How Does Distribution Matching Help Domain Generalization: An Information-theoretic Analysis

Yuxin Dong, Tieliang Gong, Hong Chen, Shuangyong Song, Weizhan Zhang, *Senior Member, IEEE*, Chen Li

***Abstract*—Domain generalization aims to learn invariance across multiple source domains, thereby enhancing generalization against out-of-distribution data. While gradient or representation matching algorithms have achieved remarkable success in domain generalization, these methods generally lack generalization guarantees or depend on strong assumptions, leaving a gap in understanding the underlying mechanism of distribution matching. In this work, we formulate domain generalization from a novel probabilistic perspective, ensuring robustness while avoiding overly conservative solutions. Through comprehensive information-theoretic analysis, we provide key insights into the roles of gradient and representation matching in promoting generalization. Our results reveal the complementary relationship between these two components, indicating that existing works focusing solely on either gradient or representation alignment are insufficient to solve the domain generalization problem. In light of these theoretical findings, we introduce IDM to simultaneously align the inter-domain gradients and representations. Integrated with the proposed PDM method for complex distribution matching, IDM achieves superior performance over various baseline methods.**

***Index Terms*—Information Theory, Domain Generalization, Generalization Analysis, Distribution Matching.**

I. Introduction

DISTRIBUTION shifts are prevalent in various real-world learning contexts, often leading to machine learning systems overfitting domain-specific correlations that may negatively impact performance when facing out-of-distribution (OOD) data [1]–[4]. Domain generalization (DG) is then introduced to address this challenge: By assuming the training data constitutes multiple domains that share some invariant underlying correlations, DG algorithms then attempt to learn this invariance so that domain-specific variations do not affect the model's performance. To this end, various DG approaches have been proposed, including invariant representation learning [5], [6], adversarial learning [7], [8], causal inference [9], [10], gradient manipulation [11]–[13], and robust optimization [14]–[16].

DG is typically formulated as an average-case [17], [18] or worst-case [9], [14] optimization problem, which however either lacks robustness against OOD data [9], [19] or leads to overly conservative solutions [16]. In this paper, we introduce a novel probabilistic formulation that ensures robustness by minimizing the gap between source and target-domain population risks with high probability. Our comprehensive generalization analysis then reveals that the input-output mutual information and the representation space covariate shift are pivotal in controlling this domain-level generalization gap, which could be achieved by aligning inter-domain gradients and representations, respectively.

Although distribution matching techniques are already well-explored in existing DG literature, these methods generally lack generalization guarantees or depend on strong assumptions, e.g. controllable invariant features [11], quadratic bowl loss landscape [13] or Lipschitz-continuous gradients [20]. In contrast, we derive instructive generalization bounds by leveraging a relaxed i.i.d domain assumption [16], which is easily satisfied in practice. Our results indicate that combining gradient and representation matching effectively minimizes the domain-level generalization gap. Crucially, we reveal the complementary nature of these two components, highlighting that neither of them alone is sufficient to solve the DG problem.

In light of these theoretical findings, we propose inter-domain distribution matching (IDM) for high-probability DG by simultaneously aligning gradients and representations across source domains. Furthermore, we point out the limitations of traditional distribution alignment techniques, especially for high-dimensional and complex probability distributions. To circumvent these issues, we further propose per-sample distribution matching (PDM) by slicing and aligning individual sorted data points. The main framework of theoretical results in this paper is summarized in Figure 1. IDM jointly working with PDM achieves superior performance on the Colored MNIST dataset [9] and the DomainBed benchmark [21]. Our primary contributions can be summarized as follows:

- **Probabilistic formulation**: We introduce a novel probabilistic perspective for evaluating DG algorithms, focusing on their ability to minimize the domain-level generalization gap with high probability. Our approach leverages milder assumptions about the domains and enables generalization analysis with information-theoretic tools.
- **Information-theoretic insights**: Our analysis comprehensively elucidates the role of gradient and representation matching in promoting domain generalization. Most importantly, we reveal the complementary relationship between these two components, indicating that neither of them alone is sufficient to solve the DG problem.

This work was supported in part by the National Natural Science Foundation of China under Grant 62172326. *(Corresponding author: Tieliang Gong.)*

Yuxin Dong, Tieliang Gong, Weizhan Zhang, and Chen Li are with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: yxdong9805@gmail.com; adidas-gtl@gmail.com; zhangwzh@xjtu.edu.cn; cli@xjtu.edu.cn).

Hong Chen is with the College of Science, Huazhong Agriculture University, Wuhan 430070, China (e-mail: chenh@mail.hzau.edu.cn).

Shuangyong is with the China Telecom Corporation, Beijing 100033, China (e-mail: songshy@chinatelecom.cn).

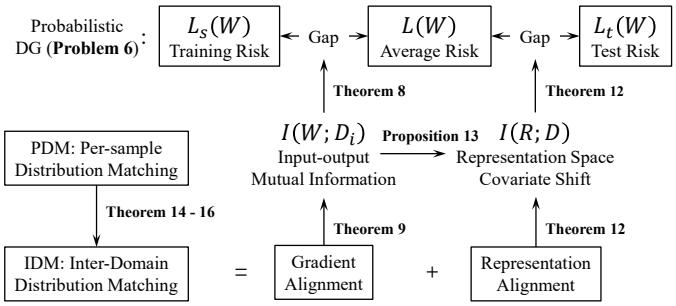

Fig. 1. The main framework of our theoretical results. We reveal that the input-output mutual information and the representation space covariate shift are pivotal in controlling the domain generalization error, and can be minimized by aligning inter-domain gradients and representations respectively.

- **Novel algorithms**: We propose IDM for high-probability DG by simultaneously aligning inter-domain gradients and representations, and PDM for complex distribution matching by slicing and aligning individual sorted data points. IDM jointly working with PDM achieves superior performance over various baseline methods.

II. Problem Setting

We denote random variables by capitalized letters (X), their realizations by lower-case letters (x), and the corresponding spaces by calligraphic letters ($\mathcal{X}$). Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be the instance space of interest, where $\mathcal{X}$ and $\mathcal{Y}$ are the input space and the label space, respectively. Let $\mathcal{W}$ be the hypothesis space, each $w \in \mathcal{W}$ characterizes a predictor f_w: $\mathcal{X} \mapsto \mathcal{Y}$, comprised of an encoder f_ϕ: $\mathcal{X} \mapsto \mathcal{R}$ and a classifier f_ψ: $\mathcal{R} \mapsto \mathcal{Y}$ with the assist of the representation space $\mathcal{R}$.

Following [16], we assume that there exists a distribution ν over the space of possible domains $\mathcal{D}$, where each domain $d \in \mathcal{D}$ corresponds to a specific data-generating distribution $\mu_d = P_{Z|D=d}$. The unconditional data distribution is $\mu = P_Z$. The source $D_s = \{D_i\}_{i=1}^m$ and target $D_t = \{D'_k\}_{k=1}^{m'}$ domains are both random variables sampled from ν. Let $S = \{S_i\}_{i=1}^m$ denote the training dataset, with each subset $S_i = \{Z_j^i\}_{j=1}^n$ containing n i.i.d data sampled from μ_{D_i}. The task is to design algorithm $\mathcal{A} : \mathcal{D}^m \mapsto \mathcal{W}$, taking D_s as the input (with proxy S) and providing possibly randomized hypothesis $W = \mathcal{A}(D_s)$. Given the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$, the general performance of some hypothesis $w \in \mathcal{W}$ in average is evaluated by the global population risk:

$$L(w) = \mathbb{E}_{D\sim\nu}[L_D(w)] = \mathbb{E}_{Z\sim\mu}[\ell(f_w(X), Y)],$$

where $L_d(w) = \mathbb{E}_{Z\sim\mu_d}[\ell(f_w(X), Y)]$ is the domain-level population risk. Since ν is unknown, only the source and target-domain population risks are tractable in practice:

$$L_s(w) = \frac{1}{m}\sum_{i=1}^{m} L_{D_i}(w), \quad L_t(w) = \frac{1}{m'}\sum_{k=1}^{m'} L_{D'_k}(w).$$

Main Assumptions. We list the assumptions considered in our theoretical analysis as follows:

Assumption 1. D_t is independent of D_s.

Assumption 2. $\ell(\cdot,\cdot)$ is bounded in $[0, M]$.

Assumption 3. $\ell(f_w(X), Y)$ is σ-subGaussian w.r.t $Z \sim \mu$ for any $w \in \mathcal{W}$.

Assumption 4. $\ell(\cdot,\cdot)$ is symmetric and satisfies the triangle inequality, i.e. for any $y_1, y_2, y_3 \in \mathcal{Y}$, $\ell(y_1, y_2) = \ell(y_2, y_1)$ and $\ell(y_1, y_2) \le \ell(y_1, y_3) + \ell(y_3, y_2)$.

Assumption 5. $\ell(f_w(X), Y)$ is β-Lipschitz w.r.t a metric c on $\mathcal{Z}$ for any $w \in \mathcal{W}$, i.e. for any $z_1, z_2 \in \mathcal{Z}$, $|\ell(f_w(x_1), y_1) - \ell(f_w(x_2), y_2)| \le \beta c(z_1, z_2)$.

Sub-Gaussianity (Assumption 3) is one of the most common assumptions for information-theoretic generalization analysis [22]–[25], and is also naturally satisfied when using Gibbs algorithms [26]. Notably, Assumption 2 is a strengthened version of Assumption 3, since any $[0, M]$-bounded random variable is always $M/2$-subGaussian. Lipschitzness (Assumption 5) is a crucial prerequisite for stability analysis and has also been utilized in deriving Wasserstein distance generalization bounds [27]–[32]. Assumption 4 is fulfilled when distance functions, such as mean absolute error (MAE) and 0-1 loss, are used as loss functions. This assumption has also been examined in previous studies [33]–[35].

High-Probability DG. The classical empirical risk minimization (ERM) technique, which minimizes the average-case risk: $\min_w L(w)$, is found ineffective in achieving invariance across different domains [9], [19]. To overcome this limitation, recent works [12], [13], [15], [36]–[38] have cast DG as a worst-case optimization problem: $\min_w \max_d L_d(w)$. However, this approach is generally impractical without strong assumptions made in the literature [16], [39], e.g. linearity of the underlying causal mechanism [9], [15], [36], or strictly separable spurious and invariant features [38]. On the contrary, we propose the following high-probability objective by leveraging the mild Assumption 1:

Problem 6. (High-Probability DG)

$$\min_{\mathcal{A}} \mathbb{E}[L_s(W)], \quad \text{s.t.} \quad \mathbb{P}\{|L_t(W) - L_s(W)| \ge \epsilon\} \le \delta.$$

In domain generalization, we are interested in the target-domain performance of a model trained with source-domain data, so an appropriate formulation of the generalization ability would be the gap between source-domain $L_s(W)$ and target-domain $L_t(W)$ population risks. This gap also measures the robustness of the model against distribution shifts. The probability is taken over both the sampled domains (D_s and D_t) and the learning algorithm (W). Notably, our Assumption 1 is significantly weaker than the previously adopted i.i.d domain assumption [16] by allowing correlations between source (or target) domains, and should be trivially satisfied in practice.

Here, we mainly focus on the domain-level generalization error between $L_s(w)$ and $L_t(w)$, while in practice the standard generalization error between the source-domain empirical risk $L'(w)$ (defined in Theorem 22) and $L_s(w)$ may also arise due to the effect of finite training samples. This is also an important research direction that has been well-studied in the literature, with various works establishing generalization bounds based

on input-output or conditional mutual information metrics under the supersample framework [22], [40], [41]. However, it is beyond the scope of solving Problem 6 and thus is not considered in our main results. We further explore the effect of finite samples in Section VII-E.

III. Generalization Analysis

The primary goal of DG is to tackle the distribution shift problem, raised by the variation in the data-generating distribution μ_d for different domains d. When using the KL divergence as a measure of distance between the data distributions of distinct domains, this inconsistency can be quantified by the mutual information $I(Z;D)$ between the data pair Z and the domain identifier D. Specifically, $I(Z;D)=0$ if and only if Z and D are independent, i.e. the data distributions μ_d are all the same for any $d\in\mathcal{D}$. This metric can be further decomposed into:

$$I(Z;D)\text{ (distribution shift)} = I(X;D)\text{ (covariate shift)} + I(Y;D|X)\text{ (concept shift)}. \quad (1)$$

While D is binary to distinguish training and test samples in [42], [43], we extend this concept to any discrete or continuous space, provided that each $d\in\mathcal{D}$ corresponds to a distinct data distribution μ_d. The right hand side (RHS) characterizes the changes in the marginal input distribution P_X (covariate shift) as well as the predictive distribution $P_{Y|X}$ (concept shift). We first show that the achievable level of average-case risk $L(w)$ is constrained by the degree of concept shift as following:

Proposition 7. *For any predictor $Q_{Y|X}$, we have*

$$\mathbb{E}_{X,D}[\mathrm{KL}(P_{Y|X,D}\,\|\,Q_{Y|X})]\geq I(Y;D|X).$$

When ℓ represents the cross-entropy loss, the population risk of predictor Q on domain d can be represented as the KL divergence $\mathbb{E}_{X|D=d}[\mathrm{KL}(P_{Y|X,D=d}\,\|\,Q_{Y|X})]$, provided that $H(Y|X,D)=0$ (i.e. the label can be entirely inferred from X and D). Therefore, the LHS expectation could be understood as the global population risk $L(w)$. When $d_t=\mathcal{D}\setminus d_s$ and $m,m'<\infty$, this implies that any model fitting well in source domains ($L_s(w)\approx 0$) will suffer from strictly positive risks in target domains ($L_t(w)\geq\Omega(I(Y;D|X))$) once concept shift is induced, which violates the goal of DG. This observation verifies the failure of ERM on the Colored MNIST dataset [9] which introduces a high concept shift, emphasizing that any algorithm must balance training and test risks (i.e. minimize $|L_t(W)-L_s(W)|$ in Problem 6) to achieve domain generalization.

A. Decomposing the Generalization Gap

We further demonstrate that by connecting source and target-domain population risks via the average-case risk $L(W)$, one can decompose the constraint of Problem 6 into source and target-domain generalization gaps. To be specific, since the predictor W is trained on the source domains D_s, it is commonly seen that W achieves lower population risks on D_s than on average, i.e. $L_s(W)\leq L(W)$. Moreover, since the sampling process of target domains is independent of the hypothesis, the target-domain population risk $L_t(W)$ is an unbiased estimate of $L(W)$. Combining these two observations, it is natural to observe that $L_s(W)\leq L(W)\approx L_t(W)$, implying that the average-case risk $L(W)$ acts as a natural bridge between the two. For any constant $\lambda\in(0,1)$, we can prove that:

$$\mathbb{P}\{|L_s(W)-L_t(W)|\geq\epsilon\}\leq\mathbb{P}\{|L_s(W)-L(W)|\geq\lambda\epsilon\} + \mathbb{P}\{|L_t(W)-L(W)|\geq(1-\lambda)\epsilon\}.$$

While the first event heavily correlates with the hypothesis W, the second event is instead hypothesis-independent. This observation inspires us to explore both hypothesis-based and hypothesis-independent bounds to address source and target-domain generalization errors, respectively.

B. Source-domain Generalization

We first provide a sufficient condition for source-domain generalization. Our results are motivated by recent advancements in generalization analysis within the information-theoretic framework [44], [45]. Specialized to our problem, we quantify the changes in the hypothesis once the source domains are observed through the input-output mutual information $I(W;D_i)$:

Theorem 8. *If Assumption 2 holds, then*

$$\mathbb{P}\{|L_s(W)-L(W)|\geq\epsilon\}\leq\frac{M}{m\epsilon\sqrt{2}}\sum_{i=1}^{m}\sqrt{I(W;D_i)} + \frac{1}{\epsilon}\mathbb{E}_{W,D}|L_D(W)-L(W)|,$$

where $D\sim\nu$ is independent of W.

Intuitively, extracting correlations between X and Y that are invariant across source domains enhances the generalization ability of machine learning models. The mutual information $I(W;D_i)$ approaches zero when the correlations that a model learns from a specific source domain D_i are also present in other source domains. This does not imply that the model learns nothing from D_s: by further assuming the independence of these domains, the summation of $I(W;D_i)$ can be relaxed to $I(W;D_s)$, which measures the actual amount of information learned by the model. By minimizing each $I(W;D_i)$ and $L_s(W)$ simultaneously, learning algorithms are encouraged to discard domain-specific correlations while preserving invariant ones and thus achieve high generalization performance.

Interestingly, the second term at the RHS of Theorem 8 is highly relevant to the target-domain generalization gap, so we will postpone related analysis to Section III-C.

Next, we demonstrate that the minimization of $I(W;D_i)$ can be achieved by matching the conditional distributions of inter-domain gradients. To see this, we assume that W is optimized by some noisy and iterative learning algorithms, e.g. stochastic gradient descent (SGD). Then the rule of updating W at step t can be formulated as:

$$W_t=W_{t-1}-\eta_t\sum_{i=1}^{m}g(W_{t-1},B_t^i),$$

$$\text{where} \quad g(w, B_t^i) = \frac{1}{m|B_t^i|} \sum_{z \in B_t^i} \nabla_w \ell(f_w(x), y),$$

providing W_0 as the initial guess. Here, η_t is the learning rate, and B_t^i is the batch of data points randomly drawn from source domain D_i to compute the gradient. Suppose that algorithm $\mathcal{A}$ finishes in T steps, we then have:

Theorem 9. *Let $G_t = -\eta_t \sum_{i=1}^m g(W_{t-1}, B_t^i)$, then*

$$I(W_T; D_i) \leq \sum_{t=1}^{T} I(G_t; D_i | W_{t-1}).$$

Although our analysis is derived from the update rule of SGD, the same conclusion applies to a variety of iterative and noisy learning algorithms, e.g. SGLD and AdaGrad. Theorem 9 suggests that minimizing $I(G_t; D_i|W_{t-1})$ in each update penalizes $I(W_T; D_i)$ and thus leads to source-domain generalization. Notably, this conditional mutual information $I(G_t; D_i|W_{t-1})$ can be rewritten as the KL divergence $\mathbb{E}_{D_i, W_{t-1}}[\mathrm{KL}(P_{G_t|D_i,W_{t-1}} \,\|\, P_{G_t|W_{t-1}})]$, which directly motivates matching the marginal and conditional gradient distributions of each source domain. This can also be done by minimizing the difference between each pair of the source-domain gradient distributions $\{P_{G_t|D_i,W_{t-1}}\}_{i=1}^m$, i.e. inter-domain gradients. Therefore:

> Gradient matching promotes source-domain generalization when $\mathbb{E}_{W,D}|L_D(W) - L(W)|$ is minimized.

Intuitively, gradient alignment enforces the model to learn common correlations shared across source domains, thus preventing overfitting to spurious features and promoting invariance [12], [13].

We further present an alternative approach by assuming Lipschitzness instead of sub-Gaussianity, which usually leads to tighter bounds beyond information-theoretic measures:

Theorem 10. *If $\ell(f_w(X), Y)$ is β'-Lipschitz w.r.t w, then*

$$|\mathbb{E}_{W,D_s}[L_s(W)] - \mathbb{E}_W[L(W)]| \leq \frac{\beta'}{m} \sum_{i=1}^{m} \mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)].$$

Besides the elegant symmetry compared to KL divergence metrics, Wasserstein distance bounds are generally considered to be tighter improvements over information-theoretic bounds. To see this, we assume that the adopted metric c is discrete, which leads to the following reductions:

$$\begin{aligned} \mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)] &= \mathbb{E}_{D_i}[\mathrm{TV}(P_{W|D_i}, P_W)] \\ &\leq \mathbb{E}_{D_i} \sqrt{\frac{1}{2}\mathrm{KL}(P_{W|D_i} \,\|\, P_W)} \\ &\leq \sqrt{\frac{1}{2} I(W; D_i)}, \end{aligned} \tag{2}$$

where TV is the total variation. These reductions confirm that the RHS of Theorem 8 also upper bounds other alternative measures of domain differences i.e. total variation and Wasserstein distance. This observation encourages us to directly penalize the mutual information $I(W; D_i)$, which is not only more stable for optimization [35], [46] but also enables simultaneous minimization of these alternative metrics.

C. Target-domain Generalization

We then investigate sufficient conditions for target-domain generalization. Since the training process is independent of the target domains, the predictor could be considered as some constant hypothesis $w \in \mathcal{W}$. It is straightforward to verify that $\mathbb{E}_{D_t}[L_t(w)] = L(w)$ due to the identical domain distribution ν. We then establish the following bound for the target-domain generalization gap:

Theorem 11. *If Assumption 3 holds, then for any fixed $w \in \mathcal{W}$,*

$$\mathbb{P}\{|L_t(w) - L(w)| \geq \epsilon\} \leq \frac{\sigma}{\epsilon} \sqrt{2I(Z; D)}.$$

The result above can be interpreted from two perspectives. Firstly, evaluating the predictor w on randomly sampled target domains reflects its ability to generalize on average, since $L_t(w)$ is an unbiased estimate of $L(w)$. Secondly, knowledge about $L(w)$ can be used to predict the ability of w to generalize on unseen domains, which complements Theorem 8 in solving Problem 6.

