlations on all datasets used in the experiments and report the results in Table 13. Compared with other datasets, the Cramér's V on Camelyon17, FMoW and RxRx1 are significantly smaller, indicating weaker spurious correlations.

Table 13. Analysis of the strength of spurious correlations on datasets with subpopulation shifts or domain shifts.

Subpopulation Shifts				Domain Shifts				
CMNIST	Waterbirds	CelebA	CivilComments	Camelyon17	FMoW	RxRx1	Amazon	MetaShift
0.6000	0.8672	0.3073	0.8723	0.0004	0.1114	0.0067	0.3377	0.4932

A.4. Results on Datasets without Spurious Correlations

In order to analyze the factors that lead to the performance gains of LISA, we conduct experiments on datasets without spurious correlations. To be more specific, we balance the number of samples for each group under the subpopulation shifts setting. The results of ERM, Vanilla mixup and LISA on CMNIST, Waterbirds and CelebA are reported in Table 14. The results show that LISA performs similarly compared with ERM when datasets do not have spurious correlations. If there exists any spurious correlation, LISA significantly outperforms ERM. Another interesting finding is that Vanilla mixup outperforms LISA and ERM without spurious correlations, while LISA achieves the best performance with spurious correlations. This finding strengthens our conclusion that the performance gains of LISA are from eliminating spurious correlations rather than simple data augmentation.

Table 14. Results on datasets without spurious correlations. LISA performs similarly to ERM when there are no spurious correlations. However, Vanilla mixup outperforms LISA and ERM when there are no spurious correlations while underperforms LISA on datasets with spurious correlations. The results further strengthen our claim that the performance gains of LISA are not from simple data augmentation.

Dataset	CMNIST	Waterbirds	CelebA
ERM	73.67%	88.07%	86.11%
Vanilla mixup	74.28%	88.23%	88.89%
LISA	73.18%	87.05%	87.22%

A.5. Additional Invariance Analysis

A.5.1. Additional Metrics of Invariant Predictor Analysis

In Table 15, we report two additional metrics to measure the invariance of predictors – Risk Variance and Gradient Norm, which is defined as:

- **Risk Variance** (IP_{var}). Motivated by Krueger et al. (2021), we use the variance of test risks across all domains to measure the invariance, which is defined as $\text{IP}_{var} = \text{Var}(\{\mathcal{R}_1(\theta), \ldots, \mathcal{R}_D(\theta)\})$, where D represents the number of test domains and $\mathcal{R}_d(\theta)$ represents the risk of domain d.

- **Gradient Norm** (IP_{norm}). Follow IRMv1 (Arjovsky et al., 2019), we use the gradient norm of the classifier to measure the optimality of the dummy classifier at each domain d. Assume the classifier is parameterized by w, IP_{norm} is defined as: $\text{IP}_{norm} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \|\nabla_{w|w=1.0} \mathcal{R}_d(\theta)\|^2$.

Table 15. Additional Invariance Metrics for Invariant Predictor Analysis. We report the results of risk variance (IP_{var}) and gradient norm (IP_{norm}), where smaller values indicate stronger invariance.

	$\text{IP}_{var} \downarrow$				$\text{IP}_{norm} \downarrow$			
	CMNIST	Waterbirds	Camelyon	MetaShift	CMNIST	Waterbirds	Camelyon	MetaShift
ERM	12.0486	0.2456	0.0150	1.8824	1.1162	1.5780	1.2959	1.0914
Vanilla mixup	0.2769	0.1465	0.0180	0.2659	1.5347	1.8631	0.3993	0.1985
IRM	0.0112	0.1243	0.0201	0.8748	0.0908	0.9798	0.5266	0.2320
IB-IRM	0.0072	0.2069	0.0329	0.5680	0.6225	0.8814	0.6890	0.1683
V-REx	0.0056	0.1257	0.0106	0.4220	0.0290	0.8329	0.9641	0.3680
LISA (ours)	**0.0012**	**0.0016**	**9.97e-5**	**0.2387**	**0.0039**	**0.0538**	**0.3081**	**0.1354**

Comparing LISA to other invariant learning methods, the results of IP_{var} and IP_{norm} further confirm that LISA does indeed improve predictor invariance.

A.5.2. Analysis of Learned Invariant Representations

In this section, we use pairwise divergence of representations (IR_{kl}) to measure representation-level invariance. Specifically, assume the representation before classifier of each sample (x_i, y_i, d) is $h_{i,d}$, we compute the KL divergence of the distribution of representations. Similarly, kernel density estimation is also used to estimate the probability density function $P(h_d^y)$ of representations from domain d with label y. Formally, IR_{kl} is defined as $\text{IR}_{kl} = \frac{1}{|\mathcal{Y}||\mathcal{D}|^2} \sum_{y \in \mathcal{Y}} \sum_{d', d \in \mathcal{D}} \text{KL}(P(h_D^y \mid D = d) | P(h_D^y \mid D = d'))$. Smaller IR_{kl} values indicate more invariant representations with respect to the labels. We report the results on CMNIST, Waterbirds, Camelyon17 and MetaShift in Table 16. Our key observations are: (1) Compared with ERM, LISA learns stronger representation-level invariance. The potential reason is that a stronger invariant predictor implicitly includes stronger invariance representation; (2) LISA provides more invariant representations than other regularization-based invariant predictor learning methods, i.e., IRM, IB-IRM, V-REx, showing its capability in learning stronger invariance.

Table 16. Results of representation-level invariance IR_{kl} ($\times 10^8$ for CMNIST), where smaller IR_{kl} value denotes stronger invariance.

	CMNIST	Waterbirds	Camelyon17	MetaShift
ERM	1.683	3.592	8.213	0.632
Vanilla mixup	4.392	3.935	7.786	0.634
IRM	1.905	2.413	8.169	0.627
IB-IRM	3.178	3.306	8.824	0.646
V-REx	3.169	3.414	8.838	0.617
LISA (ours)	**0.421**	**1.912**	**7.570**	**0.585**

Besides the quantitative analysis, follow Appendix C in Lee et al. (2019), we visualize the hidden representations for all test samples and the decision boundary on Waterbirds and illustrate the results in Figure 3. Compared with other methods, the representations of samples with the same label that learned by LISA are closer regardless of their domain information, which further demonstrates the promise of LISA in producing invariant representations.

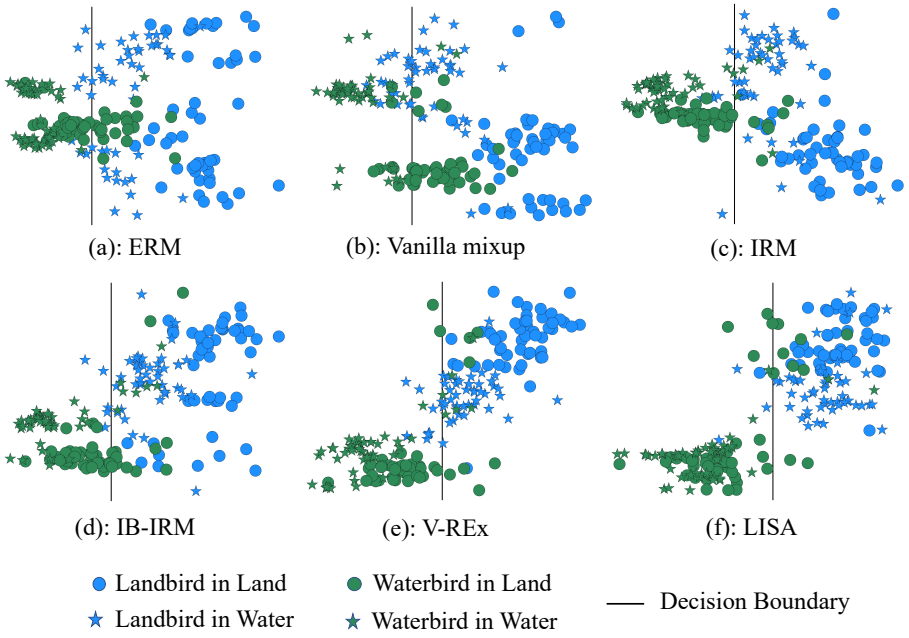

Figure 3. Visualization of sample representations and decision boundaries on Waterbirds dataset.

A.6. Full Results of WILDS data

Follow Koh et al. (2021), we reported more results on WILDS datasets in Table 17 - Table 20, including validation performance and the results of other metrics. According to these additional results, we could see that LISA outperforms other baseline approaches in all scenarios. Particularly, we here discuss two additional findings: (1) In Camelyon dataset, the test data is much more visually distinctive compared with the validation data, resulting in the large gap ($\sim 10\%$) between validation and test performance of ERM (see Table 17). However, LISA significantly reduces the performance gap between the validation and test sets, showing its promise in improving OOD robustness; (2) In Amazon dataset, though LISA performs worse than ERM in average accuracy, it achieves the best accuracy at the 10th percentile, which is regarded as a more common and important metric to evaluate whether models perform consistently well across all users (Koh et al., 2021).

Table 17. Full Results of Camelyon17. We report both validation accuracy and test accuracy.

