LEARNING THE ROBUSTNESS MECHANISM WITH BILEVEL OPTIMIZATION

Yiyang Shen
Department of Informatics
University of Iowa
yiyang-shen@uiowa.edu

Qihang Lin
Tippie College of Business
University of Iowa
qihang-lin@uiowa.edu

Weiran Wang
Department of Computer Science
University of Iowa
weiran-wang@uiowa.edu

ABSTRACT

We propose a distributionally robust learning framework where parameters defining the robustness mechanism are learned from held-out data instead of extensively tuned. Using bilevel optimization with both upper and lower level minimax problems, we create two instances of our framework to tackle setups with and without group labels in the training set. Theoretically, we provide sample complexity analysis for our robustness mechanism learning paradigm, showing that it achieves generalization guarantees comparable to exhaustive grid search while being more computationally efficient. Empirically, we evaluate our framework under a challenging setup when both intra-group and inter-group test distribution shifts occur at the same time, thereby demonstrating the efficacy and scalability of our method.

1 INTRODUCTION

Many machine learning methods require optimizing model parameters to minimize the empirical risk or average sample loss. The empirical risk minimization (ERM) paradigm assumes that unseen data are sampled from the same distribution as seen training data are. Since ERM weighs all samples equally, it is particularly vulnerable to *subpopulation shift*, where the training samples consist of several groups divided by spurious attributes whose proportions are different from those of the test data (Sagawa et al., 2020; Shen et al., 2021; Cai et al., 2021; Yang et al., 2023; Yu et al., 2024). In real-world applications such as healthcare (Zech et al., 2018; Badgeley et al., 2019), fairness (Buolamwini & Gebru, 2018; Mehta et al., 2024; Lei et al., 2024), robotics (Ryu & Mehr, 2024), and autonomous driving (Zhang et al., 2017; Azizi et al., 2025), the classifier parameters may inadvertently depend on spurious attributes, causing failures when the testing environment is different.

Distributionally robust optimization (DRO, Duchi et al. (2021)) along with many out-of-distribution generalization algorithms (Arjovsky et al., 2019; Sohoni et al., 2020; Krueger et al., 2021) are developed to address this issue. A simplified setup for DRO is Group DRO (GDRO, Sagawa et al. (2020)), which assumes grouping among samples and robustifies the model by minimizing the training loss of the group with the highest loss or worst training accuracy. Take the widely used CelebrityAttributes (CelebA) benchmark as an example, an ERM-trained model tend to correlate golden hair color (class labels) with female (attribute) celebrities, leading to severe performance degradation on minority groups, male celebrities with blond hair. Thus, GDRO explicitly minimizes the loss of the group incurring high loss. This has inspired a growing line of research which aims at developing out-of-distribution (OOD) generalization algorithms for the GDRO setup(Ahmed et al., 2021; Creager et al., 2021; Piratla et al., 2022; Izmailov et al., 2022; Nam et al., 2022; Seo et al., 2022; Asgari et al., 2022; Zhang et al., 2022a; Ghosal & Li, 2023; Paranjape et al., 2023; Wu et al., 2023; Deng et al., 2023; Han & Zou, 2024; Jain et al., 2024; LaBonte et al., 2023; Pezeshki et al., 2024; Jeong et al., 2025; Jo et al., 2026). Most methods fall under two general categories of setup. One stream focuses

on the more ideal setup where all attribute labels are known, so the algorithms know the ground truth group membership of samples during training. The other focuses on a weakly group-supervised or group-unsupervised setup which is more realistic since attribute labels that define groups are often unavailable. Typically, a worst-group identification model is trained (e.g., worst loss samples in ERM, attribute prediction) during the first stage with or without a small amount of group-labeled data; in the second stage, robust training is done using those pseudo-labeled groups, e.g., GDRO.

GDRO is suited to address shifts in group distributions, i.e., inter-group subpopulation shift, which leads to failures concentrated in minority groups. However, treating groups as fixed distributions overlooks another type of uncertainty: the conditional distribution *within* a group, i.e., **intra-group subpopulation shift** (Ben-Tal & Nemirovski, 2002; Devroye et al., 2013; Duchi & Namkoong, 2021). For example, a minority group at training time may contain a small variation of environment, while the test distribution gives underrepresented variants of the same group. DRO with hierarchical ambiguity set (HDRO, Jo et al. (2026)) addresses this issue by introducing an adversarial perturbation within each group, which controls how much intra-group variation the model protects against. However, the appropriate amount of robustness signals per group is generally unknown, may not be uniform across groups, and requires extensive tuning. Similarly, when **group membership is unknown at training time**, worst group membership predictor typically is crucial to robustness and requires tuning as well.

As such, while effective at tackling various adversarial conditions, advanced robustness mechanisms require users to manually specify the type of uncertainty, and *the level of uncertainty* the trained model should be robust against under different evaluation distributions. This introduced many more potentially sensitive models and hyper-parameters, making exhaustive tuning expensive and naturally raises the following question:

> *Can we treat the robustness mechanism itself as an active, integral component to be learned during active training?*

Our affirmative answer contributes to DRO research in three aspects:

1. We propose a bilevel adaptive tuning framework that *directly learns a given robustness mechanism* based on the validation set, with a minimax problem for tuning robustness parameters on the upper level and a minimax problem for robust training on the lower level. With two instantiations, we show such bilevel problem can be tractably solved by a first-order method proposed by Shen et al. (2026).

2. While most existing theory is concerned with the optimization complexity of solving such empirical problem to certain optimality conditions, such as saddle point and ϵ-KKT point (Lu & Mei, 2024), we derive generalization guarantees for our bilevel framework, providing the sample complexity for learning robust model which is new to the best of our knowledge. The guarantee shows advantage of continuous optimization of the hyperparameters avoids discretization error compared to grid search.

3. We validate and enhance the evaluation setting proposed by Jo et al. (2026) where both inter-group and intra-group minority group test distribution shifts are manifest, showing significant improvement on worst-group prediction using our methods both with and without group labels at training time.

2 A Bilevel Adaptive Framework for GDRO

2.1 Group Distributionally Robust Optimization

Suppose there are G groups in the data distribution. Let (x, y, a) be a data point, where x is the feature vector, y is the target variable, and $a \in \mathcal{A}$ is an attribute label that defines the group the data point belongs to together with known y. We consider a predictive task where the goal is to predict y based on x through a model $f_{W,\theta}(x) := Wh_\theta(x)$, where $h_\theta(x)$ is a mapping parameterized by θ that produces a representation of x while W is a matrix that defines a linear model that produces the prediction $Wh_\theta(x)$ for y.

Let $D_{\text{tr}}^{(g)} = \{(x_{g,i}^{\text{tr}}, y_{g,i}^{\text{tr}})\}_{i=1}^{n_g^{\text{tr}}}$ be a set of n_g^{tr} training samples from group g for $g \in \{1, \ldots, G\}$, and $\ell(f_{W,\theta}(x), y)$ be a loss function that measures the discrepancy between the prediction $f_{W,\theta}(x)$ and the target y. The average training loss on group g is $L_g^{\text{tr}}(W;\theta) := \frac{1}{n_g^{\text{tr}}} \sum_{i=1}^{n_g^{\text{tr}}} \ell\left(f_{W,\theta}(x_{g,i}^{\text{tr}}), y_{g,i}^{\text{tr}}\right)$. The GDRO model for learning W and θ can be formulated as

$$(W^\star, \theta^\star) \in \arg\min_{W,\theta} \max_{q \in \Delta_G} \left\{ \sum_{g=1}^{G} q_g L_g^{\text{tr}}(W;\theta) - \frac{\eta}{2} \left\| q - \frac{1}{G}\mathbf{1} \right\|_2^2 + \frac{\lambda}{2}\|W\|_F^2 + \frac{\lambda}{2}\|\theta\|_2^2 \right\}, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\Delta_G := \{q = (q_1, \ldots, q_G) | q_g \geq 0, 1 \leq g \leq G, \mathbf{1}^\top q = 1\}$, $\eta \geq 0$ is the *robustness parameter* (Huang et al., 2021; Zhang et al., 2022b), and $\lambda \geq 0$ is the regularization parameter. Here, the goal is to achieve a robust performance of $f_{W,\theta}$ by minimizing a weighted loss over groups with more weight put on the groups of larger average losses. Note that η controls how far the group weight q may deviate from uniform weighting, i.e, $\mathbf{1}/G$. Naturally, it is critical to select η in (1) to achieve the best out-of-sample performance. Typically, a validation set is used, denoted by $D_{\text{val}}^{(g)} = \{(x_{g,i}^{\text{val}}, y_{g,i}^{\text{val}})\}_{i=1}^{n_g^{\text{val}}}$ for group $g \in \{1, \ldots, G\}$, to evaluate the robustness of the performance of $f_{W,\theta}$, for example, in its largest validation loss among the groups, i.e., [1]

$$\max_{p \in \Delta_G} \sum_{g=1}^{G} p_g L_g^{\text{val}}(W^\star;\theta^\star), \quad \text{where} \quad L_g^{\text{val}}(W;\theta) := \frac{1}{n_g^{\text{val}}} \sum_{i=1}^{n_g^{\text{val}}} \ell\left(f_{W,\theta}(x_{g,i}^{\text{val}}), y_{g,i}^{\text{val}}\right). \tag{2}$$

Then a value of η is selected from a grid to minimize (2). While widely used, this approach requires training a model for each candidate of η and does not directly extend to the setting where the attribute label a is missing from most of the training data.