In Theorem 11, the probability of generalization is mainly controlled by the extent of distribution shift $I(Z; D)$. Notably, $I(Z; D)$ is an intrinsic property of the data collection procedure, and thus cannot be penalized from the perspective of learning algorithms. Fortunately, the encoder ϕ can be considered as part of the data preprocessing procedure, enabling learning algorithms to minimize the representation space distribution shift. Under the same conditions as Theorem 11, we have that for any fixed classifier ψ:

$$\mathbb{P}\{|L_t(\psi) - L(\psi)| \geq \epsilon\} \leq \frac{\sigma}{\epsilon} \sqrt{2I(R, Y; D)},$$

where $L(\psi) = \mathbb{E}_D[L_D(\psi)]$, $L_t(\psi) = \frac{1}{m'} \sum_{k=1}^{m'} L_{D'_k}(\psi)$, $L_d(\psi) = \mathbb{E}_{R,Y}[\ell(f_\psi(R), Y)]$ and $P_{R,Y}$ is the joint distribution by pushing forward P_Z via the encoder as $R = f_\phi(X)$. The representation space distribution shift can then be decomposed into:

$$\begin{aligned} I(R, Y; D) \text{ (distribution shift)} = I(R; D) \text{ (covariate shift)} \\ + I(Y; D|R) \text{ (concept shift)}. \end{aligned}$$

This motivates us to simultaneously minimize the representation space covariate shift and concept shift to achieve target-domain generalization. We further demonstrate that bounding the covariate shift $I(R; D)$ solely is sufficient for target-domain generalization with Assumption 4:

Theorem 12. *If Assumptions 2 and 4 hold, then for any fixed classifier ψ,*

$$\mathbb{P}\{L_t(\psi) - L(\psi) \geq \epsilon\} \leq \frac{M}{\epsilon\sqrt{2}} \sqrt{I(R; D)} + \frac{2}{\epsilon} L^*,$$

where $L^* = \min_{f^*: \mathcal{R} \mapsto \mathcal{Y}}[L(f^*)]$ and $L(f) = \mathbb{E}_{R,Y}[\ell(f(R), Y)]$.

Similarly, we further refine these target-domain generalization bounds by incorporating the more stringent Assumption

5 in Section VII. Theorem 12 indicates that target-domain generalization is mainly controlled by the amount of covariate shift. Notably, the optimal classifier f^* is chosen from the entire space of functions mapping from $\mathcal{R}$ to $\mathcal{Y}$, so L^* could be regarded as the minimum population risk that the optimal classifier can achieve. In the noiseless case where there exists a ground-truth labeling function h^* such that $Y = h^*(R)$, we will have $L^* = 0$. Moreover, the representation space covariate shift $I(R;D)$ is equivalent to the KL divergence $\mathbb{E}_D[\mathrm{KL}(P_{R|D} \,\|\, P_R)]$, which directly motivates matching the representation distributions across different domains:

> Representation matching promotes target-domain generalization when the best achievable population risk L^* is low.

A byproduct of the proof of Theorem 12 is an upper bound on the expected absolute target-domain generalization error:

$$\mathbb{E}_{W,D}|L_D(W) - L(W)| \leq \frac{M}{\sqrt{2}}\sqrt{I(R;D)} + 2L^*.$$

This result complements the source-domain generalization bound in Theorem 8, confirming that penalizing $I(W;D_i)$ and $I(R;D)$ simultaneously is sufficient to minimize the source-domain generalization gap.

While minimizing $I(R;D)$ guarantees target-domain generalization, this operation requires matching target-domain representations since the classifier Ψ and target domains D_t must be independent to apply Theorem 12. However, knowledge about target-domain samples is not available and we only have source-domain samples during the entire training process. To this end, many existing DG algorithms are matching the source-domain representations as an alternative, utilizing $I(R_i;D_i)$ for each $D_i \in D_s$ as a proxy to penalize $I(R;D)$. The following proposition verifies the feasibility of this approach:

Proposition 13. *Let $W = \mathcal{A}(D_s)$. Assume that $P_{R,D} \ll P_{R_i,D_i}$ and $P_{R_i,D_i} \ll P_{R,D}$, we then have*

$$\mathrm{SKL}(P_{R,D} \,\|\, P_{R_i,D_i}) \leq \log(B)\sqrt{2I(W;D_i)},$$

where $\mathrm{SKL}(P \,\|\, Q) = \mathrm{KL}(P \,\|\, Q) + \mathrm{KL}(Q \,\|\, P)$ and $B = \sup_{r\in\mathcal{R},d\in\mathcal{D}}\left\{\max\left(\frac{P_{R,D}(r,d)}{P_{R_i,D_i}(r,d)}, \frac{P_{R_i,D_i}(r,d)}{P_{R,D}(r,d)}\right)\right\}$.

Interestingly, the discrepancy between the target-domain joint distribution of $P_{R,D}$ and its source-domain counterpart P_{R_i,D_i} can be upper bounded by the input-output mutual information $I(W;D_i)$. This verifies that by letting $I(W;D_i) \to 0$, we will have $P_{R,D} \approx P_{R_i,D_i}$, so one may use $I(R_i;D_i)$ as a proxy to penalize $I(R;D)$ and achieve target-domain generalization.

While our analysis does not necessitate the independence condition between source domains or target domains, such a condition is also naturally satisfied in most learning scenarios and can lead to tighter generalization bounds. Specifically, Theorem 11 and 12 can be further tightened by a factor of $\frac{1}{m'}$ when target domains are i.i.d. We refer the readers to the Appendix for the proof of these results.

IV. Inter-domain Distribution Matching

Motivated by our theoretical analysis in Section III, we propose inter-domain distribution matching (IDM) to achieve high-probability DG (Problem 6). Recall that the average-case risk $L(W)$ serves as a natural bridge to connect $L_s(W)$ and $L_t(W)$, the regularization in Problem 6 directly indicates an objective for optimization by combining the high-probability concentration bounds in Theorem 8 and 11. Specifically, for any $\lambda \in (0,1)$, if Assumption 1 holds, we have:

$$\mathbb{P}\{|L_t(W) - L_s(W)| \geq \epsilon\} \leq \frac{M}{m\epsilon\lambda\sqrt{2}}\sum_{i=1}^{m}\sqrt{I(W;D_i)} + \frac{1}{\epsilon\lambda(1-\lambda)}\left(\frac{M}{\sqrt{2}}\sqrt{I(R;D)} + 2L^*\right). \quad (3)$$

This observation directly motivates aligning inter-domain distributions of the gradients and representations simultaneously. While the idea of distribution matching is not new, we are the first to explore the complementary relationship between gradient and representation matching:

> Gradient and representation matching together minimize $\mathbb{P}\{|L_t(W) - L_s(W)| \geq \epsilon\}$ in Problem 6.

Specifically, source-domain generalization requires the minimization of the target-domain generalization gap (Theorem 8), and target-domain generalization requires the minimization of the input-output mutual information (Proposition 13). Therefore, existing works focusing exclusively on either gradient or representation alignment are insufficient to fully address the domain-level generalization gap. To our best knowledge, this is the first exploration in the literature where gradient and representation matching are combined to yield a sufficient solution for the DG problem.

A. Per-sample Distribution Matching

While various distribution matching methods have been proposed in the literature, these techniques are generally either ineffective or insufficient for high-dimensional and complex distributions. Typically, learning algorithms have no knowledge about the underlying distribution of either the representation or the gradient, and the only available way is to align them across batched data points. We first provide an impossibility theorem for high-dimensional distribution matching in the cases of limited number of samples:

Theorem 14. *Let n and b be the dimension and the number of data points respectively. Then*

- *If $n > b+1$, then given an arbitrarily group of data points $s = \{x_i\}_{i=1}^b$, there exists infinite domains d_1, d_2, $\cdots$ that satisfies $P(S = s|D = d_1) = P(S = s|D = d_2) = \cdots$.*
- *If $n > 2b+1$, then for any two groups of sampled data points $s_1 = \{x_i^1\}_{i=1}^b$ and $s_2 = \{x_i^2\}_{i=1}^b$, there exists infinite domains d_1, d_2, $\cdots$ such that for any $j \in [1,\infty)$, $P(S = s_1|D = d_j) = P(S = s_2|D = d_j)$.*

Intuitively speaking, Theorem 14 states that different domains are theoretically indistinguishable with finite samples if $n \gg b$. In real-world scenarios, the dimensionality of the

feature or the gradient easily exceeds that of the batch size, making algorithms that aim to align the entire distribution (e.g. CORAL [5] and MMD [6]) generally ineffective since distribution alignment is basically impossible given such few data points. This observation is also verified by [13] that aligning the entire covariance matrix achieves no better performance than aligning the diagonal elements only. Furthermore, prior distribution alignment techniques mainly focus on aligning the directions [12], [20], [47] or low-order moments [5], [11], [13], which are insufficient for complex probability distributions. For example, while the standard Gaussian distribution $N(0,1)$ and the uniform distribution $U(-\sqrt{3},\sqrt{3})$ share the same expectation and variance, they are fundamentally different to one another. To address these issues, we propose the per-sample distribution matching (PDM) technique that aligns distributions in a per-dimension manner, by minimizing an upper bound of the KL divergence between probability density estimators.

Let $\{x_i^1\}_{i=1}^b$ and $\{x_i^2\}_{i=1}^b$ be two groups of 1-dimensional data points drawn from probability distributions P and Q respectively. Let p_i denote the density of Gaussian distribution with expectation x_i^1 and variance σ^2, then the kernel density estimator $\bar{P}$ for P can be written as $\bar{p}(x) = \frac{1}{b}\sum_i p_i(x)$ (respectively for q_i, $\bar{Q}$ and $\bar{q}$). The following theorem suggests a computable upper bound for the KL divergence (Wasserstein distance) between probability density estimators:

Theorem 15. *Let f be a bijection: $[1,b] \leftrightarrow [1,b]$ and P_i be the probability measure defined by p_i (respectively for Q_i and q_i), then $\mathrm{KL}(\bar{P} \,\|\, \bar{Q}) \le \frac{1}{b}\sum_{i=1}^b \mathrm{KL}(P_i \,\|\, Q_{f(i)})$, and $\mathbb{W}(\bar{P},\bar{Q}) \le \frac{1}{b}\sum_{i=1}^b \mathbb{W}(P_i, Q_{f(i)})$.*

Hence, distribution matching can be achieved by minimizing the KL divergence or Wasserstein distances between point Gaussian densities, which can be achieved by aligning individual data points. The following theorem suggests an optimal bijection for choosing the order of alignment:

Theorem 16. *Suppose that $\{x_i^1\}_{i=1}^b$ and $\{x_i^2\}_{i=1}^b$ are both sorted in the same order, then $f(j) = j$ is the minimizer of both $\sum_{i=1}^b \mathrm{KL}(P_i \,\|\, Q_{f(i)})$ and $\sum_{i=1}^b \mathbb{W}(P_i, Q_{f(i)})$.*

To summarize, the procedure of PDM is to slice the data points into separate dimensions, sort the data points in ascending (or descending) order for each dimension, and then match the sorted data points across different source domains. PDM improves over previous distribution matching techniques by simultaneously capturing multiple orders of moments, avoiding ineffective high-dimensional distribution matching, as well as enabling straightforward implementation and efficient computation.

The pseudo-code for PDM is provided in Algorithm 1, where we adopted the moving average trick from [13] to enlarge the equivalent batch size for accurate probability density estimation. This does not invalidate our analysis in Theorem 14, as the maximum equivalent batch size ($b/(1-\gamma) \approx 640$) remains significantly smaller than the dimensionality of the representation (2048 for ResNet-50 in DomainBed) or the gradient ($2048 \times c$, the number of classes) and satisfies $d > 2b+1$. Therefore, it is still impossible to distinguish different distributions as indicated by Theorem 14. However, this moving average technique indeed helps to improve the empirical performance, as shown by our ablation studies. In Algorithm 1, the moving averages X_{ma}^i are initialized with 0, and the input data points are represented as matrices $X^i \in \mathbb{R}^{b\times d}$, where b and d denote the batch size and dimensionality respectively. Each row of X then corresponds to an individual data point.

Algorithm 1 PDM for distribution matching.

Input: Data matrices $\{X^i\}_{i=1}^m$, moving average γ.
Output: Penalty of distribution matching.
for i **from** 1 **to** m **do**
  Sort the elements of X^i in each column in ascending order.
  Calculate moving average $X_{ma}^i = \gamma X_{ma}^i + (1-\gamma)X^i$.
end for
Calculate the mean of data points across domains: $X_{ma} = \frac{1}{m}\sum_{i=1}^m X_{ma}^i$.
Output: $\mathcal{L}_{\mathrm{PDM}} = \frac{1}{mdb}\sum_{i=1}^m \|X_{ma} - X_{ma}^i\|_F^2$.

B. Algorithm Design

Combining the methods discussed above, we finally propose the IDM algorithm for high-probability DG by simultaneously aligning inter-domain gradients and representations. Recall that Problem 6 incorporates an additional regularization based on ERM, we adopt the following Lagrange multipliers to optimize the IDM objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{IDM}} &= \mathcal{L}_{\mathrm{E}} + \lambda_1\mathcal{L}_{\mathrm{G}} + \lambda_2\mathcal{L}_{\mathrm{R}} \\ &= L_s(W) + \lambda_1\mathcal{L}_{\mathrm{PDM}}(\{G_i\}_{i=1}^m) + \lambda_2\mathcal{L}_{\mathrm{PDM}}(\{R_i\}_{i=1}^m). \quad (4)\end{aligned}$$

Here $\mathcal{L}_{\mathrm{E}}$ is the risk of ERM, $\mathcal{L}_{\mathrm{G}}$ and $\mathcal{L}_{\mathrm{R}}$ denote the penalty of distribution matching for the gradients and representations respectively, implemented with the proposed PDM method. To cooperate representation alignment which regards the classifier ψ as the true predictor and also for memory and time concerns, we only apply gradient alignment for the classifier ψ as in [13]. Furthermore, λ_1 and λ_2 should be adaptively chosen according to the extent of covariate and concept shifts respectively: On the one hand, representation matching is closely connected to covariate shift, as it aims to minimize the representation space covariate shift $I(R;D)$, which is induced by the input space covariate shift $I(X;D)$. According to the Markov chain $D \to X \to R$, the representations are naturally aligned when $I(X;D) = 0$. On the other hand, concept shift $I(Y;D|X)$ implies the existence of domain-specific correlations between X and Y, i.e. the predictive distribution $P_{Y|X}$ is different for each domain. This will cause models to overfit these source domain-specific features and thus generalize poorly. When $I(Y;D|X) = 0$, gradient alignment is not required since the distribution shift can solely be addressed by aligning the representations. In general, gradient and representation matching reduces the impact of concept shift and covariate shift respectively to achieve domain generalization, and λ_1, λ_2 should scale with the amount of the two shifts respectively.

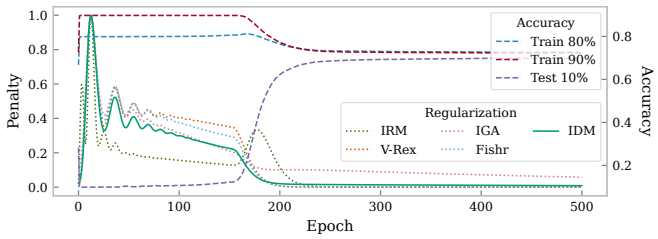

Fig. 2. Dynamics of different DG regularization penalties (dotted line) and prediction accuracies of each domain (dashed line) on Colored MNIST, optimized by the IDM objective.

TABLE I
THE COLORED MNIST TASK.

Method	Train Acc	Test Acc	Gray Acc
ERM	86.4 ± 0.2	14.0 ± 0.7	71.0 ± 0.7
IRM	71.0 ± 0.5	65.6 ± 1.8	66.1 ± 0.2
V-REx	71.7 ± 1.5	67.2 ± 1.5	68.6 ± 2.2
IGA	68.9 ± 3.0	67.7 ± 2.9	67.5 ± 2.7
Fishr	69.6 ± 0.9	71.2 ± 1.1	70.2 ± 0.7
IDM	70.2 ± 1.4	70.6 ± 0.9	70.5 ± 0.7

The pseudo-code for IDM is presented in Algorithm 2 for completeness.

Algorithm 2 IDM for high-probability DG.

Input: Initial model W, training dataset S, hyper-parameters λ_1, λ_2, t_1, t_2, γ_1, γ_2.
for t **from** 1 **to** #steps **do**
  for i **from** 1 **to** m **do**
    Randomly sample a batch $B_t^i = (X_t^i, Y_t^i)$ from S_i of size b.
    Compute individual representations: $(R_t^i)_j = f_\Phi\big((X_t^i)_j\big)$, for $j \in [1, b]$.
    Compute individual risks: $(L_t^i)_j = \ell\big(f_\Psi\big((R_t^i)_j\big), (Y_t^i)_j\big)$, for $j \in [1, b]$.
    Compute individual gradients: $(G_t^i)_j = \nabla_\Psi (L_t^i)_j$, for $j \in [1, b]$.
  end for
  Compute total empirical risk: $\mathcal{L}_{\text{IDM}} = \frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n (L_t^i)_j$.
  if $t \geq t_1$ **then**
    Compute gradient alignment penalty: $\mathcal{L}_{\text{G}} = \mathcal{L}_{\text{PDM}}(\{G_t^i\}_{i=1}^m, \gamma_1)$.
    $\mathcal{L}_{\text{IDM}} \leftarrow \mathcal{L}_{\text{IDM}} + \lambda_1 \mathcal{L}_{\text{G}}$.
  end if
  if $t \geq t_2$ **then**
    Compute representation alignment penalty: $\mathcal{L}_{\text{R}} = \mathcal{L}_{\text{PDM}}(\{R_t^i\}_{i=1}^m, \gamma_2)$.
    $\mathcal{L}_{\text{IDM}} \leftarrow \mathcal{L}_{\text{IDM}} + \lambda_2 \mathcal{L}_{\text{R}}$.
  end if
  Back-propagate gradients $\nabla_W \mathcal{L}_{\text{IDM}}$ and update the model W.
end for

V. RELATED WORKS

Domain Generalization. In the literature, various approaches have been proposed by incorporating external domain information to achieve OOD generalization. Most recent works achieve invariance by employing additional regularization criteria based on ERM. These methods differ in the choice of the statistics used to match across source domains and can be categorized by the corresponding objective of 1) gradient, 2) representation, and 3) predictor, as follows:

- *Invariant Gradients:* Gradient alignment enforces batched data points from different domains to cooperate and promotes OOD generalization by finding loss minima shared across source domains. Specifically, IGA [11] aligns the empirical expectations, Fish [12] maximizes the dot-product of inter-domain gradients, AND-mask [20] and SAND-mask [47] only update weights when the gradients share the same direction, and Fishr [13] matches the gradient variance. These gradient-based objectives are generally restricted to aligning the directions or low-order moments, resulting in substantial information loss in more granular statistics. Besides, these works either lack generalization guarantees or rely on strong assumptions including the existence of invariant and controllable features (IGA), the shape of loss landscape around the local minima (Fishr), Lipschitz continuous gradients and co-diagonalizable Hessian matrix (AND-mask). In contrast, our analysis successfully connects gradient alignment and source-domain generalization by leveraging the mild assumption of identical domain distributions.
- *Invariant Representations:* Extracting domain-invariant features has been extensively studied to solve both DG and domain adaptation (DA) problems. DANN [7] and CDANN [8] align inter-domain representations via adversarial learning, MMD [6] uses kernel methods for distribution alignment, and CORAL [5] matches low-order moments of the representations. Still, these methods are insufficient for complex probability distributions [48], ineffective for high-dimensional distributions (Theorem 14), and incapable of addressing the concept shift. Besides, the viability of minimizing the representation shift $I(R; D)$ through the source-domain proxy $I(R_i; D_i)$ remains questionable without gradient matching (Proposition 7). Our analysis sheds light on understanding how representation alignment enhances target-domain generalization by minimizing the variance of target-domain risks.
- *Invariant Predictors:* A recent line of works proposes to explore the connection between invariance and causality. IRM [9] and subsequent works [37], [38] learn an invariant classifier that is simultaneously optimal for all source domains. However, later works have shown that IRM may fail on non-linear data and lead to sub-optimal predictors [36], [49]. Parallel works include: V-REx [15] which minimizes the variance of source-domain risks, GroupDRO [14] which minimizes the worst-domain training risk, and QRM [16] which optimizes a quantile of the risk distribution. As shown in the next section, IDM also promotes domain-invariant predictors and ensures

optimality across different source domains.