	Validation Acc.	Test Acc.
ERM	84.9 ± 3.1%	70.3 ± 6.4%
IRM	**86.2 ± 1.4%**	64.2 ± 8.1%
IB-IRM	80.5 ± 0.4%	68.9 ± 6.1%
V-REx	82.3 ± 1.3%	71.5 ± 8.3%
Coral	**86.2 ± 1.4%**	59.5 ± 7.7%
GroupDRO	85.5 ± 2.2%	68.4 ± 7.3%
DomainMix	83.5 ± 1.1%	69.7 ± 5.5%
Fish	83.9 ± 1.2%	74.7 ± 7.1%
LISA (ours)	81.8 ± 1.3%	**77.1 ± 6.5%**

B. Proofs of Theorem 1 and Theorem 2

Outline of the proof. We will first find the mis-classification errors based on the population version of OLS with different mixup strategies. Next, we will develop the convergence rate of the empirical OLS based on n samples towards its population version. These two steps together give us the empirical mis-classification errors of different methods. We will separately show that the upper bounds in Theorem 1 and Theorem 2 hold for two selective augmentation strategies of LISA and hence hold for any $p_{sel} \in [0, 1]$. Let LL denote intra-label LISA and LD denote intra-domain LISA.

Let $\pi_1 = \mathbb{P}(y_i = 1)$ and $\pi_0 = \mathbb{P}(y_i = 0)$ denote the marginal class proportions in the training samples. Let $\pi_R = \mathbb{P}(d_i = R)$ and $\pi_G = \mathbb{P}(d_i = G)$ denote the marginal subpopulation proportions in the training samples. Let $\pi_{G|1} = \mathbb{P}(d_i = G|y_i = 1)$ and define $\pi_{G|0}$, $\pi_{R|1}$, and $\pi_{R|0}$ similarly.

Table 18. Full Results of FMoW. Here, we report the average accuracy and the worst-domain accuracy on both validation and test sets.

	Validation		Test	
	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.
ERM	**59.5 ± 0.37%**	48.9 ± 0.62%	**53.0 ± 0.55%**	32.3 ± 1.25%
IRM	57.4 ± 0.37%	47.5 ± 1.57%	50.8 ± 0.13%	30.0 ± 1.37%
IB-IRM	56.1 ± 0.48%	45.0 ± 0.62%	49.5 ± 0.49%	28.4 ± 0.90%
V-REx	55.3 ± 1.75%	44.7 ± 1.31%	48.0 ± 0.64%	27.2 ± 0.78%
Coral	56.9 ± 0.25%	47.1 ± 0.43%	50.5 ± 0.36%	31.7 ± 1.24%
GroupDRO	58.8 ± 0.19%	46.5 ± 0.25%	52.1 ± 0.50%	30.8 ± 0.81%
DomainMix	58.6 ± 0.29%	48.9 ± 1.15%	51.6 ± 0.19%	34.2 ± 0.76%
Fish	57.8 ± 0.15%	**49.5 ± 2.34%**	51.8 ± 0.32%	34.6 ± 0.18%
LISA (ours)	58.7 ± 0.92%	48.7 ± 0.74%	**52.8 ± 0.94%**	**35.5 ± 0.65%**

Table 19. Full Results of RxRx1. ID: in-distribution; OOD: out-of-distribution

	Validation Acc.	Test ID Acc.	Test OOD Acc.
ERM	19.4 ± 0.2%	35.9 ± 0.4%	29.9 ± 0.4%
IRM	5.6 ± 0.4%	9.9 ± 1.4%	8.2 ± 1.1%
IB-IRM	4.3 ± 0.7%	7.9 ± 0.5%	6.4 ± 0.6%
V-REx	5.2 ± 0.6%	9.3 ± 0.9%	7.5 ± 0.8%
Coral	18.5 ± 0.4%	34.0 ± 0.3%	28.4 ± 0.3%
GroupDRO	15.2 ± 0.1%	28.1 ± 0.3%	23.0 ± 0.3%
DomainMix	19.3 ± 0.7%	39.8 ± 0.2%	30.8 ± 0.4%
Fish	7.5 ± 0.6%	12.7 ± 1.9%	10.1 ± 1.5%
LISA (ours)	**20.1 ± 0.4%**	**41.2 ± 1.0%**	**31.9 ± 0.8%**

We consider the setting where $\alpha := \pi^{(1,G)} = \pi^{(0,R)}$ is relatively small and $\pi^{(1)} = \pi^{(0)} = \pi^{(G)} = \pi^{(R)} = 1/2$.

B.1. Decomposing the loss function

Recall that $\Delta = \mu^{(1,G)} - \mu^{(0,G)} = \mu^{(1,R)} - \mu^{(0,R)}$. We further define $\widetilde{\Delta} = \mu^{(1)} - \mu^{(0)}$, $\theta^{(G)} = \mu^{(0,G)} - \mathbb{E}[x_i]$, and $\theta^{(R)} = \mu^{(0,R)} - \mathbb{E}[x_i]$.

For the mixup estimators, we will repeatedly use the fact that λ_i has a symmetric distribution with support $[0, 1]$.

For ERM estimator based on (X, y), where $b_0 = \frac{1}{2} - \mathbb{E}[x_i]^T b$, we have

$$\begin{aligned}(\mu^{(0,G)})^T b + b_0 &= (\mu^{(0,G)} - \mathbb{E}[x_i])^T b + \frac{1}{2} \\ &= (\theta^{(G)})^T b + \mathbb{E}[y_i] \\ (\mu^{(1,G)})^T b + b_0 &= (\mu^{(1,G)} - \mathbb{E}[x_i])^T b + \frac{1}{2} \\ &= \Delta^T b + (\theta^{(G)})^T b + \mathbb{E}[y_i],\end{aligned}$$

Notice that based on the estimator b, b_0, for $d \in \{G, R\}$,

$$E^{(1,d)}(b, b_0) = \Phi(\frac{-\Delta^T b - (\theta^{(d)})^T b}{\sqrt{b^T \Sigma b}}) \text{ and } E^{(0,d)}(b, b_0) = \Phi(\frac{(\theta^{(d)})^T b}{\sqrt{b^T \Sigma b}}). \tag{7}$$

In the extreme case where $\pi_{0,R} = \pi_{1,G} = 0$, we have

$$\widetilde{\Delta} = \mu^{(1,R)} - \mu^{(0,G)}, \theta^{(G)} = -\frac{1}{2}\widetilde{\Delta}, \theta^{(R)} = \frac{1}{2}\widetilde{\Delta} - \Delta, \text{ and } \Delta_0 := \mu^{(0,G)} - \mu^{(0,R)} = \Delta - \widetilde{\Delta}.$$

Table 20. Full Results of Amazon. Both the average accuracy and the 10th Percentile accuracy are reported.

	Validation		Test	
	Avg. Acc.	10-th Per.	Avg. Acc.	10-th Per. Acc.
ERM	**72.7 ± 0.1%**	**55.2 ± 0.7%**	71.9 ± 0.1%	53.8 ± 0.8%
IRM	71.5 ± 0.3%	54.2 ± 0.8%	70.5 ± 0.3%	52.4 ± 0.8%
IB-IRM	72.4 ± 0.4%	**55.1 ± 0.6%**	**72.2 ± 0.3%**	53.8 ± 0.7%
V-REx	72.7 ± 1.2%	53.8 ± 0.7%	71.4 ± 0.4%	53.3 ± 0.0%
Coral	72.0 ± 0.3%	54.7 ± 0.0%	70.0 ± 0.6%	52.9 ± 0.8%
GroupDRO	70.7 ± 0.6%	54.7 ± 0.0%	70.0 ± 0.6%	53.3 ± 0.0%
DomainMix	71.9 ± 0.2%	54.7 ± 0.0%	71.1 ± 0.1%	53.3 ± 0.0%
Fish	72.5 ± 0.0%	54.7 ± 0.0%	71.7 ± 0.1%	53.3 ± 0.0%
LISA (ours)	71.6 ± 0.4%	**55.1 ± 0.6%**	70.8 ± 0.3%	**54.7 ± 0.0%**

Hence,

$$E_0^{(wst)} = \max\{\Phi\left(\frac{(\frac{1}{2}\widetilde{\Delta} - \Delta)^T b}{\sqrt{b^T \Sigma b}}\right), \Phi\left(\frac{-\frac{1}{2}\widetilde{\Delta}^T b}{\sqrt{b^T \Sigma b}}\right)\}. \tag{8}$$

B.2. Classification errors of four methods with infinite training samples

We first provide the limit of the classification errors when $n \to \infty$.

