Can Variance-Based Regularization Improve Domain Generalization?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Without prior information, domain generalization with only access to multi-domain training data relies on guessing what the test data is. In this work, we consider mild assumptions that there is a distribution over domains and the out-of-distribution data is generated by the shift of the domain distribution. We study a domain-level variance-based regularizer. We show that the variance-regularized method locally approximates the group distributionally robust optimization and embeds the local information into the objective function as a weighting scheme. By taking the empirical domain distribution as an anchor of the location, we propose a weighting correction scheme and provide guarantees of in-distribution generalization. Compared to the Empirical Risk Minimization, we prove the potential benefits of our proposed method but do not observe consistent improvements in general.

1 Introduction

Domain generalization [12, 28] is an out-of-distribution (OOD) generalization problem and has drawn much attention recently [39, 44, 35]. Some recent works consider an ambitious goal that generalizes to "absolutely" unseen domain by learning domain-invariant features. From the perspective of theory, the price of such invariant learning methods is the requirement for harsh assumptions or strong prior information, which is necessary to guarantee that the invariance exists and is identifiable. In this work, we assume that there exits a distribution of domains and the OOD test data is generated by the shift of the domain distribution. Then domain generalization is formulated into a distributionally robust optimization problem (DRO, [9, 11, 10]).

Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ be a data point consisting of an input vector $\mathbf{x} \in \mathcal{X}$ and the target label $\mathbf{y} \in \mathcal{Y}$. Suppose the training data is structured with respect to a latent domain label:

$$\mathcal{D}_{tr} = \{\mathbf{z}_l, 1 \le l \le m\} = \{\{\mathbf{z}_{i,j}, 1 \le j \le m_i\}, 1 \le i \le n\}, \tag{1}$$

where m is the total sample size , m_i is the sample size of the i-th domain and n is the number of domains. We assume that the training domains are randomly drawn from possible domains with a domain distribution Q, i.e. $\mathcal{E}_{tr} = \{e_1, e_2, \dots, e_n\} \subseteq \mathcal{E}$ with $e_i \sim Q$ and the data points under domain e is sampled from the distribution P_e. Let $\mathcal{H}$ be the hypothetical space and $h \in \mathcal{H}$ be a model that maps $\mathbf{x} \in \mathcal{X}$ to $h(\mathbf{x}) \in \mathcal{Y}$. The loss function $\ell(\hat{\mathbf{y}}, \mathbf{y}) : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ measures how poorly the output $\hat{\mathbf{y}} = h(\mathbf{x})$ predicts the target $\mathbf{y}$. Denote $\mathcal{F}$ as the collection of the functions $f = \ell(h(\cdot), \cdot) : \mathbf{z} \to [0, M]$ with $h \in \mathcal{H}$. The in-domain expected risk and its sample average approximation ([34]) are denoted by

$$R(f|e_i) = \mathbb{E}_{\mathbf{z} \sim P_{e_i}}[f(\mathbf{z})] \quad \text{and} \quad \hat{R}(f|e_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} f(\mathbf{z}_{i,j}) \tag{2}$$

respectively. The distribution shift between training and test data is characterized by the change of Q, while the data distributions P_e, $e \in \mathcal{E}$ are fixed.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We study the group distributionally robust optimization problem (group DRO, [21, 30, 33]):

$$\min_{f\in\mathcal{F}} \max_{Q} \mathbb{E}_{\mathbf{z}\sim P}[f(\mathbf{z})], \quad s.t. \quad P = \int P_e Q(\mathrm{de}),\ D_\phi(Q\|Q_0) \le \rho, \tag{3}$$

where Q_0 is a selected domain distribution, $D_\phi(\cdot\|\cdot)$ stands for the ϕ-divergence ([3, 14]) and the tuning parameter ρ modulates the distribution shift. Throughout this paper, $D_\phi(\cdot\|\cdot)$ is the χ^2-divergence, i.e., $\phi(t) = \frac{1}{2}(t-1)^2$. Sagawa* et al. [33] consider the empirical optimization problem,

$$\min_{f\in\mathcal{F}} \max_{\boldsymbol{q}\in\Delta_n} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) \quad \text{with} \quad \Delta_n = \left\{(q_1,\ldots,q_n) : q_i \ge 0, \sum_{i=1}^{n} q_i = 1\right\}.$$

Here Δ_n is the $(n-1)$-dimensional probability simplex. In this case, the parameter ρ is fixed and sufficiently large. For more ambitious goals, Krueger et al. [25] propose the minimax risk extrapolation (MM-REx) that extends the uncertainty region Δ_n into

$$\tilde{\Delta}_n(\alpha) = \left\{\boldsymbol{q} = (q_1,\ldots,q_n) : q_i \ge \alpha, \sum_{i=1}^{n} q_i = 1\right\},$$

where the parameter $\alpha \in (-\infty, 1/n]$ modulates the uncertainty region. The negative value of α extrapolates risks and encourages robustness to large distribution shifts.

At the sample level, the DRO loss can be asymptotically approximated by the sum of the ERM loss [38] and a variance-based regularizer [17], where the negligible error term converges to zero almost surely. Section 7 in [17] gives general results when the DRO objective is a Hadamard differentiable functional to P and $\mathcal{F}$ is a P_0-Donsker class. From the perspective of generalization, the upper bound of the prediction risk may also have a variance-based regularization term that trades between approximation error and estimation error [5, 6, 13, 24]. Sample variance penalization [27] replaces the variance-based regularization with its empirical estimator and gives theoretical guarantees on the prediction performance. To address the computationally intractable problem caused by the non-convexity of the regularizer, Namkoong and Duchi [29] and Duchi and Namkoong [16] investigate the robustly regularized risk, that provides a convex surrogate for variance-regularized loss, and prove finite-sample and asymptotic results characterizing prediction performance. Back to domain generalization problem, Krueger et al. [25] develop a variance-regularized empirical loss (V-REx): $\tilde{R}(f) + \lambda\tilde{V}_{out}(f)$, where

$$\tilde{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \hat{R}(f|e_i) \quad \text{and} \quad \tilde{V}_{out}(f) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{R}(f|e_i) - \tilde{R}(f)\right)^2.$$

Xie et al. [41] prove that with high probability, optimizing the regularized loss $\tilde{R}(f) + \lambda\sqrt{\tilde{V}_{out}(f)}$ is equivalent to solve a MM-REx problem.

In this work, we refine $\tilde{R}(f)$ and $\tilde{V}_{out}(f)$ based on the intuitive understanding of generalization and distribution estimation. Recall the problem in (3). In general, Q_0 is the ground-truth domain distribution and Q belongs to a neighborhood of Q_0. Therefore, the empirical version of (3) should replace Q_0 with its empirical approximation over $\mathcal{E}_{tr}$, i.e.,

$$\hat{\boldsymbol{q}} = (\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_n) = (\frac{m_1}{m}, \ldots, \frac{m_n}{m}).$$

However, the existing variance-regularized methods directly replace Q_0 with a discrete uniform distribution (the center of $\tilde{\Delta}_n(\alpha)$) without considering a consistent and efficient estimator $\hat{\boldsymbol{q}}$. In the sample variance penalization, this problem does not exist because the discrete uniform distribution on sample points (no tie), i.e. the empirical distribution, is a consistent estimator of the ground-truth data distribution. Consider a new uncertainty region:

$$\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}}) = \tilde{\Delta}_n(\alpha) \cap \left\{\boldsymbol{q} : D_\phi(\boldsymbol{q}\|\hat{\boldsymbol{q}}) \le \rho\right\}.$$

Specifically, any $\boldsymbol{q} = (q_1,\ldots,q_n) \in \mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$ satisfies

$$q_i \ge \alpha, \quad \sum_{i=1}^{n} q_i = 1, \quad \sum_{i=1}^{n} \frac{1}{2}(\frac{q_i}{\hat{q}_i} - 1)^2 \hat{q}_i \le \rho.$$

In Section 3.2, we prove that with high probability, the MM-REx problem on $\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$ can be uniformly equivalent to minimize the variance-regularized empirical loss $\hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)}$ where

$$\hat{R}(f) = \sum_{i=1}^{n} \hat{q}_i \hat{R}(f|e_i) \quad \text{and} \quad \hat{V}_{out}(f) = \sum_{i=1}^{n} \hat{q}_i \big(\hat{R}(f|e_i) - \hat{R}(f)\big)^2. \tag{4}$$

Comparing to $\tilde{R}(f)$ and $\tilde{V}_{out}(f)$, the two terms $\hat{R}(f)$ and $\hat{V}_{out}(f)$ just introduce a weighting scheme derived from the empirical domain distribution $\hat{\boldsymbol{q}}$. In addition, the term $\hat{R}(f)$ is the exact ERM loss [38]. In Section 3.1, we investigate the generalization guarantee of the variance-regularized estimator,

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \ \hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)}, \tag{5}$$

via the covering number of the function class $\mathcal{F}$. Appendix C also provides a version of the generalization guarantee with localized Rademacher complexities, which may provide tighter generalization bounds in some cases. In Section 4, we consider a general uncertainty region $\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)$, where the choice of $\boldsymbol{q}_0$ represents a kind of prior knowledge. Similar to the arguments in Section 3, we can also write $\boldsymbol{q}_0$ as a weight assignment and embed it into the variance-regularized loss function. We present a general form of the proposed method and prove that the optimization equivalence in Section 3.2 still holds when we replace $\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$ with $\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)$.

Our results clearly show that

- From the perspective of generalization, we propose a weighting correction scheme for variance-regularized domain generalization methods. The proposed method can outperform ERM under some cases, which shows the potential competitive edge of the proposed weighting correction method.
- We do not observe that our method consistently improves ERM under general cases.
- The proposed method is robust to the change of the domain distribution Q. From an optimization perspective, it is equivalent to solve a group DRO problem.

2 Preliminaries

In this section, we present the rationale for using variance-based regularization to improve the robustness of generalization. Section 2.1 gives two domain adaptation examples that the test data is known. We prove that the standard deviation of risk can bound the generalization gap between training and test data. In Section 2.2, we formulate an invariant learning principle as a hypothesis testing problem. We point out that penalizing the risk variance can protect the null hypothesis: the model is invariant across domains.

2.1 Risk variance bounds generalization gap

We present two simple examples to show that penalizing the standard deviation of risk is a natural strategy to improve robustness to the domain distribution shift.

Risk Interpolation. In the first example, we assume the test distribution belongs to the convex hull of training domains. This is a typical risk interpolation case. Let P^* be the test distribution. Suppose there exists $\boldsymbol{q}^* = (q_1^*, \cdots, q_n^*) \in \Delta_n$ such that $P^* = \sum_{i=1}^n q_i^* P_{e_i}$, where $P_{e_i}, 1 \le i \le n$ are training domains. Then the generalization gap between the training and test data is

$$\text{err}_f = \sum_{i=1}^{n} q_i^* R(f|e_i) - \sum_{i=1}^{n} q_i R(f|e_i) = \sum_{i=1}^{n} (q_i^* - q_i)\Big(R(f|e_i) - \sum_{i=1}^{n} q_i R(f|e_i)\Big),$$

where $q_i = Q(\mathrm{d}e_i)/Q(\mathrm{d}\mathcal{E}_{tr})$ is the proportion of the training domain e_i in the training data. We write $\boldsymbol{q} = (q_1, \ldots, q_n)$. By the Cauchy–Schwarz inequality, we have

$$\text{err}_f \le \sqrt{2D_\phi(\boldsymbol{q}^*\|\boldsymbol{q})} \times \sqrt{V_{out}(f)}, \tag{6}$$

where $V_{out}(f)$ is the between-domain risk variance over the training domains:

$$V_{out}(f) = \sum_{i=1}^{n} q_i \Big(R(f|e_i) - \sum_{i=1}^{n} q_i R(f|e_i) \Big)^2.$$

Notice that $V_{out}(f)$ only depends the training data. Therefore, it is natural to penalize $\sqrt{V_{out}(f)}$ to obtain a tight upper bound of the test error. The principle here is that if for $\forall e_i \in \mathcal{E}_{tr}$, $R(f|e_i)$ is a constant that only depends on f, i.e. $V_{out}(f) = 0$, then changes from $\boldsymbol{q}$ to $\boldsymbol{q}^*$ cannot cause any generalization gap.