Information-theoretic Generalization Analysis. The utilization of information-theoretic tools to acquire generalization guarantees for randomized learning algorithms has gained significant attention after the seminal works of [22], [50]. These works differ in the choice of information-theoretic metrics and can be categorized as follows:

- *Mutual Information:* The works of [22], [50] successfully connect generalization error and the mutual information between the hypothesis and the training dataset. This approach is also shown to be highly effective in characterizing the learning dynamics of noisy and iterative algorithms such as SGD [23]–[25]. However, this line of research mainly focuses on the traditional in-distribution learning settings, with minimal investigations into OOD generalization. To the best of our knowledge, we are the first to provide a comprehensive analysis of the domain-level generalization error. Moreover, our proof techniques shed new light on upper bounding the expectation of the absolute generalization gap, which is intrinsically more difficult and may be of independent interest in new deriving high-probability generalization bounds.
- *KL Divergence:* As an alternative approach, [51] explores the possibility of establishing generalization bounds with KL divergence between average joint distributions, which demonstrates potential in deriving tighter bounds. [52] shows that the generalization error of Gibbs algorithms can be exactly characterized by the symmetrized KL information. [35] investigates the generalization of domain adaptation through the KL divergence between source and target-domain data distributions. Although Proposition 13 share certain similarities in the usage of SKL metrics, their problem settings are fundamentally different to ours and thus cannot be directly applied to acquire the insights provided by Proposition 13, that gradient matching complements representation matching in target-domain generalization.
- *Wasserstein Distance:* Recently, the Wasserstein distance has been found as a tighter improvement over KL divergence to quantify distances between probability distributions. To this end, multiple works [30], [51] establish generalization bounds via Wasserstein distance metrics by adopting a more stringent Lipschitz continuity assumption. While Theorem 10 shares certain similarities with these results, we successfully extend Wasserstein distance analysis to characterize the source-domain generalization error and motivate the minimization of $I(W; D_i)$.

VI. Experimental Results

In this section, we evaluate the proposed IDM algorithm on the Colored MNIST task [9] and the DomainBed benchmark [21] to demonstrate its capability of generalizing against various types of distribution shifts[1]. Detailed settings of these experiments and further empirical results including ablation studies are reported in Appendix C-A - D-C.

[1]The source code is available at https://github.com/Yuxin-Dong/IDM.

A. Colored MNIST

The Colored MNIST task [9] is carefully designed to create high correlations between image colors and the true labels, leading to spurious features that possess superior predictive power (90% and 80% accuracy) over the actual digits (75%). However, this correlation is reversed in the target domain (10%), causing any learning algorithm that solely minimizes training errors to overfit the color information and fail when testing. As such, Colored MNIST is an ideal task to evaluate the capability of learning algorithms to achieve invariance across source domains.

Following the settings of [9], we adopt a two-stage training technique, where the penalty strength λ is set low initially and higher afterward. We visualize the learning dynamics of relevant DG penalties, including IRM, V-Rex, IGA, and Fishr, using the IDM objective for optimization in Figure 2. The penalty values are normalized for better clarity. This visualization confirms Theorem 8 that IDM promotes source-domain generalization by minimizing the gap between training risks, thus ensuring the optimality of the predictor across different source domains. Moreover, it verifies the superiority of PDM by showing that penalizing the IDM objective solely is sufficient to minimize other types of invariance penalties.

Table I presents the performance comparison on Colored MNIST across 10 independent runs. Following the hyper-parameter tuning technique as [9], we select the best model by $\max_w \min(L_s(w), L_t(w))$. As can be seen, IDM achieves the best trade-off between source and target-domain accuracies (70.2%), and near-optimal gray-scale accuracy (70.5%) compared to the Oracle predictor (71.0%, ERM trained with gray-scale images).

B. DomainBed Benchmark

The DomainBed Benchmark [21] comprises multiple synthetic and real-world datasets for assessing the performance of both DA and DG algorithms. To ensure a fair comparison, DomainBed limits the number of attempts for hyper-parameter tuning to 20, and the results are averaged over 3 independent trials. Therefore, DomainBed serves as a rigorous and comprehensive benchmark to evaluate different DG strategies. We compare the performance of our method with 20 baselines in total for a thorough evaluation. Table II summarizes the results using target-domain model selection, which is a common choice for validation purposes [13], [15] and highly motivated by our discussion in Appendix D-C.

As can be seen, IDM achieves top-1 accuracy (72.0%) on CMNIST which is competitive with the Oracle (75.0%), outperforming all previous distribution alignment techniques by aligning the directions (AND-mask, SAND-mask, Fish) or low-order moments (Fishr). This verifies the superiority of the proposed PDM method as well as the complementary relationship between gradient and representation alignment. On the contrary, algorithms that only align the representations (CORAL, MMD, DANN, CDANN) are incapable of addressing the concept shift, thus performing poorly on CMNIST. Moreover, IDM achieves the highest accuracy among all distribution matching algorithms on RMNIST / PACS,

TABLE II
THE DOMAINBED BENCHMARK. WE FORMAT **BEST**, SECOND BEST AND WORSE THAN ERM RESULTS.

Algorithm	Accuracy (↑)								Ranking (↓)		
	CMNIST	**RMNIST**	**VLCS**	**PACS**	**OffHome**	**TerraInc**	**DomNet**	**Avg**	**Mean**	**Median**	**Worst**
ERM	57.8 ± 0.2	97.8 ± 0.1	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	53.0 ± 0.3	41.3 ± 0.1	68.7	12.3	11	20
IRM	67.7 ± 1.2	97.5 ± 0.2	76.9 ± 0.6	84.5 ± 1.1	63.0 ± 2.7	50.5 ± 0.7	28.0 ± 5.1	66.9	18.3	20	22
GroupDRO	61.1 ± 0.9	97.9 ± 0.1	77.4 ± 0.5	87.1 ± 0.1	66.2 ± 0.6	52.4 ± 0.1	33.4 ± 0.3	67.9	11.7	10	19
Mixup	58.4 ± 0.2	98.0 ± 0.1	78.1 ± 0.3	86.8 ± 0.3	68.0 ± 0.2	**54.4** ± 0.3	39.6 ± 0.1	69.0	7.3	6	15
MLDG	58.2 ± 0.4	97.8 ± 0.1	77.5 ± 0.1	86.8 ± 0.4	66.6 ± 0.3	52.0 ± 0.1	41.6 ± 0.1	68.7	12.6	13	18
CORAL	58.6 ± 0.5	98.0 ± 0.0	77.7 ± 0.2	87.1 ± 0.5	**68.4** ± 0.2	52.8 ± 0.2	41.8 ± 0.1	69.2	6.4	5	14
MMD	63.3 ± 1.3	98.0 ± 0.1	77.9 ± 0.1	87.2 ± 0.1	66.2 ± 0.3	52.0 ± 0.4	23.5 ± 9.4	66.9	10.0	10	22
DANN	57.0 ± 1.0	97.9 ± 0.1	79.7 ± 0.5	85.2 ± 0.2	65.3 ± 0.8	50.6 ± 0.4	38.3 ± 0.1	67.7	15.0	18	22
CDANN	59.5 ± 2.0	97.9 ± 0.0	**79.9** ± 0.2	85.8 ± 0.8	65.3 ± 0.5	50.8 ± 0.6	38.5 ± 0.2	68.2	12.4	14	18
MTL	57.6 ± 0.3	97.9 ± 0.1	77.7 ± 0.5	86.7 ± 0.2	66.5 ± 0.4	52.2 ± 0.4	40.8 ± 0.1	68.5	11.7	10	21
SagNet	58.2 ± 0.3	97.9 ± 0.0	77.6 ± 0.1	86.4 ± 0.4	67.5 ± 0.2	52.5 ± 0.4	40.8 ± 0.2	68.7	11.3	9	17
ARM	63.2 ± 0.7	**98.1** ± 0.1	77.8 ± 0.3	85.8 ± 0.2	64.8 ± 0.4	51.2 ± 0.5	36.0 ± 0.2	68.1	13.0	16	21
VREx	67.0 ± 1.3	97.9 ± 0.1	78.1 ± 0.2	87.2 ± 0.6	65.7 ± 0.3	51.4 ± 0.5	30.1 ± 3.7	68.2	10.6	8	20
RSC	58.5 ± 0.5	97.6 ± 0.1	77.8 ± 0.6	86.2 ± 0.5	66.5 ± 0.6	52.1 ± 0.2	38.9 ± 0.6	68.2	13.4	13	19
AND-mask	58.6 ± 0.4	97.5 ± 0.0	76.4 ± 0.4	86.4 ± 0.4	66.1 ± 0.2	49.8 ± 0.4	37.9 ± 0.6	67.5	17.0	16	22
SAND-mask	62.3 ± 1.0	97.4 ± 0.1	76.2 ± 0.5	85.9 ± 0.4	65.9 ± 0.5	50.2 ± 0.1	32.2 ± 0.6	67.2	17.9	19	22
Fish	61.8 ± 0.8	97.9 ± 0.1	77.8 ± 0.6	85.8 ± 0.6	66.0 ± 2.9	50.8 ± 0.4	**43.4** ± 0.3	69.1	11.3	11	18
Fishr	68.8 ± 1.4	97.8 ± 0.1	78.2 ± 0.2	86.9 ± 0.2	68.2 ± 0.2	53.6 ± 0.4	41.8 ± 0.2	70.8	5.4	**3**	16
SelfReg	58.0 ± 0.7	**98.1** ± 0.7	78.2 ± 0.1	**87.7** ± 0.1	68.1 ± 0.3	52.8 ± 0.9	43.1 ± 0.1	69.4	5.0	**3**	19
CausIRL$_{\text{CORAL}}$	58.4 ± 0.3	98.0 ± 0.1	78.2 ± 0.1	87.6 ± 0.1	67.7 ± 0.2	53.4 ± 0.4	42.1 ± 0.1	69.4	5.0	**3**	15
CausIRL$_{\text{MMD}}$	63.7 ± 0.8	97.9 ± 0.1	78.1 ± 0.1	86.6 ± 0.7	65.2 ± 0.6	52.2 ± 0.3	40.6 ± 0.2	69.2	10.4	10	20
IDM	**72.0** ± 1.0	98.0 ± 0.1	78.1 ± 0.4	87.6 ± 0.3	68.3 ± 0.2	52.8 ± 0.5	41.8 ± 0.2	**71.2**	**3.3**	**3**	**6**

competitive performances to the best algorithm on RMNIST (98.0% v.s. 98.1%), PACS (87.6% v.s. 87.7%), OfficeHome (68.3% v.s. 68.4%), the highest average accuracy (71.2%) and best rankings (mean, median and worst rankings on 7 datasets) among all baseline methods. IDM also enables efficient computation, such that the running-time overhead is only 5% compared to ERM on the largest DomainNet dataset, and negligible for other smaller datasets. Notably, IDM is the only algorithm that consistently achieves top rankings (Top 6 of 22), while any other method failed to outperform most of the competitors on at least 1 dataset.

While the overall performance is promising, we notice that IDM is not very effective on TerraIncognita. There are several possible reasons: Firstly, the number of hyper-parameters in IDM exceeds most competing methods, which is critical to model selection since the number of tuning attempts is limited in DomainBed. Recall that the value of λ_1 and λ_2 should adapt to the amount of covariate and concept shifts respectively: While CMNIST manually induces high concept shift, covariate shift is instead dominant in other datasets, raising extra challenges for hyper-parameter tuning. Secondly, representation space distribution alignment may not always help since $L_t(w) \leq L(w)$ is possible by the randomized nature of target domains. These factors together result in sub-optimal hyper-parameter selection results.

VII. FURTHER DISCUSSIONS

A. Information Bottleneck for Target-domain Generalization

Alternatively, one can also decompose the representation space distribution shift from an anti-causal perspective:

$$I(R, Y; D) \text{ (distribution shift)} = I(Y; D) \text{ (label shift)} + I(R; D|Y) \text{ (concept shift)}.$$

While label shift $I(Y; D)$ is an intrinsic property of the data distribution and cannot be optimized by learning algorithms, the anti-causal concept shift $I(R; D|Y)$ is closely connected to the information bottleneck (IB) principle: Notice the Markov chain $(D, Y) - X - R$, we then have that by applying the data-processing inequality,

$$I(R; D|Y) = I(R; D, Y) - I(R; Y) \leq I(R; X) - I(R; Y).$$

Recall that the spirit of IB is to minimize $I(X; R)$ while maximizing $I(R; Y)$, this target can be achieved by solely penalizing $I(R; X|Y)$ [53]–[55]. Therefore, when there is no label imbalance issues (i.e. $I(Y; D) \to 0$), the anti-causal concept shift $I(R; D|Y)$ can be minimized by the IB principle:

> Information bottleneck promotes target-domain generalization if $I(Y; D) \to 0$.

Notably, our analysis facilitates previous works [36] utilizing IB to enhance the performance of OOD generalization. While the analysis of [36] is primarily restricted to linear models, our results apply to any encoder-classifier type network. Similar to representation matching, directly minimizing $I(R; X|Y)$ requires knowledge about target-domain samples which is inaccessible in practice. Following Proposition 13, we provide the following result on the feasibility of utilizing the empirical IB $I(R_i; X_i|Y_i)$ as a proxy to optimize $I(R; X|Y)$, where $Z_i = (X_i, Y_i)$ is a training sample of domain D_i.

Proposition 17. *Let $W = \mathcal{A}(D_s)$. Assume that $P_{R,Z} \ll P_{R_i,Z_i}$ and $P_{R_i,Z_i} \ll P_{R,Z}$, then*

$$\mathrm{SKL}(P_{R,Z} \,\|\, P_{R_i,Z_i}) \leq \log(B)\sqrt{2I(W; D_i)}.$$

where $B = \sup_{r\in\mathcal{R}, z\in\mathcal{Z}} \left\{ \max\left(\frac{P_{R,Z}(r,z)}{P_{R_i,Z_i}(r,z)}, \frac{P_{R_i,Z_i}(r,z)}{P_{R,Z}(r,z)} \right) \right\}$.

Proof. The proof follows the same development as Proposition 13. □

This result indicates that with the assistance of gradient matching (i.e. minimizing $I(W; D_i)$), one can achieve target-domain generalization by optimizing the source-domain IB objective $I(R_i; X_i|Y_i)$.

B. Leveraging the Independence Assumption

In the analysis above, we only assume that the target domains are independent of the source domains (Assumption 1), while the source (or target) domains are not necessarily independent of each other. This assumption is much weaker than the i.i.d domains assumption adopted in [16] by allowing correlations between source domains, e.g. sampling from a finite set without replacement. While this weaker assumption is preferable, we highlight that the target-domain generalization bounds in Theorem 11 and 12 can be further tightened by a factor of $1/m'$ when the i.i.d condition is incorporated. Therefore, one can now guarantee better generalization by increasing the number of domains, which is consistent with real-world observations.

Theorem 18. *If Assumption 3 holds and the target domains D_t are independent, then for any fixed $w \in \mathcal{W}$,*

$$\mathbb{P}\{|L_t(w) - L(w)| \geq \epsilon\} \leq \frac{2\sigma^2}{m'\epsilon^2} I(Z; D).$$

Theorem 19. *If Assumption 2 and 4 hold and the target domains D_t are independent, then for any fixed classifier ψ,*

$$\mathbb{P}\{L_t(\psi) - L(\psi) \geq \epsilon + L^*\} \leq \frac{2\sigma^2}{m'\epsilon^2} I(R; D),$$

where $L^* = \min_{f^*:\mathcal{R}\mapsto\mathcal{Y}} L_t(f^*) + L(f^*)$.

Furthermore, when the source domains satisfy the i.i.d condition, it can be proved that $\sum_{i=1}^{m} I(W; D_i) \leq I(W; D_s)$. Otherwise, we can only guarantee $I(W; D_i) \leq I(W; D_s)$ for any $i \in [1, m]$. This indicates that while the model achieves source-domain generalization by letting $I(W; D_i) \to 0$, it can still learn from source domains D_s. That is, having $I(W; D_i) = 0$ for all $i \in [1, m]$ does not necessarily lead to $I(W; D_s) = 0$. To see this, one can take D_i as independent Bernoulli variables $\mathrm{Bern}(1/2)$, and let $W = D_1 \oplus \cdots \oplus D_m$, where $\oplus$ is the XOR operator. Then it is easy to verify that W is independent of each D_i since $P_{W|D_i} = P_W$, implying $I(W; D_i) = 0$. However, $I(W; D_s) = H(W)$ is strictly positive.

C. High-probability Problem Formulation

An alternative high-probability formulation of the DG problem is presented by [16], named Quantile Risk Minimization (QRM). Under our notations, the QRM objective can be expressed as:

$$\min_{w} \epsilon \quad s.t. \quad \mathbb{P}\{L_t(w) \geq \epsilon\} \leq \delta.$$

The main difference between our formulation (Problem 6) and QRM is that we not only consider the randomness of D_t, but also that of D_s and W. The randomized nature of domains and the hypothesis serve as the foundation for our information-theoretic generalization analysis. While our formulation inspires an algorithm by matching inter-domain gradients and representations, the optimization of the objective in [16] is not directly tractable and requires further kernel density estimation to approximate the quantile of the risk distribution. It is also questionable to use training risks as a surrogate to optimize the quantile of test risk distribution, as source domains do not satisfy the independence assumption between W and D_s required by the QRM objective. Furthermore, kernel density estimation would be challenging when the number of source domains is not sufficiently large. On the contrary, IDM could be easily implemented when there are at least 2 source domains.

Additionally, Problem 6 aims to find the optimal learning algorithm instead of the optimal hypothesis. This would be essential to analyze the correlations between the hypothesis W and source domains D_s, and is more suitable in robust learning settings when measuring the error bar. The trade-off between optimization and generalization is also more explicitly and intuitively characterized in our formulation.

D. Tighter Bounds for Target-Domain Population Risk

In a similar vein, we provide the following bounds for target-domain generalization error in terms of Wasserstein distances.

Theorem 20. *If Assumption 5 holds, then for any $w \in \mathcal{W}$,*

$$\mathbb{P}\{|L_t(w) - L(w)| \geq \epsilon\} \leq \frac{\beta^2}{m'\epsilon^2} \mathbb{E}_D[\mathbb{W}^2(P_{Z|D=d}, P_Z)].$$

Theorem 21. *If Assumption 4 holds, and $\ell(f_w(X), f_{w'}(X))$ is β-Lipschitz for any $w, w' \in \mathcal{W}$, then for any $w \in \mathcal{W}$,*

$$\mathbb{P}\{L_t(w) - L(w) \geq \epsilon + L^*\} \leq \frac{\beta^2}{m'\epsilon^2} \mathbb{E}_D[\mathbb{W}^2(P_{X|D=d}, P_X)],$$

where $L^* = \min_{w^*\in\mathcal{W}}(L_t(w^*) + L(w^*))$.

Proof. The proof follows the same development as Theorem 10. □

The expected Wasserstein distance metrics above serve as analogs to the extent of covariate shift $I(Z; D)$ and $I(X; D)$, through a similar reduction as depicted in equation (2).

Inspired by recent advancements in information-theoretic generalization analysis which incorporate network predictions or losses to derive tighter bounds [41], [56], we further tighten these target-domain generalization bounds by considering the distribution shift of the risk. For any hypothesis $w \in \mathcal{W}$ and domain $d \in \mathcal{D}$, let $L = \ell(f_w(X), Y)$ be the risk of predicting a randomly given sample $Z \sim P_{Z|D=d}$. Through a similar sketch as the proof of Theorem 11, we can prove that

$$\mathbb{P}\{|L_t(w) - L(w)| \geq \epsilon\} \leq \frac{M^2}{2\epsilon^2} I(L; D).$$

According to the Markov chain relationship $D \to (X, Y) \to (f_\phi(X), Y) \to (f_w(X), Y) \to L$, this bound is strictly tighter than Theorem 11 which uses sample space $I(Z; D)$ or representation space $I(R, Y; D)$ distribution shifts. Also, notice that the mutual information $I(L; D)$ could be rewritten

as $\mathbb{E}_D[\mathrm{KL}(P_{L|D} \,\|\, P_L)]$, this suggests that matching the inter-domain distributions of the risks helps to generalize on target domains. This observation facilitates the work of [15], which proposes to align the empirical risks of distinct source domains. However, there still exists a gap between target-domain risk shift $I(L;D)$ and its empirical counterpart $I(L_i;D_i)$, which further requires minimizing $I(W;D_i)$. Considering that L is a scalar while R is a vector, aligning the distributions of the risks avoids high-dimensional distribution matching and may enable efficient implementation. We will leave this method for future research.

E. Generalization Bounds for Source-Domain Empirical Risk

In the analysis above, we are mainly focusing on the domain-level generalization error, which corresponds to the case when we have an infinite number of samples for each domain to evaluate the population risk. However, the standard generalization error raised by a finite number of samples cannot be simply ignored in practice, and requires additional regularization (e.g. weight decay, dropout) to tackle. In this section, we further consider the effect of finite training samples to address the generalization gap between domain-level population and empirical risks:

Theorem 22. *If Assumption 2 holds, then*

$$|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \le \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{M^2}{2}I(W;D_i)} + \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{\frac{M^2}{2}I(W;Z_j^i|D_i)},$$

$$\mathbb{P}\{|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \ge \epsilon\} \le \frac{M^2}{mn\epsilon^2}(I(W;S)+\log 3),$$

where $L'(W) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\ell(f_W(X_j^i), Y_j^i)$.