B.2.1. Baseline method: ERM

For the training data, it is easy to show that

$$\begin{aligned}
var(x) &= \mathbb{E}[var(x|y)] + var(\mathbb{E}[x|y]) \\
&= \Sigma + \mathbb{E}[var(\mathbb{E}[x|y, D]|y)] + var((\mu^{(1)} - \mu^{(0)})y) \\
&= \Sigma + \mathbb{E}[var(\mu^{(0,R)} - \mu^{(0,G)})\mathbb{1}(D = R)|y)] + \widetilde{\Delta}^{\otimes 2}\pi^{(1)}\pi^{(0)} \\
&= \Sigma + \frac{1}{2}(\mu^{(0,R)} - \mu^{(0,G)})^{\otimes 2}(\pi_{R|1}\pi_{G|1} + \pi_{R|0}\pi_{G|0}) + \widetilde{\Delta}^{\otimes 2}\pi^{(1)}\pi^{(0)} \\
cov(x, y) &= cov(\mathbb{E}[x|y], y) \\
&= cov(\mu^{(0)} + \widetilde{\Delta}y, y) \\
&= cov(\widetilde{\Delta}y, y) = \widetilde{\Delta}\pi^{(1)}\pi^{(0)}
\end{aligned}$$

For $a_0 = \frac{1}{2}(\pi_{R|1}\pi_{G|1} + \pi_{R|0}\pi_{G|0})$ and $\Delta_0 = \mu^{(0,G)} - \mu^{(0,R)}$, the ERM has slope and intercept being

$$\begin{aligned}
b &= var(x)^{-1} cov(x, y) \\
&\propto (\Sigma + a_0\Delta_0^{\otimes 2})^{-1}\widetilde{\Delta} \\
&= \Sigma^{-1}\widetilde{\Delta} - \Sigma^{-1}\Delta_0 \cdot \frac{a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}{1 + a_0\Delta_0^T\Sigma^{-1}\Delta_0} \\
b_0 &= \mathbb{E}[y] - \mathbb{E}[x^T b].
\end{aligned}$$

B.2.2. Baseline method: Vanilla mixup

The vanilla mixup does not use the group information. Let i_1 be a random draw from $\{1, \ldots, n\}$. Let i_2 be a random draw from $\{1, \ldots, n\}$ independent of i_1. Let

$$\tilde{y}_i = \lambda_i y_{i_1} + (1 - \lambda_i)y_{i_2}$$

and

$$\tilde{x}_i = \lambda_i x_{i_1} + (1 - \lambda_i)x_{i_2}.$$

We can find that

$$\begin{aligned} cov(\tilde{x}_i, \tilde{y}_i) &= cov(\lambda_i x_{i_1} + (1-\lambda_i)x_{i_2}, \lambda_i y_{i_1} + (1-\lambda_i)y_{i_2}) \\ &= cov(\lambda_i x_{i_1}, \lambda_i y_{i_1}) + cov((1-\lambda_i)x_{i_2}, (1-\lambda_i)y_{i_2}) \\ &= (\mathbb{E}[\lambda_i^2] + \mathbb{E}[(1-\lambda_i)^2])cov(x_i, y_i). \\ cov(\tilde{x}_i) &= (\mathbb{E}[\lambda_i^2] + \mathbb{E}[(1-\lambda_i)^2])cov(x_i). \end{aligned}$$

Hence, the population-level slope is the same as the slope in the benchmark method. It is easy to show that the population-level intercept is also the same. Hence,

$$E_{\text{mix}}^{(wst)} = E_0^{(wst)}.$$

B.3. Intra-label LISA (LISA-L): mixup across domain

Define

$$x_i^{(\lambda)} = \lambda_i x_{i_1}^{(y_i,G)} + (1-\lambda_i)x_{i_2}^{(y_i,R)},$$

where i_1 is a random draw from $\{l : y_l = y_i, D_l = G\}$ and i_2 is a random draw from $\{l : y_l = y_i, D_l = R\}$. Then we perform OLS based on $(x_i^{(\lambda)}, y_i), i = 1, \ldots, n$.

We can calculate that

$$\begin{aligned} cov(x_i^{(\lambda)}, y_i) &= cov(\mathbb{E}[x_i^{(\lambda)}|y_i], y_i) = cov(\frac{1}{2}\mu^{(y_i,G)} + \frac{1}{2}\mu^{(y_i,R)}, y_i) \\ &= var(y_i)\Delta = \pi^{(1)}\pi^{(0)}\Delta \\ cov(x_i^{(\lambda)}) &= \mathbb{E}[cov(x_i^{(\lambda)}|y_i, \lambda_i)] + cov(\mathbb{E}[x_i^{(\lambda)}|y_i, \lambda_i]) \\ &= 2\mathbb{E}[\lambda_i^2]\Sigma + cov(\lambda_i(\mu^{(0,G)} - \mu^{(0,R)}) + \Delta y_i) \\ &= 2\mathbb{E}[\lambda_i^2]\Sigma + var(\lambda_i)(\mu^{(0,G)} - \mu^{(0,R)})^{\otimes 2} + \pi^{(1)}\pi^{(0)}\Delta^{\otimes 2}. \end{aligned}$$

B.4. Intra-domain LISA (LISA-D): mixup within each domain

The interpolated sample can be written as

$$\begin{aligned} (\tilde{y}_i, \tilde{x}_i) &= (\lambda_i, \lambda_i x_{i_1}^{(1,G)} + (1-\lambda_i)x_{i_2}^{(0,G)}) \text{ if } d_i = G \\ (\tilde{y}_i, \tilde{x}_i) &= (\lambda_i, \lambda_i x_{i_1}^{(1,R)} + (1-\lambda_i)x_{i_2}^{(0,R)}) \text{ if } d_i = R, \end{aligned}$$

where i_1 is a random draw from $\{l : d_l = d_i, y_i = 1\}$ and i_2 is a random draw from $\{l : d_l = d_i, y_i = 0\}$.

We consider regress $\tilde{y}_i$ on $\tilde{x}_i$.

$$\begin{aligned} cov(\tilde{x}_i, \tilde{y}_i|d_i = G) &= cov(\mathbb{E}[\tilde{x}_i|\tilde{y}_i, d_i = G], \tilde{y}_i|d_i = G) = var(\tilde{y}_i)(\mu^{(1,G)} - \mu^{(0,G)}) \\ var(\tilde{x}_i|d_i = G) &= \mathbb{E}[var(\tilde{x}_i|, \lambda_i, D_i = G)|d_i = G] + var(\mathbb{E}[\tilde{x}_i|, \lambda_i, d_i = G]|D_i = G] \\ &= 2\mathbb{E}[\lambda_i^2]\Sigma + var(\lambda_i\mu^{(1,G)} + (1-\lambda_i)\mu^{(0,G)}|d_i = G) \\ &= 2\mathbb{E}[\lambda_i^2]\Sigma + var(\tilde{y}_i)\Delta^{\otimes 2}. \end{aligned}$$

We further have

$$\begin{aligned} cov(\tilde{x}_i, \tilde{y}_i) &= \mathbb{E}[cov(\tilde{x}_i, \tilde{y}_i|d_i)] + cov(\mathbb{E}[\tilde{x}_i|d_i], \mathbb{E}[\tilde{y}_i|d_i]) \\ &= cov(\tilde{x}_i^{(G)}, \tilde{y}_i^{(G)})\pi^{(G)} + cov(\tilde{x}_i^{(R)}, \tilde{y}_i^{(R)})\pi^{(R)} \\ &= var(\tilde{y}_i)(\mu^{(1,G)} - \mu^{(0,G)})\pi^{(G)} + var(\tilde{y}_i)(\mu^{(1,R)} - \mu^{(0,R)})\pi^{(R)} \\ &= var(\tilde{y}_i)\Delta. \end{aligned}$$

Moreover,

$$\begin{aligned} var(\tilde{x}_i) &= \mathbb{E}[var(\tilde{x}_i|d_i)] + var(\mathbb{E}[\tilde{x}_i|d_i]) \\ &= var(\tilde{x}_i^{(G)})\pi^{(G)} + var(\tilde{x}^{(R)})\pi^{(R)} + (\mathbb{E}[\tilde{x}^{(G)}] - \mathbb{E}[\tilde{x}^{(R)}])^{\otimes 2}\pi^{(G)}\pi^{(R)} \\ &= 2\mathbb{E}[\lambda_i^2]\Sigma + var(\lambda_i)\Delta^{\otimes 2} + (\mu^{(0,G)} - \mu^{(0,R)})^{\otimes 2}\pi^{(G)}\pi^{(R)}. \end{aligned}$$

Slope:

$$\begin{aligned} b &= var(\tilde{x}_i)^{-1} cov(\tilde{x}_i, \tilde{y}_i) \\ &\propto (\Sigma + a_{\text{LD}}\Delta_0^{\otimes 2})^{-1}\Delta \\ &= \Sigma^{-1}\Delta - \Sigma^{-1}\Delta_0 \cdot \frac{a_{\text{LD}}(\Delta_0)^T\Sigma^{-1}\Delta}{1 + a_{\text{LD}}(\Delta_0)^T\Sigma^{-1}\Delta_0} \\ &\propto \Sigma^{-1}\tilde{\Delta} + c_{\text{LD}}\Sigma^{-1}\Delta, \end{aligned}$$

where $a_{\text{LD}} = \frac{\pi^{(R)}\pi^{(G)}}{2\mathbb{E}[\lambda_i^2]}$ and

$$c_{\text{LL}} = \frac{1 + a_{\text{LD}}\Delta_0^T\Sigma^{-1}\Delta_0 - a_{\text{LD}}\Delta^T\Sigma^{-1}\Delta_0}{a_{\text{LD}}\Delta^T\Sigma^{-1}\Delta_0}.$$

Moreover, $b_0 = \mathbb{E}[\tilde{y}_i] - \mathbb{E}[\tilde{x}_i]^T b = \frac{1}{2} - \mathbb{E}[\tilde{x}_i]^T b$. Notice that

$$\begin{aligned} \mathbb{E}[\tilde{x}_i] &= \frac{1}{4}(\mu^{(0,G)} + \mu^{(1,G)} + \mu^{(0,R)} + \mu^{(1,R)}) \\ &= \frac{1}{4}(2\mu^{(0,G)} + \Delta + 2\mu^{(1,R)} - \Delta) \\ &= \frac{1}{2}(\mu^{(0,G)} + \mu^{(1,R)}) = \mathbb{E}[x_i]. \end{aligned}$$

Method comparison. We only need to compare $E_{\text{ERM}}^{(wst)}, E_{LL}^{(wst)}, E_{LD}^{(wst)}$.