Sub-population Shift. Recall that the training data in (1) is structured with respect to a latent domain label. In this example, the domain label is the class label $\mathbf{y}$. Therefore, the marginal distribution of $\mathbf{y}$ is different in the training and test data, and the conditional distribution $P(\mathbf{x}|\mathbf{y})$ is the same. Let $\mathcal{Y} = \{1, 2, \ldots, K\}$. Then the generalization gap between the training and test data is

$$\begin{aligned} \text{err}_f &= \sum_{k=1}^{K} \mathbb{E}[f(\mathbf{z})|\mathbf{y}=k] \times \big(P_{e'}(\mathbf{y}=k) - P_e(\mathbf{y}=k)\big) \\ &\leq \sqrt{2D_\phi(P_{e'}(\mathbf{y}) \| P_e(\mathbf{y}))} \times \sqrt{V_{out}(f)}, \end{aligned}$$

where

$$V_{out}(f) = \sum_{k=1}^{K} P_e(\mathbf{y}=k) \Big(\mathbb{E}[f(\mathbf{z})|\mathbf{y}=k] - \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[f(\mathbf{z})|\mathbf{y}=k] \Big)^2$$

is the between-class risk variance over the training data. Therefore, the generalization gap is also bounded above by the between-domain risk variance. If the in-class risks are equal, i.e., $\mathbb{E}[f(\mathbf{z})|\mathbf{y} = k] = \mathbb{E}[f(\mathbf{z})|\mathbf{y} = k'], \ \forall k, k' \in \mathcal{Y}$, then the sub-population shift cannot cause generalization gap.

2.2 Penalizing risk variance protects invariant models

In this section, we heuristically discuss the relationship between variance-based regularization and invariant learning. The REx principle [25] presents two training goals: **Reducing training risks** and **Increasing the similarity of training risks**. Krueger et al. [25] heuristically explain the utility of V-REx as enforcing the equality of training risks in the limit case $\lambda \to +\infty$. In some experiments, V-REx with small λ also shows robust generalization and may outperform ERM. Here we understand this phenomenon by extending the REx principle to the population level:

(i) Minimizing the expected risk $R(f)$;

(ii) Cannot reject the null hypothesis of the test:

$$H_0 : R(f|e) = R(f|e'), \forall e, e' \in \mathcal{E} \quad \text{vs} \quad H_1 : R(f|e) \neq R(f|e'), \exists e, e' \in \mathcal{E}. \tag{7}$$

In general, Principle (i) is achieved by minimizing the ERM loss. Next we show that variance-based regularization is related to the hypothesis testing problem in Principle (ii). Under regular assumptions, one can use the one-way ANOVA F-test to check the hypothesis testing in (7). The F-test statistic is the ratio of the between-domain variance to the in-domain variance, i.e.,

$$F = \frac{\hat{V}_{out}(f)}{\hat{V}(f) - \hat{V}_{out}(f)} \quad \text{with} \quad \hat{V}(f) = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \big(f(\mathbf{z}_{ij}) - \hat{R}(f) \big)^2.$$

Here $\hat{V}_{out}(f)$ and $\hat{R}(f)$ are defined in (4). If F is larger than a threshold, e.g. the $(1 - 5\%)$-quantile of a F distribution, one should reject the null hypothesis. Here 5% is the significance level. If the in-domain variance of a well-trained model is approximately stable, then *penalizing $\hat{V}_{out}(f)$ is equivalent to a constraint that H_0 cannot be rejected.* Therefore our proposed method that penalizes $\hat{V}_{out}(f)$ is consistent with the REx principle and the regularization term $\hat{V}_{out}(f)$ is a generalized version of V-REx.

3 Variance-Based Regularization

Motivated by Section 2, we study a variance-based regularization method for domain generalization, which minimizes the following empirical loss function:

$$\hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)}, \tag{8}$$

where λ is a tuning parameter and $\hat{V}_{out}(f)$ is an empirical estimator of the between-domain risk variance. The proposed loss (8) directly optimizes the ERM principle $\hat{R}(f)$, which is different to the recent invariant learning methods that minimize $\tilde{R}(f)$, e.g. Invariant Risk Minimization [1]. The regularization term is slightly different to V-REx: (i) The square-root operator is derived from generalization gap; (ii) *Different to the empirical variance of $R(f|e)$, we penalize the between-domain variance of $f(\mathbf{z})$.*

We consider Q_0 in (3) as the training domain distribution and denote the training distribution as $P_0 = \int P_e Q_0(\mathrm{d}e)$. To proceed further, we denote more notations as follows:

$$R(f) = \mathbb{E}_{e\sim Q_0}[R(f|e)] = \mathbb{E}_{\mathbf{z}\sim P_0}[f(\mathbf{z})], \quad V(f) = \mathbb{E}_{\mathbf{z}\sim P_0}\big[(f(\mathbf{z}) - R(f))^2\big],$$
$$V_{in}(f|e) = \mathbb{E}_{\mathbf{z}\sim P_e}\big[(f(\mathbf{z}) - R(f|e))^2\big], \quad V_{out}(f) = \mathbb{E}_{e\sim Q_0}\big[(R(f|e) - R(f))^2\big],$$

where $R(f|e)$ is defined in (2). Here $V_{out}(f)$ is the between-domain variance and $V_{in}(f|e)$ is the in-domain variance of the domain $e \in \mathcal{E}$. According to the decomposition of the total variance, we have

$$V(f) = \mathrm{Var}\big(\mathbb{E}[f(\mathbf{z})|e]\big) + \mathbb{E}\big[\mathrm{Var}(f|e)\big] = V_{out}(f) + \mathbb{E}_{e\sim Q_0}\big[V_{in}(f|e)\big].$$

When Q_0 and P_e are replaced by the corresponding empirical distributions, we rewrite $V(f)$, $V_{in}(f|e)$ and $V_{out}(f)$ as $\hat{V}(f)$, $\hat{V}_{in}(f|e)$ and $\hat{V}_{out}(f)$ respectively. In the finite-sample setup, the decomposition of the total variance also holds:

$$\hat{V}(f) = \hat{V}_{out}(f) + \sum_{i=1}^{n} \frac{m_i}{m}\hat{V}_{in}(f|e).$$

3.1 Generalization

Since the empirical loss (8) is derived from generalization bounds, we present two versions of the generalization guarantee. The first result depends on the covering number of the function class $\mathcal{F}$. In the Appendix, we also derive a version of the generalization bound with localized Rademacher complexities, which can provide more refined uniform generalization bounds in some cases.

We start with the definition of the covering number. Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \times \mathcal{Y} \to [0, M]$. Suppose $\mathcal{F}$ is a subset of a metric space with a norm $\|\cdot\|$. We say a collection $\{f^1, \ldots, f^N\} \subseteq \mathcal{F}$ is an ϵ-cover of $\mathcal{F}$ if for each $f \in \mathcal{F}$, there exists f^i such that $\|f - f^i\| \le \epsilon$. The covering number of $\mathcal{F}$ is

$$N\big(\mathcal{F}, \epsilon, \|\cdot\|\big) \;:=\; \inf\Big\{N \in \mathbb{N} : \text{there exists a collection } \{f^1, \ldots, f^N\} \text{ which is an } \epsilon\text{-cover of } \mathcal{F} \text{ with respect to } \|\cdot\|\Big\}.$$

In the following, we use the ℓ^∞ norm: $\|f - g\|_\infty = \sup_{z\in\mathcal{X}\times\mathcal{Y}} |f(z) - g(z)|$. Now we are ready to present the following theorem:

Theorem 1 *Let $n \ge 2$ and $\{\mathbf{z}_{i,j}, 1 \le i \le n, 1 \le j \le m_i\}$ is an i.i.d sample drawn from P_0. Suppose $f(z) \in [0, M]$ for any $f \in \mathcal{F}$ and $z \in \mathcal{X} \times \mathcal{Y}$ and the function class $\mathcal{F}$ has the over number: $N_\epsilon = N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})})$. Let $0 < \delta < 1$ and*

$$t = \log\frac{(n+2)N_\epsilon}{\delta}, \quad \lambda = \sqrt{\frac{2t}{m-1}}.$$

Then we have, with probability at least $1-\delta$*,*

$$\begin{aligned} R(f) \quad \le \quad & \hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)} + \sum_{i=1}^{n} \lambda\sqrt{\frac{(m_i-1)V_{in}(f|e_i)}{m}} \\ & + \sum_{i=1}^{n} \frac{\sqrt{(m-1)m_i}M\lambda^2}{\sqrt{m(m_i-1)}} + \frac{(4m-1)M\lambda^2}{3m} \\ & + \Big(2+\lambda+\sum_{i=1}^{n}\lambda\sqrt{\frac{(m_i-1)}{m}}\Big)\epsilon, \end{aligned}$$

holds for every $f \in \mathcal{F}$*.*

The proof of Theorem 1 is presented in the Appendix B. In some cases, the covering number-based analysis cannot provide a tight generalization bound [4, 5, 36]. Therefore, we also use the local Rademacher complexity [5] to present the generalization of the proposed variance-based regularization. The details and proof are postponed into Appendix C.

Why we study In-Distribution generalization? Theorem 1 provides the generalization guarantee for the in-distribution (ID) generalization rather than the OOD generalization. But its result gives important insights into the OOD generalization. First, the ID error provides a lower bound for the worst-case OOD error since $\mathcal{E}_{tr}$ is a subset of $\mathcal{E}$. Second, some empirical studies of OOD generalization have observed a linear relationship between the ID and OOD test error [31, 20, 22]. Third, some OOD generalization bounds are derived from a domain adaptation framework [8, 7, 2, 42, 43], e.g.,

$$\text{OOD error} \le \text{ID error} + \text{error gap} + O(\cdot), \tag{9}$$

which starts from the ID error and then depicts the error gap. *Most recent works focus on minimising the error gap and ignore how their robust (or invariant) methods increase the ID test error.* Fourth, our assumptions are mild and general. We do not impose strong constraint on the test data, e.g. structured generative mechanism, and only assume the domain distribution shift. Therefore, we analyze the ID error of the proposed robust method under mild assumptions.

We denote f^* as the optimal function and let $\hat{f}$ be a solution:

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)}.$$

Next we study the excess risk of $\hat{f}$. According to Theorem 1, we obtain the following result.