Theorem 23. *If $\ell(f_w(X), Y)$ is β'-Lipschitz w.r.t w, then*

$$|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \le \frac{\beta'}{m}\sum_{i=1}^{m}\mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)] + \frac{\beta'}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{D_i,Z_j^i}[\mathbb{W}(P_{W|D_i,Z_j^i}, P_{W|D_i})].$$

The theorems above provide upper bounds for the empirical generalization risk by exploiting the mutual information between the hypothesis and the samples (or the Wasserstein distance counterparts). Compared to Theorems 8 and 10, these results additionally consider the randomness of the data samples Z_j^i, indicating that traditional techniques for improving the generalization of classical supervised learning algorithms by minimizing $I(W;S)$, such as gradient clipping [25], [35] and stochastic gradient perturbation [57], [58] methods, also enhance the capability of domain generalization algorithms $\mathcal{A}$ to generalize on target domains under our high-probability problem setting by preventing overfitting to training samples. This observation is also verified in [35]. Relevant analysis may also motivate information-theoretic generalization analysis for meta-learning tasks [59]–[62]. Although addressing the standard generalization error is also a promising research direction, it is beyond the scope of solving Problem 6 and is thus not our main focus in this paper.

F. High-probability Generalization Bounds

Our main results in Section III generally have a linear dependence $\frac{1}{\delta}$ on the probability factor, which may lead to sub-optimal bounds when $\delta \to 0$. In this section, we demonstrate that these results can be further improved to achieve a logarithm dependence of $\log(\frac{1}{\delta})$, which will be more intriguing under high-probability scenarios.

Theorem 24. *If Assumption 2 holds, then for any $\lambda \in [0,1)$, with probability at least $1-\delta$ over the draw of D_s and W,*

$$|L(W) - L_s(W)| \le \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{M^2}{2\lambda}\left(\imath(W;D_i) + \log\frac{m}{\delta\sqrt{1-\lambda}}\right)}.$$

Theorem 25. *If assumption 3 holds, then for any $w \in \mathcal{W}$ and $\lambda \in [0,1)$, with probability at least $1-\delta$ over the draw of D_t,*

$$|L(w) - L_t(w)| \le \frac{1}{m'}\sum_{k=1}^{m'}\sqrt{\frac{2\sigma^2}{\lambda}\left(\mathrm{KL}\Big(P_{Z|D_k'}\Big\|P_Z\Big) + \log\frac{m'}{\delta\sqrt{1-\lambda}}\right)}.$$

The above Theorem 24 and 25 serve as high-probability analogs for Theorem 8 and 11 respectively. Here, the information density $\imath(W;D_i)$ acts as an alternative measure of the extent of correlation between the hypothesis and source domains, by noticing that $\mathbb{E}_{W,D_i}[\imath(W;D_i)] = I(W;D_i)$. Similarly, the KL divergence $\mathrm{KL}\big(P_{Z|D}\,\big\|\,P_Z\big)$ substitutes the original distribution shift mutual information and satisfies $\mathbb{E}_D[\mathrm{KL}\big(P_{Z|D}\,\big\|\,P_Z\big)] = I(Z;D)$. One can also extend Theorem 25 to accommodate the representation space covariate shift through similar procedures as the proof of Theorem 12. Therefore, these high-probability generalization bounds also motivate the design of new DG algorithms via gradient and representation matching techniques. However, Theorem 24 and 25 do not necessarily converge to 0 even if $\imath(W;D_i) \to 0$ and $\mathrm{KL}\big(P_{Z|D}\,\big\|\,P_Z\big) \to 0$ because of the existence of an additive high-probability term. In contrast, Theorem 8 and 11 guarantee that $L_s(W) = L(W)$ or $L_t(w) = L(w)$ once $I(W;D_i) = 0$ or $I(Z;D) = 0$ is satisfied.

VIII. Conclusion

In this work, we explore a novel perspective for DG by minimizing the domain-level generalization gap with high probability, which facilitates information-theoretic analysis for the generalization behavior of learning algorithms. Our analysis sheds light on understanding how gradient or representation matching enhances generalization and unveils the complementary relationship between these two elements. These theoretical insights inspire us to design the IDM algorithm by simultaneously aligning inter-domain gradients and representations, which then achieves superior performance on the DomainBed benchmark.

APPENDIX A
PREREQUISITE DEFINITIONS AND LEMMAS

Definition 26. (Sub-Gaussian) A random variable X is σ-subGaussian if for any $\rho \in \mathbb{R}$, $\mathbb{E}[\exp(\rho(X - \mathbb{E}[X]))] \leq \exp(\rho^2\sigma^2/2)$.

Definition 27. (Kullback-Leibler Divergence) Let P and Q be probability measures on the same space $\mathcal{X}$, the KL divergence from P to Q is defined as $\mathrm{KL}(P \,\|\, Q) \triangleq \int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \,\mathrm{d}x$.

Definition 28. (Mutual Information) Let (X, Y) be a pair of random variables with values over the space $\mathcal{X} \times \mathcal{Y}$. Let their joint distribution be $P_{X,Y}$ and the marginal distributions be P_X and P_Y respectively, the mutual information between X and Y is defined as $I(X;Y) = \mathrm{KL}(P_{X,Y} \,\|\, P_X P_Y)$.

Definition 29. (Wasserstein Distance) Let $c(\cdot,\cdot)$ be a metric and let P and Q be probability measures on $\mathcal{X}$. Denote $\Gamma(P,Q)$ as the set of all couplings of P and Q (i.e. the set of all joint distributions on $\mathcal{X} \times \mathcal{X}$ with two marginals being P and Q), then the Wasserstein distance of order p between P and Q is defined as $\mathbb{W}_p(P,Q) \triangleq \left(\inf_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(x,x')^p \,\mathrm{d}\gamma(x,x')\right)^{1/p}$.

Unless otherwise noted, we use $\log$ to denote the logarithmic function with base e, and use $\mathbb{W}(\cdot,\cdot)$ to denote the Wasserstein distance of order 1.

Definition 30. (Total Variation) The total variation between two probability measures P and Q is $\mathrm{TV}(P,Q) \triangleq \sup_E |P(E) - Q(E)|$, where the supremum is over all measurable set E.

Lemma 31 (Lemma 1 in [45])**.** *Let (X,Y) be a pair of random variables with joint distribution $P_{X,Y}$ and let $\bar{Y}$ be an independent copy of Y. If $f(x,y)$ is a measurable function such that $E_{X,Y}[f(X,Y)]$ exists and $f(X,\bar{Y})$ is σ-subGaussian, then*

$$\left|\mathbb{E}_{X,Y}[f(X,Y)] - \mathbb{E}_{X,\bar{Y}}[f(X,\bar{Y})]\right| \leq \sqrt{2\sigma^2 I(X;Y)}.$$

Furthermore, if $f(x,Y)$ is σ-subGaussian for each x and the expectation below exists, then

$$\mathbb{E}_{X,Y}\left[\left(f(X,Y) - \mathbb{E}_{\bar{Y}}[f(X,\bar{Y})]\right)^2\right] \leq 4\sigma^2(I(X;Y) + \log 3),$$

and for any $\epsilon > 0$, we have

$$\mathbb{P}\left\{\left|f(X,Y) - \mathbb{E}_{\bar{Y}}[f(X,\bar{Y})]\right| \geq \epsilon\right\} \leq \frac{4\sigma^2(I(X;Y) + \log 3)}{\epsilon^2}.$$

Lemma 32 (Lemma 2 in [45])**.** *Let X be σ-subGaussian and $\mathbb{E}[X] = 0$, then for any $\lambda \in [0, 1/4\sigma^2)$:*

$$\mathbb{E}_X\left[e^{\lambda X^2}\right] \leq 1 + 8\lambda\sigma^2.$$

Lemma 33. *(Donsker-Varadhan formula) Let P and Q be probability measures defined on the same measurable space, where P is absolutely continuous with respect to Q. Then*

$$\mathrm{KL}(P \,\|\, Q) = \sup_X \left\{\mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X]\right\},$$

where X is any random variable such that e^X is Q-integrable and $\mathbb{E}_P[X]$ exists.

Lemma 34. *Let P, and Q be probability measures defined on the same measurable space. Let $X \sim P$ and $X' \sim Q$. If $f(X)$ is σ-subGaussian w.r.t X and the following expectations exists, then*

$$|\mathbb{E}_{X'}[f(X')] - \mathbb{E}_X[f(X)]| \leq \sqrt{2\sigma^2 \mathrm{KL}(Q \,\|\, P)},$$
$$\mathbb{E}_{X'}\left[\left(f(X') - \mathbb{E}_X[f(X)]\right)^2\right] \leq 4\sigma^2(\mathrm{KL}(Q \,\|\, P) + \log 3).$$

Furthermore, by combining the results above and Markov's inequality, we have that for any $\epsilon > 0$:

$$\mathbb{P}\{|f(X') - \mathbb{E}_X[f(X)]| \geq \epsilon\} \leq \frac{4\sigma^2}{\epsilon^2}(\mathrm{KL}(Q \,\|\, P) + \log 3).$$

Proof. Let $\lambda \in \mathbb{R}$ be any non-zero constant, then by the subGaussian property of $f(X)$:

$$\log \mathbb{E}_X\left[e^{\lambda(f(X) - \mathbb{E}_X[f(X)])}\right] \leq \frac{\lambda^2\sigma^2}{2},$$
$$\log \mathbb{E}_X\left[e^{\lambda f(X)}\right] - \lambda \mathbb{E}_X[f(X)] \leq \frac{\lambda^2\sigma^2}{2}.$$

By applying Lemma 33 with $X = \lambda f(X)$ we have

$$\begin{aligned}
\mathrm{KL}(Q \,\|\, P) &\geq \sup_\lambda \left\{\mathbb{E}_{X'}[\lambda f(X')] - \log \mathbb{E}_X\left[e^{\lambda f(X)}\right]\right\} \\
&\geq \sup_\lambda \left\{\mathbb{E}_{X'}[\lambda f(X')] - \lambda \mathbb{E}_X[f(X)] - \frac{\lambda^2\sigma^2}{2}\right\} \\
&= \frac{1}{2\sigma^2}(\mathbb{E}_{X'}[f(X')] - \mathbb{E}_X[f(X)])^2,
\end{aligned}$$

where the supremum is taken by setting $\lambda = \frac{1}{\sigma^2}(\mathbb{E}_{X'}[f(X')] - \mathbb{E}_X[f(X)])$. This completes the proof of the first inequality.

To prove the second inequality, let $g(x) = (f(x) - \mathbb{E}_X[f(X)])^2$ and $\lambda \in [0, 1/4\sigma^2)$. Apply Lemma 33 again with $X = \lambda g(X)$, we have

$$\begin{aligned}
\mathrm{KL}(Q \,\|\, P) &\geq \sup_\lambda \left\{\mathbb{E}_{X'}[\lambda g(X')] - \log \mathbb{E}_X\left[e^{\lambda g(X)}\right]\right\} \\
&= \sup_\lambda \left\{\mathbb{E}_{X'}\left[\lambda(f(X') - \mathbb{E}_X[f(X)])^2\right]\right. \\
&\qquad \left. - \log \mathbb{E}_X\left[e^{\lambda(f(X) - \mathbb{E}_X[f(X)])^2}\right]\right\} \\
&\geq \sup_\lambda \left\{\mathbb{E}_{X'}\left[\lambda(f(X') - \mathbb{E}_X[f(X)])^2\right]\right. \\
&\qquad \left. - \log(1 + 8\lambda\sigma^2)\right\} \\
&\geq \frac{1}{4\sigma^2}\mathbb{E}_{X'}\left[\left(f(X') - \mathbb{E}_X[f(X)]\right)^2\right] - \log 3,
\end{aligned}$$

where the second inequality follows by applying Lemma 32 and the last inequality follows by taking $\lambda \to \frac{1}{4\sigma^2}$. This finishes the proof of the second inequality.

Furthermore, by applying Markov's inequality, we can get:

$$\begin{aligned}
&\mathbb{P}\{|f(X') - \mathbb{E}_X[f(X)]| \geq \epsilon\} \\
&= \mathbb{P}\left\{\left(f(X') - \mathbb{E}_X[f(X)]\right)^2 \geq \epsilon^2\right\} \\
&\leq \frac{1}{\epsilon^2}\mathbb{E}_{X'}\left[\left(f(X') - \mathbb{E}_X[f(X)]\right)^2\right]
\end{aligned}$$

$$\leq \frac{4\sigma^2}{\epsilon^2}(\mathrm{KL}(Q \,\|\, P) + \log 3),$$

which completes the proof. □

Lemma 35 (Proposition 5.2 in [63])**.** *Assume that almost surely under P_Y, $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is a function satisfying $\mathbb{E}_{P_{X|Y}}[f(X,Y)] < \infty$ and $P_{X|Y} \ll P_X$. Let $\bar{Y}$ be an independent copy of Y. Then with probability at least $1-\delta$,*

$$\mathbb{E}_{X|Y}[f(X,Y)] \leq \log \mathbb{E}_{X,\bar{Y}}\left[\frac{e^{f(X,\bar{Y})}}{\delta}\right] + \mathrm{KL}(P_{X|Y} \,\|\, P_X).$$

Lemma 36 (Proposition 5.10 in [63])**.** *Assume that $P_{X,Y} \ll P_X P_Y$ and $P_X P_Y \ll P_{X,Y}$. Let $\bar{Y}$ be an independent copy of Y. Then for any function $f(\cdot,\cdot)$, with probability at least $1-\delta$,*

$$f(X,Y) \leq \log \mathbb{E}_{X,\bar{Y}}\left[\frac{e^{f(X,\bar{Y})}}{\delta}\right] + \imath(X;Y),$$

where $\imath(X;Y) = \log \frac{\mathrm{d}P_{X,Y}}{\mathrm{d}P_X P_Y}$ is the information density.

Lemma 37 (Proposition 3.25 in [63])**.** *Let X be a σ-subGaussian variable and let $S = \frac{1}{n}\sum_{i=1}^n X_i$ be the average of n independent instances of X. Then for any $\lambda \in [0,1)$,*

$$\mathbb{E}\left[e^{\frac{n\lambda(S-\mathbb{E}[x])^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}}.$$

Lemma 38. *(Kantorovich-Rubinstein Duality) Let P and Q be probability measures defined on the same measurable space $\mathcal{X}$, then*

$$\mathbb{W}(P,Q) = \sup_{f \in Lip_1}\left\{\int_{\mathcal{X}} f \,\mathrm{d}P - \int_{\mathcal{X}} f \,\mathrm{d}Q\right\},$$

where Lip_1 denotes the set of 1-Lipschitz functions in the metric c, i.e. $|f(x) - f(x')| \leq c(x,x')$ for any $f \in Lip_1$ and $x, x' \in \mathcal{X}$.

Lemma 39. *(Pinsker's Inequality) Let P and Q be probability measures defined on the same space, then* $\mathrm{TV}(P,Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(Q \,\|\, P)}$.

Lemma 40 (Theorem 4.1 in [64])**.** *For any Q_Y, we have* $\mathbb{E}_X[\mathrm{KL}(P_{Y|X} \,\|\, Q_Y)] = I(X;Y) + \mathrm{KL}(P_Y \,\|\, Q_Y)$.

Proposition 41. *For any constant $\lambda \in (0,1)$, we have*

$$\begin{aligned}&\mathbb{P}\{|L_s(W) - L_t(W)| \geq \epsilon\}\\ &\quad\leq \mathbb{P}\{|L_s(W) - L(W)| \geq \lambda\epsilon\}\\ &\qquad + \mathbb{P}\{|L_t(W) - L(W)| \geq (1-\lambda)\epsilon\}.\end{aligned}$$

Proof. Notice that $|L_s(W) - L(W)| \leq \lambda\epsilon$ and $|L_t(W) - L(W)| \leq (1-\lambda)\epsilon$ together implies $|L_s(W) - L_t(W)| \leq \epsilon$, we then have

$$\begin{aligned}&\mathbb{P}\{|L_s(W) - L_t(W)| \leq \epsilon\}\\ &\geq \mathbb{P}\{|L_s(W) - L(W)| \leq \lambda\epsilon \wedge |L_t(W) - L(W)| \leq (1-\lambda)\epsilon\}.\end{aligned}$$

This implies that

$$\begin{aligned}&\mathbb{P}\{|L_s(W) - L_t(W)| \geq \epsilon\}\\ &\leq \mathbb{P}\{|L_s(W) - L(W)| \geq \lambda\epsilon \vee |L_t(W) - L(W)| \geq (1-\lambda)\epsilon\}.\end{aligned}$$

By applying Boole's inequality, we then have

$$\begin{aligned}&\mathbb{P}\{|L_s(W) - L_t(W)| \geq \epsilon\}\\ &\quad\leq \mathbb{P}\{|L_s(W) - L(W)| \geq \lambda\epsilon\}\\ &\qquad + \mathbb{P}\{|L_t(W) - L(W)| \geq (1-\lambda)\epsilon\}.\end{aligned}$$

The proof is complete. □

APPENDIX B
OMITTED PROOFS

Proof of Proposition 7. For any $x \in \mathcal{X}$, by applying Lemma 40 with $X = D|_{X=x}$ and $Y = Y|_{X=x}$, we have

$$\begin{aligned}&\mathbb{E}_{D|X=x}[\mathrm{KL}(P_{Y|D,X=x} \,\|\, Q_{Y|X=x})]\\ &\quad= I(D;Y|X=x) + \mathrm{KL}(P_{Y|X=x} \,\|\, Q_{Y|X=x})\\ &\qquad\geq I(D;Y|X=x).\end{aligned}$$

The last inequality is by the positiveness of the KL divergence. It holds with equality if and only if $Q_{Y|X} = P_{Y|X}$. Taking expectation over X, we then have

$$\mathbb{E}_{D,X}[\mathrm{KL}(P_{Y|D,X} \,\|\, Q_{Y|X})] \geq I(D;Y|X).$$ □

Proof of Theorem 8. For any $D \in D_s$, one can verify that $|L_{\bar{D}}(W) - L(W)| \in [0,M]$ and is thus $\frac{M}{2}$-subGaussian. By applying Lemma 31 with $X = W$, $Y = D$ and $f(W,D) = |L_D(W) - L(W)|$, we obtain

$$\begin{aligned}&\mathbb{E}_{W,D}|L_D(W) - L(W)| - \mathbb{E}_{W,\bar{D}}|L_{\bar{D}}(W) - L(W)|\\ &\qquad\leq \sqrt{\frac{M^2}{2}I(W;D)}.\end{aligned}$$

Summing up this inequality over each source domain, we then have

$$\begin{aligned}&\mathbb{E}_{W,D_s}|L_s(W) - L(W)|\\ &= \mathbb{E}_{W,D_s}\left|\frac{1}{m}\sum_{i=1}^m L_{D_i}(W) - L(W)\right|\\ &\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W,D_i}|L_{D_i}(W) - L(W)|\\ &\leq \frac{1}{m}\sum_{i=1}^m \sqrt{\frac{M^2}{2}I(W;D_i)} + \mathbb{E}_{W,\bar{D}}|L_{\bar{D}}(W) - L(W)|.\end{aligned}$$

By applying Markov's inequality, we finally have

$$\begin{aligned}\mathbb{P}\{|L_s(W) - L(W)| \geq \epsilon\} &\leq \frac{M}{m\epsilon\sqrt{2}}\sum_{i=1}^m \sqrt{I(W;D_i)}\\ &\quad + \frac{1}{\epsilon}\mathbb{E}_{W,\bar{D}}|L_{\bar{D}}(W) - L(W)|.\end{aligned}$$

Additionally, by assuming that the source domains are independent, we have

$$\begin{aligned}I(W;D_s) &= I(W;\{D_i\}_{i=1}^m)\\ &= I(W;D_1) + I(W;\{D_i\}_{i=2}^m|D_1)\\ &= I(W;D_1) + I(W;\{D_i\}_{i=2}^m)\\ &\quad - I(\{D_i\}_{i=2}^m;D_1) + I(\{D_i\}_{i=2}^m;D_1|W)\\ &= I(W;D_1) + I(W;\{D_i\}_{i=2}^m)\end{aligned}$$

$$
\begin{aligned}
&\quad + I(\{D_i\}_{i=2}^m; D_1|W) \\
&\geq I(W; D_1) + I(W; \{D_i\}_{i=2}^m) \\
&\geq \cdots \\
&\geq \sum_{i=1}^m I(W; D_i). \qquad \square
\end{aligned}
$$