2.2 Warm-up: Bilevel Group DRO (Bi-GDRO)

An adaptive bilevel group DRO method can be used to address the challenges caused by tuning. Instead of training the full model as in (1) and selecting η based on (2), we integrate the training and the parameter tuning into a bilevel optimization model as follows

$$\min_{W,\theta,\eta \geq 0} \max_{p \in \Delta_G} \sum_{g=1}^{G} p_g L_g^{\text{val}}(W;\theta) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{3}$$

$$\text{s.t. } W \in \arg\min_{W'} \max_{q \in \Delta_G} \left\{ \sum_{g=1}^{G} q_g L_g^{\text{tr}}(W';\theta) - \frac{\eta}{2} \left\| q - \frac{1}{G}\mathbf{1} \right\|_2^2 + \frac{\lambda}{2}\|W'\|_F^2 \right\}. \tag{4}$$

Different from (1) and (2), η becomes a continuous upper-level decision variable in (3) without being limited in a finite grid. Parameter θ is another upper-level decision variable while the linear model W is optimized in the lower-level problem (4). This way, W is learned from the training data for any given θ and η, while θ and η are learned by optimizing the performance of $f_{W,\theta}$ on the validation set. Note that, with this design, the lower-level problem (4) becomes convex in W' and concave in q, which is required by most algorithms for bilevel optimization[2], although the upper-level objective function in (3) can be nonconvex jointly in W and θ.

This bilevel minimax hyperparameter optimization model has been studied by Shen et al. (2026). However, as we show below, the framework like (3) can be extended to learn additional elements of a robustness mechanism that is more general than (1), such as the radius of an ambiguity set and the latent group structure itself when group attribute labels are unavailable during training.

2.3 Bilevel-Hierarchical DRO (Bi-HDRO)

GDRO can be extended into hierarchical DRO (HDRO) whose ambiguity set has a hierarchical structure (Jo et al., 2026). Let P_g be the empirical distribution on $D_{\text{tr}}^{(g)}$. The HDRO in our notation

[1] Other robustness metrics can be used here as well, such as a truncated simplex that replaces Δ_G in (2).

[2] As noted by Kang et al. (2020); Kirichenko et al. (2023), tuning of W alone is sufficient for robustness. Furthermore, while closed form of q is available, it is costly to compute as we show in Appendix B.

can be formulated as

$$\min_{W,\theta} \max_{Q \in \mathcal{Q}} \mathbb{E}_{(X,Y)\sim Q}[\ell\left(f_{W,\theta}(X), Y\right)], \tag{5}$$

where (X, Y) denotes a random data point, $\mathbb{E}_{(X,Y)\sim Q}$ denotes the expectation taken over (X, Y) when (X, Y) follows distribution Q, and

$$\mathcal{Q} = \left\{ \sum_{g=1}^{G} q_g Q_g : \ q \in \Delta_G, \ W_\infty(Q_g, P_g) \leq \epsilon_g, \ \text{ for } g = 1, \ldots, G \right\}$$

is the hierachical ambiguity set, where Q_g is a distribution of (X, Y), $W_\infty(Q_g, P_g)$ is the ∞-Wasserstein distance between Q_g and P_g, and ϵ_g is a radius. As in (1), weights q model the changes in group proportions, i.e., inter-group shift, while Q_g is used to further to accommodate shifts within each group, i.e., intra-group shift. Note that (5) is reduced to (1) when $\epsilon_g = 0$ since $Q_g = P_g$.

Direct optimization over the distributions of Q_g is generally intractable. Therefore, Jo et al. (2026) (Theorem 4.1) proposed solving the following upper approximation of (5)

$$(W^\star, \theta^\star) \in \arg\min_{W,\theta} \max_{q \in \Delta_G} \sum_{g=1}^{G} q_g L_g^{\mathrm{tr},\epsilon_g}(W; \theta) \tag{6}$$

$$\text{where} \quad L_g^{\mathrm{tr},\epsilon_g}(W; \theta) := \frac{1}{n_g^{\mathrm{tr}}} \sum_{i=1}^{n_g^{\mathrm{tr}}} \left[\max_{z : \|z - h_\theta(x_{g,i}^{\mathrm{tr}})\| \leq \epsilon_g} \ell\big(Wz, y_{g,i}^{\mathrm{tr}}\big) \right]. \tag{7}$$

Note that $L_g^{\mathrm{tr}}(W; \theta) \leq L_g^{\mathrm{tr},\epsilon_g}(W; \theta)$. In (6), we minimize the largest loss over groups when the latent representation of each sample can be adversarially perturbed within a ball of radius ϵ_g. The perturbation makes the model more robust to test-distribution intra-group shift.

Problems with HDRO **(1)** Solving the inner maximization over z for each data point to evaluate $L_g^{\mathrm{tr},\epsilon_g}(W; \theta)$ is computationally challenging when n_g^{tr} is large. Jo et al. (2026) proposed a heuristic method that performs one step of gradient ascent over z from $h_\theta(x_{g,i}^{\mathrm{tr}})$, which only solves the inner maximization suboptimally and thus approximates $L_g^{\mathrm{tr},\epsilon_g}(W; \theta)$ and its gradient poorly. **(2)** ϵ_g requires additional tuning for each g. Although Jo et al. (2026) (Appendix D.3) proposed tuning a scalar ϵ with $\epsilon_g = \epsilon / \sqrt{n_g^{\mathrm{tr}}}$, this remains heuristic and still requires grid search over ϵ based on a performance metric on the validation set such as (6).

Our Solution **(1)** For each g, we propose a modification of $L_g^{\mathrm{tr},\epsilon_g}(W; \theta)$, denoted by $\tilde{L}_g^{\mathrm{tr},\epsilon_g}(W; \theta; u)$ with an additional variable u. The specific form of $\tilde{L}_g^{\mathrm{tr},\epsilon_g}$ depends on the prediction task and the loss function $\ell(\cdot, \cdot)$. We show that, for a binary classification problem where ℓ is either the hinge loss or the logistic loss, $\tilde{L}_g^{\mathrm{tr},\epsilon_g}(W; \theta; u)$ is jointly convex in W and u, and (6) equals

$$\min_{W,\theta,u} \max_{q \in \Delta_G} \sum_{g=1}^{G} q_g \tilde{L}_g^{\mathrm{tr},\epsilon_g}(W; \theta; u) \ \text{ s.t. } \ r(W, u) \leq 0, \tag{8}$$

where $r(W, u)$ is a jointly convex function of W and u. For a multiclass classification problem where ℓ is the cross-entropy loss, we show that the corresponding $\tilde{L}_g^{\mathrm{tr},\epsilon_g}(W; \theta; u_g)$ and $r(W, u)$ are still jointly convex in W and u but (8) is only an upper bound of (6). Therefore, we propose solving (8) as the gradient of $\tilde{L}_g^{\mathrm{tr},\epsilon_g}$ can be evaluated exactly without solving the inner maximization problems. Furthermore, in all the aforementioned cases, we can show that projection to the constraint set defined by the inequality $r(W, u) \leq 0$ has a closed form, meaning that (8) is not computationally more difficult than (6). We present the details in Appendix C. **(2)** Similar to (3), we can tune ϵ_g in a bilevel optimization model based on the performance on the validation set after adding $\{\epsilon_g\}_{g=1}^{G}$ as

upper-level decision variables just like η:

$$\min_{W,\theta,u,\eta\geq 0,\{\epsilon_g\}_{g=1}^G} \max_{p\in\Delta_G} \sum_{g=1}^G p_g L_g^{\mathrm{val}}(W;\theta) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{9}$$

$$\text{s.t. } W \in \arg\min_{W',u'} \max_{q\in\Delta_G} \left\{ \sum_{g=1}^G q_g \tilde{L}_g^{\mathrm{tr},\epsilon_g}(W';\theta;u') - \frac{\eta}{2}\left\| q - \frac{1}{G}\mathbf{1}\right\|_2^2 + \frac{\lambda}{2}\|W'\|_F^2 \right\},$$

$$\text{s.t. } r(W',u') \leq 0,$$

where we've replaced the loss $L^{\mathrm{tr},\epsilon_g}$ in (4) to be $\tilde{L}^{\mathrm{tr},\epsilon_g}$.

2.4 Bilevel-Probabilistic Group DRO (Bi-PG-DRO)

In real-world scenarios, it is possible that only a very small portion of data has attribute label a so we are not able to formulate $L_g^{\mathrm{tr}}(W;\theta)$ using all data points due to the lack of group information. To address this issue, Ghosal & Li (2023) proposed PG-DRO, a robustness mechanism that uses a small amount of attribute-labeled training data to train a soft group predictor to generate pseudo-membership labels before using a robust model such as GDRO for training (Sagawa et al., 2020).