For the ERM, $0 \le a_0 \le 2\alpha$ and

$$\begin{aligned} b_{\text{ERM}} &= (1 + \frac{a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}{1 + a_0\Delta_0^T\Sigma^{-1}\Delta_0})\Sigma^{-1}\widetilde{\Delta} - \frac{a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}{1 + a_0\Delta_0^T\Sigma^{-1}\Delta_0}\Sigma^{-1}\Delta \\ &\propto \Sigma^{-1}\widetilde{\Delta} - \frac{a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}{1 + a_0\Delta_0^T\Sigma^{-1}\Delta_0 + a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}\Sigma^{-1}\Delta \\ &\propto \Sigma^{-1}\widetilde{\Delta} - \frac{a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}{1 + a_0\Delta^T\Sigma^{-1}\Delta_0}\Sigma^{-1}\Delta. \end{aligned}$$

Let $c_0 = \frac{a_0\widetilde{\Delta}^T\Sigma^{-1}\Delta_0}{1 + a_0\Delta^T\Sigma^{-1}\Delta_0}$ and $c_1 = |c_0|\|\Delta\|_\Sigma / \|\widetilde{\Delta}\|_\Sigma$. For simplicity, let $\|v\|_\Sigma = v^T\Sigma^{-1}v$. We first lower bound it via

$$\begin{aligned} cor(b_{\text{ERM}}, \widetilde{\Delta}) = \frac{b^T\widetilde{\Delta}}{\|\widetilde{\Delta}\|_\Sigma\sqrt{b^T\Sigma b}} &= \frac{\widetilde{\Delta}^T\Sigma^{-1}\widetilde{\Delta} - c_0\Delta^T\Sigma^{-1}\widetilde{\Delta}}{\|\widetilde{\Delta}\|_\Sigma\sqrt{b^T\Sigma b}} \\ &\ge \frac{\widetilde{\Delta}^T\Sigma^{-1}\widetilde{\Delta}}{\|\widetilde{\Delta}\|_\Sigma(\|\widetilde{\Delta}\|_\Sigma + |c_0|\|\Delta\|_\Sigma)} - \frac{|c_0\Delta^T\Sigma^{-1}\widetilde{\Delta}|}{\|\widetilde{\Delta}\|_\Sigma\sqrt{b^T\Sigma b}} \\ &\ge \frac{1}{1 + |c_0|\|\Delta\|_\Sigma/\|\widetilde{\Delta}\|_\Sigma} - \frac{c_0\xi\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma - c_0\|\Delta\|_\Sigma} \\ &\ge \frac{1 - (1+\xi)c_1 - c_1^2}{1 - c_1^2} = 1 - C\alpha. \end{aligned}$$

Similarly, we have

$$\begin{aligned} cor(b_{\text{ERM}}, \Delta) = \frac{b^T\Delta}{\|\Delta\|_\Sigma\sqrt{b^T\Sigma b}} &= \frac{\Delta^T\Sigma^{-1}\widetilde{\Delta} - c_0\Delta^T\Sigma^{-1}\Delta}{\|\Delta\|_\Sigma\sqrt{b^T\Sigma b}} \\ &\le \frac{\widetilde{\Delta}^T\Sigma^{-1}\Delta}{\|\Delta\|_\Sigma(\|\widetilde{\Delta}\|_\Sigma \pm c_0\|\Delta\|_\Sigma)} + \frac{|c_0\Delta^T\Sigma^{-1}\Delta|}{(\|\widetilde{\Delta}\|_\Sigma - c_0\|\Delta\|_\Sigma)\|\Delta\|_\Sigma} \\ &\le \frac{1}{1 \pm c_0\|\Delta\|_\Sigma/\|\widetilde{\Delta}\|_\Sigma}\xi + \frac{c_0\|\Delta\|_\Sigma/\|\widetilde{\Delta}\|_\Sigma}{1 - c_0\|\Delta\|_\Sigma/\|\widetilde{\Delta}\|_\Sigma} \\ &\le (\frac{\xi}{1 \pm c_1} - \frac{c_1}{1 - c_1})\|\Delta\|_\Sigma. \end{aligned}$$

Hence,

$$E_{\text{ERM}}^{(wst)} \ge \max\left\{\Phi((\frac{1}{2} - C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi + C\alpha)\|\Delta\|_\Sigma), \Phi((-\frac{1}{2} - C\alpha)\|\widetilde{\Delta}\|_\Sigma)\right\} \tag{9}$$

for some constant C depending on the true parameters.

For method LISA-L, using the fact that $\Delta_0 = \Delta - \widetilde{\Delta}$, for $a_{\text{LL}} = var(\lambda_i)/(2\mathbb{E}[\lambda_i^2])$,

$$\begin{aligned} b_{\text{LL}} &\propto \Sigma^{-1}\Delta + \frac{-a_{\text{LL}}\Delta^T\Sigma^{-1}\Delta_0}{1 + a_{\text{LL}}\Delta_0^T\Sigma^{-1}\Delta_0}\Sigma^{-1}\widetilde{\Delta} \\ &\propto \Sigma^{-1}\widetilde{\Delta} + c_{\text{LL}}\Sigma^{-1}\Delta \end{aligned}$$

for

$$c_{\text{LL}} = \frac{1 + a_{\text{LL}}\Delta_0^T\Sigma^{-1}\Delta_0 - a_{\text{LL}}\Delta^T\Sigma^{-1}\Delta_0}{a_{\text{LL}}\Delta^T\Sigma^{-1}\Delta_0} = \frac{1 - a_{\text{LL}}\tilde{\Delta}^T\Sigma^{-1}\Delta_0}{a_{\text{LL}}\Delta^T\Sigma^{-1}\Delta_0}.$$

Hence,

$$\begin{aligned} cor(b_{\text{LL}}, \widetilde{\Delta}) &= \frac{\widetilde{\Delta}^T b_{\text{LL}}}{\|\widetilde{\Delta}\|_\Sigma\sqrt{b_{\text{LL}}^T\Sigma b_{\text{LL}}}} = \frac{\|\widetilde{\Delta}\|_\Sigma + c_{\text{LL}}\xi\|\Delta\|_\Sigma}{\|\widetilde{\Delta} + c_{\text{LL}}\Delta\|_\Sigma} \\ cor(b_{\text{LL}}, \Delta) &= \frac{b_{\text{LL}}^T\Delta}{\|\Delta\|_\Sigma\sqrt{b_{\text{LL}}^T\Sigma b_{\text{LL}}}} = \frac{\xi\|\widetilde{\Delta}\|_\Sigma + c_{\text{LL}}\|\Delta\|_\Sigma}{\|\widetilde{\Delta} + c_{\text{LL}}\Delta\|_\Sigma}. \end{aligned}$$

To have $E_{\text{LL}}^{(wst)} < E_{\text{ERM}}^{(wst)}$, it suffices to require that $(-\frac{1}{2} - C\alpha)\|\widetilde{\Delta}\|_\Sigma < (\frac{1}{2} - C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi + C\alpha)\|\Delta\|_\Sigma$ and

$$\begin{aligned} &\frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta})\|\widetilde{\Delta}\|_\Sigma - cor(b_{\text{LL}}, \Delta)\|\Delta\|_\Sigma \le (\frac{1}{2} - C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi + C\alpha)\|\Delta\|_\Sigma \\ &-\frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta})\|\widetilde{\Delta}\|_\Sigma \le (\frac{1}{2} - C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi + C\alpha)\|\Delta\|_\Sigma. \end{aligned}$$

A sufficient condition is

$$\xi < (\frac{1}{2} + \frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta}))\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha,\ cor(b_{\text{LL}}, \Delta) \ge \xi + C\alpha,\ \ cor(b_{\text{LL}}, \widetilde{\Delta}) \le 1 - 2C\alpha.$$