Proof of Theorem 10. For any $D_i \in D_s$, let $P = P_{W|D_i=d}$, $Q = P_W$ and $f(w) = L_{D_i}(w)$ in Lemma 38, then

$$
\begin{aligned}
&|\mathbb{E}_{W,D_s}[L_s(W)] - \mathbb{E}_W[L(W)]| \\
&\leq \frac{1}{m}\mathbb{E}_{D_s}\left[\sum_{i=1}^m \left|\mathbb{E}_{W|D_i}[L_{D_i}(W)] - \mathbb{E}_W[L(W)]\right|\right] \\
&= \frac{1}{m}\mathbb{E}_{D_s}\left[\sum_{i=1}^m |\mathbb{E}_{W|D_i}[L_{D_i}(W)] - \mathbb{E}_W[L_{D_i}(W)]|\right] \\
&\leq \frac{1}{m}\mathbb{E}_{D_s}\left[\sum_{i=1}^m \beta' \mathbb{W}(P_{W|D_i}, P_W)\right] \\
&= \frac{\beta'}{m}\sum_{i=1}^m \mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)].
\end{aligned}
$$

When the metric d is discrete, the Wasserstein distance is equal to the total variation. Combining with Lemma 39, we have the following reductions:

$$
\begin{aligned}
\mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)] &= \mathbb{E}_{D_i}[\mathrm{TV}(P_{W|D_i}, P_W)] \\
&\leq \mathbb{E}_{D_i}\left[\sqrt{\frac{1}{2}\mathrm{KL}(P_{W|D_i} \,\|\, P_W)}\right] \\
&\leq \sqrt{\frac{1}{2} I(W; D_i)},
\end{aligned}
$$

where the last inequality follows by applying Jensen's inequality on the concave square root function. $\square$

Proof of Theorem 11. By the identical marginal distribution of the target domains $\mathcal{D}_t = \{D'_k\}_{k=1}^{m'}$, we have

$$
\begin{aligned}
\mathbb{E}_{D_t}[L_t(w)] &= \frac{1}{m'}\sum_{k=1}^{m'} \mathbb{E}_{D'_k}[L_{D'_k}(w)] = \frac{1}{m'}\sum_{k=1}^{m'} \mathbb{E}_D[L_D(w)] \\
&= \mathbb{E}_D[L_D(w)] = L(w).
\end{aligned}
$$

For any $d \in \mathcal{D}$, by applying Lemma 34 with $P = P_Z$, $Q = P_{Z|D=d}$ and $f(Z) = \ell(f_w(X), Y)$, we can get

$$
\begin{aligned}
|L_d(w) - L(w)| &= \left|\mathbb{E}_{Z|D=d}[\ell(f_w(X), Y)] - \mathbb{E}_Z[\ell(f_w(X), Y]\right| \\
&\leq \sqrt{2\sigma^2 \mathrm{KL}(P_{Z|D=d} \,\|\, P_Z)}.
\end{aligned}
$$

Taking the expectation over $D \sim \nu$, we can get

$$
\begin{aligned}
\mathbb{E}_D|L_d(w) - L(w)| &\leq \mathbb{E}_D\sqrt{2\sigma^2 \mathrm{KL}(P_{Z|D} \,\|\, P_Z)} \\
&\leq \sqrt{2\sigma^2 \mathbb{E}_D[\mathrm{KL}(P_{Z|D} \,\|\, P_Z)]} \\
&= \sigma\sqrt{2I(Z; D)}.
\end{aligned}
$$

By summing up the inequality above over each target domain, we can get

$$
\begin{aligned}
\mathbb{E}_{D_t}|L_t(w) - L(w)| &= \mathbb{E}_{D_t}\left|\frac{1}{m'}\sum_{k=1}^{m'} L_{D'_k}(w) - L(W)\right| \\
&\leq \frac{1}{m'}\sum_{k=1}^{m'} \mathbb{E}_{D'_k}\left|L_{D'_k}(w) - L(W)\right| \\
&\leq \frac{1}{m'}\sum_{k=1}^{m'} \sigma\sqrt{2I(Z; D'_k)} \\
&= \sigma\sqrt{2I(Z; D)}.
\end{aligned}
$$

By applying Markov's inequality, we finally have

$$
\mathbb{P}\{|L_t(w) - L(w)| \geq \epsilon\} \leq \frac{\sigma}{\epsilon}\sqrt{2I(Z; D)}. \qquad \square
$$

Proof of Theorem 12. For any domain $d \in \mathcal{D}$, classifier ψ and $f^* : \mathcal{R} \mapsto \mathcal{Y}$, denote

$$
\begin{aligned}
L_d(\psi, f^*) &= \mathbb{E}_{R|D=d}[\ell(f_\psi(R), f^*(R))], \\
L(\psi, f^*) &= \mathbb{E}_R[\ell(f_\psi(R), f^*(R))].
\end{aligned}
$$

By setting $P = P_R$, $Q = P_{R|D=d}$ and $f(R) = \ell(f_\psi(R), f^*(R))$ and applying Lemma 34, we have

$$
|L_d(\psi, f^*) - L(\psi, f^*)| \leq \sqrt{2\sigma^2 \mathrm{KL}(P_{R|D=d} \,\|\, P_R)}.
$$

By taking the expectation over $D \sim \nu$, we get

$$
\begin{aligned}
\mathbb{E}_D|L_D(\psi, f^*) - L(\psi, f^*)| &\leq \mathbb{E}_D\sqrt{2\sigma^2 \mathrm{KL}(P_{R|D} \,\|\, P_R)} \\
&\leq \sqrt{2\sigma^2 I(R; D)}.
\end{aligned}
$$

If Assumption 4 holds, then by the symmetry and triangle inequality of $\ell(\cdot, \cdot)$, we have

$$
\begin{aligned}
L_d(\psi, f^*) &= \mathbb{E}_{R|D=d}[\ell(f_\psi(R), f^*(R))] \\
&\leq \mathbb{E}_{R,Y|D=d}[\ell(f_\psi(R), Y) + \ell(f^*(R), Y)] \\
&= L_d(\psi) + L_d(f^*). \\
L(\psi, f^*) &\leq L(\psi) + L(f^*).
\end{aligned}
$$

Similarly, we can prove that

$$
\begin{aligned}
L_d(\psi, f^*) &= \mathbb{E}_{R|D=d}[\ell(f_\psi(R), f^*(R))] \\
&\geq \mathbb{E}_{R,Y|D=d}[\ell(f_\psi(R), Y) - \ell(f^*(R), Y)] \\
&= L_d(\psi) - L_d(f^*). \\
L(\psi, f^*) &\geq L(\psi) - L(f^*).
\end{aligned}
$$

Combining the results above, we have

$$
\begin{aligned}
L_d(\psi) - L(\psi) &\leq L_d(\psi, f^*) + L_d(f^*) - L(\psi, f^*) + L(f^*), \\
L(\psi) - L_d(\psi) &\leq L(\psi, f^*) + L(f^*) - L_d(\psi, f^*) + L_d(f^*).
\end{aligned}
$$

Combining the two inequalities above, we then get

$$
|L_d(\psi) - L(\psi)| \leq |L_d(\psi, f^*) - L(\psi, f^*)| + L_d(f^*) + L(f^*).
$$

By taking the expectation over $D \sim \nu$, we obtain

$$
\begin{aligned}
\mathbb{E}_D|L_D(\psi) - L(\psi)| &\leq \mathbb{E}_D|L_D(\psi, f^*) - L(\psi, f^*)| \\
&\quad + \mathbb{E}_D[L_D(f^*) + L(f^*)] \\
&\leq \sqrt{2\sigma^2 I(R; D)} + 2L(f^*).
\end{aligned}
$$

Summing up the inequality above over each target domain, we can get

$$
\mathbb{E}_{D_t}|L_t(\psi) - L(\psi)| \leq \mathbb{E}_{D_t}\left|\frac{1}{m'}\sum_{k=1}^{m'} L_{D'_k}(\psi) - L(\psi)\right|
$$

$$\leq \frac{1}{m'}\sum_{k=1}^{m'}\mathbb{E}_{D'_k}\Big|L_{D'_k}(\psi)-L(\psi)\Big|$$
$$\leq \frac{1}{m'}\sum_{k=1}^{m'}\Big(\sqrt{2\sigma^2 I(R;D)}+2L(f^*)\Big)$$
$$=\sqrt{2\sigma^2 I(R;D)}+2L(f^*).$$

By applying Markov's inequality, we finally have

$$\mathbb{P}\{L_t(\psi)-L(\psi)\geq\epsilon\}\leq\frac{\sigma}{\epsilon}\sqrt{2I(R;D)}+\frac{2}{\epsilon}L(f^*).$$

The proof is complete by taking the minimizer of $\min_{f^*}[L(f^*)]$. □

Proof of Proposition 13. The condition $P_{R,D}\ll P_{R_i,D_i}$ and $P_{R_i,D_i}\ll P_{R,D}$ implies that there exists $B>1$, such that for any $r\in\mathcal{R}$, $d\in\mathcal{D}$, we have $\frac{P_{R,D}(r,d)}{P_{R_i,D_i}(r,d)}\in[\frac{1}{B},B]$. Therefore, $\log\frac{P_{R,D}}{P_{R_i,D_i}}\in[-\log(B),\log(B)]$ and is $\log(B)$-subGaussian.

By applying Lemma 31 with $X=W$, $Y=D_i$ and $f(W,D)=\mathbb{E}_{X|D}[\log\frac{P_{R,D}(f_\Phi(X),D)}{P_{R_i,D_i}(f_\Phi(X),D)}]$, we have

$$\begin{aligned}&\mathrm{SKL}(P_{R,D}\,\|\,P_{R_i,D_i})\\&=|\mathrm{KL}(P_{R,D}\,\|\,P_{R_i,D_i})+\mathrm{KL}(P_{R_i,D_i}\,\|\,P_{R,D})|\\&=\left|\mathbb{E}_{W,D_i,R_i}\left[\log\frac{P_{R,D}(R_i,D_i)}{P_{R_i,D_i}(R_i,D_i)}\right]\right.\\&\qquad\left.-\mathbb{E}_{W,D,R}\left[\log\frac{P_{R,D}(R,D)}{P_{R_i,D_i}(R,D)}\right]\right|\\&\leq\sqrt{2\log^2(B)I(W;D_i)}.\end{aligned}$$ □

Proof of Theorem 9. Noticing the Markov chain relationship $D_i\to(W_{T-1},G_T)\to W_{T-1}+G_T$, then by the data processing inequality

$$\begin{aligned}I(W_T;D_i)&=I(W_{T-1}+G_T;D_i)\\&\leq I(W_{T-1},G_T;D_i)\\&=I(W_{T-1};D_i)+I(G_T;D_i|W_{T-1}).\end{aligned}$$

where the last equality is by the chain rule of conditional mutual information. By applying the reduction steps above recursively, we can get

$$\begin{aligned}I(W_T;D_i)&\leq I(W_{T-1};D_i)+I(G_T;D_i|W_{T-1})\\&\leq I(W_{T-2};D_i)+I(G_{T-1};D_i|W_{T-2})\\&\qquad+I(G_T;D_i|W_{T-1})\\&\leq\cdots\\&\leq\sum_{t=1}^{T}I(G_t;D_i|W_{t-1}).\end{aligned}$$

When the source domains are independent, we additionally have

$$\begin{aligned}I(W_T;D_i)&\leq I(W_{T-1};D_i)+I(G_T;D_i|W_{T-1})\\&\leq I(W_{T-1};D_i)+I\big(\{G_T^k\}_{k=1}^m;D_i|W_{T-1}\big)\\&=I(W_{T-1};D_i)+I\big(G_T^i;D_i|W_{T-1}\big)\\&\qquad+I\big(\{G_T^k\}_{k=1}^m\setminus G_T^i;D_i|W_{T-1},G_T^i\big)\\&=I(W_{T-1};D_i)+I\big(G_T^i;D_i|W_{T-1}\big).\end{aligned}$$

Then by the same scheme of recursive reduction, we can prove that

$$I(W_T;D_i)\leq\sum_{t=1}^{T}I\big(G_t^i;D_i|W_{t-1}\big).$$ □

Proof of Theorem 14. Without loss of generality, we assume that the data-generating distributions $p(X|D)$ are Gaussian with zero means for simplicity, i.e.

$$p(x|D=d)=\frac{1}{\sqrt{(2\pi)^n|\Sigma_d|}}\exp\left(-\frac{1}{2}x^\top\Sigma_d x\right),$$

where Σ_d is the corresponding covariance matrix of domain d. Let $X\in\mathbb{R}^{n\times b}$ be the data matrix of S such that the i-th column of X equals x_i, we then have

$$p(S=s|D=d)=\frac{1}{\sqrt{(2\pi)^{bn}|\Sigma_d|^b}}\exp\left(-\frac{1}{2}\mathrm{tr}(X^\top\Sigma_d X)\right).$$

Since the rank of X is at most b, one can decompose $\Sigma_d=\Sigma_d^1+\Sigma_d^2$ with $\mathrm{rank}(\Sigma_d^1)=b$ and $\mathrm{rank}(\Sigma_d^2)=n-b\geq 2$ through eigenvalue decomposition, and let the eigenvector space of Σ_d^1 cover the column space of X. Then we have

$$\mathrm{tr}(X^\top\Sigma_d^1X)=\mathrm{tr}(X^\top\Sigma_dX),\quad\text{and}\quad\mathrm{tr}(X^\top\Sigma_d^2X)=0.$$

Therefore, one can arbitrarily modify the eigenvector space of Σ_d^2 as long as keeping it orthogonal to that of Σ_d^1, without changing the value of $\mathrm{tr}(X^\top\Sigma_dX)$. This finishes the proof of the first part.

To prove the second part, similarly we decompose Σ_d by $\Sigma_d^1+\Sigma_d^2$ such that $\mathrm{rank}(\Sigma_d^1)=2b+1$ and $\mathrm{rank}(\Sigma_d^2)=n-2b-1\geq 1$, and make the eigenvector space of Σ_d^1 cover the column space of both X_1 and X_2, where X_1 and X_2 are the data matrix of S_1 and S_2 respectively. We then have

$$\begin{aligned}\mathrm{tr}(X_1^\top\Sigma_d^1X_1)&=\mathrm{tr}(X_1^\top\Sigma_dX_1),\\\mathrm{tr}(X_2^\top\Sigma_d^1X_2)&=\mathrm{tr}(X_2^\top\Sigma_dX_2),\\\text{and}\quad\mathrm{tr}(X_1^\top\Sigma_d^2X_1)&=\mathrm{tr}(X_2^\top\Sigma_d^2X_2)=0.\end{aligned}$$

Let $\Sigma_d^1=U_d^\top\Lambda_dU_d$ be the eigenvalue decomposition of Σ_d^1, where $U_d\in\mathbb{R}^{(2b+1)\times n}$ and $\Lambda_d=\mathrm{diag}(\lambda_1^d,\cdots,\lambda_{2b+1}^d)$. Notice that for any $x\in\mathbb{R}^n$, we have $x^\top\Sigma_dx=(U_dx)^\top\Lambda_d(U_dx)=\sum_{i=1}^{2b+1}(U_dx)_i^2\lambda_i$. By assuming that $p(S=s_1|D=d)=p(S=s_2|D=d)$, we have the following homogeneous linear equations:

$$\begin{aligned}a_1^1\lambda_1+a_1^2\lambda_2+\cdots+a_1^{2b+1}\lambda_{2b+1}&=0,\\a_2^1\lambda_1+a_2^2\lambda_2+\cdots+a_2^{2b+1}\lambda_{2b+1}&=0,\\&\cdots\\a_b^1\lambda_1+a_b^2\lambda_2+\cdots+a_b^{2b+1}\lambda_{2b+1}&=0,\end{aligned}$$

where $a_i^j=(U_dx_i^1)_j^2-(U_dx_i^2)_j^2$. Since $2b+1>b$, the linear system above has infinite non-zero solutions, which finishes the proof of the second part. □

Proof of Theorem 15. Recall that $\bar{p}(x)=\frac{1}{b}\sum_{i=1}^b p_i(x)$ and $\bar{q}(x)=\frac{1}{b}\sum_{i=1}^b q_i(x)$, we then have

$$\mathrm{KL}(\bar{P}\,\|\,\bar{Q})=\int_{\mathcal{X}}\bar{p}(x)\log\left(\frac{\bar{p}(x)}{\bar{q}(x)}\right)\mathrm{d}x$$

$$
\begin{aligned}
&= -\int_{\mathcal{X}} \bar{p}(x) \log\left(\frac{1}{b}\sum_{i=1}^{b} \frac{p_i(x)}{\bar{p}(x)} \cdot \frac{q_{f(i)}(x)}{p_i(x)}\right) \mathrm{d}x \\
&\le -\int_{\mathcal{X}} \bar{p}(x) \frac{1}{b}\sum_{i=1}^{b} \frac{p_i(x)}{\bar{p}(x)} \log\left(\frac{q_{f(i)}(x)}{p_i(x)}\right) \mathrm{d}x \\
&= -\frac{1}{b}\sum_{i=1}^{b} \int_{\mathcal{X}} p_i(x) \log\left(\frac{q_{f(i)}(x)}{p_i(x)}\right) \mathrm{d}x \\
&= \frac{1}{b}\sum_{i=1}^{b} \mathrm{KL}(P_i \,\|\, Q_{f(i)}),
\end{aligned}
$$

where the only inequality follows by applying Jensen's inequality on the concave logarithmic function. This finishes the proof of the upper bound for KL divergence.

To prove the counterpart for Wasserstein distance, we apply Lemma 38 on $\bar{P}$ and $\bar{Q}$:

$$
\begin{aligned}
&\mathbb{W}(\bar{P}, \bar{Q}) \\
&= \sup_{f \in Lip_1} \left\{ \int_{\mathcal{X}} f \,\mathrm{d}\bar{P} - \int_{\mathcal{X}} f \,\mathrm{d}\bar{Q} \right\} \\
&= \sup_{f \in Lip_1} \left\{ \int_{\mathcal{X}} f \,\mathrm{d}\left(\frac{1}{b}\sum_{i=1}^{b} P_i\right) - \int_{\mathcal{X}} f \,\mathrm{d}\left(\frac{1}{b}\sum_{i=1}^{b} Q_{f(i)}\right) \right\} \\
&\le \frac{1}{b}\sum_{i=1}^{b} \sup_{f \in Lip_1} \left\{ \int_{\mathcal{X}} f \,\mathrm{d}P_i - \int_{\mathcal{X}} f \,\mathrm{d}Q_{f(i)} \right\} \\
&= \frac{1}{b}\sum_{i=1}^{b} \mathbb{W}(P_i, Q_{f(i)}).
\end{aligned}
$$

The proof is complete. □

Proof of Theorem 16. For simplicity, we assume that all data points of $\{x_i^1\}_{i=1}^b$ and $\{x_i^2\}_{i=1}^b$ are different from each other. Since P_i and Q_i are Gaussian distributions with the same variance, the KL divergence and Wasserstein distance between them could be analytically acquired:

$$
\mathrm{KL}(P_i \,\|\, Q_j) = \frac{(x_i^1 - x_j^2)^2}{2\sigma^2}, \quad \text{and} \quad \mathbb{W}(P_i, Q_j) = |x_i^1 - x_j^2|.
$$

Suppose there exists $i \in [1, b]$ such that $f(i) \neq i$. Without loss of generality, we assume that $f(i) > i$. Then by the pigeonhole principle, there exists $j \in (i, b]$ that satisfies $f(j) < f(i)$. Suppose that $\{x_i^1\}_{i=1}^b$, $\{x_i^2\}_{i=1}^b$ are both sorted in ascending order, we have $x_i^1 < x_j^1$ and $x_{f(i)}^2 > x_{f(j)}^2$. For any $p \in \{1, 2\}$, the following 3 cases cover all possible equivalent combinations of the order of x_i^1, x_j^1, $x_{f(j)}^2$ and $x_{f(i)}^2$:

- When $x_i^1 < x_j^1 < x_{f(j)}^2 < x_{f(i)}^2$ and $p = 2$, we have

$$
\begin{aligned}
&(x_i^1 - x_{f(i)}^2)^2 + (x_j^1 - x_{f(j)}^2)^2 \\
&\quad - (x_i^1 - x_{f(j)}^2)^2 - (x_j^1 - x_{f(i)}^2)^2 \\
&= (2x_i^1 - x_{f(i)}^2 - x_{f(j)}^2)(x_{f(j)}^2 - x_{f(i)}^2) \\
&\quad - (2x_j^1 - x_{f(j)}^2 - x_{f(i)}^2)(x_{f(j)}^2 - x_{f(i)}^2) \\
&= (x_{f(j)}^2 - x_{f(i)}^2)(2x_i^1 - 2x_j^1) > 0.
\end{aligned}
$$

  Elsewise when $p = 1$, we have

$$
|x_i^1 - x_{f(i)}^2| + |x_j^1 - x_{f(j)}^2| = |x_i^1 - x_{f(j)}^2| + |x_j^1 - x_{f(i)}^2|.
$$

- When $x_i^1 < x_{f(j)}^2 < x_j^1 < x_{f(i)}^2$, we have

$$
|x_i^1 - x_{f(i)}^2|^p > |x_i^1 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p.
$$