Corollary 2 *Suppose the assumptions in Theorem 1 hold. Let* $0<\delta<1$ *and*

$$t = \log\frac{2N_\epsilon+2}{\delta}, \quad \lambda = \sqrt{\frac{2t}{m-1}}.$$

Then, with probability at least $1-\delta$*,*

$$\begin{aligned} R(\hat{f}) - R(f^*) \quad \le \quad & 2\lambda\sqrt{\frac{(m-1)V(f^*)}{m}} + \sum_{i=1}^{n}\lambda\sqrt{\frac{m_i\hat{V}_{in}(\hat{f}|e_i)}{m}} \\ & + \Big(2+\lambda+\sum_{i=1}^{n}\lambda\sqrt{\frac{m_i}{m}}\Big)\epsilon + \lambda^2\frac{4(4m-1)M}{3m}. \end{aligned}$$

Parametric Example. Suppose the hypothetical space $\mathcal{F}$ is a class of parametric functions:

$$\mathcal{F} = \{f_\theta(z) : z \in \mathcal{X}\times\mathcal{Y},\ \theta\in\Theta\subseteq\mathbb{R}^d\},$$

where the parameter set Θ is bounded. Further, for any data point $\mathbf{z}$, $f_\theta(\mathbf{z})$ is a L-Lipschitz function of θ with respect to ℓ^2 norm on Θ. Then the covering number is bounded above:

$$N_\epsilon \le \Big(1+\text{diam}(\Theta)\cdot L\cdot\frac{1}{\epsilon}\Big)^d, \quad \text{with} \quad \text{diam}(\Theta) = \sup_{\theta,\theta'\in\Theta}\|\theta-\theta'\|_2.$$

Then we take

$$\epsilon = \frac{1}{m}, \quad \log N_\epsilon = O(\log m), \quad \lambda = O(\sqrt{\frac{\log m}{m}}).$$

Therefore, by Corollary 2, with probability at least $1-\delta$,

$$R(\hat{f}) - R(f^*) \le 2\lambda\sqrt{\frac{(m-1)V(f^*)}{m}} + \sum_{i=1}^{n}\lambda\sqrt{\frac{m_i\hat{V}_{in}(\hat{f}|e_i)}{m}} + O(\frac{\log m}{m}). \tag{10}$$

Potential competitive edge. The second term on the RHS of (10) contains the empirical in-domain variance $\hat{V}_{in}(\hat{f}|e_i)$. For over-parameterized model, the empirical in-domain variance of $\hat{f}$ can be close to zero. If there exists an optimal function $f^* \in \arg\min_f R(f)$ such that $V(f^*) = 0$, then the term $O(\log m/m)$ dominants the convergence rate of the excess risk.[1] For ERM, the convergence rate of the excess risk is $1/\sqrt{m}$, which is slower than $\log m/m$. *Due to the fast convergence rate, our proposed method can outperform ERM when the sample size m is large enough.*

Cannot consistently outperform ERM. If there is no optimal function $f^* \in \arg\min_f R(f)$ satisfies $V(f^*) = 0$, then the first term on the RHS (10) can dominant the excess risk. In this case, the convergence rate of the the excess risk of our method is $\sqrt{\log m/m}$, which is slower than ERM. This implies that *if $V(f^*) > 0$ for $\forall f^* \in \arg\min_f R(f)$, ERM can outperform our method when m is large enough.*

OOD generalization. According to Eq. (9), OOD error can be rewritten as a sum of ID error and error gap. The distance between the training and test domain distribution can determine the error gap term under our setup. Furthermore, we only assume that the training and test domain distributions are close but different, and do not impose any structured generative models, such as structural equation models [40] or probabilistic graphical models [23]. In other words, we do not introduce prior information and use mild and general assumptions. Due to the uncertainty of the test data, the error gap should be the worst-case error gap for the domain distribution shift and hypothetical space, which is independent of the estimator. This implies that without prior information, the ID error is a reliable metric to infer the OOD error.

Non-convexity. Similar to the Sample Variance Penalization [27], the proposed objective function (8) is in general non-convex and computationally intractable. The proposed regularization term is non-convex even if the loss function is convex. It is still unclear how to actually minimize the variance-regularized objective function. Krueger et al. [25] use a penalty annealing scheme to obtain a good pre-train model. In the Appendix A, we empirically show that our method can use random initialization without dropping generalization performance.

3.2 Optimization

In this section, we show that minimizing (8) is equivalent to solving a group DRO problem concerning a local neighbourhood of the empirical domain distribution. Let $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$ be a discrete distributions defined on the domain set $\mathcal{E}_{tr} = \{e_1, e_2, \ldots, e_n\}$. We consider the following optimization problem that minimizes

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i), \tag{11}$$

which is slightly different to group DRO problem because $\mathcal{Q}_\alpha(\hat{\boldsymbol{q}}, \rho)$ is not centered at the uniform discrete distribution. We denote $\lambda = \sqrt{2\rho}$ and rewrite the empirical loss in (8) as

$$\mathcal{L}(f;\rho) = \hat{R}(f) + \sqrt{2\rho\hat{V}_{out}(f)}. \tag{12}$$

The following theorem shows that the objective (11) is bounded by two variance-regularized functions in the form of (12).

[1] The factor $\log m$ comes frome the covering number N_ϵ. If the hypothetical space $\mathcal{F}$ only contains finite models, N_ϵ is a constant and is independent to m. Then the convergence rate of the excess risk is $1/m$.

Theorem 3 *Suppose the training dataset $\mathcal{D}$ and a function $f \in \mathcal{F}$ are given. Let ρ_+ be the largest distance between $\hat{\boldsymbol{q}}$ and $\boldsymbol{q} \in \mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}})$ and*

$$\rho_- = \frac{\min_i (\alpha/\hat{q}_i - 1)^2 \hat{V}_{out}(f)}{2\big(\min_i \hat{R}(f|e_i) - \hat{R}(f)\big)^2},$$

then we have

$$\mathcal{L}(f;\rho_-) \leq \max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) \leq \mathcal{L}(f;\rho_+). \tag{13}$$

This second inequality in (13) implies that the optimization problem (11) with $\rho = +\infty$ is always bounded above by the variance-regularized loss with the tuning parameter ρ_+. On the other hand, we can also derive a tuning parameter ρ_- depends on the training data and a given model f, and then prove that $\mathcal{L}(f;\rho_-)$ is a lower boundary of (11). According to the proof of Theorem 3, one can find that the equality holds:

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) = \hat{R}(f) + \sqrt{2\rho \hat{V}_{out}(f)},$$

when the radius ρ satisfies $\rho \leq \rho_-$. If $\hat{V}_{out}(f)$ is nonzero and ρ is given, the equality holds if and only if $\forall e_i \in \mathcal{E}_{tr}$,

$$\alpha \leq \hat{q}_i \Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f)}} \big(\hat{R}(f|e_i) - \hat{R}(f)\big) + 1 \Big). \tag{14}$$

Therefore, the parameter α and the radius ρ govern each other.

Sketch of Proof: We start with a preliminary result: for any α and ρ,

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) \leq \mathcal{L}(f;\rho),$$

which is directly derived from the Cauchy-Schwarz inequality. By checking the conditions for the equality, we obtain the constraints in (14). Note that $\mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}}) \subseteq \mathcal{Q}_{+\infty,\rho_+}(\hat{\boldsymbol{q}})$ since ρ_+ is the largest distance between $\hat{\boldsymbol{q}}$ and $\boldsymbol{q} \in \mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}})$. Hence the second inequality in (13) is trivial. Let $\boldsymbol{q}_-^*$ be

$$\boldsymbol{q}_-^* \quad = \quad \arg\max_{\boldsymbol{q}\in\mathcal{Q}_{+\infty,\rho_-}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i).$$

Furthermore, $\boldsymbol{q}_-^* \in \mathcal{Q}_{\alpha,\rho_-}(\hat{\boldsymbol{q}}) \subseteq \mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}})$. Hence the first inequality in (13) holds. □

Theorem 3 shows the equivalence between group DRO and the variance-based regularization in (12). However, the lower bound $\mathcal{L}(f;\rho_-)$ still depends on the training data and model. Next we use concentration inequalities and the covering numbers of $\mathcal{F}$ to derive the uniform results.

Theorem 4 *Suppose that α is a non-positive scalar and $V'_{out}(f) = V(f) - \sum_i q_i V_{in}(f|e_i) > 0$. For each training domain, both $\hat{q}_i$ and q_i are larger than $\delta > 0$. We write*

$$\rho' = \frac{V'_{out}(f)}{16M^2}\left(\frac{\alpha}{1-(n-1)\delta} - 1\right)^2.$$

Let $\tau > 0$ and $0 < \eta < 1$ be two constants. Define

$$\mathcal{F}_{\tau,\eta} \quad = \quad \{f \in \mathcal{F} : V(f) \geq \tau, \text{ and } \frac{V_{in}(f|e_i)}{V(f)} \leq \eta, \forall e_i \in \mathcal{E}_{tr}\}.$$

For any $f \in \mathcal{F}_{\tau,\eta}$, the following expansion uniformly holds:

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho'}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) = \mathcal{L}(f;\rho'),$$

with probability at least $1-N_{\tau,\eta}\times p$*, where* $N_{\tau,\eta}=N\Big(\mathcal{F}_{\tau,\eta},\sqrt{\frac{1}{10}(1-\eta)\tau},\|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}\Big)$ *is the covering number of* $\mathcal{F}_{\tau,\eta}$ *and*

$$\begin{aligned} p &= \exp\Big(-\frac{m(1-\eta)^2\tau}{32M^2}+\frac{1}{16}\Big)+\begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix}\exp\Big(-\frac{m(1-\eta)^2\tau^2}{M^4}\Big) \\ &\quad +\sum_{i=1}^{n}\exp\Big\{-\frac{1}{2M^2m_i}\Big(\frac{m_i(1-\eta)+4\eta}{1+3\eta}\Big)^2\Big\}. \end{aligned}$$

4 General Version

Recall the uncertainty region in (3): $\{Q: D_\phi(Q\|Q_0)\le\rho\}$. In Section 3, Q_0 is the ground-truth domain distribution. In fact, it can be a selected anchor distribution closed to the target test domain. The choice of Q_0 can be regarded as a kind of prior knowledge and the hyperparameter ρ represents how strong is the confidence in the prior. We formulate the finite-sample optimization problem as

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)}\sum_{i=1}^{n}q_i\hat{R}(f|e_i), \tag{15}$$

where $\boldsymbol{q}_0$ is the conditional distribution of e given $e\in\mathcal{E}_{tr}$, which is derived from Q_0. In this problem, the uncertainty region $\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)$ is centered at a discrete distribution $\boldsymbol{q}_0$ rather than the uniform distribution or the empirical distribution $\hat{\boldsymbol{q}}$. Therefore, we can manually select $\boldsymbol{q}_0$ to introduce the prior information.

According to the proof of Theorem 3 in the , the optimization equivalence in Section 3.2 also holds when we replace $\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$ with $\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)$. To proceed further, we rewrite $\mathcal{L}(f;\rho,\boldsymbol{q}_0)=\hat{R}(f,\boldsymbol{q}_0)+\sqrt{2\rho\hat{V}_{out}(f,\boldsymbol{q}_0)}$ with

$$\hat{R}(f,\boldsymbol{q}_0)=\sum_{i=1}^{n}q_{0,i}\hat{R}(f|e_i)\quad\text{and}\quad\hat{V}_{out}(f,\boldsymbol{q}_0)=\sum_{i=1}^{n}q_{0,i}\big(\hat{R}(f|e_i)-\hat{R}(f,\boldsymbol{q}_0)\big)^2.$$

Then we restate Theorem 3 as the following general version.