We can find that a further sufficient condition is

$$\xi < \frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha, c_{\text{LL}} > 0, \xi \le \frac{\|\widetilde{\Delta} + c_{\text{LL}}\Delta\|_\Sigma - \|\widetilde{\Delta}\|_\Sigma}{c_{\text{LL}}\|\Delta\|_\Sigma} - \epsilon_1\alpha \tag{10}$$

$$\|\widetilde{\Delta} + c_{\text{LL}}\Delta\|_\Sigma \ge \|\widetilde{\Delta}\|_\Sigma,\ \xi \le \frac{c_{\text{LL}}\|\Delta\|_\Sigma}{\|\widetilde{\Delta} + c_{\text{LL}}\Delta\|_\Sigma - \|\widetilde{\Delta}\|_\Sigma} - \epsilon_1\alpha \tag{11}$$

$$\xi \le (\frac{1}{2} + \frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta}))\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha. \tag{12}$$

We first find sufficient conditions for the statements in (10) and (11). Parameterizing $t = c_{\text{LL}}\|\Delta\|_\Sigma/\|\widetilde{\Delta}\|_\Sigma$, we further simplify the condition in (10) and (11) as

$$\xi < \min\{\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, 1\} - C\alpha, t(t+2\xi) > 0$$
$$\xi \leq \frac{\sqrt{1+t^2+2t\xi}-1}{t} - \epsilon_1\alpha, \;\; \xi \leq \frac{1+\sqrt{1+t^2+2t\xi}}{t+2\xi} - \epsilon_1\alpha.$$

We only need to require

$$t \geq \max\{0, -2\xi\} \text{ and } \xi \leq \min\{\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, 1\} - C\alpha.$$

Some tedious calculation shows that $t \geq \max\{0, -2\xi\}$ can be guaranteed by

$$a_{\text{LL}} \leq \frac{1}{\|\tilde{\Delta}\|_\Sigma^2 + \|\tilde{\Delta}\|_\Sigma\|\Delta\|_\Sigma} \text{ and } \xi \leq \frac{\|\Delta\|_\Sigma}{\|\tilde{\Delta}\|_\Sigma}$$

It is left to consider the constraint in (12). Notice that it holds for any $\xi \leq 0$. When $\xi > 0$, we can see

$$cor(b_{\text{LL}}, \widetilde{\Delta}) = \frac{\|\widetilde{\Delta}\|_\Sigma + \xi c_{\text{LL}}\|\Delta\|_\Sigma}{\|\widetilde{\Delta} + c_{\text{LL}}\Delta\|_\Sigma} = \frac{1+t\xi}{\sqrt{1+t^2+2t\xi}}$$
$$\geq \frac{1+t\xi}{1+t} \geq \xi.$$

Hence, it suffices to guarantee that

$$(1 - \frac{1}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma})\xi < \frac{1}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha.$$

If $\|\widetilde{\Delta}\|_\Sigma/\|\Delta\|_\Sigma \geq 2$, then LHS is negative and it holds. If $1 \leq \|\widetilde{\Delta}\|_\Sigma/\|\Delta\|_\Sigma < 2$, then the inequality becomes $\xi < 1 - C\alpha$. If $\|\widetilde{\Delta}\|_\Sigma/\|\Delta\|_\Sigma < 1$, then the inequality becomes $\xi \leq \frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha$. Because we have required $\xi < \min\{\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, 1\} - C\alpha$ for some large enough C, the constraint (12) always holds. To summarize, $E_{\text{LL}} < E_{\text{ERM}}$ given that $\xi \leq \min\{\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, \frac{\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma}\} - C\alpha$ for some large enough C and $a_{\text{LL}} \leq 1/(\|\widetilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma\|\Delta\|_\Sigma)$.

For **method LISA-D**, we can similarly show that $E_{\text{LD}} \leq E_{\text{ERM}}$ given that $\xi < \min\{\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, \frac{\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma}\} - C\alpha$ for some large enough C and $a_{\text{LD}} \leq 1/(\|\widetilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma\|\Delta\|_\Sigma)$.

B.5. Finite sample analysis

The empirical loss can be written as

$$\mathbb{P}(\mathbb{1}((x_i^{G)})^T\hat{b} + \hat{b}_0 > \frac{1}{2}) \neq y_i^{(G)}) \tag{13}$$
$$= \frac{1}{2}\mathbb{P}((x_i^{G)})^T\hat{b} + \hat{b}_0 > \frac{1}{2}|y_i^{(G)} = 0) + \frac{1}{2}\mathbb{P}((x_i^{G)})^T\hat{b} + \hat{b}_0 < \frac{1}{2}|y_i^{(G)} = 1),$$

where

$$\mathbb{P}((x_i^{G)})^T\hat{b} + \hat{b}_0 > \frac{1}{2}|y_i^{(G)} = 0) = \Phi(-\frac{\frac{1}{2} - (\mu^{(0,G)})^T\hat{b} - \hat{b}_0}{\sqrt{\hat{b}^T\Sigma\hat{b}}}).$$
$$\mathbb{P}((x_i^{G)})^T\hat{b} + \hat{b}_0 < \frac{1}{2}|y_i^{(G)} = 1) = \Phi(\frac{\frac{1}{2} - (\mu^{(1,G)})^T\hat{b} - \hat{b}_0}{\sqrt{\hat{b}^T\Sigma\hat{b}}}).$$

First notice that

$$\hat{b}_0 = \bar{y} - \bar{x}^T\hat{b}.$$

We have

$$\begin{aligned}(\mu^{(0,G)})^T\hat{b}+\hat{b}_0 &= (\mu^{(0,G)}-\bar{x})^T\hat{b}+\bar{y}\\ &= (\mu^{(0,G)}-\mathbb{E}[x_i])^T\hat{b}+\frac{1}{2}+\underbrace{\{(\bar{y}-\bar{x}^T\hat{b})-(\mathbb{E}[y_i]-\mathbb{E}[x_i]^T\hat{b})\}}_{R_1}\\ (\mu^{(1,G)})^T\hat{b}+\hat{b}_0 &= (\mu^{(1,G)}-\bar{x})^T\hat{b}+\bar{y}\\ &= \Delta^T\hat{b}+(\mu^{(0,G)}-\mathbb{E}[x_i])^T\hat{b}+\frac{1}{2}+R_1.\end{aligned}$$

Therefore, according to (13),

$$\begin{aligned}&\frac{1}{2}\Phi(-\frac{\frac{1}{2}-(\mu^{(0,G)})^T\hat{b}-\hat{b}_0}{\sqrt{\hat{b}^T\Sigma\hat{b}}})+\frac{1}{2}\Phi(\frac{\frac{1}{2}-(\mu^{(1,G)})^T\hat{b}-\hat{b}_0}{\sqrt{\hat{b}^T\Sigma\hat{b}}})\\ =&\frac{1}{2}\Phi(\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})+\frac{1}{2}\Phi(-\frac{\Delta+(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})\\ =&\frac{1}{2}\Phi(\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})+\frac{1}{2}\Phi(-\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})\\ &-\left\{\frac{1}{2}\Phi(-\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})-\frac{1}{2}\Phi(-\frac{\Delta+(\Theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})\right\}\\ =&\frac{1}{2}-\left\{\frac{1}{2}\Phi(-\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})-\frac{1}{2}\Phi(-\frac{\Delta^T\hat{b}+(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})\right\}.\end{aligned}$$

Then the mis-classification error can be written as

$$\frac{1}{2}-\frac{1}{2}\underbrace{\left\{\Phi(\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})-\Phi(\frac{(\theta^{(G)})^T\hat{b}-\Delta^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}})\right\}}_{\widehat{L}(\hat{b})}. \tag{14}$$

Larger the $\widehat{L}(\hat{b})$, smaller the mis-classification error.

We first find that

$$\widehat{L}(\hat{b})-L(b)\le C\underbrace{|\frac{(\theta^{(G)})^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}}-\frac{(\theta^{(G)})^Tb}{\sqrt{b^T\Sigma b}}|}_{T_1}+C\underbrace{|\frac{(\theta^{(G)})^T\hat{b}-\Delta^T\hat{b}+R_1}{\sqrt{\hat{b}^T\Sigma\hat{b}}}-\frac{(\theta^{(G)})^Tb-\Delta^Tb}{\sqrt{b^T\Sigma b}}|}_{T_2}.$$

In the event that

$$\|\Sigma^{1/2}(\hat{b}-b)\|_2=o(1)\ \max_{y,d}\|\mu^{(y,d)}\|_2\le C,\ \Sigma \text{ is positive definite.}$$

for the denominator, we have

$$\begin{aligned}|b^T\Sigma b-\hat{b}^T\Sigma\hat{b}| &\le (2\|\Sigma^{1/2}b\|_2+\|\Sigma^{1/2}(\hat{b}-b)\|_2)\|\Sigma^{1/2}(\hat{b}-b)\|_2\\ &\le 2(1+o(1))\|\Sigma^{1/2}b\|_2\|\Sigma^{1/2}(\hat{b}-b)\|_2\\ |\sqrt{\hat{b}^T\Sigma\hat{b}}-\sqrt{b^T\Sigma b}| &\le \frac{|\hat{b}^T\Sigma\hat{b}-b^T\Sigma b|}{\sqrt{\hat{b}^T\Sigma\hat{b}}+\sqrt{b^T\Sigma b}}\\ &\le 2(1+o(1))\|\Sigma^{1/2}(\hat{b}-b)\|_2.\end{aligned}$$