Formally, let $D^{\mathrm{ul}} = \{(x_i^{\mathrm{ul}}, y_i^{\mathrm{ul}})\}_{i=1}^{n^{\mathrm{ul}}}$ be a separate subset without group labels. We assume $n_g^{\mathrm{tr}} \ll n^{\mathrm{ul}}$ for any $g = (a, y)$, where attribute label a and class label y jointly determines the group. PG-DRO introduces another classification model $\tilde{f}_\phi(x)$ parameterized by ϕ and train $\tilde{f}_\phi(x)$ on $D_{\mathrm{tr}}^{(g)}$ to predict the attribute label $a \in \mathcal{A}$ based on x. For each data sample $(x_i^{\mathrm{ul}}, y_i^{\mathrm{ul}})$ from D^{ul}, we assume $\tilde{f}_\phi(x_i^{\mathrm{ul}}) = (\gamma_{i1}(\phi), \gamma_{i2}(\phi), \ldots, \gamma_{iG}(\phi))^\top$ where $\gamma_{ig}(\phi)$ is the predicted probability of $(x_i^{\mathrm{ul}}, y_i^{\mathrm{ul}})$ being in group g for each g, since class label is known. Using this conditional probability as soft group labels, we can assign a fraction of $(x_i^{\mathrm{ul}}, y_i^{\mathrm{ul}})$ to each group, yielding the probabilistic loss

$$L_g^{\mathrm{PG}}(W;\theta,\phi) = \frac{\sum_{i=1}^{n^{\mathrm{ul}}} \gamma_{ig}(\phi)\ell\left(f_{W,\theta}(x_i^{\mathrm{ul}}), y_i^{\mathrm{ul}}\right)}{\sum_{i=1}^{n^{\mathrm{ul}}} \gamma_{ig}(\phi) + \epsilon},$$

where ϵ is a smoothing parameter to avoid a zero denominator. Then PG-DRO solves

$$(W^\star, \theta^\star) \in \arg\min_{W,\theta} \max_{q\in\Delta_G} \sum_{g=1}^G q_g L_g^{\mathrm{PG}}(W;\theta,\phi). \tag{10}$$

However, PG-DRO requires additional training for $\tilde{f}_\phi(x)$. For a more efficient training approach, we propose to integrate the training of $f_{W,\theta}(x)$ and $\tilde{f}_\phi(x)$ as well as the tuning of the robustness parameter into a bilevel optimization model below

$$\min_{W,\theta,\phi,\eta\geq 0} \max_{p\in\Delta_G} \sum_{g=1}^G p_g L_g^{\mathrm{val}}(W;\theta) + \frac{\lambda}{2}\|\theta\|_2^2 + \beta\,\mathrm{KL}(\pi_{\mathrm{tr}}||\pi_\phi) \tag{11}$$

$$\text{s.t. } W \in \arg\min_{W'} \max_{q\in\Delta_G} \left\{ \sum_{g=1}^G q_g L_g^{\mathrm{PG}}(W';\theta,\phi) - \frac{\eta}{2}\left\| q - \frac{1}{G}\mathbf{1}\right\|_2^2 + \frac{\lambda}{2}\|W'\|_F^2 \right\}. \tag{12}$$

Here, $\pi_\phi = (\sum_{i=1}^{n^{\mathrm{ul}}} \gamma_{ig}(\phi)/n^{\mathrm{ul}})_{g=1}^G$ is the proportion of data points predicted to be in group g by model $\tilde{f}_\phi$, $\pi_{\mathrm{tr}} = (n_g^{\mathrm{tr}}/(\sum_{g'=1}^G n_{g'}^{\mathrm{tr}}))_{g=1}^G$ is an estimation of the prior distribution of the group labels, and $\mathrm{KL}(\pi_{\mathrm{tr}}||\pi_\phi)$ is the Kullback-Leibler (KL) divergence between π_ϕ and π_{tr}. Different from PG-DRO, $\tilde{f}_\phi$ in (11) is not trained separately on a binary cross-entropy (BCE) loss to predict a. Instead, ϕ optimized in the upper-level in (11) such that the produced γ_{ig} helps ensure a good performance of the resulting $f_{W,\theta}(x)$ on the validation set, and a high prediction accuracy of $\tilde{f}_\phi$ is not necessary. One may replace $\mathrm{KL}(\pi_{\mathrm{tr}}||\pi_\phi)$ in (11) to the training loss of $\tilde{f}_\phi$ in predicting the group label g. Empirical findings (Appendix E.3) show that this has little impact on the numerical performance but using $\mathrm{KL}(\pi_{\mathrm{tr}}||\pi_\phi)$ as the regularizer makes the optimization more lightweight.

2.5 BILEVEL MINIMAX ALGORITHM

Despite their different formulations, Problems (3), (9), and (11) can be solved using the first-order method proposed by Shen et al. (2026). We provide the algorithm's pseudocode in Appendix A and summarize it here. Firstly, these problems are instances of the following bilevel minimax problem

$$\min_{\alpha,W,q}\left\{\max_{p} F(\alpha,p,W,q)\;\middle|\;(W,q)\in\arg\min_{\widetilde{W}}\max_{\widetilde{q}}\widetilde{F}(\alpha,\widetilde{w},\widetilde{q})\right\}, \tag{13}$$

where α denotes the collection of all primal upper-level decision variables, including θ, η, ϵ_g and ϕ in the three bilevel models above, p is the upper-level group weight, q is the lower-level group weight, and W is the parameter of the linear classifier within model $f_{W,\theta}$. Here, F and $\widetilde{F}$ are different objectives. The primal and dual value functions of the lower-level minimax problem of (13) are denoted by

$$V_{\mathrm{P}}(\alpha,W):=\max_{\widetilde{q}}\widetilde{F}(\alpha,W,\widetilde{q})\quad\text{and}\quad V_{\mathrm{D}}(\alpha,q):=\min_{\widetilde{W}}\widetilde{F}(\alpha,\widetilde{W},q),$$

respectively. Therefore, Problem (13) can be equivalently written as the single-level constrained problem with a primal-dual gap constraint

$$\min_{\alpha,W,q}\left\{\max_{p} F(\alpha,p,W,q)\;\middle|\;V_{\mathrm{P}}(\alpha,W)-V_{\mathrm{D}}(\alpha,q)\le 0\right\}.$$

Lastly, introducing a penalty parameter $\rho>0$ yields the penalized minimax problem

$$\min_{\alpha,W,q}\max_{p}\left\{F(\alpha,p,W,q)+\rho\left[V_{\mathrm{P}}(\alpha,W)-V_{\mathrm{D}}(\alpha,q)\right]\right\}=\min_{\alpha,W,q}\max_{p,\widetilde{W},\widetilde{q}}P_\rho\left(\alpha,p,W,q,\widetilde{W},\widetilde{q}\right), \tag{14}$$

where P_ρ denotes the corresponding penalized objective. To compute a ϵ-primal-dual stationary point of P_ρ which is nonconvex-concave, we apply the inexact proximal-point method as in Shen et al. (2026). The original nonconvex-concave problem is thereby reduced to a sequence of approximately solved strongly-convex-strongly-concave minimax subproblems, each of which solved using the stochastic accelerated primal-dual (SAPD) algorithm by treating minimizing and maximizing variables as separate blocks (Zhang et al., 2022b).

2.6 GENERALIZATION THEORY

We establish a generalization theory for our continuous bilevel hyperparameter tuning framework. For clarity of exposition, we present the results in this section without the encoder θ; the full extensions and analysis details are provided in Appendix F. Formally, we analyze the problem of finding the optimal multidimensional hyperparameters $\hat{\boldsymbol{\psi}}\in\Psi$ (e.g., $\boldsymbol{\psi}=(\lambda,\eta)$ for Group DRO, or $\boldsymbol{\psi}=(\lambda,\eta,\boldsymbol{\epsilon})$ for Bi-HDRO) that minimize the worst-group validation loss:

$$\hat{\boldsymbol{\psi}}=\arg\min_{\boldsymbol{\psi}\in\Psi}\max_{p\in\Delta_G}\sum_{g=1}^{G}p_g L_g^{\mathrm{val}}(\widehat{W}_{\boldsymbol{\psi}}), \tag{15}$$

where the robust model parameters $\widehat{W}_{\boldsymbol{\psi}}$ are trained via the lower-level robust objective:

$$\widehat{W}_{\boldsymbol{\psi}}=\arg\min_{W}\max_{q\in\Delta_G}\left[\sum_{g=1}^{G}q_g L_g^{\mathrm{tr}}(W)-\frac{\eta}{2}\left\|q-\frac{1}{G}\mathbf{1}\right\|^2+\frac{\lambda}{2}\|W\|^2\right]. \tag{16}$$

Under standard assumptions on the learning problem (e.g., Lipschitz and bounded convex loss, bounded continuous hyperparameter space, and strongly convex lower-level regularization), we establish that the lower-level optimization algorithm induces a bounded, Lipschitz-continuous hypothesis space with respect to the continuous hyperparameters.