Theorem 5 *Given the training dataset and a function* $f\in\mathcal{F}$*, then for any distribution* $\boldsymbol{q}_0$*, the inequality always holds:*

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)}\sum_{i=1}^{n}q_i\hat{R}(f|e_i)\le\mathcal{L}(f;\rho,\boldsymbol{q}_0).$$

If the between-domain variance $\hat{V}_{out}(f,\boldsymbol{q}_0)$ *is non-zero, the equality holds if and only if* $\forall e_i\in\mathcal{E}_{tr}$*,*

$$\alpha\le q_{0,i}\Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f,\boldsymbol{q}_0)}}\big(\hat{R}(f|e_i)-\hat{R}(f,\boldsymbol{q}_0)\big)+1\Big).$$

On the other hand, if α *is fixed, the equality holds when the radius of* $\mathcal{Q}_{\alpha,\rho}(\boldsymbol{q}_0)$ *satisfies*

$$\rho\le\frac{\min_i(\alpha/q_{0,i}-1)^2\hat{V}_{out}(f,\boldsymbol{q}_0)}{2\big(\min_i\hat{R}(f|e_i)-\hat{R}(f,\boldsymbol{q}_0)\big)^2}.$$

This result shows the equivalence between the optimization problem (15) and the variance-regularized loss $\mathcal{L}(f;\rho,\boldsymbol{q}_0)$. Therefore, for unbalanced domains and any given prior $\boldsymbol{q}_0$, we can still use the variance-based regularization to approximate the DRO problem. Please refer to the Appendix for the complete proof of Theorem 5.

5 Conclusion

In this work, we study a variance-based regularization method for domain generalization. We prove the guarantees for in-distribution generalization and figure out the potential benefits of our proposed method compared to ERM. Our proposed objective function is non-convex and the optimization procedure is computationally intractable. The learnt model can be highly dependent on initialization or pretraining. In future work, we will consider combining generalization bounds with specific optimization algorithms to seek fine-grained generalization guarantees.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

[2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Adversarial target-invariant representation learning for domain generalization. 2020.

[3] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[4] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[5] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[6] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[7] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[8] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[9] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[10] Aharon Ben-Tal, Elad Hazan, Tomer Koren, and Shie Mannor. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, 2015.

[11] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

[12] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24:2178–2186, 2011.

[13] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[14] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[15] Imre Csiszár. The method of types [information theory]. *IEEE Transactions on Information Theory*, 44(6): 2505–2523, 1998.

[16] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[17] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike 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*, pages 8340–8349, 2021.

[21] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[23] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[24] Vladimir Koltchinskii et al. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/krueger21a.html`.

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[27] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[28] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[29] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/5a142a55461d5fef016acfb927fee0bd-Paper.pdf`.

[30] Yonatan Oren, Shiori Sagawa, Tatsunori Hashimoto, and Percy Liang. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4218–4228, 2019.

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[33] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

[34] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[35] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[36] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

[37] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[38] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

[39] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/628. URL `https://doi.org/10.24963/ijcai.2021/628`. Survey Track.

[40] Sewall Wright. Correlation and causation. 1921.

[41] Chuanlong Xie, Haotian Ye, Fei Chen, Yue Liu, Rui Sun, and Zhenguo Li. Risk variance penalization. *arXiv preprint arXiv:2006.07544*, 2020.

[42] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[43] Guojun Zhang, Han Zhao, Yaoliang Yu, and Pascal Poupart. Quantifying and improving transferability in domain generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[44] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

A Experiments

In this section, we present empirical evidence to verify our theoretical results that under mild and general assumptions, *the proposed weighting correction scheme in (5) has better ID generalization guarantees than the existing variance-regularized domain generalization methods.* By Theorem 4, we reformulate the objective function in (5) into a batch version. Then we consider the following two settings:

- **Balanced batch.** This is the standard operation in DomainBed [18] and is commonly used by the existing variance-based regularization methods. In each iteration, the same number of data points are randomly drawn from each training domain to form a batch.
- **Unbalanced batch.** (Our method) In this setting, we randomly draw data points from each training domain with equal proportions, such that the proportion of each domain in one batch is the same as the proportion of the domains in the entire training data.

We consider three variance-regularized domain generalization methods.

- **Variance.** The V-REx regularization [25] penalizes the domain-level variance of risk without considering the empirical domain distribution. The V-REx estimator is

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \tilde{R}(f) + \lambda \tilde{V}_{out}(f),$$

  where

$$\tilde{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \hat{R}(f|e_i) \quad \text{and} \quad \tilde{V}_{out}(f) = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{R}(f|e_i) - \tilde{R}(f)\Big)^2.$$

- **Standard Deviation.** We also consider the RVP Regularization [41], which slightly changes the penalty term of V-REx into the domain-level standard deviation of risk. Xie et al. [41] provides an understanding of generalization from the perspective of quantile regression and shows that RVP locally approximates DRO. We *remove the scheme of penalty annealing* since it is not involved in group DRO. The RVP estimator is

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \tilde{R}(f) + \lambda\sqrt{\tilde{V}_{out}(f)}.$$

- **Weighting Correction.** Our proposed method introduces the empirical domain distribution as a weighting scheme into the objective function. We also *remove the scheme of penalty annealing*. Our proposed estimator is

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)},$$

  where

$$\hat{R}(f) = \sum_{i=1}^{n} \hat{q}_i \hat{R}(f|e_i) \quad \text{and} \quad \hat{V}_{out}(f) = \sum_{i=1}^{n} \hat{q}_i\big(\hat{R}(f|e_i) - \hat{R}(f)\big)^2.$$

Implementation Details. We consider two datasets: PACS [26] and VLCS [37] We use ResNet50 as neural network architecture [19] and start from a pretrained model on ImageNet [32]. In order to fairly evaluate the different regularization, we follow the DomainBed Benchmark to randomly select 20 groups of hyperparameter combinations and repeated the experiment three times for each hyperparameter group. The model is selected according to the training domain validation accuracy, that is the in-distribution validation accuracy. The hyper-parameters includes batch size, learning rate, weight decay, iterations of penalty annealing, and the regularization parameter λ. The other experimental settings are the same as those in Gulrajani and Lopez-Paz [18].

The results are reported in Table 1. One can find that the improvement of the weighting correction scheme is statistically significant compared to the original VREx method.

Table 1: The ID prediction accuracy on PACS and VLCS.

PACS						
Balance	**Regularization Method**	**A**	**C**	**P**	**S**	**Avg**
×	Weighting Correction	96.9 ± 0.1	**97.0 ± 0.2**	**96.4 ± 0.1**	**97.2 ± 0.2**	**96.9 ± 0.1**
×	Standard Deviation	67.4 ± 0.3	54.9 ± 1.0	51.0 ± 1.4	82.4 ± 0.4	63.9 ± 0.2
×	Variance	20.5 ± 0.5	19.6 ± 0.2	20.5 ± 0.5	36.7 ± 5.9	24.3 ± 1.7
✓	Standard Deviation	82.0 ± 0.6	81.3 ± 0.1	77.2 ± 0.6	86.1 ± 0.5	81.7 ± 0.4
✓	Variance	**97.0 ± 0.1**	96.6 ± 0.1	96.2 ± 0.2	97.0 ± 0.2	96.7 ± 0.1
VLCS						
Balance	**Regularization Method**	**C**	**L**	**S**	**V**	**Avg**
×	Weighting Correction	**81.9 ± 0.2**	**87.6 ± 0.1**	**85.7 ± 0.3**	**83.4 ± 0.0**	**84.7 ± 0.0**
×	Standard Deviation	43.7 ± 0.0	45.0 ± 0.2	48.2 ± 0.7	46.1 ± 0.5	45.8 ± 0.3
×	Variance	53.8 ± 1.5	49.1 ± 0.9	53.3 ± 0.9	52.7 ± 1.7	52.2 ± 0.5
✓	Standard Deviation	48.0 ± 2.3	47.5 ± 0.7	48.9 ± 0.5	53.6 ± 0.4	49.5 ± 0.7
✓	Variance	81.3 ± 0.2	87.3 ± 0.2	85.4 ± 0.5	83.1 ± 0.2	84.3 ± 0.2

B Proof of Theorem 1

Theorem. *Let $n \geq 2$ and $\{\mathbf{z}_{i,j}, 1 \leq i \leq n, 1 \leq j \leq m_i\}$ is an i.i.d sample drawn from P_0.. Suppose the function set $\mathcal{F}$ has cover numbers*

$$N_\epsilon = N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}),$$

and for any $f \in \mathcal{F}$ and $z \in \mathcal{X}\times\mathcal{Y}$, $f(z) \in [0, M]$. Then we have for every $f \in \mathcal{F}$,

$$\begin{aligned} R(f) \quad \leq \quad & \hat{R}(f) + \sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}} + \sum_{i=1}^{n}\sqrt{\frac{2(m_i-1)V_{in}(f|e_i)t}{m(m-1)}} \\ & + \sum_{i=1}^{n}\frac{2\sqrt{m_i}Mt}{\sqrt{m(m_i-1)(m-1)}} + \frac{2(4m-1)Mt}{3m(m-1)} \\ & + \Big(2 + \sqrt{\frac{2t}{m-1}} + \sum_{i=1}^{n}\sqrt{\frac{2(m_i-1)t}{m(m-1)}}\Big)\epsilon, \end{aligned}$$

with probability at least $1-(n+2)N_\epsilon\exp(-t)$.

Proof: By the Bernstein inequality, with probability at least $1-\exp(-t)$,

$$R(f) \quad \leq \quad \hat{R}(f) + \sqrt{\frac{2V(f)t}{m}} + \frac{2Mt}{3m},$$

holds for any given function f. Furthermore, by Theorem 10 in [27],

$$\sqrt{V(f)} \quad \leq \quad \sqrt{\frac{m}{m-1}\hat{V}(f)} + \frac{\sqrt{2mM^2t}}{m-1},$$

holds with probability larger than $1-\exp(-t)$. Then,

$$\begin{aligned} R(f) &\leq \hat{R}(f)+\sqrt{\frac{2\hat{V}(f)t}{m-1}}+\frac{2Mt}{m-1}+\frac{2Mt}{3m}, \\ &= \hat{R}(f)+\sqrt{\frac{2\hat{V}(f)t}{m-1}}+\frac{2(4m-1)Mt}{3m(m-1)}, \end{aligned}$$

holds with probability larger than $1-2\exp(-t)$. According to the decomposition of the total variance, we know that

$$\begin{aligned} \hat{V}(f) &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\big(f(\mathbf{z}_{i,j})-\hat{R}(f|e_i)+\hat{R}(f|e_i)-\hat{R}(f)\big)^2 \\ &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\big(f(\mathbf{z}_{i,j})-\hat{R}(f|e_i)\big)^2+\big(\hat{R}(f|e_i)-\hat{R}(f)\big)^2 \\ &= \sum_{i=1}^{n}\frac{m_i}{m}\hat{V}_{in}(f|e_i)+\sum_{i=1}^{n}\hat{q}_i\big(\hat{R}(f|e_i)-\hat{R}(f)\big)^2 \\ &= \sum_{i=1}^{n}\frac{m_i}{m}\hat{V}_{in}(f|e_i)+\hat{V}_{out}(f). \end{aligned}$$

Hence, with probability larger than $1-2\exp(-t)$,

$$R(f) \leq \hat{R}(f)+\sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}}+\sum_{i=1}^{n}\sqrt{\frac{2m_i\hat{V}_{in}(f|e_i)t}{m(m-1)}}+\frac{2(4m-1)Mt}{3m(m-1)}.$$

By applying Theorem 10 in [27],

$$\sqrt{V_{in}(f|e_i)} \geq \sqrt{\frac{m_i}{m_i-1}\hat{V}_{in}(f|e_i)}-\frac{\sqrt{2m_iM^2t}}{m_i-1},$$

holds with probability smaller than $\exp(-t)$. Hence we have,

$$\begin{aligned} R(f) \leq\ & \hat{R}(f)+\sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}}+\sum_{i=1}^{n}\sqrt{\frac{2(m_i-1)V_{in}(f|e_i)t}{m(m-1)}} \\ &+\sum_{i=1}^{n}\frac{2\sqrt{m_i}Mt}{\sqrt{m(m_i-1)(m-1)}}+\frac{2(4m-1)Mt}{3m(m-1)}, \end{aligned}$$

holds with probability larger than $1-(2+n)\exp(-t)$.