- When $x_i^1 < x_{f(j)}^2 < x_{f(i)}^2 < x_j^1$, we have

$$
\begin{aligned}
&|x_i^1 - x_{f(i)}^2|^p + |x_j^1 - x_{f(j)}^2|^p \\
&\ge |x_i^1 - x_{f(j)}^2|^p + |x_{f(i)}^2 - x_{f(j)}^2|^p \\
&\quad + |x_j^1 - x_{f(i)}^2|^p + |x_{f(i)}^2 - x_{f(j)}^2|^p \\
&> |x_i^1 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p.
\end{aligned}
$$

In conclusion, under all possible circumstances, we have $|x_i^1 - x_{f(i)}^2|^p + |x_j^1 - x_{f(j)}^2|^p \ge |x_i^1 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p$, which implies that by setting $f'(i) = f(j)$, $f'(j) = f(i)$ and $f'(k) = f(k)$ for $k \notin \{i, j\}$, f' will be a better choice over f to minimize $\mathrm{KL}(\bar{P} \,\|\, \bar{Q})$ or $\mathbb{W}(\bar{P}, \bar{Q})$. The proof is complete since the existence of a minimizer is obvious. □

Proof of Theorem 18. If Assumption 3 holds, then for any $d \in \mathcal{D}$, by setting $P = P_Z$, $Q = P_{Z|D=d}$ and $f(z) = \ell(f_w(x), y)$ in Lemma 34, we have

$$
\begin{aligned}
&(L_d(w) - L(w))^2 \\
&= \left(\mathbb{E}_{Z|D=d}[\ell(f_w(X), Y)] - \mathbb{E}_Z[\ell(f_w(X), Y]\right)^2 \\
&\le \left(\sqrt{2\sigma^2 \mathrm{KL}(P_{Z|D=d} \,\|\, P_Z)}\right)^2 \\
&\le 2\sigma^2 \mathrm{KL}(P_{Z|D=d} \,\|\, P_Z).
\end{aligned}
$$

Taking the expectation over $D \sim \nu$, we can get

$$
\begin{aligned}
\mathbb{E}_D[(L_D(w) - L(w))^2] &\le \mathbb{E}_D\left[2\sigma^2 \mathrm{KL}(P_{Z|D=d} \,\|\, P_Z)\right] \\
&= 2\sigma^2 \mathrm{KL}(P_{Z|D} \,\|\, P_Z) \\
&= 2\sigma^2 I(Z; D).
\end{aligned}
$$

When the target domains $\{D_k'\}_{i=k}^{m'}$ are independent of each other, they can be regarded as i.i.d copies of D, i.e.

$$
\begin{aligned}
\mathrm{Var}_{D_t}[L_t(w)] &= \frac{1}{m'^2}\sum_{k=1}^{m'} \mathrm{Var}_{D_k'}[L_{D_k'}(w)] \\
&= \frac{1}{m'^2}\sum_{k=1}^{m'} \mathbb{E}_{D_k'}[(L_{D_k'}(w) - L(w))^2] \\
&\le \frac{1}{m'^2}\sum_{k=1}^{m'} 2\sigma^2 I(Z; D) \\
&= \frac{1}{m'} 2\sigma^2 I(Z; D).
\end{aligned}
$$

Finally, by applying Chebyshev's inequality, we can prove that

$$
\mathbb{P}\{|L_t(w) - L(w)| \ge \epsilon\} \le \frac{2\sigma^2}{m'\epsilon^2} I(Z; D). \qquad \square
$$

Proof of Theorem 19. For any domain $d \in \mathcal{D}$, classifier ψ and $f^* : \mathcal{R} \mapsto \mathcal{Y}$, denote

$$
\begin{aligned}
L_d(\psi, f^*) &= \mathbb{E}_{R|D=d}[\ell(f_\psi(R), f^*(R))], \\
L(\psi, f^*) &= \mathbb{E}_R[\ell(f_\psi(R), f^*(R))].
\end{aligned}
$$

By setting $P = P_R$, $Q = P_{R|D=d}$ and $f(R) = \ell(f_\psi(R), f^*(R))$ and applying Lemma 34, we have

$$
(L_d(\psi, f^*) - L(\psi, f^*))^2 \le 2\sigma^2 \mathrm{KL}(P_{R|D=d} \,\|\, P_R).
$$

By taking the expectation over $D \sim \nu$, we get

$$\mathbb{E}_D\Big[\big(L_d(\psi, f^*) - L(\psi, f^*)\big)^2\Big] \le 2\sigma^2 \mathrm{KL}(P_{R|D} \,\|\, P_R) = 2\sigma^2 I(R; D).$$

Through a similar procedure of proving Theorem 11, we have

$$\mathbb{P}\{|L_d(\psi, f^*) - L(\psi, f^*)| \ge \epsilon\} \le \frac{2\sigma^2}{m'\epsilon^2} I(R; D).$$

Recall that in Theorem 12 we proved

$$\begin{aligned} L_d(\psi, f^*) &\le L_d(\psi) + L_d(f^*). \\ L(\psi, f^*) &\le L(\psi) + L(f^*). \\ L_d(\psi, f^*) &\ge L_d(\psi) - L_d(f^*). \\ L(\psi, f^*) &\ge L(\psi) - L(f^*). \end{aligned}$$

Combining the results above, we have that with probability at least $1 - \frac{2\sigma^2}{m'\epsilon^2} I(R; D)$,

$$\begin{aligned} L_t(\psi) &\le L_t(\psi, f^*) + L_t(f^*) \\ &\le L(\psi, f^*) + \epsilon + L_t(f^*) \\ &\le L(\psi) + L(f^*) + \epsilon + L_t(f^*). \end{aligned}$$

By minimizing $L_t(f^*) + L(f^*)$, it follows that

$$\mathbb{P}\Big\{L_t(\psi) - L(\psi) \ge \epsilon + \min_{f^*}(L_t(f^*) + L(f^*))\Big\} \le \frac{2\sigma^2}{m'\epsilon^2} I(R; D),$$

which finishes the proof. □

Proof of Theorem 22. Recall that any random variables bounded by $[0, M]$ are $\frac{M}{2}$-subGaussian. From assumption 2, we know that $\ell(f_W(X), Y)$ is $\frac{M}{2}$-subGaussian w.r.t $P_W \circ P_Z$. Then by applying Lemma 34, we have

$$\begin{aligned} &|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \\ &= \left|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{W,D_i,Z_j^i}[\ell(f_W(X_j^i), Y_j^i)] - \mathbb{E}_{W,D,Z}[\ell(f_W(X), Y)]\right| \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|\mathbb{E}_{W,D_i,Z_j^i}[\ell(f_W(X_j^i), Y_j^i)] - \mathbb{E}_{W,D,Z}[\ell(f_W(X), Y)]\right| \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{\frac{M^2}{2}\mathrm{KL}\Big(P_{W,D_i,Z_j^i}\Big\|P_W P_{D_i,Z_j^i}\Big)}. \end{aligned}$$

Notice that for any $D \in D_s$ and $Z \in S_D$,

$$\begin{aligned} \mathrm{KL}(P_{W,D,Z} \,\|\, P_W P_{D,Z}) &= \mathbb{E}_{W,D,Z}\left[\log\frac{P_{W,D,Z}}{P_W P_{D,Z}}\right] \\ &= I(W; D, Z) \\ &= I(W; D) + I(W; Z|D). \end{aligned}$$

Combining our results above, we then get

$$\begin{aligned} &|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{\frac{M^2}{2}(I(W; D_i) + I(W; Z_j^i|D_i))} \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\sqrt{\frac{M^2}{2}I(W; D_i)} + \sqrt{\frac{M^2}{2}I(W; Z_j^i|D_i)}\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{M^2}{2}I(W; D_i)} + \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{\frac{M^2}{2}I(W; Z_j^i|D_i)}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} &\mathbb{E}_{W,S}\big[(L'(W) - \mathbb{E}_W[L(W)])^2\big] \\ &= \mathbb{E}_{W,S}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\ell(f_W(X_j^i), Y_j^i) - \mathbb{E}_W[L(W)]\right)^2\right] \\ &= \frac{M^2}{mn}(I(W; S) + \log 3). \end{aligned}$$

This further implies by Lemma 31 that

$$\mathbb{P}\{|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \ge \epsilon\} \le \frac{M^2}{mn\epsilon^2}(I(W; S) + \log 3),$$

which completes the proof. □

Proof of Theorem 23. Recall that in the proof of Theorem 22, we have

$$\begin{aligned} &|\mathbb{E}_{W,D_s,S}[L'(W)] - \mathbb{E}_W[L(W)]| \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Big|\mathbb{E}_{W,D_i,Z_j^i}[\ell(f_W(X_j^i), Y_j^i)] - \mathbb{E}_{W,D,Z}[\ell(f_W(X), Y)]\Big| \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Big|E_{W,D_i,Z_j^i}[\ell(f_W(X_j^i), Y_j^i)] - \mathbb{E}_{W,D_i,Z}[\ell(f_W(X), Y)]\Big| \\ &\quad + \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Big|\mathbb{E}_{W,D_i,Z}[\ell(f_W(X), Y)] - \mathbb{E}_{W,D,Z}[\ell(f_W(X), Y)]\Big| \\ &\le \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{D_i,Z_j^i}\Big|\mathbb{E}_{W|D_i,Z_j^i}[\ell(f_W(X_j^i), Y_j^i)] - \mathbb{E}_{W|D_i}[\ell(f_W(X_j^i), Y_j^i)]\Big| \\ &\quad + \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{D_i}\Big|\mathbb{E}_{W|D_i}[L_{D_i}(W)] - \mathbb{E}_W[L_{D_i}(W)]\Big| \\ &\le \frac{\beta'}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Big(\mathbb{E}_{D_i,Z_j^i}[\mathbb{W}(P_{W|D_i,Z_j^i}, P_{W|D_i})] + \mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)]\Big) \\ &= \frac{\beta'}{m}\sum_{i=1}^{m}\mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)] \end{aligned}$$

$$+\frac{\beta'}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{D_i,Z_j^i}[\mathbb{W}(P_{W|D_i,Z_j^i},P_{W|D_i})],$$

where the last inequality is by applying Lemma 38. The proof is complete. □

Proof of Theorem 24. For any $D\in D_s$, let $\bar{D}$ be an independent copy of D. $L_D(W)\in[0,M]$ and is thus $\frac{M}{2}$-subGaussian. By applying Lemma 37 with $X=L_{\bar{D}}(W)$, we have that for any $\lambda\in[0,1)$,

$$\mathbb{E}_{W,\bar{D}}\left[e^{\frac{2\lambda(L(W)-L_{\bar{D}}(W))^2}{M^2}}\right]\le\frac{1}{\sqrt{1-\lambda}}.$$

Next, by applying Lemma 36 with $X=W$, $Y=D$ and $f(W,D)=\frac{2\lambda(L(W)-L_D(W))^2}{M^2}$, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}\frac{2\lambda(L(W)-L_{\bar{D}}(W))^2}{M^2}&\le\log\mathbb{E}_{W,\bar{D}}\left[\frac{e^{\frac{2\lambda(L(W)-L_{\bar{D}}(W))^2}{M^2}}}{\delta}\right]\\&\quad+\imath(W;D)\\&\le\log\frac{1}{\delta\sqrt{1-\lambda}}+\imath(W;D).\end{aligned}$$

Rearranging the terms above, with probability at least $1-\delta$,

$$|L(W)-L_D(W)|\le\sqrt{\frac{M^2}{2\lambda}\left(\imath(W;D)+\log\frac{1}{\delta\sqrt{1-\lambda}}\right)}.$$

Taking the union bound over every $D\in D_s$, we have that with probability at least $1-\delta$,

$$\begin{aligned}|L(W)-L_s(W)|&\le\frac{1}{m}\sum_{i=1}^{m}|L(W)-L_{D_i}(W)|\\&\le\frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{M^2}{2\lambda}\left(\imath(W;D_i)+\log\frac{m}{\delta\sqrt{1-\lambda}}\right)}.\end{aligned}$$ □

Proof of Theorem 25. By applying Lemma 37 with $X=\ell(f_w(X),Y)$, we have that for any $\lambda\in[0,1)$,

$$\mathbb{E}_Z\left[e^{\frac{\lambda(L(w)-\ell(f_w(X),Y))^2}{2\sigma^2}}\right]\le\frac{1}{\sqrt{1-\lambda}}.$$

Next, for any $D\in D_t$, by applying Lemma 35 with $X=Z$, $Y=D$ and $f(Z,D)=\frac{\lambda(L(w)-\ell(f_w(X),Y))^2}{2\sigma^2}$, we have that with probability at least $1-\delta$,

$$\begin{aligned}&\mathbb{E}_{Z|D}\left[\frac{\lambda(L(w)-\ell(f_w(X),Y))^2}{2\sigma^2}\right]\\&\le\log\mathbb{E}_{Z,\bar{D}}\left[\frac{e^{\frac{\lambda(L(w)-\ell(f_w(X),Y))^2}{2\sigma^2}}}{\delta}\right]+\mathrm{KL}(P_{Z|D}\,\|\,P_Z)\\&\le\log\frac{1}{\delta\sqrt{1-\lambda}}+\mathrm{KL}(P_{Z|D}\,\|\,P_Z).\end{aligned}$$

Finally, by applying Jensen's inequality, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}&|L(w)-L_D(w)|=\mathbb{E}_{Z|D}\sqrt{(L(w)-\ell(f_w(X),Y))^2}\\&\le\sqrt{\mathbb{E}_{Z|D}[L(w)-\ell(f_w(X),Y)]^2}\\&\le\sqrt{\frac{2\sigma^2}{\lambda}\left(\mathrm{KL}(P_{Z|D}\,\|\,P_Z)+\log\frac{1}{\delta\sqrt{1-\lambda}}\right)}.\end{aligned}$$

Taking the union bound over every $D\in D_t$, we have that with probability at least $1-\delta$,

$$\begin{aligned}|L(w)-L_t(w)|&\le\frac{1}{m'}\sum_{k=1}^{m'}\left|L(w)-L_{D'_k}(w)\right|\\&\le\frac{1}{m'}\sum_{k=1}^{m'}\sqrt{\frac{2\sigma^2}{\lambda}\left(\mathrm{KL}\left(P_{Z|D'_k}\,\middle\|\,P_Z\right)+\log\frac{m'}{\delta\sqrt{1-\lambda}}\right)}.\end{aligned}$$ □

APPENDIX C
EXPERIMENT DETAILS

In this paper, deep learning models are trained with an Intel Xeon CPU (2.10GHz, 48 cores), 256GB memory, and 4 Nvidia Tesla V100 GPUs (32GB).

A. Colored MNIST

The Colored MNIST dataset is a binary classification task introduced by IRM [9]. The main difference between Colored MNIST and the original MNIST dataset is the manually introduced strong correlation between the label and image colors. Colored MNIST is generated according to the following procedure:

- Give each sample an initial label by whether the digit is greater than 4 (i.e. label 0 for 0-4 digits and label 1 for 5-9 digits.
- Randomly flip the label with probability 0.25, so an oracle predictor that fully relies on the shape of the digits would achieve a 75% accuracy.
- Each domain is assigned a probability P_e, which characterizes the correlation between the label and the color: samples with label 0 have P_e chance to be red, and $1-P_e$ chance to be green, while samples with label 1 have P_e chance to be green, and $1-P_e$ chance to be red.

The original domain setting of [9] includes two source domains $D_s=\{P_1=90\%,P_2=80\%\}$, such that the predictive power of the color superiors that of the actual digits. This correlation is reversed in the target domain $D_t=\{P_3=10\%\}$, thus fooling algorithms without causality inference abilities to overfit the color features and generalize poorly on target domains.

The original implementation[2] uses a 3-layer MLP network with ReLU activation. The model is trained for 501 epochs in a full gradient descent scheme, such that the batch size equals the number of training samples $25,000$. We follow the hyper-parameter selection strategy of [9] through a random search over 50 independent trials, as reported in Table III along with the parameters selected for IDM. Considering that the covariate shift is not prominent according to the dataset construction procedure, we only apply gradient alignment without feature alignment in this experiment.

[2]https://github.com/facebookresearch/InvariantRiskMinimization

TABLE III
THE HYPER-PARAMETERS OF COLORED MNIST.

Parameter	Random Distribution	Selected Value
dimension of hidden layer	$2^{\text{Uniform}(6,9)}$	433
weight decay	$10^{\text{Uniform}(-2,-5)}$	0.00034
learning rate	$10^{\text{Uniform}(-2.5,-3.5)}$	0.000449
warmup iterations	$\text{Uniform}(50, 250)$	154
regularization strength	$10^{\text{Uniform}(4,8)}$	2888595.180638

B. DomainBed Benchmark

DomainBed [21] is an extensive benchmark for both DA and DG algorithms, which involves various synthetic and real-world datasets mainly focusing on image classification:

- Colored MNIST [9] is a variant of the MNIST dataset. As discussed previously, Colored MNIST includes 3 domains $\{90\%, 80\%, 10\%\}$, $70,000$ samples of dimension $(2, 28, 28)$ and 2 classes.
- Rotated MNIST [65] is a variant of the MNIST dataset with 7 domains $\{0, 15, 30, 45, 60, 75\}$ representing the rotation degrees, $70,000$ samples of dimension $(28, 28)$ and 10 classes.
- VLCS [66] includes 4 domains $\{\text{Caltech101}, \text{LabelMe}, \text{SUN09}, \text{VOC2007}\}$, $10,729$ samples of dimension $(3, 224, 224)$ and 5 classes.
- PACS [67] includes 4 domains $\{\text{art}, \text{cartoons}, \text{photos}, \text{sketches}\}$, $9,991$ samples of dimension $(3, 224, 224)$ and 7 classes.
- OfficeHome [68] includes 4 domains $\{\text{art}, \text{clipart}, \text{product}, \text{real}\}$, $15,588$ samples of dimension $(3, 224, 224)$ and 65 classes.
- TerraIncognita [69] includes 4 domains $\{\text{L100}, \text{L38}, \text{L43}, \text{46}\}$ representing locations of photographs, $24,788$ samples of dimension $(3, 224, 224)$ and 10 classes.
- DomainNet [70] includes 6 domains $\{\text{clipart}, \text{infograph}, \text{painting}, \text{quickdraw}, \text{real}, \text{sketch}\}$, $586,575$ samples of dimension $(3, 224, 224)$ and 345 classes.

We list all competitive DG approaches below. Note that some recent progress is omitted [16], [71]–[75], which either contributes complementary approaches, does not report full DomainBed results, or does not report the target-domain validation scores. Due to the limitation of computational resources, we are not able to reproduce the full results of these works on DomainBed.

- ERM: Empirical Risk Minimization.
- IRM: Invariant Risk Minimization [9].
- GroupDRO: Group Distributionally Robust Optimization [14].
- Mixup: Interdomain Mixup [76].
- MLDG: Meta Learning Domain Generalization [77].
- CORAL: Deep CORAL [5].
- MMD: Maximum Mean Discrepancy [6].
- DANN: Domain Adversarial Neural Network [7].
- CDANN: Conditional Domain Adversarial Neural Network [8].
- MTL: Marginal Transfer Learning [17].
- SagNet: Style Agnostic Networks [78].
- ARM: Adaptive Risk Minimization [18].
- V-REx: Variance Risk Extrapolation [15].
- RSC: Representation Self-Challenging [79].
- AND-mask: Learning Explanations that are Hard to Vary [20].
- SAND-mask: Smoothed-AND mask [47].
- Fish: Gradient Matching for Domain Generalization [12].
- Fishr: Invariant Gradient Variances for Out-of-distribution Generalization [13].
- SelfReg: Self-supervised Contrastive Regularization [80].
- CausIRL: Invariant Causal Mechanisms through Distribution Matching [10].

The same fine-tuning procedure is applied to all approaches: The network is a multi-layer CNN for synthetic MNIST datasets and is a pre-trained ResNet-50 for other real-world datasets. The hyper-parameters are selected by a random search over 20 independent trials for each target domain, and each evaluation score is reported after 3 runs with different initialization seeds[3]. The hyper-parameter selection criteria are shown in Table IV. Note that warmup iterations and moving average techniques are not adopted for representation alignment.

Note that although the same Colored MNIST dataset is adopted by DomainBed, the experimental settings are completely different from the previous one [9]. The main difference is the batch size (25000 for IRM, less than 512 for DomainBed), making it much harder to learn invariance for causality inference and distribution matching methods since fewer samples are available for probability density estimation. This explains the huge performance drop between these two experiments using the same DG algorithms.

APPENDIX D
ADDITIONAL EXPERIMENTAL RESULTS

A. Component Analysis

In this section, we conduct ablation studies to demonstrate the effect of each component of the proposed IDM algorithm. Specifically, we analyze the effect of gradient alignment (GA), representation alignment (RA), warmup iterations (WU), moving average (MA), and the proposed PDM method for distribution matching.