Informally, we prove a *Continuous Oracle Inequality* for our robust tuning frameworks, demonstrating that tuning continuous multidimensional hyperparameters (such as the robustness penalty η, the L_2 regularization coefficient λ, and group-specific perturbation radii $\boldsymbol{\epsilon}$) on a validation set allows our algorithm to achieve an optimal bias-variance trade-off without suffering from discretization error or grid-search penalties. Our main technical tools rely on the uniform stability of the lower-level predictor (due to strong convexity) and the Rademacher complexity of the algorithmic hypothesis class (due to algorithmic Lipschitzness) (Shalev-Shwartz & Ben-David, 2014).

Theorem 2.1 (Informal Continuous Oracle Inequality for Bilevel DRO)**.** *Let $\hat{\boldsymbol{\psi}}$ be the multidimensional continuous hyperparameters tuned via the upper-level continuous validation process. Let $\widehat{W}_{\hat{\boldsymbol{\psi}}}$ be the corresponding robust model trained in the lower level. With high probability over the training and validation sets, the true worst-group risk $L_{\mathcal{D}}^{worst}(\widehat{W}_{\hat{\boldsymbol{\psi}}}) := \max_{g\in[G]} L_{\mathcal{D},g}(\widehat{W}_{\hat{\boldsymbol{\psi}}})$ is bounded by:*

$$L_{\mathcal{D}}^{worst}(\widehat{W}_{\hat{\boldsymbol{\psi}}}) \le \underbrace{L_{\mathcal{D}}^{worst}(W^*)}_{\textit{Reference Risk}} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\log G}{\min_g n_g^{\mathrm{tr}}}} + \sqrt{\frac{\log G}{\min_g n_g^{\mathrm{val}}}}\right)}_{\textit{Statistical Gaps}}$$
$$+ \min_{\boldsymbol{\psi}} \left(\underbrace{\mathcal{O}(\lambda+\eta)}_{\textit{Approximation Bias}(W^*;\boldsymbol{\psi})} + \underbrace{\mathcal{O}\left(\frac{1}{\lambda\eta \min_g n_g^{\mathrm{tr}}}\right)}_{\textit{Stability Gap}(\boldsymbol{\psi})} \right) + \underbrace{\mathcal{O}\left(\sqrt{\frac{k}{\min_g n_g^{\mathrm{val}}} \log(\rho_{\mathcal{A},\boldsymbol{\psi}})}\right)}_{\textit{Continuous Tuning Penalty}}$$

where $\min_g n_g^{\mathrm{tr}}$ and $\min_g n_g^{\mathrm{val}}$ are the sizes of the smallest groups in the training and validation sets respectively, G is the number of groups, k is the dimensionality of the hyperparameter space (e.g., $k = 2$ for standard GDRO, $k = G + 2$ for HDRO), and $\rho_{\mathcal{A},\boldsymbol{\psi}}$ is the algorithmic Lipschitz constant of the lower-level optimization. The Approximation Bias scales with the algorithmic regularization penalties λ and η, while the Stability Gap (derived via uniform stability) shrinks as regularization increases, fundamentally capturing the bias-variance trade-off parameterized by $\boldsymbol{\psi}$.

Remark 2.2 (Algorithmic Lipschitz Constant). For standard Group DRO where $\boldsymbol{\psi} = (\lambda, \eta)$, the Lipschitz constant is bounded by $\rho_{\mathcal{A}} = \mathcal{O}\big(\frac{1}{\lambda_{\min}^2} + \frac{1}{\lambda_{\min}\eta_{\min}^2}\big)$ (see Corollary F.6), where $\lambda_{\min} > 0$ and $\eta_{\min} > 0$ are lower bounds of search spaces. When extending to Bi-HDRO with tunable perturbation radii $\boldsymbol{\psi} = (\lambda, \eta, \boldsymbol{\epsilon})$, the mapping remains Lipschitz continuous with the constant expanding by $\mathcal{O}\big(\frac{1}{\lambda_{\min}} + \frac{1}{\lambda_{\min}\eta_{\min}}\big)$ due to the norm-bounded inner perturbations (see Lemma F.9). Furthermore, as detailed in the appendix, this framework extends to the scenario where the weights of a deep neural network encoder (upper level decision variables) are also treated as tuning parameters. Similar bounds on uniform stability and algorithmic Lipschitz continuity hold in this high-dimensional regime (see Corollaries F.7 and F.12).

This result confirms that continuous bilevel tuning discovers the theoretically optimal configuration for any unknown reference predictor W^*. Crucially, the statistical penalty for tuning over a continuous space scales logarithmically with the algorithmic Lipschitz constant. This constant dictates an "effective grid size"—the finite number of distinguishable models within the search space—allowing us to bypass discretization error while paying a statistical penalty no worse than a discrete grid search. Furthermore, unlike exhaustive grid search which suffers from exponential computational complexity in high dimensions, our scheme enables efficient continuous optimization over multidimensional hyperparameter spaces (full assumptions and proofs are provided in Appendix F.5).

3 Related Works

GDRO GDRO (Sagawa et al., 2020) aims at minimizing the training loss on the group with least training signals due to the spurious attribute. Empirically, minimizing the worst group training loss per se does not translate to robustness over test data, necessitating a tuned weight decay term (Sagawa et al., 2020). DFR (Kirichenko et al., 2023) and AFR (Qiu et al., 2023) improve on GDRO by retraining the convex classifier (last layer of a deep neural network) using a group-balanced set (Ren et al., 2018), which they show to improve the worst-group robustness even with a ERM-trained model. When both inter-group and intra-group uncertainty exist, HDRO (Jo et al., 2026) perturbs the latent representation with group-dependent radii before performing classification in the last layer. The perturbation radii is fixed and sensitive, so extensive tuning is necessary.

Other methods focus on using a small amount of group-labeled data to achieve similar worst-case oracle performance. SSA (Nam et al., 2022) first trains a hard group predictor before running GDRO on pseudolabeled data, and PG-DRO (Ghosal & Li, 2023) instead use soft group prediction and robust training on soft labels. Notably, AGRO (Paranjape et al., 2023) jointly trains an adversarial soft group prediction model by changing group assignments to increase robust classifier's group-wise prediction loss. CnC (Zhang et al., 2022a) aligns samples with the same class but different

attributes in a two-stage contrastive learning framework. DISC (Wu et al., 2023) partitions data with a "concept bank" which consists of candidate spurious attributes. GIC (Han & Zou, 2024) has three stages and identifies spurious features by comparing the training set with a carefully selected reference dataset. D3M (Jain et al., 2024) removes examples that disproportionately degrade worst-group accuracy. Remarkably, XRM (Pezeshki et al., 2024) requires no auxiliary datasets whatsoever and instead identifies spurious attributes by training twin classifiers with mutually exclusive training splits and discover attributes using worst-performing samples before running GDRO.

Bilevel Optimization Bilevel optimization methods has been widely applied to hyperparameter tuning (Bennett et al., 2008; Franceschi et al., 2018), meta-learning (Franceschi et al., 2018; Bertinetto et al., 2019; Rajeswaran et al., 2019), reinforcement learning (Hong et al., 2023; Yang et al., 2024; Li et al., 2024a;b), and neural architecture search (Liu et al., 2019).

Among hyperparameter tuning applications, validation set performance is optimized in the upper level to improve model generalizability over the training set (Domke, 2012; Maclaurin et al., 2015; Franceschi et al., 2017; 2018; Shaban et al., 2019; Feurer & Hutter, 2019; Lorraine et al., 2020). In our work, we treat the robustness mechanism, which may contain non-convex neural networks, as hyperparameters to be optimized. Inspired by Lu & Mei (2024) and Lu & Mei (2026) which solved bilevel optimization problem via deterministic minimax optimization, Shen et al. (2026) addressed a more general bilevel optimization problem when both upper and lower level are minimax problems and extended it to a stochastic case. The fact that GDRO itself is a minimax problem naturally makes such bilevel-minimax algorithm good solver candidates.

4 Experiments

In this section, we describe the modified datasets under both inter-group and intra-group subpopulation shift before showing the performance on two instantiations of our bilevel robust mechanism learning framework for DRO, namely **Bi-HDRO** and **Bi-PG-DRO**.

4.1 Datasets with Test Distribution Shift

We use datasets where both intra-group distribution and inter-group distribution shifts exist. Since baseline methods have reached similar performance under inter-group subpopulation shift, and further tuning offers no improvements, we do not investigate it here. Dataset details in Appendix D.1.

Shifted CMNIST Class label is digit (0-4 or 5-9), and the spurious attribute is color. We rotate minority group samples (red, lebel 1 (digit 5-9)) by 90° in the validation and test sets.

Shifted CelebA Class label is hair color, and spurious attribute is gender. We include only no-glasses images in training and validation, and only with-glasses images at test time for minority group (male with blond hair).