Next, we consider a set of functions $\{f^1,\ldots,f^{N_\epsilon}\}$, which is a minimal ϵ-cover of the function space $\mathcal{F}$ of size

$$N_\epsilon = N(\mathcal{F},\epsilon,\|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}).$$

Then, for any $f\in\mathcal{F}$, there exists f^j, $1\leq j\leq N_\epsilon$ such that $\|f-f^j\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}\leq\epsilon$. Therefore,

$$\begin{aligned} R(f) &\leq R(f^j)+\epsilon \\ &\leq \hat{R}(f^j)+\sqrt{\frac{2\hat{V}_{out}(f^j)t}{m-1}}+\sum_{i=1}^{n}\sqrt{\frac{2(m_i-1)V_{in}(f^j|e_i)t}{m(m-1)}} \\ &\quad+\sum_{i=1}^{n}\frac{2\sqrt{m_i}Mt}{\sqrt{m(m_i-1)(m-1)}}+\frac{2(4m-1)Mt}{3m(m-1)}+\epsilon, \end{aligned}$$

with probability larger than $1-(2+n)\exp(-t)$. Notice that $\hat{R}(f^j)\leq\hat{R}(f)+\epsilon$ and

$$\begin{aligned} \sqrt{\hat{V}_{out}(f^j)} &\leq \sqrt{\hat{V}_{out}(f)}+\sqrt{\hat{V}_{out}(f^j-f)}\leq\sqrt{\hat{V}_{out}(f)}+\epsilon, \\ \sqrt{\hat{V}_{in}(f^j|e_i)} &\leq \sqrt{\hat{V}_{in}(f|e_i)}+\sqrt{\hat{V}_{in}(f^j-f|e_i)}\leq\sqrt{\hat{V}_{in}(f|e_i)}+\epsilon. \end{aligned}$$

Therefore, for every $f \in \mathcal{F}$,

$$\begin{aligned} R(f) &\leq \hat{R}(f) + \sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}} + \sum_{i=1}^{n} \sqrt{\frac{2(m_i-1)V_{in}(f|e_i)t}{m(m-1)}} \\ &\quad + \sum_{i=1}^{n} \frac{2\sqrt{m_i}Mt}{\sqrt{m(m_i-1)(m-1)}} + \frac{2(4m-1)Mt}{3m(m-1)} \\ &\quad + \Big(2 + \sqrt{\frac{2t}{m-1}} + \sum_{i=1}^{n} \sqrt{\frac{2(m_i-1)t}{m(m-1)}}\Big)\epsilon, \end{aligned}$$

holds with probability at least $1-(n+2)N_\epsilon \exp(-t)$.

$\square$

Let

$$t = \log \frac{(n+2)N_\epsilon}{\delta}, \quad \text{and} \quad \lambda = \sqrt{\frac{2t}{m-1}}.$$

Then we have, with probability at least $1-\delta$,

$$\begin{aligned} R(f) &\leq \hat{R}(f) + \lambda\sqrt{\hat{V}_{out}(f)} + \sum_{i=1}^{n} \lambda \sqrt{\frac{(m_i-1)V_{in}(f|e_i)}{m}} \\ &\quad + \sum_{i=1}^{n} \frac{\sqrt{(m-1)m_i}M\lambda^2}{\sqrt{m(m_i-1)}} + \frac{(4m-1)M\lambda^2}{3m} \\ &\quad + \Big(2 + \sqrt{\frac{2t}{m-1}} + \sum_{i=1}^{n} \sqrt{\frac{2(m_i-1)t}{m(m-1)}}\Big)\epsilon, \end{aligned}$$

for every $f \in \mathcal{F}$. Hence Theorem 1 is proved.

$\square$

C More results for Theorem 1

In this section, we use the local Rademacher complexity [5] to present the generalization of the proposed variance-based regularization. We start with the definition of the sub-root function, which can be used to bound the local Rademacher complexity.

Definition 6 ([5], Definition 3.1) *A function $\psi : [0, \infty) \to [0, \infty)$ is sub-root if it is non-negative, non-decreasing and if $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r > 0$.*

For any nontrivial sub-root function ψ, i.e., not the constant function $\psi \equiv 0$, it is continuous and has a unique positive fix point $r^* = \psi(r^*)$. In addition, for all $r > 0$, $r \geq \psi(r)$ if and only if $r^* \leq r$. We consider the local Rademacher complexity:

$$\mathbb{E}\big[\mathcal{R}_n(\{cf(z) : f \in \mathcal{F}, c \in [0,1], \text{ and } \mathbb{E}_{\mathbf{z}\sim P}[c^2 f^2(\mathbf{z})] \leq r\})\big] \tag{16}$$

which is also used by [16]. Here the notation $\mathbb{E}$ in (16) takes the expectations with respect to the Rademacher random variables. We denote a sub-root function $\psi_m(r)$ as an upper bound of the localized Rademacher complexity:

$$\psi_m(r) \quad \geq \quad \mathbb{E}\big[\mathcal{R}_n(\{cf(z) : f \in \mathcal{F}, c \in [0,1], \text{ and } \mathbb{E}_{\mathbf{z}\sim P}[c^2 f^2(\mathbf{z})] \leq r\})\big]. \tag{17}$$

The solution of $\psi_m(r) = r$ is denoted as r_m^*. When the distribution P is replaced by P_{e_i}, the upper bound sub-root function and the corresponding fixed point are written as $\psi_{m,i}$ and $r_{m,i}^*$.

Theorem 7 *Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \times \mathcal{Y} \to [0, M]$ satisfying the localization inequality (17) for some sub-root function $\psi_m(r)$ ($\psi_{m,i}(r)$) with root r_m^* ($r_{m,i}^*$). Let*

$$B_m = \frac{1}{m}(t + \log\lceil \log \frac{m}{t} \rceil) \quad \textit{and} \quad C_m = 2((2e + 84M)B_m + 36 r_m^*),$$

where $\lceil \cdot \rceil$ stands for the ceiling function. Then, for every $f \in \mathcal{F}$,

$$\begin{aligned}
R(f) \quad \leq \quad & (1+\sqrt{2C_m})\Big(\hat{R}(f) + \frac{\sqrt{2C_m}}{1+\sqrt{2C_m}}\sqrt{\hat{V}_{out}(f)}\Big) \\
& + \sqrt{\frac{3}{2}C_m \sum_{i=1}^{n} \frac{m_i}{m} \mathbb{E}[f^2(\mathbf{z})|e_i]} \\
& + \sqrt{144 C_m M^2 \sum_{i=1}^{n} \frac{m_i}{m} r^*_{m,i} + C_m \frac{nMt}{m}(4+\frac{7}{3}M)} \\
& + \sqrt{144 C_m M^2 r^*_m + \frac{C_m M t}{m}(4+\frac{7}{3}M)} + 6r^*_m + 14MB_m,
\end{aligned}$$

with probability at least $1-(2+n)\exp(-t)$.

Proof: Before proving the theorem, we state a useful lemma that provides a version of uniform Bernstein's inequality by measuring the complexity of the localized functions that near the optimum of an empirical risk.

Lemma 8 [[16], Lemma 17 and Lemma 18] *Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \times \mathcal{Y} \to [0, M]$ satisfying the localization inequality (17) for some sub-root function $\psi_m(r)$ with root r^*_m. Let*

$$B_m = \frac{1}{m}(t + \log\lceil \log \frac{m}{t} \rceil) \quad \text{and} \quad \eta > 0.$$

Then with probability at least $1-\exp(-t)$, for every $f \in \mathcal{F}$,

$$R(f) - \hat{R}(f)| \leq (\sqrt{2eB_m} + 6\sqrt{r^*_m + \frac{7}{3}MB_m})\sqrt{\mathbb{E}[f^2]} + 6r^*_m + 14MB_m, \tag{18}$$

$$\mathbb{E}[f^2] \leq \hat{\mathbb{E}}[f^2] + \frac{1}{\eta}\hat{\mathbb{E}}[f^2] + 72M^2(1+\eta)r^*_m + \frac{Mt}{m}(4+\frac{7}{3}M), \tag{19}$$

and

$$\hat{\mathbb{E}}[f^2] \leq \mathbb{E}[f^2] + \frac{\eta}{\eta+1}\mathbb{E}[f^2] + 72M^2(1+\eta)r^*_m + \frac{Mt}{m}(4+\frac{7}{3}M), \tag{20}$$

where $\hat{\mathbb{E}}[f^2(\mathbf{z})] = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i} f^2(\mathbf{z}_{i,j})$.

Proof: Please refer to Appendix D.1 and D.2 of [16] for a complete proof of the lemma.

□

Now we turn to the proof of Theorem 7. We denote $C_m = 2((2e+84M)B_m + 36r^*_m)$. It is easy to see that

$$\begin{aligned}
& (\sqrt{2eB_m} + 6\sqrt{r^*_m + 7MB_m/3})^2 \\
= \quad & 2eB_m + (36r^*_m + 84MB_m) + 2\sqrt{2eB_m}\sqrt{36r^*_m + 84MB_m} \\
\leq \quad & 2(2eB_m + 36r^*_m + 84MB_m) = 2C_m.
\end{aligned}$$

Then the inequality (18) implies that for every $f \in \mathcal{F}$,

$$R(f) \quad \leq \quad \hat{R}(f) + \sqrt{C_m \mathbb{E}[f^2(\mathbf{z})]} + 6r^*_m + 14MB_m,$$

holds with probability at least $1-\exp(-t)$. By (19) with $\eta = 1$, we have for all $f \in \mathcal{F}$,

$$R(f) \leq \hat{R}(f) + \sqrt{2C_m\hat{\mathbb{E}}[f^2(\mathbf{z})]} + \sqrt{144C_m M^2 r^*_m + \frac{C_m Mt}{m}(4+\frac{7}{3}M)} + 6r^*_m + 14MB_m,$$

with probability at least $1-2\exp(-t)$. Notice that

$$\begin{aligned}\hat{\mathbb{E}}[f^2(\mathbf{z})] &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\big(f(\mathbf{z}_{i,j})-\hat{R}(f|e_i)+\hat{R}(f|e_i)-\hat{R}(f)+\hat{R}(f)\big)^2\\ &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\big(f(\mathbf{z}_{i,j})-\hat{R}(f|e_i)\big)^2+\big(\hat{R}(f|e_i)-\hat{R}(f)\big)^2+\hat{R}(f)^2\\ &= \hat{R}(f)^2+\hat{V}_{out}(f)+\sum_{i=1}^{n}\frac{m_i}{m}\hat{V}_{in}(f|e_i).\end{aligned}$$

Then we have

$$\begin{aligned}R(f) &\le (1+\sqrt{2C_m})\hat{R}(f)+\sqrt{2C_m\hat{V}_{out}(f)}+\sqrt{C_m\sum_{i=1}^{n}\frac{m_i}{m}\hat{V}_{in}(f|e_i)}\\ &\quad+\sqrt{144C_mM^2r_m^*+\frac{C_mMt}{m}(4+\frac{7}{3}M)}+6r_m^*+14MB_m.\end{aligned}$$

Next we deal with the upper bound of $\hat{V}_{in}(f|e_i)$. We denote

$$\hat{\mathbb{E}}[f^2|e_i]=\frac{1}{m_i}\sum_{j=1}^{m_i}f^2(\mathbf{z}_{i,j}).$$

According to (20), for every $f\in\mathcal{F}$,

$$\hat{\mathbb{E}}[f^2(\mathbf{z})|e_i] \le \frac{2\eta+1}{\eta+1}\mathbb{E}[f^2(\mathbf{z})|e_i]+72M^2(1+\eta)r_{m,i}^*+\frac{Mt}{m_i}(4+\frac{7}{3}M),$$

holds with probability at least $1-\exp(-t)$. Let $\eta=1$. Then, for every $f\in\mathcal{F}$,

$$\begin{aligned}R(f) &\le (1+\sqrt{2C_m})\hat{R}(f)+\sqrt{2C_m\hat{V}_{out}(f)}+\sqrt{\frac{3}{2}C_m\sum_{i=1}^{n}\frac{m_i}{m}\mathbb{E}[f^2(\mathbf{z})|e_i]}\\ &\quad+\sqrt{144C_mM^2\sum_{i=1}^{n}\frac{m_i}{m}r_{m,i}^*+C_m\frac{nMt}{m}(4+\frac{7}{3}M)}\\ &\quad+\sqrt{144C_mM^2r_m^*+\frac{C_mMt}{m}(4+\frac{7}{3}M)}+6r_m^*+14MB_m,\end{aligned}$$

with probability at least $1-(2+n)\exp(-t)$.