For the numerator, we have

$$|\frac{1}{2}\widetilde{\Delta}^T\hat{b}+R_1-\frac{1}{2}\widetilde{\Delta}^Tb|\le|R_1|+\frac{1}{2}\|\Sigma^{-1/2}\widetilde{\Delta}\|_2\|\Sigma^{1/2}(\hat{b}-b)\|_2.$$

We arrive at

$$T_1 \leq (1+o(1))\frac{|R_1| + \frac{1}{2}\|\Sigma^{-1/2}\widetilde{\Delta}\|_2\|\Sigma^{1/2}(\hat{b}-b)\|_2}{\|\Sigma^{1/2}b\|_2} + (1+o(1))\frac{|\widetilde{\Delta}^T b|}{\sqrt{b^T\Sigma b}}\frac{\|\Sigma^{1/2}(\hat{b}-b)\|_2}{\sqrt{b^T\Sigma b}}.$$

$$\begin{aligned} T_2 \leq & (1+o(1))\frac{|R_1| + \frac{1}{2}(\|\Sigma^{-1/2}\widetilde{\Delta}\|_2 + \|\Sigma^{-1/2}\Delta\|_2)\|\Sigma^{1/2}(\hat{b}-b)\|_2}{\|\Sigma^{1/2}b\|_2} \\ & + (1+o(1))\frac{|\frac{1}{2}\widetilde{\Delta}^T b - \Delta^T b|}{\sqrt{b^T\Sigma b}}\frac{\|\hat{b}-b\|_2}{\sqrt{b^T\Sigma b}}. \end{aligned}$$

Moreover $R_1 \leq \|\hat{b}-b\|_2 + O_P(\frac{1}{\sqrt{n}})$. To summarize,

$$|\widehat{L}(\hat{b}) - L(b)| \lesssim (1+o(1))(\|\hat{b}-b\|_2 + \frac{1}{\sqrt{n}}).$$

In the following, we will upper bound $\|\hat{b}-b\|_2$ for each method. For the **ERM method**,

$$\hat{b} = \{(X-\bar{X})^T(X-\bar{X})\}^{-1}(X-\bar{X})^T(y-\bar{y}).$$

It is easy to show that

$$\|\hat{b}-b\|_2^2 = O_P(\frac{p\sum_{i=1}^N var(y_i|x_i)}{N^2}) = O_P(\frac{p}{N}).$$

For the **vanilla mixup method**, we first see that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \tilde{x}_i &= \frac{1}{n}\sum_{i=1}^n(\lambda_i x_{i_1} + (1-\lambda_i)x_{i_2}) = \bar{x} + O_P(n^{-1/2}) = \mu + O_P(n^{-1/2}) \\ \frac{1}{n}\sum_{i=1}^n \tilde{y}_i &= \pi^{(1)} + O_P(n^{-1/2}). \end{aligned}$$

Next,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{y}_i &= \frac{1}{n}\sum_{i=1}^n \left\{\lambda_i^2 x_{i_1}y_{i_1} + (1-\lambda_i)^2 x_{i_2}y_{i_2} + \lambda_i(1-\lambda_i)x_{i_1}y_{i_2} + \lambda_i(1-\lambda_i)x_{i_2}y_{i_i}\right\} \\ \frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{y}_i - \mathbb{E}[\tilde{x}_i\tilde{y}_i] &= \underbrace{\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{y}_i - \mathbb{E}[\tilde{x}_i\tilde{y}_i|X,y]}_{E_1} + \underbrace{\mathbb{E}[\tilde{x}_i\tilde{y}_i|X,y] - \mathbb{E}[\tilde{x}_i\tilde{y}_i]}_{E_2}. \end{aligned}$$

For E_2,

$$E_2 = \frac{2\mathbb{E}[\lambda_i^2]}{n}\sum_{i=1}^n x_i y_i - \mathbb{E}[\tilde{x}_i\tilde{y}_i] = 2\mathbb{E}[\lambda_i^2]\mathbb{E}[x_i y_i].$$

Hence,

$$\|E_2\|_2^2 = O_P(\frac{p}{n}).$$

For E_1, conditioning on (X,y), $\lambda_i^2 x_{i_1}y_{i_1} - \frac{\mathbb{E}[\lambda_i^2]}{n}\sum_{i=1}^n x_i y_i$ are independent sub-Gaussian vectors. The sub-Gaussian norm of $\frac{1}{N}\sum_{i=1}^n \lambda_i^2 x_{i_1,j}y_{i_1} - \frac{\mathbb{E}[\lambda_i^2]}{n}\sum_{i=1}^n x_{i,j}y_i$ (conditioning on (X,y)) can be upper bounded by $c\max_{i\leq N}|x_{i,j}|/\sqrt{n}$. Hence

$$\mathbb{P}(\|E_1\|_2 \geq t|X,y) \leq 2\exp\{-\frac{c_2 n t^2}{\max_{j=1}^p \max_{i\leq N} x_{i,j}^2}\}.$$

As $x_{i,j}$ are Gaussian distributed, we know that

$$\mathbb{P}(\sum_{j=1}^{p} \max_{i\leq n} x_{i,j}^2 \geq p\log n) \leq \exp\{-c_3 \log n\}.$$

Hence, with probability at least $1-\exp(-c_1\log n)$,

$$E_1 \leq \frac{Cp\log n}{n}.$$

To summarize,

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\tilde{y}_i - (\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i)(\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i) - cov(\tilde{x}_i,\tilde{y}_i) \right\|_2^2 = O_P(\frac{p\log n}{n}).$$

Similarly, we can show that

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\tilde{x}_i^T - (\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i)(\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i)^T - cov(\tilde{x}_i) \right\|_2^2 = O_P(\frac{p\log n}{n}).$$

Hence,

$$\|\hat{b}-b\|_2^2 = O_P(\frac{p\log n}{n}).$$

For the **LISA-L**, we first see that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)} &= \frac{1}{n}\sum_{y_i=1}(\lambda_i x_{i_1}^{(1,G)} + (1-\lambda_i)x_{i_2}^{(1,R)}) + \frac{1}{n}\sum_{y_i=0}(\lambda_i x_{i_1}^{(0,G)} + (1-\lambda_i)x_{i_2}^{(0,R)}) \\ &= \frac{1}{2}(\bar{x}^{(1,G)} + \bar{x}^{(1,R)})\hat{\pi}_1 + \frac{1}{2}(\bar{x}^{(0,G)} + \bar{x}^{(0,R)})\hat{\pi}_0\end{aligned}$$

We have

$$\begin{aligned}\frac{1}{n}(X^{(\lambda)})^T y - \bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)} - cov(x_i^{(\lambda)}, y_i) = &\underbrace{\frac{1}{n}(X^{(\lambda)})^T y - \bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)} - cov(x_i^{(\lambda)}, y_i|X,y)}_{E_1} \\ &+ \underbrace{cov(x_i^{(\lambda)}, y_i|X,y) - cov(x_i^{(\lambda)}, y_i)}_{E_2}\end{aligned}$$

For E_2,

$$\begin{aligned}E_2 &= \frac{\hat{\pi}_1}{2}(\bar{x}^{(1,G)} + \bar{x}^{(1,R)}) - \hat{\pi}_1(\frac{1}{2}(\bar{x}^{(1,G)} + \bar{x}^{(1,R)})\hat{\pi}_1 + \frac{1}{2}(\bar{x}^{(0,G)} + \bar{x}^{(0,R)})\hat{\pi}_0) - cov(x_i^{(\lambda)}, y_i) \\ &= \frac{1}{2}(\bar{x}^{(1,G)} + \bar{x}^{(1,R)} - \bar{x}^{(0,G)} - \bar{x}^{(0,R)})\hat{\pi}_1\hat{\pi}_0 - \pi^{(1)}\pi^{(0)}\Delta.\end{aligned}$$

It is easy to show that

$$\|E_2\|_2^2 = O_P\left(\frac{p}{\min_{y,e} n^{(y,e)}}\right).$$

For E_1, conditioning on X and y, $x_i^{(\lambda)}y_i - \mathbb{E}[x_i^{(\lambda)}y_i|X,y]$ are independent sub-Gaussian vectors with mean zero. The sub-Gaussian norm of $\frac{1}{n}\sum_{i=1}^{n} x_{i,j}^{(\lambda)}y_i$ (conditioning on X and y) can be upper bounded by $c\max_{i\leq n}|x_{i,j}|/\sqrt{N}$.

$$\begin{aligned}&\mathbb{P}(\|E_1\|_2 \geq t|X,y) = \mathbb{P}\left(\sum_{j=1}^{p} |\frac{1}{n}\sum_{i=1}^{n}\{x_{i,j}^{(\lambda)}y_i - \mathbb{E}[x_{i,j}^{(\lambda)}y_i|X,y]\}|^2 \geq t^2|X,y\right) \\ &\leq 2\exp\left\{-\frac{c_2 n t^2}{\sum_{j=1}^{p}\max_{i\leq n} x_{i,j}^2}\right\}.\end{aligned}$$