Shifted CivilComments Class label is toxicity and spurious attribute is black/white. Since "black" and "white" attributes may be co-mentioned, for the minority toxic/black group, we include no "white" attribute in the training set, whereas in the test set all entries have the "white" attribute.

4.2 Results on Bi-HDRO

In this setup, group membership is known at all times, so we use **GDRO** (Sagawa et al., 2020), **DFR**$^{\text{Tr}}$ (Kirichenko et al., 2023) and **PDE** (Deng et al., 2023) as baselines since they need access to group information at training time. See details in Appendix D.3.

Validity of our enhanced setup As seen in the top half of Table 1, intra-group shift on top of existing inter-group shift indeed decimates the performance of strong baselines, making the doubly-shifted datasets valuable benchmarks. We further validated the improvement of HDRO compared to existing baselines and show that employing ambiguity sets in such setup provides tangible benefits across all datasets. For example, HDRO improves from 59.2 of GDRO to 72.4 on CelebA. However,

Table 1: Accuracy over 3 runs under shifted distributions for Bi-HDRO and its baselines

Method	CMNIST		CelebA		CivilComments	
	Worst	Avg	Worst	Avg	Worst	Avg
GDRO	71.3$\pm$1.3	72.7$\pm$1.6	59.2$\pm$0.1	92.7$\pm$0.1	34.8$\pm$6.0	87.8$\pm$1.4
DFR$^{\text{Tr}}$	62.9$\pm$8.3	68.9$\pm$4.3	65.5$\pm$4.8	89.4$\pm$0.3	40.8$\pm$2.7	87.9$\pm$0.9
PDE	62.8$\pm$7.8	69.1$\pm$3.8	35.9$\pm$3.4	92.0$\pm$0.6	39.0$\pm$3.9	81.8$\pm$0.8
HDRO	72.7$\pm$0.2	76.3$\pm$3.1	72.4$\pm$3.0	91.4$\pm$0.2	40.8$\pm$3.1	88.0$\pm$1.8
Fixed ϵ						
– Bi-GDRO ($\epsilon = 0$)	73.9$\pm$0.9	74.8$\pm$0.8	77.0$\pm$4.3	90.1$\pm$0.5	57.9$\pm$5.1	79.1$\pm$1.3
– Grid search ϵ	72.3$\pm$2.1	73.9$\pm$1.8	82.5$\pm$2.0	90.7$\pm$1.3	54.7$\pm$2.3	80.0$\pm$1.2
– Grid search η and ϵ	73.7$\pm$0.7	75.0$\pm$1.2	82.1$\pm$2.5	90.4$\pm$0.8	57.9$\pm$6.5	78.0$\pm$3.4
Bi-HDRO (learnable ϵ)	74.0$\pm$0.3	76.4$\pm$0.4	81.8$\pm$3.5	89.2$\pm$1.0	**59.5$\pm$3.5**	78.5$\pm$2.6
Bi-HDRO (learnable ϵ_g)	**74.2$\pm$0.3**	76.9$\pm$1.6	**82.8$\pm$3.1**	90.1$\pm$1.2	57.4$\pm$2.0	78.5$\pm$0.6

we used 4 ϵ and 5 C values (part of the scaling term C/n_g^{tr} in Sagawa et al., 2020, Eq (5)) for HDRO tuning, yielding 16 combinations in grid search.

Benefit of parameter tuning In the bottom half of Table 1, we show that (1) Bi-HDRO performs better than HDRO (lower level) alone; and (2) ablating individual ϵ_g and η as manually tuned and fixed components shows the superior performance of Bi-HDRO.

4.3 Results on Bi-PG-DRO

In this setting, group labels are not available at training time, so we uniformly sample a small fraction of group-labeled validation data (5% from CMNIST, 15% from CelebA, and 3% from CivilComments) and create two splits, where one is used to tune the attribute prediction model and the other is used to tune the validation set. For comparison fairness among baseline methods, we made sure attribute prediction training sees the same split, and the other split is for manual model tuning. We use **ERM**, **GIC**^{C_y} (Han & Zou, 2024), **XRM** (Pezeshki et al., 2024), **AGRO** (Paranjape et al., 2023), and **SSA** (Nam et al., 2022) as our baselines since they need minimal or no group-labeled data at training time. See details in Appendix D.4.

Group-labeled validation data boost our performance With validation tuning, fixed η, and no attribute prediction regularization, our method already performs better than all baseline methods.

Implicitly tuning auxiliary model yields superior performance Compared to explicitly using a cross-entropy loss to train the attribute predictor, implicitly tuning the attribute predictor by minimizing KL divergence works just as well. Finally, we show that the complete bilevel method (Bi-PG-DRO) tuning η works the best when combined with KL, outperforming all baseline methods.

5 Conclusion

We proposed a DRO framework to learn the robustness mechanism that requires minimal tuning and provided two instantiantions. There are two future directions from this work. First, our bilevel framework is still somewhat restrictive in that the algorithm proposed by Shen et al. (2026) requires lower level training to be convex-concave. By assuming Kurdyka-Łojasiewicz (KL) condition so that lower level becomes a non-convex-concave problem, a more general class of applications ensues, but no such algorithm exists yet. Second, this class of algorithms may be further extended to machine unlearning and language model alignment tasks (Fan et al., 2025; Wu et al., 2025; Asif & Amiri, 2026).

Table 2: Accuracy over 3 runs under shifted subpopulations for Bi-PG-DRO and its baselines

Method	CMNIST		CelebA		CivilComments	
	Worst	Avg	Worst	Avg	Worst	Avg
ERM	1.6$\pm$1.7	15.7$\pm$6.1	25.0$\pm$2.6	95.4$\pm$0.1	36.8$\pm$4.6	91.3$\pm$0.8
GIC^{C_y}	25.7$\pm$8.8	48.0$\pm$13.4	47.1$\pm$9.2	90.8$\pm$0.8	54.2$\pm$4.9	86.8$\pm$1.4
XRM	68.8$\pm$3.8	72.2$\pm$3.2	51.4$\pm$2.8	89.7$\pm$0.2	25.9$\pm$6.4	89.6$\pm$1.5
AGRO	22.8$\pm$4.9	32.4$\pm$2.2	22.1$\pm$1.8	95.4$\pm$0.2	24.5$\pm$3.5	91.4$\pm$0.4
SSA	70.0$\pm$2.9	72.3$\pm$1.7	58.6$\pm$2.3	89.7$\pm$0.1	27.4$\pm$8.6	87.1$\pm$1.9
PG-DRO	64.3$\pm$2.0	70.7$\pm$0.7	66.7$\pm$1.0	91.6$\pm$0.4	16.1$\pm$3.1	82.3$\pm$3.0
Fixed $\eta = 1.0$	67.8$\pm$1.5	74.1$\pm$1.6	69.8$\pm$6.9	92.6$\pm$0.2	66.2$\pm$1.1	84.3$\pm$1.1
+ BCE	69.5$\pm$1.4	75.3$\pm$4.0	73.3$\pm$5.4	91.8$\pm$2.2	65.0$\pm$1.3	79.3$\pm$1.4
+ KL	71.0$\pm$0.8	80.1$\pm$0.9	73.3$\pm$4.8	92.5$\pm$0.3	68.0$\pm$0.3	84.7$\pm$1.1
Bi-PG-DRO	**71.7**$\pm$**0.8**	79.5$\pm$1.5	**76.7**$\pm$**3.4**	92.2$\pm$0.3	**68.2**$\pm$**0.2**	83.6$\pm$1.9

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, clean and reformat dataset, and assist in the writing of proofs.

We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, or interpret results.

Generating synthetic datasets is not applicable to this work.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. LLM-generated proof steps are judiciously reviewed and revised by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Faruk Ahmed, Yoshua Bengio, Harm Van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

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

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, 2022.

Sadia Asif and Mohammad Mohammadi Amiri. OFMU: OPTIMIZATION-DRIVEN FRAMEWORK FOR MACHINE UNLEARNING. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZDuyNJI56H.

Kasra Azizi, Kumar Anurag, and Wenbin Wan. Towards resilient tracking in autonomous vehicles: A distributionally robust input and state estimation approach. *IFAC-PapersOnLine*, 59(3):31–36, 2025. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2025.07.006. 12th IFAC Symposium on Intelligent Autonomous Vehicles IAV 2025.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2(1):31, 2019.

Aharon Ben-Tal and Arkadi Nemirovski. Robust optimization-methodology and applications. *Mathematical Programming*, 92:453–480, 2002.

Kristin P. Bennett, Gautam Kunapuli, Jing Hu, and Jong-Shi Pang. Bilevel optimization and machine learning. In *IEEE World Congress on Computational Intelligence*, pp. 25–47. Springer, 2008.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *Proceedings of the International Conference on Machine Learning*, pp. 1170–1182, 2021.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. In *Journal of machine learning research*, 2001.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the International Conference on Machine Learning*, pp. 2189–2200, 2021.

John M Danskin. The theory of max-min and its application to weapons allocation problems. *Econometrics and Operations Research*, 5, 1967.

Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9QEVJ9qm46.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Science & Business Media, 2013.

Justin Domke. Generic methods for optimization-based modeling. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2012.

J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *Proceedings of the International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zZjLv6F0Ks.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated Machine Learning*, pp. 3–33. Springer, 2019.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the International Conference on Machine Learning*, pp. 1165–1173, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1568–1577, 2018.

Soumya Suvra Ghosal and Yixuan Li. Distributionally robust optimization with probabilistic group. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11809–11817, 2023.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *Proceedings of the International Conference on Machine Learning*, pp. 17480–17504, 2024. URL https://openreview.net/forum?id=KycvgOCBBR.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10431–10443. Curran Associates, Inc., 2021.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, volume 31, 2018.

Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Madry. Improving subgroup robustness via data selection. In *Advances in Neural Information Processing Systems*, 2024.

Jinyong Jeong, Hyungu Kahng, and Seoung Bum Kim. Multi-expert distributionally robust optimization for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 101081–101116, 2025. doi: 10.52202/085713-3384.

Sung Ho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mruL1LDjzV.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021.

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 *Proceedings of the International Conference on Machine Learning*, pp. 5815–5826, 2021.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Advances in Neural Information Processing Systems*, 2023.

Haoyu Lei, Amin Gohari, and Farzan Farnia. On the inductive biases of demographic parity-based fair learning algorithms. In Negar Kiyavash and Joris M. Mooij (eds.), *Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2205–2225. PMLR, 15–19 Jul 2024.

Chenliang Li, Siliang Zeng, Zeyi Liao, Jiaxiang Li, Dongyeop Kang, Alfredo Garcia, and Mingyi Hong. Learning reward and policy jointly from demonstration and preference improves alignment. *arXiv preprint arXiv:2406.06874*, 2024a.

Jiaxiang Li, Siliang Zeng, Hoi To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Getting more juice out of the SFT data: Reward learning from human demonstration improves SFT for LLM alignment. In *Advances in Neural Information Processing Systems*, 2024b.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024. doi: 10.1137/23M1566753. URL https://doi.org/10.1137/23M1566753.

Zhaosong Lu and Sanyou Mei. Solving bilevel optimization via sequential minimax optimization. *Mathematics of Operations Research*, 2026. doi: 10.1287/moor.2024.0521. URL https://doi.org/10.1287/moor.2024.0521. Published online January 6, 2026.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the International Conference on Machine Learning*, 2015.

Raghav Mehta, Changjian Shui, and Tal Arbel. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 1453–1492. PMLR, 10–12 Jul 2024.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_F9xpOrqyX9.

Bhargavi Paranjape, Pradeep Dasigi, Vivek Srikumar, Luke Zettlemoyer, and Hannaneh Hajishirzi. AGRO: Adversarial discovery of error-prone groups for robust optimization. In *International Conference on Learning Representations*, 2023.

Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with XRM. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gPStP3FSY9.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Focus on the common good: Group distributional robustness follows. In *International Conference on Learning Representations*, 2022.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proceedings of the International Conference on Machine Learning*, pp. 28448–28467, 2023. URL https://openreview.net/forum?id=s5F1a6s1HS.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

Kanghyun Ryu and Negar Mehr. Integrating predictive motion uncertainties with distributionally robust risk-aware control for safe robot navigation in crowds. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2410–2417, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16742–16751, 2022.

Amir Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Yiyang Shen, Yutian He, Weiran Wang, and Qihang Lin. Penalty-based first-order methods for bilevel optimization with minimax and constrained lower-level problems. *arXiv preprint arXiv:2605.08006*, 2026.

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.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, 2020.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CbfsKHiWEn.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *Proceedings of the International Conference on Machine Learning*, pp. 37765–37786, 2023.

Yan Yang, Bin Gao, and Ya-xiang Yuan. Bilevel reinforcement learning via the development of hyper-gradient without lower-level convexity. *arXiv preprint arXiv:2405.19697*, 2024.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *Proceedings of the International Conference on Machine Learning*, pp. 39584–39622, 2023.

Han Yu, Jiashuo Liu, Xingxuan Zhang, Jiayun Wu, and Peng Cui. A survey on evaluation of out-of-distribution generalization. *ArXiv*, 2024.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-N-Contrast: A contrastive approach for improving robustness to spurious correlations. In *Proceedings of the International Conference on Machine Learning*, pp. 26484–26516, 2022a.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21668–21681, 2022b.

Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2020–2030, 2017.

A A First-Order Method for Bilevel Minimax Problems

Algorithm 1 A First-Order Method for (13)

Require: Initial iterates (α^0, W^0, q^0) and $(p^0, \widetilde{W}^0, \widetilde{q}^0)$ and number of proximal iterations K.

1: Set ρ, ρ_1, ρ_2 and relevant parameters according to Shen et al., 2026, Algorithm 1.
2: **for** $k = 0, \ldots, K-1$ **do**
3: Construct the proximal penalized objective

$$\bar{\mathcal{P}}_k = P_\rho(\alpha, p, W, q, \widetilde{W}, \widetilde{q}) + \frac{\rho_1}{2}\|(\alpha, W, q) - (\alpha^k, W^k, q^k)\|^2 - \frac{\rho_2}{2}\|(p, \widetilde{W}, \widetilde{q}) - (p^k, \widetilde{W}^k, \widetilde{q}^k)\|^2$$

4: Solve the resulting strongly-convex-strongly-concave problem with SAPD (Zhang et al., 2022b):

$$((\alpha^{k+1}, W^{k+1}, q^{k+1}), (p^{k+1}, \widetilde{W}^{k+1}, \widetilde{q}^{k+1})) \leftarrow \mathrm{SAPD}(\bar{\mathcal{P}}_k)$$

5: **end for**
6: **return** $(\alpha^{k'}, W^{k'}, q^{k'})$ with k' sampled uniformly from $\{1, \ldots, K\}$.

While the original work has many hyperparameters on the algorithmic level, we stress that those are tuned once and can then be applied to all datasets used in this paper and both of our proposed methods (See Appendix D.2 for details). We further note that validation accuracy is used for model selection rather than objective convergence, which is prohibitively expensive in deep learning experiments.

B Closed Form of Lower Level Adversary

We can use the definition of training loss in Section 2.1,

$$L^{\mathrm{tr}}(W; \theta) = \left(L_1^{\mathrm{tr}}(W; \theta), \ldots, L_G^{\mathrm{tr}}(W; \theta)\right)^\top,$$

to obtain the adversarial subproblem,

$$q^\star \in \arg\max_{q \in \Delta_G} \left\{ q^\top L^{\mathrm{tr}}(W; \theta) - \frac{\eta}{2}\left\| q - \frac{1}{G}\mathbf{1} \right\|_2^2 \right\}.$$

For $\eta > 0$, we can complete the square:

$$q^\top L^{\mathrm{tr}} - \frac{\eta}{2}\left\| q - \frac{1}{G}\mathbf{1} \right\|_2^2 = -\frac{\eta}{2}\left\| q - \left(\frac{1}{G}\mathbf{1} + \frac{L^{\mathrm{tr}}}{\eta}\right) \right\|_2^2 + C,$$

where C does not depend on q. Therefore,

$$q^\star(W, \theta, \eta) = \Pi_{\Delta_G}\left(\frac{1}{G}\mathbf{1} + \frac{L^{\mathrm{tr}}(W; \theta)}{\eta}\right),$$

where component-wise $q_g^\star = \left[\frac{1}{G} + \frac{L_g^{\mathrm{tr}}(W;\theta)}{\eta} - \nu\right]_+$, and ν is chosen such that $\sum_{g=1}^G q_g^\star = 1$.

While there is a closed form solution, computing it requires the losses of all groups and hence a full pass of the entire dataset, which is prohibitively expensive during stochastic training.

C Perturbation Formulations

C.1 Closed-form Perturbation for Binary Classification

We show below that closed form can be derived for latent representation perturbation when the task is binary classification. We use the binary cross entropy (BCE) loss for all experiments but show the formulation for hinge loss as well.