□

D Proof of Corollary 2

Corollary. *Let $n\ge 2$ and $\{\mathbf{z}_{i,j},1\le i\le n,1\le j\le m_i\}$ is an i.i.d sample. Suppose the function set $\mathcal{F}$ has cover numbers*

$$N_\epsilon=N(\mathcal{F},\epsilon,\|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}),$$

and for any $f\in\mathcal{F}$ and $z\in\mathcal{X}\times\mathcal{Y}$, $f(z)\in[0,M]$. Let

$$t=\log\frac{2N_\epsilon+2}{\delta},\quad \textit{and}\quad \lambda=\sqrt{\frac{2t}{m-1}}.$$

Then, with probability at least $1-\delta$,

$$\begin{aligned}R(\hat{f})-R(f^*) &\le 2\lambda\sqrt{\frac{(m-1)V(f^*)}{m}}+\sum_{i=1}^{n}\lambda\sqrt{\frac{m_i\hat{V}_{in}(\hat{f}|e_i)}{m}}\\ &\quad+\Big(2+\lambda+\sum_{i=1}^{n}\lambda\sqrt{\frac{m_i}{m}}\Big)\epsilon+\lambda^2\frac{4(4m-1)M}{3m}.\end{aligned}$$

Proof: According to the proof of Theorem 1,

$$R(f) \quad \leq \quad \hat{R}(f) + \sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}} + \sum_{i=1}^{n}\sqrt{\frac{2m_i\hat{V}_{in}(f|e_i)t}{m(m-1)}} + \frac{2(4m-1)Mt}{3m(m-1)},$$

holds with probability larger than $1-2\exp(-t)$. Next, we consider a set of functions $\{f^1,\ldots,f^{N_\epsilon}\}$, which is a minimal ϵ-cover of the function space $\mathcal{F}$ of size $N_\epsilon = N(\mathcal{F},\epsilon,\|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})})$. Then, for any $f\in\mathcal{F}$, there exists f^j, $1\leq j\leq N_\epsilon$ such that $\|f-f^j\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}\leq\epsilon$. Therefore,

$$\begin{aligned} R(f) \quad &\leq \quad R(f^j)+\epsilon \\ &\leq \quad \hat{R}(f^j) + \sqrt{\frac{2\hat{V}_{out}(f^j)t}{m-1}} + \sum_{i=1}^{n}\sqrt{\frac{2m_i\hat{V}_{in}(f^j|e_i)t}{m(m-1)}} \\ &\quad\quad + \frac{2(4m-1)Mt}{3m(m-1)} + \epsilon, \end{aligned}$$

with probability larger than $1-(2+n)\exp(-t)$. Notice that $\hat{R}(f^j)\leq\hat{R}(f)+\epsilon$ and

$$\begin{aligned} \sqrt{\hat{V}_{out}(f^j)} \quad &\leq \quad \sqrt{\hat{V}_{out}(f)} + \sqrt{\hat{V}_{out}(f^j-f)} \leq \sqrt{\hat{V}_{out}(f)}+\epsilon, \\ \sqrt{\hat{V}_{in}(f^j|e_i)} \quad &\leq \quad \sqrt{\hat{V}_{in}(f|e_i)} + \sqrt{\hat{V}_{in}(f^j-f|e_i)} \leq \sqrt{\hat{V}_{in}(f|e_i)}+\epsilon. \end{aligned}$$

Therefore, for every $f\in\mathcal{F}$,

$$\begin{aligned} R(f) \quad &\leq \quad \hat{R}(f) + \sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}} + \sum_{i=1}^{n}\sqrt{\frac{2m_i\hat{V}_{in}(f|e_i)t}{m(m-1)}} \\ &\quad\quad + \frac{2(4m-1)Mt}{3m(m-1)} + \Big(2+\sqrt{\frac{2t}{m-1}}+\sum_{i=1}^{n}\sqrt{\frac{2m_it}{m(m-1)}}\Big)\epsilon, \end{aligned} \tag{21}$$

holds with probability at least $1-2N_\epsilon\exp(-t)$. Since

$$\hat{f} \quad = \quad \arg\min_{f\in\mathcal{F}} \hat{R}(f) + \sqrt{\frac{2\hat{V}_{out}(f)t}{m-1}},$$

then we have

$$\hat{R}(\hat{f}) + \sqrt{\frac{2\hat{V}_{out}(\hat{f})t}{m-1}} \quad \leq \quad \hat{R}(f^*) + \sqrt{\frac{2\hat{V}_{out}(f^*)t}{m-1}}.$$

By the Bernstein inequality, with probability at least $1-\exp(-t)$,

$$\hat{R}(f^*) \quad \leq \quad R(f^*) + \sqrt{\frac{2V(f^*)t}{m}} + \frac{2Mt}{3m}. \tag{22}$$

By Theorem 10 in [27],

$$\sqrt{\hat{V}_{out}(f^*)} \quad \leq \quad \sqrt{\hat{V}(f^*)} \leq \sqrt{\frac{m-1}{m}V(f^*)} + \sqrt{\frac{2M^2t}{m-1}}, \tag{23}$$

holds with probability larger than $1-\exp(-t)$. Combining (21), (22) and (23),

$$\begin{aligned} R(\hat{f}) \quad &\leq \quad R(f^*) + 2\sqrt{\frac{2V(f^*)t}{m}} + \sum_{i=1}^{n}\sqrt{\frac{2m_i\hat{V}_{in}(\hat{f}|e_i)t}{m(m-1)}} \\ &\quad\quad + \frac{2Mt}{3m} + \frac{2Mt}{m-1} + \frac{2(4m-1)Mt}{3m(m-1)} \\ &\quad\quad + \Big(2+\sqrt{\frac{2t}{m-1}}+\sum_{i=1}^{n}\sqrt{\frac{2m_it}{m(m-1)}}\Big)\epsilon, \end{aligned}$$

holds with probability larger than $1-(2N_\epsilon+2)\exp(-t)$.

□

E Proof of Theorem 3

Theorem *Suppose the training dataset $\mathcal{D}$ and a function $f \in \mathcal{F}$ are given. Let ρ_+ be the largest distance between $\boldsymbol{q} \in \mathcal{Q}_\alpha(\hat{\boldsymbol{q}}, +\infty)$ and*

$$\rho_- = \frac{\min_i (\alpha/\hat{q}_i - 1)^2 \hat{V}_{out}(f)}{2\big(\min_i \hat{R}(f|e_i) - \hat{R}(f)\big)^2}.$$

Then we have

$$\mathcal{L}(f;\rho_-) \leq \max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}})} \sum_{i=1}^n q_i \hat{R}(f|e_i) \leq \mathcal{L}(f;\rho_+).$$

In this section, we decompose the complete proof into the three steps.

Step 1. Given a training dataset and a function $f \in \mathcal{F}$, the following inequality always holds:

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})} \sum_{i=1}^n q_i \hat{R}(f|e_i) \quad \leq \quad \hat{R}(f) + \sqrt{2\rho \hat{V}_{out}(f)}. \tag{24}$$

Proof: Since $\sum_{i=1}^n q_i = 1$ and $\sum_{i=1}^n \hat{q}_i = 1$, we have

$$\begin{aligned}\sum_{i=1}^n q_i \hat{R}(f|e_i) &= \sum_{i=1}^n (q_i - \hat{q}_i)\hat{R}(f|e_i) + \sum_{i=1}^n \hat{q}_i \hat{R}(f|e_i) \\ &= \hat{R}(f) + \sum_{i=1}^n (q_i - \hat{q}_i)\hat{R}(f|e_i).\end{aligned}$$

Since $\sum_{i=1}^n (q_i - \hat{q}_i)C = 0$ holds for any constant C,

$$\sum_{i=1}^n (q_i - \hat{q}_i)\hat{R}(f|e_i) \quad = \quad \sum_{i=1}^n (q_i - \hat{q}_i)\big(\hat{R}(f|e_i) - \hat{R}(f)\big).$$

Thus the max problem in (24) is equivalent to maximize

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})} \sum_{i=1}^n (q_i - \hat{q}_i)\big(\hat{R}(f|e_i) - \hat{R}(f)\big).$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}&\sum_{i=1}^n (q_i - \hat{q}_i)\big(\hat{R}(f|e_i) - \hat{R}(f)\big) \\ = \quad & \sum_{i=1}^n \frac{q_i - \hat{q}_i}{\sqrt{\hat{q}_i}} \sqrt{\hat{q}_i}\big(\hat{R}(f|e_i) - \hat{R}(f)\big) \\ \leq \quad & \sqrt{\sum_{i=1}^n \frac{(q_i - \hat{q}_i)^2}{\hat{q}_i}} \times \sqrt{\sum_{i=1}^n \hat{q}_i \big(\hat{R}(f|e_i) - \hat{R}(f)\big)^2} \\ \leq \quad & \sqrt{2\rho} \times \sqrt{\hat{V}_{out}(f)}\end{aligned}$$

□

Step 2. If the between-domain variance $\hat{V}_{out}(f)$ is non-zero, the equality holds if and only if $\forall e_i \in \mathcal{E}_{tr}$,

$$\alpha \leq \hat{q}_i \Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f)}} \big(\hat{R}(f|e_i) - \hat{R}(f)\big) + 1 \Big).$$

On the other hand, if α is fixed, the equality holds when the radius of $\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$ satisfies

$$\rho \leq \frac{(\alpha/\hat{q}_i - 1)^2 \hat{V}_{out}(f)}{2\big(\min_i \hat{R}(f|e_i) - \hat{R}(f)\big)^2}.$$

Proof: The equality in (24) is attained if and only if the following requirements hold at the same time:

(i) There exists a constant c such that $\forall 1 \leq i \leq n$,

$$\frac{q_i - \hat{q}_i}{\sqrt{\hat{q}_i}} = c\sqrt{\hat{q}_i}\big(\hat{R}(f|e_i) - \hat{R}(f)\big).$$

(ii) The χ^2 divergence between $\boldsymbol{q}$ and $\hat{\boldsymbol{q}}$ achieves ρ, that is

$$\sum_{i=1}^{n} \frac{(q_i - \hat{q}_i)^2}{\hat{q}_i} = 2\rho.$$

It is easy to see

$$c^2 \sum_{i=1}^{n} \hat{q}_i \big(\hat{R}(f|e_i) - \hat{R}(f)\big)^2 = 2\rho \quad \Rightarrow \quad c = \sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}.$$

Then the discrete distribution $\boldsymbol{q}$ satisfies (i) is

$$q_i = \sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}\hat{q}_i\big(\hat{R}(f|e_i) - \hat{R}(f)\big) + \hat{q}_i.$$

Since $\boldsymbol{q}$ belongs to $\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$, the only constraint here is $q_i \geq \alpha$, $\forall e_i$ which holds if and only if $\forall e_i \in \mathcal{E}_{tr}$,

$$\alpha \leq \hat{q}_i\Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}\big(\hat{R}(f|e_i) - \hat{R}(f)\big) + 1\Big).$$

On the other hand, if α is fixed and non-positive, the constraint implies that the radius ρ of $\mathcal{Q}_{\alpha,\rho}(\hat{\boldsymbol{q}})$ should be sufficiently small:

$$\begin{aligned}
& \alpha \leq \hat{q}_i\Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}\big(\hat{R}(f|e_i) - \hat{R}(f)\big) + 1\Big), \\
\Leftrightarrow \quad & \frac{\alpha}{\hat{q}_i} - 1 \leq \sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}\big(\hat{R}(f|e_i) - \hat{R}(f)\big), \\
\Leftarrow \quad & \frac{\alpha/\hat{q}_i - 1}{\min_i \big(\hat{R}(f|e_i) - \hat{R}(f)\big)} \geq \sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}, \\
\Leftarrow \quad & \rho \leq \frac{\min_i(\alpha/\hat{q}_i - 1)^2 \hat{V}_{out}(f)}{2\big(\min_i \hat{R}(f|e_i) - \hat{R}(f)\big)^2}.
\end{aligned}$$

Hence the proof is finished.