1) Gradient Alignment: According to our theoretical analysis, gradient alignment promotes source-domain generalization, especially when concept shift is prominent. As can be seen in Table V, IDM without gradient alignment (57.7%) performs similarly to ERM (57.8%), which is unable to learn invariance across source domains. Gradient alignment also significantly boosts the performance on VLCS (77.4% to 78.1%) and PACS (86.8% to 87.6%), as seen in Table VII and VIII. However, for datasets where concept shift is not prominent e.g. OfficeHome, gradient alignment cannot help to improve performance as shown in Table VI. It is worth noting that gradient alignment also penalizes a lower bound for the representation

[3]https://github.com/facebookresearch/DomainBed

TABLE IV
THE HYPER-PARAMETERS OF DOMAINBED.

Condition	Parameter	Default Value	Random Distribution
MNIST Datasets	learning rate	0.001	$10^{\text{Uniform}(-4.5,-3.5)}$
	batch size	64	$2^{\text{Uniform}(3,9)}$
Real-world Datasets	learning rate	0.00005	$10^{\text{Uniform}(-5,-3.5)}$
	batch size	32	$2^{\text{Uniform}(3,5)}$ (DomainNet) / $2^{\text{Uniform}(3,5.5)}$ (others)
	weight decay	0	$10^{\text{Uniform}(-6,-2)}$
	dropout	0	$\text{Uniform}(\{0, 0.1, 0.5\})$
-	steps	5000	5000
IDM	gradient penalty	1000	$10^{\text{Uniform}(1,5)}$
	gradient warmup	1500	$\text{Uniform}(0, 5000)$
	representation penalty	1	$10^{\text{Uniform}(-1,1)}$
	moving average	0.95	$\text{Uniform}(0.9, 0.99)$

TABLE V
COMPONENT ANALYSIS ON COLOREDMNIST OF DOMAINBED.

Algorithm	**GA**	**RA**	**WU**	**MA**	**90%**	**80%**	**10%**	**Average**
ERM			-		71.8 ± 0.4	72.9 ± 0.1	28.7 ± 0.5	57.8
IDM	✗	✓	✗	✗	71.9 ± 0.4	72.5 ± 0.0	28.8 ± 0.7	57.7
	✓	✗	✓	✓	73.1 ± 0.2	72.7 ± 0.3	67.4 ± 1.6	71.1
	✓	✓	✗	✓	72.9 ± 0.2	72.7 ± 0.1	60.8 ± 2.1	68.8
	✓	✓	✓	✗	72.0 ± 0.1	71.5 ± 0.3	48.7 ± 7.1	64.0
	✓	✓	✓	✓	**74.2** ± 0.6	**73.5** ± 0.2	**68.3** ± 2.5	**72.0**

TABLE VI
COMPONENT ANALYSIS ON OFFICEHOME OF DOMAINBED.

Algorithm	**GA**	**RA**	**WU**	**MA**	**A**	**C**	**P**	**R**	**Average**
ERM			-		61.7 ± 0.7	53.4 ± 0.3	74.1 ± 0.4	76.2 ± 0.6	66.4
IDM	✗	✓	✗	✗	**64.7** ± 0.5	**54.6** ± 0.3	76.2 ± 0.4	**78.1** ± 0.5	**68.4**
	✓	✗	✓	✓	61.9 ± 0.4	53.0 ± 0.3	75.5 ± 0.2	77.9 ± 0.2	67.1
	✓	✓	✗	✓	62.5 ± 0.1	53.0 ± 0.7	75.0 ± 0.4	77.2 ± 0.7	66.9
	✓	✓	✓	✗	64.2 ± 0.3	53.5 ± 0.6	76.1 ± 0.4	**78.1** ± 0.4	68.0
	✓	✓	✓	✓	64.4 ± 0.3	54.4 ± 0.6	**76.5** ± 0.3	78.0 ± 0.4	68.3

space distribution shift: In the t-th step of gradient descent, the Markov chain relationship $D_i \to B_t^i \to (R_t^i, Y_t^i) \to G_t^i$ holds conditioned on the current predictor W_{t-1}, which implies the lower bound $I(G_t^i; D_i|W_{t-1}) \leq I(R_t^i, Y_t^i; D_i|W_{t-1})$ by the data processing inequality. This indicates that gradient alignment also helps to address the covariate shift, which explains the promising performance of gradient-based DG algorithms e.g. Fish and Fishr. However, since this is a lower bound rather than an upper bound, gradient manipulation is insufficient to fully address representation space covariate shifts, as seen in the following analysis for representation alignment.

2) Representation Alignment: Representation alignment promotes target-domain generalization by minimizing the representation level covariate shift. As shown in Table V - IX, representation alignment is effective in OfficeHome (67.1% to 68.3%) and RotatedMNIST (97.8% to 98.0%), and still enhances the performance even though covariate shift is not prominent in ColoredMNIST (71.1% to 72.0%). This verifies our claim that representation alignment complements gradient alignment in solving Problem 6, and is necessary for achieving high-probability DG.

3) Warmup Iterations: Following the experimental settings of [9], [13], we do not apply the penalties of gradient or representation alignment until the number of epochs reaches a certain value. This is inspired by the observation that forcing invariance in early steps may hinder the models from extracting useful correlations. By incorporating these warmup iterations, predictors are allowed to extract all possible correlations between the inputs and the labels at the beginning, and then discard spurious ones in later updates. As can be seen in Table V and VI, this strategy helps to enhance the final performances on ColoredMNIST (68.8% to 72.0%) and OfficeHome (66.9% to 68.3%).

4) Moving Average: Following [13], [81], we use an exponential moving average when computing the gradients or the

TABLE VII
EFFECT OF GRADIENT ALIGNMENT (GA) ON VLCS OF DOMAINBED.

Algorithm	GA	A	C	P	S	Average
ERM	-	**97.6** ± 0.3	**67.9** ± 0.7	70.9 ± 0.2	74.0 ± 0.6	77.6
IDM	✗	97.1 ± 0.7	67.2 ± 0.4	69.9 ± 0.4	75.6 ± 0.8	77.4
IDM	✓	**97.6** ± 0.3	66.9 ± 0.3	**71.8** ± 0.5	**76.0** ± 1.3	**78.1**

TABLE VIII
EFFECT OF GRADIENT ALIGNMENT (GA) ON PACS OF DOMAINBED.

Algorithm	GA	A	C	P	S	Average
ERM	-	86.5 ± 1.0	81.3 ± 0.6	96.2 ± 0.3	**82.7** ± 1.1	86.7
IDM	✗	87.8 ± 0.6	81.6 ± 0.3	97.4 ± 0.2	80.6 ± 1.3	86.8
IDM	✓	**88.0** ± 0.3	**82.6** ± 0.6	**97.6** ± 0.4	82.3 ± 0.6	**87.6**

TABLE IX
EFFECT OF REPRESENTATION ALIGNMENT (RA) ON ROTATEDMNIST OF DOMAINBED.

Algorithm	RA	0	15	30	45	60	75	Average
ERM	-	95.3 ± 0.2	**98.7** ± 0.1	98.9 ± 0.1	98.7 ± 0.2	**98.9** ± 0.0	96.2 ± 0.2	97.8
IDM	✗	95.6 ± 0.1	98.4 ± 0.1	98.7 ± 0.2	**99.1** ± 0.0	98.7 ± 0.1	**96.6** ± 0.4	97.8
IDM	✓	**96.1** ± 0.3	**98.7** ± 0.1	**99.1** ± 0.1	98.9 ± 0.1	**98.9** ± 0.1	**96.6** ± 0.1	**98.0**

representations. This strategy helps when the batch size is not sufficiently large to sketch the probability distributions. In the IRM experiment setup where the batch size is 25000, Fishr (70.2%) and IDM (70.5%) both easily achieve near-optimal accuracy compared to Oracle (71.0%). In the DomainBed setup, the batch size $2^{\mathrm{Uniform}(3,9)}$ is significantly diminished, resulting in worse target-domain accuracy of Fishr (68.8%). As shown in Table V and VI, this moving average strategy greatly enhances the performance of IDM on ColoredMNIST (64.0% to 72.0%) and OfficeHome (68.0% to 68.3%).

5) PDM for Distribution Matching: We then demonstrate the superiority of our PDM method over moment-based distribution alignment techniques. Specifically, we compare IGA [11] which matches the empirical expectation of the gradients, Fishr [13] which proposes to align the gradient variance, the combination of IGA + Fishr (i.e. aligning the expectation and variance simultaneously), and our approach IDM (without representation space alignment). The performance gain of IDM on the Colored MNIST task in [9] is not significant, since it is relatively easier to learn invariance with a large batch size (25000). In the DomainBed setting, the batch size is significantly reduced (8-512), making this learning task much harder. The results are reported in Table X.

TABLE X
SUPERIORITY OF PDM ON COLORED MNIST OF DOMAINBED.

Algorithm	90%	80%	10%	Average
ERM	71.8 ± 0.4	72.9 ± 0.1	28.7 ± 0.5	57.8
IGA	72.6 ± 0.3	72.9 ± 0.2	50.0 ± 1.2	65.2
Fishr	74.1 ± 0.6	73.3 ± 0.1	58.9 ± 3.7	68.8
IGA + Fishr	73.3 ± 0.0	72.6 ± 0.5	66.3 ± 2.9	70.7
IDM	**74.2** ± 0.6	**73.5** ± 0.2	**68.3** ± 2.5	**72.0**

As can be seen, IDM achieves significantly higher performance on Colored MNIST (72.0%) even compared to the combination of IGA + Fishr (70.7%). This verifies our conclusion that matching the expectation and the variance is not sufficient for complex probability distributions, and demonstrates the superiority of the proposed PDM method for distribution alignment.

B. Running Time Comparison

Since IDM only stores historical gradients and representations for a single batch from each source domain, the storage and computation overhead is marginal compared to training the entire network. As shown in Table XI, the training time is only 5% longer compared to ERM on the largest DomainNet dataset.

C. Full DomainBed Results

In this paper, we focus on the target-domain model selection criterion, where the validation set follows the same distribution as the target domains. Our choice is well-motivated for the following reasons:

- Target-domain validation is provided by the DomainBed benchmark as one of the default model-selection methods, and is also widely adopted in the literature in many significant works like IRM [9], V-Rex [15], and Fishr [13].
- As suggested by Proposition 7, any algorithm that fits well on source domains will suffer from strictly positive risks in target domains once concept shift is induced. Therefore, source-domain validation would result in sub-optimal selection results.
- Source-domain validation may render efforts to address concept shift useless, as spurious features are often more

TABLE XI
COMPUTATIONAL OVERHEAD OF IDM USING DEFAULT BATCH SIZE.

Dataset	Training Time (h)			Memory Requirement (GB)		
	ERM	IDM	Overhead	ERM	IDM	Overhead
ColoredMNIST	0.076	0.088	14.6%	0.138	0.139	0.2%
RotatedMNIST	0.101	0.110	9.3%	0.338	0.342	1.0%
VLCS	0.730	0.744	2.0%	8.189	8.199	0.1%
PACS	0.584	0.593	1.5%	8.189	8.201	0.1%
OfficeHome	0.690	0.710	2.9%	8.191	8.506	3.8%
TerraIncognita	0.829	0.840	1.3%	8.189	8.208	0.2%
DomainNet	2.805	2.947	5.0%	13.406	16.497	23.1%

predictive than invariant ones. This is particularly unfair for algorithms that aim to tackle the concept shift. As shown in Table 9 in [13], no algorithm can significantly outperform ERM on Colored MNIST using source-domain validation (an exception is ARM which uses test-time adaptation, and thus cannot be directly compared), even though ERM is shown to perform much worse than random guessing (10% v.s. 50% accuracy) for the last domain (see Table 1 in [9] and Appendix D.4.1 in [13]). As a result, models selected by source-domain validation may not generalize well when concept shift is substantial.

- As mentioned by [82], source-domain validation suffers from underspecification, where predictors with equivalently strong performances in source domains may behave very differently during testing. It is also emphasized by [83] that OOD performance cannot, by definition, be performed with a validation set from the same distribution as the training data. This further raises concerns about the validity of using source-domain accuracies for validation purposes.
- Moreover, target-domain validation is also applicable in practice, as it is feasible to label a few target-domain samples for validation purposes. It is also unrealistic to deploy models in target domains without any form of verification, making such efforts necessary in practice.

Due to the reasons listed above, we believe that the target-domain validation results are sufficient to demonstrate the effectiveness of our approach in real-world learning scenarios. We report detailed results of IDM for each domain in each dataset of the DomainBed benchmark under target-domain model selection for a complete evaluation in Table XII - XVIII. Note that detailed scores of certain algorithms (Fish, CausIRL) are not available.

TABLE XII
DETAILED RESULTS ON COLORED MNIST IN DOMAINBED.

Algorithm	**90%**	**80%**	**10%**	**Average**
ERM	71.8 ± 0.4	72.9 ± 0.1	28.7 ± 0.5	57.8
IRM	72.0 ± 0.1	72.5 ± 0.3	58.5 ± 3.3	67.7
GroupDRO	73.5 ± 0.3	73.0 ± 0.3	36.8 ± 2.8	61.1
Mixup	72.5 ± 0.2	73.9 ± 0.4	28.6 ± 0.2	58.4
MLDG	71.9 ± 0.3	73.5 ± 0.2	29.1 ± 0.9	58.2
CORAL	71.1 ± 0.2	73.4 ± 0.2	31.1 ± 1.6	58.6
MMD	69.0 ± 2.3	70.4 ± 1.6	50.6 ± 0.2	63.3
DANN	72.4 ± 0.5	73.9 ± 0.5	24.9 ± 2.7	57.0
CDANN	71.8 ± 0.5	72.9 ± 0.1	33.8 ± 6.4	59.5
MTL	71.2 ± 0.2	73.5 ± 0.2	28.0 ± 0.6	57.6
SagNet	72.1 ± 0.3	73.2 ± 0.3	29.4 ± 0.5	58.2
ARM	84.9 ± 0.9	76.8 ± 0.6	27.9 ± 2.1	63.2
V-REx	72.8 ± 0.3	73.0 ± 0.3	55.2 ± 4.0	67.0
RSC	72.0 ± 0.1	73.2 ± 0.1	30.2 ± 1.6	58.5
AND-mask	71.9 ± 0.6	73.6 ± 0.5	30.2 ± 1.4	58.6
SAND-mask	79.9 ± 3.8	75.9 ± 1.6	31.6 ± 1.1	62.3
Fishr	74.1 ± 0.6	73.3 ± 0.1	58.9 ± 3.7	68.8
SelfReg	71.3 ± 0.4	73.4 ± 0.2	29.3 ± 2.1	58.0
IDM	74.2 ± 0.6	73.5 ± 0.2	68.3 ± 2.5	72.0

REFERENCES

[1] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." in *International Conference on Learning Representations*, 2019.

[2] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[3] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *Journal of Machine Learning Research*, vol. 20, pp. 1–25, 2019.

[4] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[5] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[6] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[8] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 624–639.

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

[10] M. Chevalley, C. Bunne, A. Krause, and S. Bauer, "Invariant causal mechanisms through distribution matching," *arXiv preprint arXiv:2206.11646*, 2022.

[11] M. Koyama and S. Yamaguchi, "When is invariance useful in an out-of-distribution generalization problem?" *arXiv preprint arXiv:2008.01883*, 2020.

[12] Y. Shi, J. Seely, P. Torr, S. N, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," in *International*

TABLE XIII
DETAILED RESULTS ON ROTATED MNIST IN DOMAINBED.

Algorithm	0	15	30	45	60	75	Average
ERM	95.3 ± 0.2	98.7 ± 0.1	98.9 ± 0.1	98.7 ± 0.2	98.9 ± 0.0	96.2 ± 0.2	97.8
IRM	94.9 ± 0.6	98.7 ± 0.2	98.6 ± 0.1	98.6 ± 0.2	98.7 ± 0.1	95.2 ± 0.3	97.5
GroupDRO	95.9 ± 0.1	99.0 ± 0.1	98.9 ± 0.1	98.8 ± 0.1	98.6 ± 0.1	96.3 ± 0.4	97.9
Mixup	95.8 ± 0.3	98.7 ± 0.0	99.0 ± 0.1	98.8 ± 0.1	98.8 ± 0.1	96.6 ± 0.2	98.0
MLDG	95.7 ± 0.2	98.9 ± 0.1	98.8 ± 0.1	98.9 ± 0.1	98.6 ± 0.1	95.8 ± 0.4	97.8
CORAL	96.2 ± 0.2	98.8 ± 0.1	98.8 ± 0.1	98.8 ± 0.1	98.9 ± 0.1	96.4 ± 0.2	98.0
MMD	96.1 ± 0.2	98.9 ± 0.0	99.0 ± 0.0	98.8 ± 0.0	98.9 ± 0.0	96.4 ± 0.2	98.0
DANN	95.9 ± 0.1	98.9 ± 0.1	98.6 ± 0.2	98.7 ± 0.1	98.9 ± 0.0	96.3 ± 0.3	97.9
CDANN	95.9 ± 0.2	98.8 ± 0.0	98.7 ± 0.1	98.9 ± 0.1	98.8 ± 0.1	96.1 ± 0.3	97.9
MTL	96.1 ± 0.2	98.9 ± 0.0	99.0 ± 0.0	98.7 ± 0.1	99.0 ± 0.0	95.8 ± 0.3	97.9
SagNet	95.9 ± 0.1	99.0 ± 0.1	98.9 ± 0.1	98.6 ± 0.1	98.8 ± 0.1	96.3 ± 0.1	97.9
ARM	95.9 ± 0.4	99.0 ± 0.1	98.8 ± 0.1	98.9 ± 0.1	99.1 ± 0.1	96.7 ± 0.2	98.1
V-REx	95.5 ± 0.2	99.0 ± 0.0	98.7 ± 0.2	98.8 ± 0.1	98.8 ± 0.0	96.4 ± 0.0	97.9
RSC	95.4 ± 0.1	98.6 ± 0.1	98.6 ± 0.1	98.9 ± 0.0	98.8 ± 0.1	95.4 ± 0.3	97.6
AND-mask	94.9 ± 0.1	98.8 ± 0.1	98.8 ± 0.1	98.7 ± 0.2	98.6 ± 0.2	95.5 ± 0.2	97.5
SAND-mask	94.7 ± 0.2	98.5 ± 0.2	98.6 ± 0.1	98.6 ± 0.1	98.5 ± 0.1	95.2 ± 0.1	97.4
Fishr	95.8 ± 0.1	98.3 ± 0.1	98.8 ± 0.1	98.6 ± 0.3	98.7 ± 0.1	96.5 ± 0.1	97.8
SelfReg	96.0 ± 0.3	98.9 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	96.8 ± 0.1	98.1
IDM	96.1 ± 0.3	98.7 ± 0.1	99.1 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	96.6 ± 0.1	98.0

TABLE XIV
DETAILED RESULTS ON VLCS IN DOMAINBED.

Algorithm	C	L	S	V	Average
ERM	97.6 ± 0.3	67.9 ± 0.7	70.9 ± 0.2	74.0 ± 0.6	77.6
IRM	97.3 ± 0.2	66.7 ± 0.1	71.0 ± 2.3	72.8 ± 0.4	76.9
GroupDRO	97.7 ± 0.2	65.9 ± 0.2	72.8 ± 0.8	73.4 ± 1.3	77.4
Mixup	97.8 ± 0.4	67.2 ± 0.4	71.5 ± 0.2	75.7 ± 0.6	78.1
MLDG	97.1 ± 0.5	66.6 ± 0.5	71.5 ± 0.1	75.0 ± 0.9	77.5
CORAL	97.3 ± 0.2	67.5 ± 0.6	71.6 ± 0.6	74.5 ± 0.0	77.7
MMD	98.8 ± 0.0	66.4 ± 0.4	70.8 ± 0.5	75.6 ± 0.4	77.9
DANN	99.0 ± 0.2	66.3 ± 1.2	73.4 ± 1.4	80.1 ± 0.5	79.7
CDANN	98.2 ± 0.1	68.8 ± 0.5	74.3 ± 0.6	78.1 ± 0.5	79.9
MTL	97.9 ± 0.7	66.1 ± 0.7	72.0 ± 0.4	74.9 ± 1.1	77.7
SagNet	97.4 ± 0.3	66.4 ± 0.4	71.6 ± 0.1	75.0 ± 0.8	77.6
ARM	97.6 ± 0.6	66.5 ± 0.3	72.7 ± 0.6	74.4 ± 0.7	77.8
V-REx	98.4 ± 0.2	66.4 ± 0.7	72.8 ± 0.1	75.0 ± 1.4	78.1
RSC	98.0 ± 0.4	67.2 ± 0.3	70.3 ± 1.3	75.6 ± 0.4	77.8
AND-mask	98.3 ± 0.3	64.5 ± 0.2	69.3 ± 1.3	73.4 ± 1.3	76.4
SAND-mask	97.6 ± 0.3	64.5 ± 0.6	69.7 ± 0.6	73.0 ± 1.2	76.2
Fishr	97.6 ± 0.7	67.3 ± 0.5	72.2 ± 0.9	75.7 ± 0.3	78.2
SelfReg	97.9 ± 0.4	66.7 ± 0.1	73.5 ± 0.7	74.7 ± 0.7	78.2
IDM	97.6 ± 0.3	66.9 ± 0.3	71.8 ± 0.5	76.0 ± 1.3	78.1

Conference on Learning Representations, 2022.