Case 1: the last layer is linear and $\mathcal{L}$ is hinge loss. Let the final layer be binary linear classification: $f_\theta^L(z) = w^\top z + b$. The hinge loss is

$$L(f_\theta^L(z), y) = \max(0, 1 - y(w^\top z + b)).$$

The inner maximization in equation becomes

$$\sup_{\|z'-z\|\le\epsilon} \max\big(0, 1 - y(w^\top z' + b)\big).$$

Write $z' = z + \delta, \|\delta\| \le \epsilon$, then

$$\sup_{\|\delta\|\le\epsilon} \max\big(0, 1 - y(w^\top z + b) - yw^\top \delta\big).$$

Since $x \mapsto \max(0, x)$ is monotonically increasing,

$$= \max\left(0, 1 - y(w^\top z + b) + \sup_{\|\delta\|\le\epsilon} (-yw^\top \delta)\right).$$

Because of $y \in \{\pm 1\}$ and the support function of the norm ball,

$$\sup_{\|\delta\|\le\epsilon} (-yw^\top \delta) = \sup_{\|\delta\|\le\epsilon} w^\top \delta = w^\top \delta^* = w^\top (\epsilon \frac{w}{\|w\|^2}) = \epsilon \|w\|_*,$$

where $\|\cdot\|_*$ is the dual norm. Therefore, the inner robust hinge loss has closed form:

$$\sup_{\|z'-z\|\le\epsilon} L(f_\theta^L(z'), y) = \max\big(0, 1 - y(w^\top z + b) + \epsilon\|w\|_*\big)$$

Case 2: the last layer is linear and $\mathcal{L}$ is logistic/binary cross entropy (BCE) loss. Using the same notations as above, the logistic loss is

$$L(f_\theta^L(z), y) = \log\big(1 + \exp(-y(w^\top z + b))\big).$$

Due to the monotonicity of the exponent of the maximized perturbation, we have

$$\sup_{\|\delta\|\le\epsilon} \log\big(1 + \exp\big(-y(w^\top(z+\delta) + b)\big)\big) \Leftrightarrow \inf_{\|\delta\|\le\epsilon} y(w^\top(z+\delta)+b) = y(w^\top z+b) + \inf_{\|\delta\|\le\epsilon} yw^\top\delta.$$

As derived in Case 1, $\inf_{\|\delta\|\le\epsilon} yw^\top\delta = -\epsilon\|w\|_*$, so we can substitute this back to the maximized perturbation, forming

$$\sup_{\|\delta\|\le\epsilon} \log\big(1 + \exp(-y(w^\top(z+\delta) + b))\big) = \log\big(1 + \exp\big(-(y(w^\top z + b) - \epsilon\|w\|_*)\big)\big).$$

$\|w\|$ makes the both losses non-smooth, so we smoothen it using a substitution scaler u. In the logistic example, it becomes

$$J(w, u) = \sum_{i=1}^{n} \log\big(1 + \exp(-y_i(w^\top x_i + b) + \epsilon u)\big) \quad \text{s.t.} \quad \|w\|_* \le u, \tag{17}$$

which is jointly convex in (w, u).

As a result, we perform projected descent when optimizing Eq (17). Define the second-order cone $\mathcal{K} = \{(w, u) : \|w\|_2 \le u\}$, the Euclidean projection is then

$$\min_{(w,u)\in\mathcal{K}} \frac{1}{2}\|w - v\|_2^2 + \frac{1}{2}(u - t)^2, \tag{18}$$

the closed form of which falls into three categories, where (v, t) falls inside, outside and "behind", and outside but near the boundary of the cone, i.e.,

$$\Pi_{\mathcal{K}}(v, t) = \begin{cases} (v, t), & \|v\|_2 \le t \\ (0, 0), & \|v\|_2 \le -t \\ \left(\frac{\|v\|_2 + t}{2\|v\|_2} v, \frac{\|v\|_2 + t}{2}\right), & \text{otherwise.} \end{cases}$$

C.2 PERTURBATION FOR MULTI-CLASS CLASSIFICATION

Next, we show that latent representation perturbation can still be done, although in a relaxed form, when the task is multi-class.

Let the final layer be linear with K classes:

$$w_k(z) = \mathbf{w}_k^\top z + b_k, \qquad k = 1, \dots, K,$$

and let the true label be $y \in \{1, \dots, K\}$.

Case 1: multi-class hinge loss. We use the multi-class hinge loss (Crammer & Singer, 2001):

$$L(w(z), y) = \max\left\{0, \max_{j \neq y} [1 + (w_j(z) + b_j) - (w_y(z) + b_y)]\right\},$$

where $w_y(z)$ is the score of the correct class, and the inner max aims to find the largest violation over all incorrect classes. For a perturbation $z' = z + \delta$, $\|\delta\| \leq \epsilon$, we have

$$w_j(z+\delta) - w_y(z+\delta) = (w_j - w_y)^\top z + (b_j - b_y) + (w_j - w_y)^\top \delta.$$

Therefore,

$$\sup_{\|\delta\| \leq \epsilon} L(w(z+\delta), y) = \sup_{\|\delta\| \leq \epsilon} \max\left\{0, \max_{j \neq y} \left[1 + (w_j - w_y)^\top z + b_j - b_y + (w_j - w_y)^\top \delta\right]\right\}.$$

Since the maximum is over finitely many (K) affine functions, the supremum can be exchanged with the finite maximum:

$$= \max\left\{0, \max_{j \neq y} \left[1 + (w_j - w_y)^\top z + b_j - b_y + \sup_{\|\delta\| \leq \epsilon} (w_j - w_y)^\top \delta\right]\right\}.$$

Using the support function of the norm ball,

$$\sup_{\|\delta\| \leq \epsilon} (w_j - w_y)^\top \delta = \epsilon \|w_j - w_y\|_*.$$

Thus the robust multi-class hinge loss has the closed form

$$\sup_{\|z' - z\| \leq \epsilon} L(w(z'), y) = \max\left\{0, \max_{j \neq y} [1 + w_j(z) - w_y(z) + \epsilon \|\mathbf{w}_j - \mathbf{w}_y\|_*]\right\}.$$

Equivalently, one may introduce auxiliary variables u_j satisfying

$$\|w_j - w_y\|_* \leq u_j, \qquad j \neq y,$$

and write the robust loss as

$$\max\left\{0, \max_{j \neq y} [1 + w_j(z) - w_y(z) + \epsilon u_j]\right\}.$$

This form is convex in the final-layer parameters for fixed features z.

Case 2: multi-class logistic / softmax cross-entropy loss. For true class y, the softmax cross-entropy loss is

$$L(w(z), y) = -\log \frac{\exp(w_y(z))}{\sum_{k=1}^K \exp(w_k(z))} = \log\left(\sum_{k=1}^K \exp(w_k(z) - w_y(z))\right).$$

Define

$$a_k = w_k(z) - w_y(z), \qquad v_k = \mathbf{w}_k - \mathbf{w}_y.$$

Then

$$w_k(z+\delta) - w_y(z+\delta) = a_k + v_k^\top \delta,$$

with $a_y = 0$ and $v_y = 0$. The robust loss is

$$\sup_{\|\delta\| \le \epsilon} \log \left(\sum_{k=1}^{K} \exp(a_k + v_k^\top \delta) \right).$$

A useful variational representation is obtained from the Fenchel form of log-sum-exp:

$$\log \sum_{k=1}^{K} \exp(r_k) = \sup_{p \in \Delta_K} \left\{ p^\top r + H(p) \right\},$$

where

$$H(p) = -\sum_{k=1}^{K} p_k \log p_k.$$

Applying this with $r_k = a_k + v_k^\top \delta$ gives

$$\sup_{\|\delta\| \le \epsilon} \log \sum_{k=1}^{K} \exp(a_k + v_k^\top \delta) = \sup_{p \in \Delta_K} \left\{ \sum_{k=1}^{K} p_k a_k + H(p) + \epsilon \left\| \sum_{k=1}^{K} p_k v_k \right\|_* \right\}. \tag{19}$$

However, this is not a closed-form perturbation of the same type as the binary case. Continuing from (19), we derive an efficient upper bound relaxation as follows

$$\begin{aligned}
&\sup_{\|\delta\| \le \epsilon} \log \sum_{k=1}^{K} \exp(a_k + v_k^\top \delta) \\
&\le \sup_{p \in \Delta_K} \left\{ \sum_{k=1}^{K} p_k a_k + H(p) + \epsilon \sum_{k=1}^{K} p_k \|v_k\|_* \right\} \quad \text{(Jensen's inequality on } \|\cdot\|_* \text{)} \\
&= \sup_{p \in \Delta_K} \left\{ \sum_{k=1}^{K} p_k \left(a_k + \epsilon \|v_k\|_* \right) + H(p) \right\} \quad \text{(Grouping linear terms)} \\
&= \log \sum_{k=1}^{K} \exp\left(a_k + \epsilon \|v_k\|_* \right) \quad \text{(Reverse Fenchel conjugate)}
\end{aligned}$$

D Experimental Details

D.1 Shifted Datasets

Shifted CMNIST Dataset statistics are presented in Table 3. In the original HDRO experiment (Jo et al., 2026), rotation is only applied to test set and its distribution significantly differs from the validation distribution, causing large variance across different random seeds (8%-10%) and tuning on such validation set would make little sense. In our setup, rotation is applied to both validation and test split of the minority group (red, label 1). This is done to reduce the extreme large variance observed during model tuning and stabilize prediction performance. Note that only rotations are applied, and samples are not moved.

Table 3: CMNIST Statistics.

Group	**Train**	**Val**	**Test**
$y = 0$, green	2,998	2,591	8,966
$y = 1$, green	11,781	2,513	1,013
$y = 0$, red	12,130	2,465	1,068
$y = 1$, red	3,091	2,431	8,953
Total	30,000	10,000	20,000

Shifted CelebA Dataset statistics are presented in Table 4. Our shifting procedure is the same as that of Jo et al. (2026). For minority group (blond male), 164 without glasses are moved from test to train, 90 with glasses are moved from train to test, and 10 with glasses are moved from validation to test.