□

Step 3. Proof of (13) in Theorem 3.

Proof: First, ρ_+ is the largest distance between $\hat{\boldsymbol{q}}$ and $\boldsymbol{q} \in \mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}})$. Therefore,

$$\mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}}) = \mathcal{Q}_{\alpha,\rho_+}(\hat{\boldsymbol{q}}) \subseteq \mathcal{Q}_{+\infty,\rho_+}(\hat{\boldsymbol{q}}).$$

Hence the second inequality in (13) is trivial. Let $\boldsymbol{q}^*$ be the solution:

$$\boldsymbol{q}^*_- \quad = \quad \operatorname*{arg\,max}_{\boldsymbol{q} \in \mathcal{Q}_{+\infty,\rho_-}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i),$$

According to **Step 2**,

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{+\infty,\rho_-}(\hat{\boldsymbol{q}})}\sum_{i=1}^{n} q_i\hat{R}(f|e_i)=\mathcal{L}(f;\rho_-),$$

and

$$\boldsymbol{q}_-^*\in\mathcal{Q}_{+\infty,\rho_-}(\hat{\boldsymbol{q}})=\mathcal{Q}_{\alpha,\rho_-}(\hat{\boldsymbol{q}})\subseteq\mathcal{Q}_{\alpha,+\infty}(\hat{\boldsymbol{q}}).$$

Hence the first inequality in (13) is proved.

□

F Proof of Theorem 4

In this section, we start with a preliminary result.

Theorem 9 *Suppose that α is a non-positive scalar and*

$$V'_{out}(f)=V(f)-\sum_{i=1}^{n}q_iV_{in}(f|e_i)>0.$$

For each training domain, both $\hat{q}_i$ and q_i are larger than $\delta>0$. Let

$$\rho'=\frac{V'_{out}(f)}{8M^2}\left(\frac{\alpha}{1-(n-1)\delta}-1\right)^2.$$

If $n>2$ and m is sufficiently large such that

$$\frac{m}{4}V'_{out}(f)>V(f),$$

then, given $f\in\mathcal{F}$, the following expansion uniformly holds:

$$\max_{\boldsymbol{q}\in\mathcal{Q}_{\alpha,\rho'}(\hat{\boldsymbol{q}})}\sum_{i=1}^{n} q_i\hat{R}(f|e_i)=\mathcal{L}(f;\rho'),$$

with probability at least

$$1-\exp\Big(-\frac{\big(\frac{m}{4}V'_{out}(f)-V(f)\big)^2}{2M^2(m-1)V(f)}\Big)-\left(\begin{array}{c} m+n-1\\ n-1\end{array}\right)\exp\Big(-\frac{mV'_{out}(f)^2}{M^4}\Big)$$
$$-\sum_{i=1}^{n}\exp\Big(-\frac{(V_{in}(f|e_i)+\frac{m_i}{4}V'_{out}(f))^2}{2M^2m_i(V_{in}(f|e_i)+\frac{1}{4}V'_{out}(f))^2}\Big).$$

Proof: Note that $\hat{R}(f|e)\in[0,M]$ for any $f\in\mathcal{F}$ and $e\in\mathcal{E}_{tr}$. Thus for any training data $\mathbb{D}$,

$$\big(\min_i\hat{R}(f|e_i)-\hat{R}(f)\big)^2\le M^2.$$

In addition, since $\alpha\le 0$ and $q_i,\hat{q}_i\ge\delta$,

$$\min_i\left(\frac{\alpha}{\hat{q}_i}-1\right)^2\le\left(\frac{\alpha}{1-(n-1)\delta}-1\right)^2.$$

Hence, to satisfying

$$\rho'\le\frac{\min_i(\alpha/\hat{q}_i-1)^2\hat{V}_{out}(f)}{2\big(\min_i\hat{R}(f|e_i)-\hat{R}(f)\big)^2},$$

it suffices to show that

$$\hat{V}_{out}(f)\ge\frac{1}{4}V'_{out}(f).$$

Notice that

$$\hat{V}_{out}(f) = \hat{V}(f) - \sum_{i=1}^{n} \hat{q}_i \hat{V}_{in}(f|e_i) = V'_{out}(f) + I_1 + I_2 + I_3,$$

where

$$\begin{aligned} I_1 &= \hat{V}(f) - V(f), \\ I_2 &= -\sum_{i=1}^{n} (\hat{q}_i - q_i)\hat{V}_{in}(f|e_i), \\ I_3 &= -\sum_{i=1}^{n} q_i\big(\hat{V}_{in}(f|e_i) - V_{in}(f|e_i)\big). \end{aligned}$$

By Theorem 10 in [27], for any $\delta > 0$,

$$P\Big(\frac{m}{m-1}\hat{V}(f) < V(f) - \delta\Big) \le \exp\Big(-\frac{(m-1)\delta^2}{2M^2 V(f)}\Big).$$

We take

$$\delta = \frac{m}{4(m-1)} V'_{out}(f) - \frac{1}{m-1} V(f).$$

Therefore,

$$\begin{aligned} P\Big(I_1 < -\frac{1}{4}V'_{out}(f)\Big) &= P\Big(\hat{V}(f) < V(f) - \frac{1}{4}V'_{out}(f)\Big) \\ &\le \exp\Big(-\frac{\big(\frac{m}{4}V'_{out}(f) - V(f)\big)^2}{2M^2(m-1)V(f)}\Big). \end{aligned}$$

For the term I_2,

$$|I_2| \le \sum_{i=1}^{n} \big|\hat{q}_i - q_i\big| \hat{V}_{in}(f|e_i) \le \|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_1 \frac{M^2}{4}.$$

According to the Pinkser's inequality and the method of types [15], for any $\delta > 0$,

$$\|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_1 \le \sqrt{2D_{KL}(\hat{\boldsymbol{q}}\|\boldsymbol{q})} < \delta$$

with probability at least

$$1 - \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(-m\delta^2\Big).$$

We take $\delta = V'_{out}(f)/M^2$. Then we know

$$P\Big(I_2 < -\frac{1}{4}V'_{out}(f)\Big) \le \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(-\frac{mV'_{out}(f)^2}{M^4}\Big).$$

Next, we deal with I_3. By Theorem 10 in [27], for any $\delta > 0$ and any $e_i \in \mathcal{E}_{tr}$,

$$P\Big(\frac{m}{m-1}\hat{V}_{in}(f|e_i) > V_{in}(f|e_i) + \delta\Big) \le \exp\Big(-\frac{(m_i-1)\delta^2}{2M^2(V_{in}(f|e_i)+\delta)}\Big).$$

Let

$$\delta = \frac{1}{m_i - 1}V_{in}(f|e_i) + \frac{m_i}{4(m_i-1)}V'_{out}(f).$$

Then we have

$$\begin{aligned} P\Big(\hat{V}_{in}(f|e_i) > V_{in}(f|e_i) + \frac{1}{4}V'_{out}(f)\Big) &= P\Big(\frac{m}{m-1}\hat{V}_{in}(f|e_i) > V_{in}(f|e_i) + \delta\Big) \\ &\le \exp\Big(-\frac{(V_{in}(f|e_i) + \frac{m_i}{4}V'_{out}(f))^2}{2M^2 m_i(V_{in}(f|e_i) + \frac{1}{4}V'_{out}(f))^2}\Big). \end{aligned}$$

Therefore $I_3 < -\frac{1}{4}V'_{out}(f)$ with probability smaller thant

$$\sum_{i=1}^{n} \exp\Big(- \frac{(V_{in}(f|e_i) + \frac{m_i}{4}V'_{out}(f))^2}{2M^2 m_i (V_{in}(f|e_i) + \frac{1}{4}V'_{out}(f))^2}\Big)$$

Combining the results of I_1, I_2 and I_3, we have

$$\hat{V}_{out}(f) \geq \frac{1}{4}V'_{out}(f),$$

with probability larger than

$$1 - \exp\Big(- \frac{\big(\frac{m}{4}V'_{out}(f) - V(f)\big)^2}{2M^2(m-1)V(f)}\Big) - \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(- \frac{mV'_{out}(f)^2}{M^4}\Big)$$
$$- \sum_{i=1}^{n} \exp\Big(- \frac{(V_{in}(f|e_i) + \frac{m_i}{4}V'_{out}(f))^2}{2M^2 m_i (V_{in}(f|e_i) + \frac{1}{4}V'_{out}(f))^2}\Big).$$

Hence the proof is finished.

$\square$

Next, we extend Theorem 9 to a more general variant with respect to the family of functions $\mathcal{F}$.