Hence,

$$E_1 = O_P(\sqrt{\frac{\sum_{j=1}^{p} \max_{i\leq n} x_{i,j}^2}{n}}) = O_P(\frac{p\log n}{n}).$$

To summarize,

$$\|\frac{1}{n}(X^{(\lambda)})^T y - \mathbb{E}[x_i^{(\lambda)} y_i]\|_2^2 = O_P(\frac{p}{\min_{y,e} n^{(y,e)}} + \frac{p\log n}{n}).$$

We can use similar analysis to bound

$$\|\frac{1}{N}(X^{(\lambda)})^T X^{(\lambda)} - \mathbb{E}[x_i^{(\lambda)}(x_i^{(\lambda)})^T]\|_2.$$

The sub-exponential norm of $\frac{1}{N}\sum_{i=1}^{N} x_{i,j}^{(\lambda)} x_{i,k}^{(\lambda)}$ (conditioning on X) can be upper bounded by $\max_{i\leq N} |x_{i,j}||x_{i,k}|/\sqrt{N}$. We can show that

$$\|\frac{1}{n}(X^{(\lambda)})^T X^{(\lambda)} - \mathbb{E}[x_i^{(\lambda)}(x_i^{(\lambda)})^T]\|_2 = O_P(\frac{p}{\min_{y,e} n^{(y,e)}} + \frac{p\log n}{n}).$$

For the **LISA-D**, we first see that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i &= \frac{1}{n}\sum_{D_i=G}(\lambda_i x_{i_1}^{(1,G)} + (1-\lambda_i)x_{i_2}^{(0,G)}) + \frac{1}{n}\sum_{D_i=R}(\lambda_i x_{i_1}^{(1,R)} + (1-\lambda_i)x_{i_2}^{(0,R)}) \\
&= \frac{1}{2}(\bar{x}^{(1,G)} + \bar{x}^{(0,G)})\hat{\pi}_G + \frac{1}{2}(\bar{x}^{(1,R)} + \bar{x}^{(0,R)})\hat{\pi}_R \\
\bar{\tilde{y}} &= \frac{1}{2}.
\end{aligned}$$

Next,

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\tilde{y}_i = \frac{1}{n}\sum_{D_i=G}\left\{\lambda_i^2 x_{i_1}^{(1,G)} + \lambda_i(1-\lambda_i)x_{i_2}^{(0,G)}\right\} + \frac{1}{n}\sum_{D_i=R}\left\{\lambda_i^2 x_{i_1}^{(1,R)} + \lambda_i(1-\lambda_i)x_{i_2}^{(0,R)}\right\}$$

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\tilde{y}_i - \bar{\tilde{x}}\bar{\tilde{y}} - cov(\tilde{x},\tilde{y}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\tilde{y}_i - \bar{\tilde{x}}\bar{\tilde{y}} - cov(\tilde{x}_i,\tilde{y}_i|X,y)}_{E_1} + \underbrace{cov(\tilde{x}_i,\tilde{y}_i|X,y) - cov(\tilde{x}_i,\tilde{y}_i)}_{E_2}.$$

For E_2,

$$\begin{aligned}
E_2 &= \hat{\pi}^{(G)}(\mathbb{E}[\lambda_i^2](\bar{x}^{(1,G)} - \bar{x}^{(0,G)}) + \frac{1}{2}\bar{x}^{(0,G)}) + \hat{\pi}^{(R)}(\mathbb{E}[\lambda_i^2](\bar{x}^{(1,R)} - \bar{x}^{(0,R)}) + \frac{1}{2}\bar{x}^{(0,R)}) - \\
&\frac{1}{4}(\bar{x}^{(1,G)} + \bar{x}^{(0,G)})\hat{\pi}_G - \frac{1}{4}(\bar{x}^{(1,R)} + \bar{x}^{(0,R)})\hat{\pi}_R - var(\lambda_i)\Delta \\
&= \hat{\pi}^{(G)} var(\lambda_i)(\bar{x}^{(1,G)} - \bar{x}^{(0,G)}) + \hat{\pi}^{(R)} var(\lambda_i)(\bar{x}^{(1,R)} - \bar{x}^{(0,R)}) - var(\lambda_i)\Delta.
\end{aligned}$$

Notice that E_2 is a sub-Gaussian vector with sub-Gaussian norm upper bounded by

$$\frac{\hat{\pi}_G^2}{n^{(1,G)}} + \frac{\hat{\pi}_G^2}{n^{(0,G)}} + \frac{\hat{\pi}_R^2}{n^{(1,R)}} + \frac{\hat{\pi}_R^2}{n^{(0,R)}} \leq \frac{4}{n}\max_{y,d}\frac{\pi_d}{\pi_{y|d}}.$$

Using sub-Gaussian concentration, we can show that

$$E_2 = O_P(\sqrt{\frac{p}{n}\max_{y,d}\frac{\pi_d}{\pi_{y|d}}}).$$

Notice that $\max_{y,d} \frac{\pi_d}{\pi_{y|d}} \geq 1$. For E_1, conditioning on X and y $\tilde{x}_i\tilde{y}_i - \mathbb{E}[\tilde{x}_i\tilde{y}_i|X,y]$ are independent sub-Gaussian vectors with mean zero. The sub-Gaussian norm of $\frac{1}{n}\sum_{i=1}^n \tilde{x}_{i,j}\tilde{y}_i$ conditioning on X and y can be upper bounded by $c\max_{i,j}|x_{i,j}|$. Similar analysis on E_1 leads to

$$\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{y}_i - \bar{\tilde{x}}\bar{\tilde{y}} - cov(\tilde{x},\tilde{y}) = O_P(\sqrt{\frac{p\log n}{n}} + \sqrt{\frac{p}{n}\max_{y,d}\frac{\pi_d}{\pi_{y|d}}}).$$

For the sample covariance matrix, we can also show that

$$\left\|\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{x}_i^T - (\frac{1}{n}\sum_{i=1}^n \tilde{x}_i)(\frac{1}{n}\sum_{i=1}^n \tilde{x}_i)^T - cov(\tilde{x}_i)\right\|_2^2 = O_P(\sqrt{\frac{p\log n}{n}} + \sqrt{\frac{p}{n}\max_{y,d}\frac{\pi_d}{\pi_{y|d}}}).$$

B.6. A ξ-dependent lower bound for $E_{\text{ERM}}^{(wst)} - E_{\text{LL}}^{(wst)}$

Next, we provide a ξ-dependent lower bound for $E_{\text{ERM}}^{(wst)} - E_{\text{LL}}^{(wst)}$. Based on our previous analysis

$$\begin{aligned}E_{\text{ERM}}^{(wst)} - E_{\text{LL}}^{(wst)} \geq c_1 \min\Big\{&(\frac{1}{2} - C\alpha - \frac{1}{2}cor(b_{\text{LL}},\tilde{\Delta}))\|\tilde{\Delta}\|_\Sigma + (cor(b_{\text{LL}},\Delta) - \xi - C\alpha)\|\Delta\|_\Sigma,\\ &(\frac{1}{2} - C\alpha + \frac{1}{2}cor(b_{\text{LL}},\tilde{\Delta}))\|\tilde{\Delta}\|_\Sigma - (\xi + C\alpha)\|\Delta\|_\Sigma\Big\},\end{aligned}$$

where c_1 is a positive constant given by the derivative of $\Phi(\cdot)$. Plugging in the expression of $cor(b_{\text{LL}},\tilde{\Delta})$ and $cor(b_{\text{LL}},\Delta)$, we have for the first term of $E_{\text{ERM}}^{(wst)} - E_{\text{LL}}^{(wst)}$, it is no smaller than

$$\begin{aligned}&\frac{1}{2}(1 - 2C\alpha - \frac{1+\xi t}{\sqrt{1+t^2+2t\xi}})\|\tilde{\Delta}\|_\Sigma + (\frac{\xi+t}{\sqrt{1+t^2+2t\xi}} - \xi - C\alpha)\|\Delta\|_\Sigma\\ &\geq \frac{1}{2}\frac{t^2}{(1+t)^2}(1-\xi^2)\|\tilde{\Delta}\|_\Sigma + \frac{t^2}{(1+t)^2}(1-\xi)\|\Delta\|_\Sigma - C\alpha(\|\Delta\|_\Sigma + \|\tilde{\Delta}\|_\Sigma),\end{aligned}$$

where the last step is due to the current constraint that $t > \max\{0, -2\xi\}$. For the second term, it is no smaller than

$$\|\tilde{\Delta}\|_\Sigma - \xi\|\Delta\|_\Sigma - C\alpha(\|\tilde{\Delta}_\Sigma + \|\Delta\|_\Sigma).$$