[13] A. Rame, C. Dancette, and M. Cord, "Fishr: Invariant gradient variances for out-of-distribution generalization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 347–18 377.

[14] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *International Conference on Learning Representations*, 2020.

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

[16] C. Eastwood, A. Robey, S. Singh, J. Von Kügelgen, H. Hassani, G. J. Pappas, and B. Schölkopf, "Probable domain generalization via quantile risk minimization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 340–17 358, 2022.

[17] G. Blanchard, A. A. Deshmukh, Ü. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 46–100, 2021.

[18] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn, "Adaptive risk minimization: Learning to adapt to domain shift," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 664–23 678, 2021.

[19] V. Nagarajan, A. Andreassen, and B. Neyshabur, "Understanding the failure modes of out-of-distribution generalization," in *International Conference on Learning Representations*, 2021.

[20] G. Parascandolo, A. Neitz, A. Orvieto, L. Gresele, and B. Schölkopf, "Learning explanations that are hard to vary," in *International Conference on Learning Representations*, 2021.

[21] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2021.

[22] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for sgld via data-dependent estimates," *Advances in Neural Information Processing Systems*, vol. 32,

TABLE XV
DETAILED RESULTS ON PACS IN DOMAINBED.

Algorithm	A	C	P	S	Average
ERM	86.5 ± 1.0	81.3 ± 0.6	96.2 ± 0.3	82.7 ± 1.1	86.7
IRM	84.2 ± 0.9	79.7 ± 1.5	95.9 ± 0.4	78.3 ± 2.1	84.5
GroupDRO	87.5 ± 0.5	82.9 ± 0.6	97.1 ± 0.3	81.1 ± 1.2	87.1
Mixup	87.5 ± 0.4	81.6 ± 0.7	97.4 ± 0.2	80.8 ± 0.9	86.8
MLDG	87.0 ± 1.2	82.5 ± 0.9	96.7 ± 0.3	81.2 ± 0.6	86.8
CORAL	86.6 ± 0.8	81.8 ± 0.9	97.1 ± 0.5	82.7 ± 0.6	87.1
MMD	88.1 ± 0.8	82.6 ± 0.7	97.1 ± 0.5	81.2 ± 1.2	87.2
DANN	87.0 ± 0.4	80.3 ± 0.6	96.8 ± 0.3	76.9 ± 1.1	85.2
CDANN	87.7 ± 0.6	80.7 ± 1.2	97.3 ± 0.4	77.6 ± 1.5	85.8
MTL	87.0 ± 0.2	82.7 ± 0.8	96.5 ± 0.7	80.5 ± 0.8	86.7
SagNet	87.4 ± 0.5	81.2 ± 1.2	96.3 ± 0.8	80.7 ± 1.1	86.4
ARM	85.0 ± 1.2	81.4 ± 0.2	95.9 ± 0.3	80.9 ± 0.5	85.8
V-REx	87.8 ± 1.2	81.8 ± 0.7	97.4 ± 0.2	82.1 ± 0.7	87.2
RSC	86.0 ± 0.7	81.8 ± 0.9	96.8 ± 0.7	80.4 ± 0.5	86.2
AND-mask	86.4 ± 1.1	80.8 ± 0.9	97.1 ± 0.2	81.3 ± 1.1	86.4
SAND-mask	86.1 ± 0.6	80.3 ± 1.0	97.1 ± 0.3	80.0 ± 1.3	85.9
Fishr	87.9 ± 0.6	80.8 ± 0.5	97.9 ± 0.4	81.1 ± 0.8	86.9
SelfReg	87.5 ± 0.1	83.0 ± 0.1	97.6 ± 0.1	82.8 ± 0.2	87.7
IDM	88.0 ± 0.3	82.6 ± 0.6	97.6 ± 0.4	82.3 ± 0.6	87.6

TABLE XVI
DETAILED RESULTS ON OFFICEHOME IN DOMAINBED.

Algorithm	A	C	P	R	Average
ERM	61.7 ± 0.7	53.4 ± 0.3	74.1 ± 0.4	76.2 ± 0.6	66.4
IRM	56.4 ± 3.2	51.2 ± 2.3	71.7 ± 2.7	72.7 ± 2.7	63.0
GroupDRO	60.5 ± 1.6	53.1 ± 0.3	75.5 ± 0.3	75.9 ± 0.7	66.2
Mixup	63.5 ± 0.2	54.6 ± 0.4	76.0 ± 0.3	78.0 ± 0.7	68.0
MLDG	60.5 ± 0.7	54.2 ± 0.5	75.0 ± 0.2	76.7 ± 0.5	66.6
CORAL	64.8 ± 0.8	54.1 ± 0.9	76.5 ± 0.4	78.2 ± 0.4	68.4
MMD	60.4 ± 1.0	53.4 ± 0.5	74.9 ± 0.1	76.1 ± 0.7	66.2
DANN	60.6 ± 1.4	51.8 ± 0.7	73.4 ± 0.5	75.5 ± 0.9	65.3
CDANN	57.9 ± 0.2	52.1 ± 1.2	74.9 ± 0.7	76.2 ± 0.2	65.3
MTL	60.7 ± 0.8	53.5 ± 1.3	75.2 ± 0.6	76.6 ± 0.6	66.5
SagNet	62.7 ± 0.5	53.6 ± 0.5	76.0 ± 0.3	77.8 ± 0.1	67.5
ARM	58.8 ± 0.5	51.8 ± 0.7	74.0 ± 0.1	74.4 ± 0.2	64.8
V-REx	59.6 ± 1.0	53.3 ± 0.3	73.2 ± 0.5	76.6 ± 0.4	65.7
RSC	61.7 ± 0.8	53.0 ± 0.9	74.8 ± 0.8	76.3 ± 0.5	66.5
AND-mask	60.3 ± 0.5	52.3 ± 0.6	75.1 ± 0.2	76.6 ± 0.3	66.1
SAND-mask	59.9 ± 0.7	53.6 ± 0.8	74.3 ± 0.4	75.8 ± 0.5	65.9
Fishr	63.4 ± 0.8	54.2 ± 0.3	76.4 ± 0.3	78.5 ± 0.2	68.2
SelfReg	64.2 ± 0.6	53.6 ± 0.7	76.7 ± 0.3	77.9 ± 0.5	68.1
IDM	64.4 ± 0.3	54.4 ± 0.6	76.5 ± 0.3	78.0 ± 0.4	68.3

2019.

[24] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy, "Information-theoretic generalization bounds for stochastic gradient descent," in *Conference on Learning Theory*. PMLR, 2021, pp. 3526–3545.

[25] Z. Wang and Y. Mao, "On the generalization of models trained with sgd: Information-theoretic bounds and implications," in *International Conference on Learning Representations*, 2021.

[26] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, 2024.

[27] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International conference on machine learning*. PMLR, 2016, pp. 1225–1234.

[28] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar, "Stability of stochastic gradient descent on nonsmooth convex losses," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4381–4391, 2020.

[29] Y. Lei, Z. Yang, T. Yang, and Y. Ying, "Stability and generalization of stochastic gradient methods for minimax problems," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6175–6186.

[30] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via wasserstein distance," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 109–19 121, 2021.

[31] Z. Yang, Y. Lei, P. Wang, T. Yang, and Y. Ying, "Simple stochastic and online gradient descent algorithms for pairwise learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 160–20 171, 2021.

[32] Z. Yang, Y. Lei, S. Lyu, and Y. Ying, "Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2026–2034.

[33] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada, 2009.

[34] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings of the*

TABLE XVII
DETAILED RESULTS ON TERRAINCOGNITA IN DOMAINBED.

Algorithm	L100	L38	L43	L46	Average
ERM	59.4 ± 0.9	49.3 ± 0.6	60.1 ± 1.1	43.2 ± 0.5	53.0
IRM	56.5 ± 2.5	49.8 ± 1.5	57.1 ± 2.2	38.6 ± 1.0	50.5
GroupDRO	60.4 ± 1.5	48.3 ± 0.4	58.6 ± 0.8	42.2 ± 0.8	52.4
Mixup	67.6 ± 1.8	51.0 ± 1.3	59.0 ± 0.0	40.0 ± 1.1	54.4
MLDG	59.2 ± 0.1	49.0 ± 0.9	58.4 ± 0.9	41.4 ± 1.0	52.0
CORAL	60.4 ± 0.9	47.2 ± 0.5	59.3 ± 0.4	44.4 ± 0.4	52.8
MMD	60.6 ± 1.1	45.9 ± 0.3	57.8 ± 0.5	43.8 ± 1.2	52.0
DANN	55.2 ± 1.9	47.0 ± 0.7	57.2 ± 0.9	42.9 ± 0.9	50.6
CDANN	56.3 ± 2.0	47.1 ± 0.9	57.2 ± 1.1	42.4 ± 0.8	50.8
MTL	58.4 ± 2.1	48.4 ± 0.8	58.9 ± 0.6	43.0 ± 1.3	52.2
SagNet	56.4 ± 1.9	50.5 ± 2.3	59.1 ± 0.5	44.1 ± 0.6	52.5
ARM	60.1 ± 1.5	48.3 ± 1.6	55.3 ± 0.6	40.9 ± 1.1	51.2
V-REx	56.8 ± 1.7	46.5 ± 0.5	58.4 ± 0.3	43.8 ± 0.3	51.4
RSC	59.9 ± 1.4	46.7 ± 0.4	57.8 ± 0.5	44.3 ± 0.6	52.1
AND-mask	54.7 ± 1.8	48.4 ± 0.5	55.1 ± 0.5	41.3 ± 0.6	49.8
SAND-mask	56.2 ± 1.8	46.3 ± 0.3	55.8 ± 0.4	42.6 ± 1.2	50.2
Fishr	60.4 ± 0.9	50.3 ± 0.3	58.8 ± 0.5	44.9 ± 0.5	53.6
SelfReg	60.0 ± 2.3	48.8 ± 1.0	58.6 ± 0.8	44.0 ± 0.6	52.8
IDM	60.1 ± 1.4	48.8 ± 1.9	57.9 ± 0.2	44.3 ± 1.2	52.8

TABLE XVIII
DETAILED RESULTS ON DOMAINNET IN DOMAINBED.

Algorithm	clip	info	paint	quick	real	sketch	Average
ERM	58.6 ± 0.3	19.2 ± 0.2	47.0 ± 0.3	13.2 ± 0.2	59.9 ± 0.3	49.8 ± 0.4	41.3
IRM	40.4 ± 6.6	12.1 ± 2.7	31.4 ± 5.7	9.8 ± 1.2	37.7 ± 9.0	36.7 ± 5.3	28.0
GroupDRO	47.2 ± 0.5	17.5 ± 0.4	34.2 ± 0.3	9.2 ± 0.4	51.9 ± 0.5	40.1 ± 0.6	33.4
Mixup	55.6 ± 0.1	18.7 ± 0.4	45.1 ± 0.5	12.8 ± 0.3	57.6 ± 0.5	48.2 ± 0.4	39.6
MLDG	59.3 ± 0.1	19.6 ± 0.2	46.8 ± 0.2	13.4 ± 0.2	60.1 ± 0.4	50.4 ± 0.3	41.6
CORAL	59.2 ± 0.1	19.9 ± 0.2	47.4 ± 0.2	14.0 ± 0.4	59.8 ± 0.2	50.4 ± 0.4	41.8
MMD	32.2 ± 13.3	11.2 ± 4.5	26.8 ± 11.3	8.8 ± 2.2	32.7 ± 13.8	29.0 ± 11.8	23.5
DANN	53.1 ± 0.2	18.3 ± 0.1	44.2 ± 0.7	11.9 ± 0.1	55.5 ± 0.4	46.8 ± 0.6	38.3
CDANN	54.6 ± 0.4	17.3 ± 0.1	44.2 ± 0.7	12.8 ± 0.2	56.2 ± 0.4	45.9 ± 0.5	38.5
MTL	58.0 ± 0.4	19.2 ± 0.2	46.2 ± 0.1	12.7 ± 0.2	59.9 ± 0.1	49.0 ± 0.0	40.8
SagNet	57.7 ± 0.3	19.1 ± 0.1	46.3 ± 0.5	13.5 ± 0.4	58.9 ± 0.4	49.5 ± 0.2	40.8
ARM	49.6 ± 0.4	16.5 ± 0.3	41.5 ± 0.8	10.8 ± 0.1	53.5 ± 0.3	43.9 ± 0.4	36.0
V-REx	43.3 ± 4.5	14.1 ± 1.8	32.5 ± 5.0	9.8 ± 1.1	43.5 ± 5.6	37.7 ± 4.5	30.1
RSC	55.0 ± 1.2	18.3 ± 0.5	44.4 ± 0.6	12.5 ± 0.1	55.7 ± 0.7	47.8 ± 0.9	38.9
AND-mask	52.3 ± 0.8	17.3 ± 0.5	43.7 ± 1.1	12.3 ± 0.4	55.8 ± 0.4	46.1 ± 0.8	37.9
SAND-mask	43.8 ± 1.3	15.2 ± 0.2	38.2 ± 0.6	9.0 ± 0.2	47.1 ± 1.1	39.9 ± 0.6	32.2
Fishr	58.3 ± 0.5	20.2 ± 0.2	47.9 ± 0.2	13.6 ± 0.3	60.5 ± 0.3	50.5 ± 0.3	41.8
SelfReg	60.7 ± 0.1	21.6 ± 0.1	49.5 ± 0.1	14.2 ± 0.3	60.7 ± 0.1	51.7 ± 0.1	43.1
IDM	58.8 ± 0.3	20.7 ± 0.2	48.3 ± 0.1	13.7 ± 0.4	59.1 ± 0.1	50.2 ± 0.3	41.8

AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[35] Z. Wang and Y. Mao, "Information-theoretic analysis of unsupervised domain adaptation," in *The Eleventh International Conference on Learning Representations*, 2023.

[36] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish, "Invariance principle meets information bottleneck for out-of-distribution generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3438–3450, 2021.

[37] Y. Lin, H. Dong, H. Wang, and T. Zhang, "Bayesian invariant risk minimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 021–16 030.

[38] X. Zhou, Y. Lin, W. Zhang, and T. Zhang, "Sparse invariant risk minimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 27 222–27 244.

[39] R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters, "A causal framework for distribution generalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6614–6630, 2021.

[40] F. Hellström and G. Durisi, "A new family of generalization bounds using samplewise evaluated cmi," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 108–10 121, 2022.

[41] Z. Wang and Y. Mao, "Tighter information-theoretic generalization bounds from supersamples," in *International Conference on Machine Learning*. PMLR, 2023, pp. 36 111–36 137.

[42] M. Federici, R. Tomioka, and P. Forré, "An information-theoretic approach to distribution shifts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 628–17 641, 2021.

[43] B. Li, Y. Wang, S. Zhang, D. Li, K. Keutzer, T. Darrell, and H. Zhao, "Learning invariant representations and risks for semi-supervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1104–1113.

[44] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[45] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning al-

gorithms," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 670–24 682, 2021.
[46] A. T. Nguyen, T. Tran, Y. Gal, P. Torr, and A. G. Baydin, "KL guided domain adaptation," in *International Conference on Learning Representations*, 2022.
[47] S. Shahtalebi, J.-C. Gagnon-Audet, T. Laleh, M. Faramarzi, K. Ahuja, and I. Rish, "Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization," *arXiv preprint arXiv:2106.02266*, 2021.
[48] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *International conference on machine learning*. PMLR, 2019, pp. 7523–7532.
[49] P. Kamath, A. Tangella, D. Sutherland, and N. Srebro, "Does invariant risk minimization capture invariance?" in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 4069–4077.
[50] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, 2019.
[51] G. Aminian, Y. Bu, G. W. Wornell, and M. R. Rodrigues, "Tighter expected generalization error bounds via convexity of information measures," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 2481–2486.
[52] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the gibbs algorithm," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8106–8118, 2021.
[53] I. Fischer, "The conditional entropy bottleneck," *Entropy*, vol. 22, no. 9, p. 999, 2020.
[54] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *International Conference on Learning Representations*, 2019.
[55] K.-H. Lee, A. Arnab, S. Guadarrama, J. Canny, and I. Fischer, "Compressive visual representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 538–19 552, 2021.
[56] Y. Dong, T. Gong, H. Chen, S. Yu, and C. Li, "Rethinking information-theoretic generalization: Loss entropy induced pac bounds," in *The Twelfth International Conference on Learning Representations*, 2024.
[57] A. Pensia, V. Jog, and P.-L. Loh, "Generalization error bounds for noisy, iterative algorithms," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 546–550.
[58] H. Wang, Y. Huang, R. Gao, and F. Calmon, "Analyzing the generalization capability of sgld using properties of gaussian channels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 222–24 234, 2021.
[59] Q. Chen, C. Shui, and M. Marchand, "Generalization bounds for meta-learning: An information-theoretic analysis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 878–25 890, 2021.
[60] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *Entropy*, vol. 23, no. 1, p. 126, 2021.
[61] F. Hellström and G. Durisi, "Evaluated cmi bounds for meta learning: Tightness and expressiveness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 648–20 660, 2022.
[62] Y. Bu, H. V. Tetali, G. Aminian, M. Rodrigues, and G. Wornell, "On the generalization error of meta learning for the gibbs algorithm," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 2488–2493.
[63] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky, "Generalization bounds: Perspectives from information theory and pac-bayes," *arXiv preprint arXiv:2309.04381*, 2023.
[64] Y. Polyanskiy and Y. Wu, *Information theory: From coding to learning*. Cambridge university press, 2024.
[65] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2551–2559.
[66] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1657–1664.
[67] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550.
[68] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5018–5027.
[69] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 456–473.
[70] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.
[71] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 405–22 418, 2021.
[72] X. Wang, M. Saxon, J. Li, H. Zhang, K. Zhang, and W. Y. Wang, "Causal balancing for domain generalization," in *The Eleventh International Conference on Learning Representations*, 2023.
[73] Z. Wang, J. Grigsby, and Y. Qi, "PGrad: Learning principal gradients for domain generalization," in *The Eleventh International Conference on Learning Representations*, 2023.
[74] A. Setlur, D. Dennis, B. Eysenbach, A. Raghunathan, C. Finn, V. Smith, and S. Levine, "Bitrate-constrained DRO: Beyond worst case robustness to unknown group shifts," in *The Eleventh International Conference on Learning Representations*, 2023.
[75] Y. Chen, K. Zhou, Y. Bian, B. Xie, B. Wu, Y. Zhang, M. KAILI, H. Yang, P. Zhao, B. Han, and J. Cheng, "Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization," in *The Eleventh International Conference on Learning Representations*, 2023.
[76] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, "Improve unsupervised domain adaptation with mixup training," *arXiv preprint arXiv:2001.00677*, 2020.
[77] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
[78] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8690–8699.
[79] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 124–140.
[80] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9619–9628.
[81] O. Pooladzandi, D. Davini, and B. Mirzasoleiman, "Adaptive second order coresets for data-efficient machine learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 848–17 869.
[82] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 10 237–10 297, 2022.
[83] D. Teney, E. Abbasnejad, S. Lucey, and A. Van den Hengel, "Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 761–16 772.

Yuxin Dong received the B.E. and Ph.D. degrees in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2019 and 2024, respectively. His research interests include information theory, statistical learning theory, and bioinformatics.

Tieliang Gong received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2018. From September 2018 to October 2020, he was a Post-Doctoral Researcher with the Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada. He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. His research interests include statistical learning theory, machine learning, and information theory.

Hong Chen received the B.S., M.S., and Ph.D. degrees from Hubei University, Wuhan, China, in 2003, 2006, and 2009, respectively. Currently, he is a Professor with the Department of Mathematics and Statistics, College of Informatics, Huazhong Agricultural University, Wuhan, China. His current research interests include machine learning, statistical learning theory, and approximation theory.

Shuangyong Song received the Ph.D. degree from the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently a Manager with the China Telecom Corporation, Beijing, China. His research interests include information retrieval, web/text mining, and natural language processing.

Weizhan Zhang (Senior Member, IEEE) received the B.S. degree in applied electronics from Zhejiang University, Hangzhou, China, in 1999, and the Ph.D. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2010. He was a Software Engineer at Datang Telecom, Beijing, China, from 1999 to 2002, and a Visiting Scholar at Pennsylvania State University, University Park, PA, USA, from 2015 to 2016. He joined the faculty of Xi'an Jiaotong University as an Assistant Professor in 2010, an Associate Professor in 2014, and a Professor in 2019. He is currently a Professor with the School of Computer Science and Technology, Xi'an Jiaotong University.

Chen Li received his Ph.D. degree from the University of Cambridge, UK, in 2014. From Jun 2014 to Mar 2016, he was a Post-Doctoral Researcher with the Massachusetts Institute of Technology, USA. He is currently a Professor with the School of Computer Science and Technology, Xi'an Jiaotong University, China. His research interests include natural language processing, biological text mining, digital pathology, and bioinformatics.