Theorem. *Suppose that α is a non-positive scalar and $V'_{out}(f) = V(f) - \sum_i q_i V_{in}(f|e_i) > 0$. For each training domain, both $\hat{q}_i$ and q_i are larger than $\delta > 0$. We write*

$$\rho' = \frac{V'_{out}(f)}{16M^2}\left(\frac{\alpha}{1-(n-1)\delta} - 1\right)^2.$$

Let $\tau > 0$ and $0 < \eta < 1$ be two constants. Define

$$\mathcal{F}_{\tau,\eta} \quad = \quad \{f \in \mathcal{F} : V(f) \geq \tau, \text{ and } \frac{V_{in}(f|e_i)}{V(f)} \leq \eta, \forall e_i \in \mathcal{E}_{tr}\}.$$

For any $f \in \mathcal{F}_{\tau,\eta}$, the following expansion uniformly holds:

$$\max_{\boldsymbol{q} \in \mathcal{Q}_{\alpha,\rho'}(\hat{\boldsymbol{q}})} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) = \mathcal{L}(f;\rho'),$$

with probability at least $1 - N_{\tau,\eta} \times p$, where $N_{\tau,\eta} = N\Big(\mathcal{F}_{\tau,\eta}, \sqrt{\frac{1}{10}(1-\eta)\tau}, \|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}\Big)$ is the covering number of $\mathcal{F}_{\tau,\eta}$ and

$$p \quad = \quad \exp\Big(- \frac{m(1-\eta)^2\tau}{32M^2} + \frac{1}{16}\Big) + \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(- \frac{m(1-\eta)^2\tau^2}{M^4}\Big)$$
$$+ \sum_{i=1}^{n} \exp\Big\{ - \frac{1}{2M^2 m_i}\Big(\frac{m_i(1-\eta) + 4\eta}{1+3\eta}\Big)^2\Big\}.$$

Proof: We consider a set of functions $\{f^1, \ldots, f^N\}$, which is a minimal ϵ-cover of $\mathcal{F}_{\tau,\eta}$ of size

$$N_{\tau,\eta} = N(\mathcal{F}_{\tau,\eta}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}).$$

Define the event

$$\mathfrak{E} = \Big\{\hat{V}_{out}(f^i) \geq \frac{1}{4}V'_{out}(f^i), \text{ for } i = 1, \ldots, N\Big\}.$$

Recall the proof of Theorem 3 and the terms I_1 to I_3. For any $f \in \mathcal{F}_{\tau,\eta}$,

$$\begin{aligned} P\Big(I_1 < -\frac{1}{4}V'_{out}(f)\Big) &\leq \exp\Big(- \frac{\big(\frac{m}{4}V'_{out}(f) - V(f)\big)^2}{2M^2(m-1)V(f)}\Big) \\ &= \exp\Big(- \frac{m^2 V'_{out}(f)^2}{32M^2(m-1)V(f)} + \frac{mV'_{out}(f)}{4M^2(m-1)} - \frac{V(f)}{2M^2(m-1)}\Big) \\ &\leq \exp\Big(- \frac{m^2 V'_{out}(f)^2}{32M^2(m-1)V(f)} + \frac{1}{16}\Big) \\ &\leq \exp\Big(- \frac{m(1-\eta)^2\tau}{32M^2} + \frac{1}{16}\Big). \end{aligned}$$

For the term I_2,

$$P\Big(I_2 < -\frac{1}{4}V'_{out}(f)\Big) \le \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(-\frac{mV'_{out}(f)^2}{M^4}\Big)$$
$$\le \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(-\frac{m(1-\eta)^2\tau^2}{M^4}\Big).$$

For any $e_i \in \mathcal{E}_{tr}$,

$$\begin{aligned}
-\frac{(V_{in}(f|e_i) + \frac{m_i}{4}V'_{out}(f))^2}{2M^2 m_i (V_{in}(f|e_i) + \frac{1}{4}V'_{out}(f))^2} &= -\frac{1}{2M^2 m_i}\Big(\frac{\frac{m_i - 1}{4}V'_{out}(f)}{V_{in}(f|e_i) + \frac{1}{4}V'_{out}(f)} + 1\Big)^2 \\
&= -\frac{1}{2M^2 m_i}\Big(\frac{m_i - 1}{4\frac{V_{in}(f|e_i)}{V'_{out}(f)} + 1} + 1\Big)^2 \\
&\le -\frac{1}{2M^2 m_i}\Big(\frac{m_i - 1}{4\frac{\eta}{1-\eta} + 1} + 1\Big)^2 \\
&= -\frac{1}{2M^2 m_i}\Big(\frac{(m_i - 1)(1-\eta)}{1+3\eta} + 1\Big)^2 \\
&= -\frac{1}{2M^2 m_i}\Big(\frac{m_i(1-\eta) + 4\eta}{1+3\eta}\Big)^2
\end{aligned}$$

Therefore, for the term I_3,

$$P\Big(I_3 < -\frac{1}{4}V'_{out}(f)\Big) \le \sum_{i=1}^{n} \exp\Big\{-\frac{1}{2M^2 m_i}\Big(\frac{m_i(1-\eta) + 4\eta}{1+3\eta}\Big)^2\Big\}.$$

Furthermore, we have for any $f \in \mathcal{F}_{\tau,\eta}$,

$$\hat{V}_{out}(f) \ge \frac{1}{4}V'_{out}(f),$$

with probability larger than $1-p$, where

$$\begin{aligned}
p &= \exp\Big(-\frac{m(1-\eta)^2\tau}{32M^2} + \frac{1}{16}\Big) + \begin{pmatrix} m+n-1 \\ n-1 \end{pmatrix} \exp\Big(-\frac{m(1-\eta)^2\tau^2}{M^4}\Big) \\
&\quad + \sum_{i=1}^{n} \exp\Big\{-\frac{1}{2M^2 m_i}\Big(\frac{m_i(1-\eta) + 4\eta}{1+3\eta}\Big)^2\Big\}.
\end{aligned}$$

Then, combining the results of $\{f^j, j = 1, \dots, N\}$,

$$P(\mathfrak{E}) \ge 1 - N(\mathcal{F}_{\tau,\eta}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}) \times p.$$

Given the event $\mathfrak{E}$, for any $f \in \mathcal{F}$, there exists f^j such that

$$\|f - f^j\|_{L^\infty(\mathcal{X}\times\mathcal{Y})} \le \epsilon.$$

Hence,

$$\begin{aligned}
\hat{V}_{out}(f) &\ge \hat{V}_{out}(f^j) - \epsilon^2 \\
&\ge \frac{1}{4}V'_{out}(f^j) - \epsilon^2 \ge \frac{1}{4}V'_{out}(f) - \frac{5}{4}\epsilon^2.
\end{aligned}$$

We take

$$\epsilon = \sqrt{\frac{1}{10}(1-\eta)\tau}.$$

Then,

$$\hat{V}_{out}(f) \ge \frac{1}{4}V'_{out}(f) - \frac{5}{4}\epsilon^2 \ge \frac{1}{8}V'_{out}(f).$$

Hence the proof is finished.

□

G Proof of Theorem 5

Theorem. *Given the training dataset and a function $f \in \mathcal{F}$, then for any anchor distribution $\boldsymbol{q}_0$, the inequality always holds:*

$$\max_{\boldsymbol{q}\in\mathcal{Q}_\alpha(\boldsymbol{q}_0,\rho)} \sum_{i=1}^{n} q_i \hat{R}(f|e_i) \le \hat{R}(f,\boldsymbol{q}_0) + \sqrt{2\rho \hat{V}_{out}(f,\boldsymbol{q}_0)},$$

where

$$\begin{aligned} \hat{R}(f,\boldsymbol{q}_0) &= \frac{1}{n}\sum_{i=1}^{n} q_{0,i}\hat{R}(f|e_i), \\ \hat{V}_{out}(f,\boldsymbol{q}_0) &= \sum_{i=1}^{n} q_{0,i}\big(\hat{R}(f|e_i) - \hat{R}(f)\big)^2. \end{aligned}$$

If the between-domain variance $\hat{V}_{out}(f,\boldsymbol{q}_0)$ is non-zero, the equality holds if and only if $\forall e_i \in \mathcal{E}_{tr}$,

$$\alpha \le q_{0,i}\Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f,\boldsymbol{q}_0)}}\big(\hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big) + 1\Big).$$

On the other hand, if α is fixed, the equality holds when the radius of $\mathcal{Q}_\alpha(\boldsymbol{q}_0,\rho)$ satisfies,

$$\rho \le \frac{\min_i(\alpha/q_{0,i}-1)^2\hat{V}_{out}(f,\boldsymbol{q}_0)}{2\big(\min_i \hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big)^2}.$$

The proof here is similar to that of Theorem 3.

Proof: Since $\sum_{i=1}^{n} q_i = 1$ and $\sum_{i=1}^{n} q_{0,i} = 1$, we have

$$\begin{aligned} \sum_{i=1}^{n} q_i\hat{R}(f|e_i) &= \sum_{i=1}^{n} q_{0,i}\hat{R}(f|e_i) + \sum_{i=1}^{n}(q_i - q_{0,i})\hat{R}(f|e_i) \\ &= \hat{R}(f,\boldsymbol{q}_0) + \sum_{i=1}^{n}(q_i - q_{0,i})\hat{R}(f|e_i). \end{aligned}$$

Since $\sum_{i=1}^{n}(q_i - \hat{q}_i) = 0$, then we have

$$\sum_{i=1}^{n}(q_i - q_{0,i})\hat{R}(f|e_i) = \sum_{i=1}^{n}(q_i - q_{0,i})\big(\hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big).$$

Thus the max problem with respect to $\boldsymbol{q}$ is equivalent to maximize

$$\max_{\boldsymbol{q}\in\mathcal{Q}_\alpha(\boldsymbol{q}_0,\rho)} \sum_{i=1}^{n}(q_i - q_{0,i})\big(\hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big).$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} & \sum_{i=1}^{n}(q_i - q_{0,i})\big(\hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big) \\ = & \sum_{i=1}^{n}\frac{q_i - q_{0,i}}{\sqrt{q_{0,i}}}\sqrt{q_{0,i}}\big(\hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big) \\ \le & \sqrt{\sum_{i=1}^{n}\frac{(q_i - q_{0,i})^2}{q_{0,i}}} \times \sqrt{\sum_{i=1}^{n} q_{0,i}\big(\hat{R}(f|e_i) - \hat{R}(f,\boldsymbol{q}_0)\big)^2} \\ \le & \sqrt{2\rho} \times \sqrt{\hat{V}_{out}(f,\boldsymbol{q}_0)} \end{aligned}$$

The equality is attained if and only if the following requirements hold at the same time:

(i) There exists a constant c such that $\forall 1 \leq i \leq n$,

$$\frac{q_i - q_{0,i}}{\sqrt{q_{0,i}}} = c\sqrt{q_{0,i}}\big(\hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big).$$

(ii) The χ^2 divergence between $\boldsymbol{q}$ and $\boldsymbol{q}_0$ achieves ρ:

$$\sum_{i=1}^{n} \frac{(q_i - q_{0,i})^2}{q_{0,i}} = 2\rho.$$

It is easy to see

$$c^2 \sum_{i=1}^{n} q_{0,i}\big(\hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big)^2 = 2\rho \quad \Rightarrow \quad c = \sqrt{\frac{2\rho}{\hat{V}_{out}(f, \boldsymbol{q}_0)}}.$$

Then the discrete distribution $\boldsymbol{q}$ satisfies (i) is

$$q_i = \sqrt{\frac{2\rho}{\hat{V}_{out}(f, \boldsymbol{q}_0)}} q_{0,i}\big(\hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big) + q_{0,i}.$$

Since $\boldsymbol{q}$ belongs to $\mathcal{Q}_\alpha(\boldsymbol{q}_0, \rho)$, the constraint here is $q_i \geq \alpha$, $\forall e_i$ which holds if and only if

$$\alpha \leq q_{0,i}\Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f)}}\big(\hat{R}(f|e_i) - \hat{R}(f)\big) + 1\Big),\ \forall e_i \in \mathcal{E}_{tr}.$$

On the other hand, if α is fixed and non-positive, the constraint implies that the radius ρ of $\mathcal{Q}_\alpha(\boldsymbol{q}_0, \rho)$ should be sufficiently small:

$$\begin{aligned}
& \alpha \leq q_{0,i}\Big(\sqrt{\frac{2\rho}{\hat{V}_{out}(f, \boldsymbol{q}_0)}}\big(\hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big) + 1\Big) \\
\Leftrightarrow \quad & \frac{\alpha}{q_{0,i}} - 1 \leq \sqrt{\frac{2\rho}{\hat{V}_{out}(f, \boldsymbol{q}_0)}}\big(\hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big), \\
\Leftarrow \quad & \frac{\alpha/q_{0,i} - 1}{\min_i \big(\hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big)} \geq \sqrt{\frac{2\rho}{\hat{V}_{out}(f, \boldsymbol{q}_0)}}, \\
\Leftarrow \quad & \rho \leq \frac{\min_i (\alpha/q_{0,i} - 1)^2 \hat{V}_{out}(f, \boldsymbol{q}_0)}{2\big(\min_i \hat{R}(f|e_i) - \hat{R}(f, \boldsymbol{q}_0)\big)^2}.
\end{aligned}$$

Hence the proof is finished.

□