Notice that $t^2/(1+t^2) \geq \min\{\frac{t^2}{4}, \frac{1}{4}\}$. We can show that $t \geq \|\tilde{\Delta}\|_\Sigma/\|\Delta\|_\Sigma$, then

$$E_{\text{ERM}}^{(wst)} - E_{\text{LL}}^{(wst)} \geq c_3\min\{(\frac{\|\tilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - \xi)\|\Delta\|_\Sigma, (1-\xi)\|\Delta\|_\Sigma, (1-\xi)\|\tilde{\Delta}\|_\Sigma^2/\|\Delta\|_\Sigma\} - c_4\alpha(\|\Delta\|_\Sigma + \|\tilde{\Delta}\|_\Sigma).$$

B.7. Domain shifts: Proof of Theorem 2

It still holds that $\widetilde{\Delta}^* = 2(\mu^{(0,*)} - \mathbb{E}[x_i^{(\lambda)}]) = 2(\mu^{(0,*)} - \mathbb{E}[\tilde{x}_i])$. It is easy to show that the worst group mis-classification error for this new environment is

$$E_A^{(wst,*)} = \max\left\{\Phi\left(-\frac{\frac{1}{2}(\widetilde{\Delta}^*)^T b_A}{\sqrt{b_A^T\Sigma b_A}}\right), \Phi\left(\frac{\frac{1}{2}(\widetilde{\Delta}^*)^T b_A - \Delta^T b_A}{\sqrt{b_A^T\Sigma b_A}}\right)\right\}, \tag{15}$$

where $A \in \{\text{ERM}, \text{mix}, \text{LL}, \text{LD}\}$. Notice that

$$\widetilde{\Delta}^* = 2\mu^{(0,*)} - (\mu^{(0,G)} + \mu^{(1,R)}) = \widetilde{\Delta} + \mu^{(0,*)} - \mu^{(0,G)}$$

We assume $\|\widetilde{\Delta}^*\|_2 = \|\widetilde{\Delta}\|_2$. Let $\xi^* = cor(\Delta, \widetilde{\Delta}^*)$ and $\gamma = cor(\widetilde{\Delta}, \widetilde{\Delta}^*)$. We have

$$\begin{aligned}cor(b_{\text{ERM}}, \widetilde{\Delta}^*) &= \frac{\gamma\|\widetilde{\Delta}\|_\Sigma\|\widetilde{\Delta}^*\|_\Sigma - c_0\xi^*\|\Delta\|_\Sigma\|\widetilde{\Delta}^*\|_\Sigma}{\|\widetilde{\Delta}^*\|_\Sigma\|\widetilde{\Delta} + c_0\Delta\|_\Sigma}\\ &= \frac{\gamma\|\widetilde{\Delta}\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma \pm \|c_0\Delta\|_\Sigma} \pm \frac{|c_0\xi^*|\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma \pm \|c_0\Delta\|_\Sigma} = \gamma \pm C\alpha.\end{aligned}$$

Hence,

$$E_{\text{ERM}}^{(wst)} \geq \max\left\{\Phi((\frac{\gamma}{2}-C\alpha)\|\widetilde{\Delta}\|_\Sigma-(\xi-C\alpha)\|\Delta\|_\Sigma), \Phi((-\frac{\gamma}{2}-C\alpha)\|\widetilde{\Delta}\|_\Sigma)\right\} \tag{16}$$

for some constant C depending on the true parameters.

Hence,

$$cor(b_{\text{LL}}, \widetilde{\Delta}^*) = \frac{(\widetilde{\Delta}^*)^T b_{\text{LL}}}{\|\widetilde{\Delta}^*\|_\Sigma \sqrt{b_{\text{LL}}^T \Sigma b_{\text{LL}}}} = \frac{\gamma\|\widetilde{\Delta}\|_\Sigma + c_{\text{LL}}\xi^*\|\Delta\|_\Sigma}{\|\widetilde{\Delta}+c_{\text{LL}}\Delta\|_\Sigma}.$$

To have $E_{\text{LL}}^{(wst*)} < E_{\text{ERM}}^{(wst*)}$, it suffices to require that $(-\frac{\gamma}{2}-C\alpha)\|\widetilde{\Delta}\|_\Sigma < (\frac{\gamma}{2}-C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi+C\alpha)\|\Delta\|_\Sigma$ and

$$\begin{aligned}&\frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta}^*)\|\widetilde{\Delta}\|_\Sigma - cor(b_{\text{LL}}, \Delta)\|\Delta\|_\Sigma \leq (\frac{\gamma}{2}-C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi+C\alpha)\|\Delta\|_\Sigma\\ &-\frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta}^*)\|\widetilde{\Delta}\|_\Sigma \leq (\frac{\gamma}{2}-C\alpha)\|\widetilde{\Delta}\|_\Sigma - (\xi+C\alpha)\|\Delta\|_\Sigma.\end{aligned}$$

A sufficient condition is

$$\xi < (\frac{\gamma}{2}+\frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta}^*))\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha,\ cor(b_{\text{LL}}, \Delta) \geq \xi + C\alpha,\ \ cor(b_{\text{LL}}, \widetilde{\Delta}^*) \leq \gamma - 2C\alpha.$$

We can find that a further sufficient condition is

$$\xi < \frac{1+\gamma}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha, c_{\text{LL}} > 0, \xi^* \leq \frac{\gamma(\|\widetilde{\Delta}+c_{\text{LL}}\Delta\|_\Sigma - \|\widetilde{\Delta}\|_\Sigma)}{c_{\text{LL}}\|\Delta\|_\Sigma} - \epsilon_1\alpha \tag{17}$$

$$\|\widetilde{\Delta}+c_{\text{LL}}\Delta\|_\Sigma \geq \|\widetilde{\Delta}\|_\Sigma,\ \xi \leq \frac{c_{\text{LL}}\|\Delta\|_\Sigma}{\|\widetilde{\Delta}+c_{\text{LL}}\Delta\|_\Sigma - \|\widetilde{\Delta}\|_\Sigma} - \epsilon_1\alpha \tag{18}$$

$$\xi \leq (\frac{\gamma}{2}+\frac{1}{2}cor(b_{\text{LL}}, \widetilde{\Delta}^*))\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha. \tag{19}$$

We first find sufficient conditions for the statements in (10) and (11). Parameterizing $t = c_{\text{LL}}\|\Delta\|_\Sigma/\|\widetilde{\Delta}\|_\Sigma$, we further simplify the condition in (17) and (18) as

$$\begin{aligned}&\xi < \frac{1+\gamma}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma} - C\alpha, t > 0,\ \xi^* \leq \frac{\gamma(\sqrt{1+t^2+2t\xi}-1)}{t} - \epsilon_1\alpha,\\ &-\frac{t}{2} \leq \xi \leq t,\ \ \xi \leq \frac{1+\sqrt{1+t^2+2t\xi}}{t+2\xi} - \epsilon_1\alpha.\end{aligned}$$

We only need to require

$$t \geq \max\{0, -2\xi\} \text{ and } \xi < \min\{\frac{1+\gamma}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, 1\} - C\alpha, \xi^* \leq \gamma\xi.$$

Some tedious calculation shows that $t \geq \max\{0, -2\xi\}$ can be guaranteed by

$$a_{\text{LL}} \leq \frac{1}{\|\widetilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma\|\Delta\|_\Sigma} \text{ and } \xi \leq \frac{\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma}$$

It is left to consider the constraint in (19). Notice that it holds for any $\xi \leq 0$. When $\xi > 0$, we can see

$$\begin{aligned}cor(b_{\text{LL}}, \widetilde{\Delta}^*) &= \frac{\gamma\|\widetilde{\Delta}\|_\Sigma + \xi^* c_{\text{LL}}\|\Delta\|_\Sigma}{\|\widetilde{\Delta}+c_{\text{LL}}\Delta\|_\Sigma} = \frac{\gamma+t\xi^*}{\sqrt{1+t^2+2t\xi}}\\ &\geq \frac{\gamma+t\xi^*}{1+t}.\end{aligned}$$

Hence, it suffices to guarantee that

$$\xi^* + \gamma \geq \frac{2\|\Delta\|_\Sigma}{\|\widetilde{\Delta}\|_\Sigma}\xi + C\alpha.$$

To summarize, it suffices to require

$$a_{\text{LL}} \leq \frac{1}{\|\tilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma\|\Delta\|_\Sigma}, 0 \leq \xi^* \leq \gamma\xi, \xi < \min\{\frac{\gamma}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, \frac{\|\Delta\|_\Sigma}{\|\tilde{\Delta}\|_\Sigma}\} - C\alpha.$$

For LISA-D, we can similarly show that $E_{\text{LD}}^{(wst*)} < E_{\text{ERM}}^{(wst*)}$ given that

$$a_{\text{LD}} \leq \frac{1}{\|\tilde{\Delta}\|_\Sigma^2 + \|\widetilde{\Delta}\|_\Sigma\|\Delta\|_\Sigma}, 0 \leq \xi^* \leq \gamma\xi, \xi < \min\{\frac{\gamma}{2}\frac{\|\widetilde{\Delta}\|_\Sigma}{\|\Delta\|_\Sigma}, \frac{\|\Delta\|_\Sigma}{\|\tilde{\Delta}\|_\Sigma}\} - C\alpha.$$