Table 4: CelebA Statistics

Group	Train	Val	Test
non-blond female	71,629	8,535	9,767
non-blond male	66,874	8,276	7,535
blond female	22,880	2,874	2,480
blond male (before shift)	1,387	182	180
blond male (after shift)	1,461	172	116
Total (before shift)	162,770	19,867	19,962
Total (after shift)	162,844	19,857	19,898

Shifted CivilComments Dataset statistic presented in Table 5. Similar to the test distribution shift of CMNIST and CelebA, we created a shifted version of the CivilComments dataset (Koh et al., 2021), which is originally used for toxic language classification. Intra-group shift on this dataset has not been studied previously to our knowledge. For the shifted version, 1,278 minority group (toxic, black) samples with white=1 are moved from train to test, and 905 minority group samples with white=0 are moved from test to train. After shifting, train minority group contains only non-white-annotated samples, and test minority contains only white annotated samples. The fact that attributes black=1 and white=1 are not mutually exclusive allows such shift to happen.

Table 5: CivilComments Statistics

Group	Train	Val	Test
non-toxic, non-black	231,738	39,006	115,223
non-toxic, black	6,785	1,119	3,335
toxic, non-black	27,404	4,522	13,687
toxic, black (before shift)	3,111	533	1,537
toxic, black (after shift)	2,738	533	1,910
Total (before shift)	269,038	45,180	133,782
Total (after shift)	268,665	45,180	134,155

Shifted Waterbirds Waterbirds (Sagawa et al., 2020) is one of the most commonly used benchmark in this line of DRO research, but we do not use it here because (1) the dataset itself has known mislabeled attributes (Asgari et al., 2022), making performance metrics less informative; (2) the above issue is compounded with the small size of the training set (4,795 total, with 56 in the minority group of waterbird with land background); and (3) the performance has saturated with or without intra-group shifts on existing methods, rendering comparisons banal.

D.2 SETTINGS

In Table 6, we provide the important settings (hyperparameters and architectures) used in this work. We emphasize that those hyperparameters are relatively insensitive to the dataset, so we only tune them on CMNIST for one method and use the same values for the other two datasets. ϵ_g is cliped [0,1] to follow the HDRO preset range.

Table 6: Settings and Hyperparameters for our methods.

Setting	CMNIST	CelebA	CivilComments
Model	ResNet-50	ResNet-50	DistilBERT
Weight decay λ	10^{-3}	10^{-1}	10^{-3}
Lower Train/Upper Validation batch size	256	128	64
$L_{\nabla h}$ in Shen et al. (2026)		10000	
Iterations of SAPD per outer iteration T		1	
η_{init}		1.0	
Gradient multiplier of η		10	
Gradient multiplier of validation simplex u		100	
ϵ_{init} for Bi-HDRO		$96/255$	
KL coefficient β for Bi-PG-DRO		10	

D.3 Bi-HDRO

D.3.1 Ablation Procedure

To show the advantage of automatically tuning both η and ϵ (or ϵ_g) in Bi-HDRO, (1) we fixed ϵ to be 0 and leave η learnable, which recovers Bi-GDRO; (2) we performed grid search with fixed ϵ over {60,72,84,96,108}/255 when $\eta = 0$, and (3) we performed grid search ϵ with the above schedule and η over {0.01, 0.1, 1.0, 10.0}.

D.3.2 Baselines

DFR$^{\text{Tr}}$ (Kirichenko et al., 2023) A two-stage method where the training set is partitioned 80-20 for different purposes. (1) The larger chunk of data is used to train ERM with uniform sampling. (2) The smaller group-balanced subset to retrain last layer. For DFR$^{\text{Val}}$, they include all minority group data in validation and sample the same amount from other groups. For fairness of group information access, we compare our method to DFR$^{\text{Tr}}$ only. Default hyperparameters from their implementation are used.

PDE (Deng et al., 2023) A two-stage method. (1) Warmup the entire model using group-balanced subset where all groups have same size as the minority group. (2) Progressively add more training data (progressive data expansion) for training the entire model as well as using existing warmup subset. We tune the warm-up epochs in {10,20} and added samples in {50,100,500}.

HDRO (Jo et al., 2026) A single-level min-max-max method described in Section 2.3. We use the implementation from the authors and tune ϵ in the range {60,72,84,96}/255 and $C \in \{0, 1, 2, 3\}$ across all three of our datasets.

D.4 Bi-PG-DRO

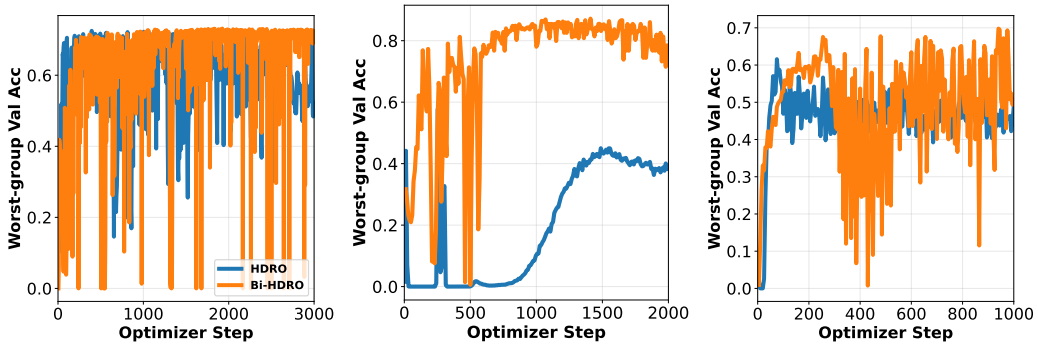
D.4.1 Ablation Procedure

We ablated two types of regularization: BCE loss for attribute predictor, KL prior regularization for attribute predictor. We performed a grid search {1,10,30} for these two terms.

D.4.2 Baselines

GIC^{C_y} (Han & Zou, 2024) A three-stage method. (1) Feature extraction. (2) Train attribute predictor maximizing spurious attribute label KL between predicted label and true label (Eq (11) in their paper, hence the superscript). (3) Train GDRO. We tune the γ (Eq (9) in their paper) in {2,5,10} for our shifted datasets.

XRM (Pezeshki et al., 2024) A two-stage method. (1) Train two auxilary models with mutually exclusive held-in and held-out split of the *training set*. Notably, it does not rely on any attribute



(a) CMNIST (b) CelebA (c) CivilComments

Figure 1: Validation Curve of HDRO and Bi-HDRO.

labels from, for example, the validation set. Instead, it flips training set labels so that minority samples are identified, since they are confidently misclassified in the held-out set. (2) Train GDRO with pseudo-group-labels. We sample three hyperparameter combination candidates, choose one with the highest flip rates, and train GDRO using 3 different seeds.

AGRO (Paranjape et al., 2023) A greedy unified method that jointly trains attribute predictor and robust model. It learns soft group memberships that makes robust training (GDRO) difficult. We use the hyperparameters described in their Table 7 (and 4 slices for CMNIST), but tune their α in $\{0.2,0.3,0.4\}$ due to our different setup (intra-group shift).

SSA (Nam et al., 2022) A two-stage method. (1) Train a attribute label prediction model and performs *hard prediction* to generate pseudo-group-label and (2) Train GDRO. Default hyperparameters are used. We tune the model with $C \in \{0, 1, 2, 3\}$ across all three of our datasets.

PG-DRO (Ghosal & Li, 2023) A two-stage method described in Section 2.4. Essentially SSA but with probabilistic group labels. We tune the model with $C \in \{0, 1, 2, 3\}$ across all three of our datasets.

E Additional Results

E.1 Efficiency

Validation performance. In Figure 1, we show that Bi-HDRO reaches optimal validation accuracy quicker than HDRO does. For CelebA, HDRO took about 6,000 optimization steps to reach level similar to Bi-HDRO. Additionally, HDRO requires extensive grid search to obtain such results, while Bi-HDRO requires only one run.

Runtime. We show below in Table 7 that Bi-HDRO takes shorter time to run. HDRO would require much longer runtime than stated if grid search is taken into account. CMNIST in general takes longer to achieve reasonable performance.

Table 7: Wall-clock time for one run in seconds.

	CMNIST	**CelebA**	**CivilComments**
HDRO	405	3145	625
Bi-HDRO (ϵ_g)	6674	129	3912

E.2 Does Perturbation Help Soft Group Assignments (Bi-PG-DRO)?

We decide not to perturb the latent space in Bi-PG-DRO because empirically, fixed perturbation alone degrades model performance, and ϵ_g converges to 0 when automatically tuned. On CMNIST, we found that perturbation ϵ_g all converged to 0 regardless of initial values (Figure 2). We attempted 3 setups: initial value is 0, initial value is 96/255, initial value is 0 but unbounded from above. Validation performance worsens in Bi-PG-DRO with active perturbation. We observed similar performance degradation on CelebA and CivilComments. Based on this result, we decide not to include perturbation with probabilistic group membership.

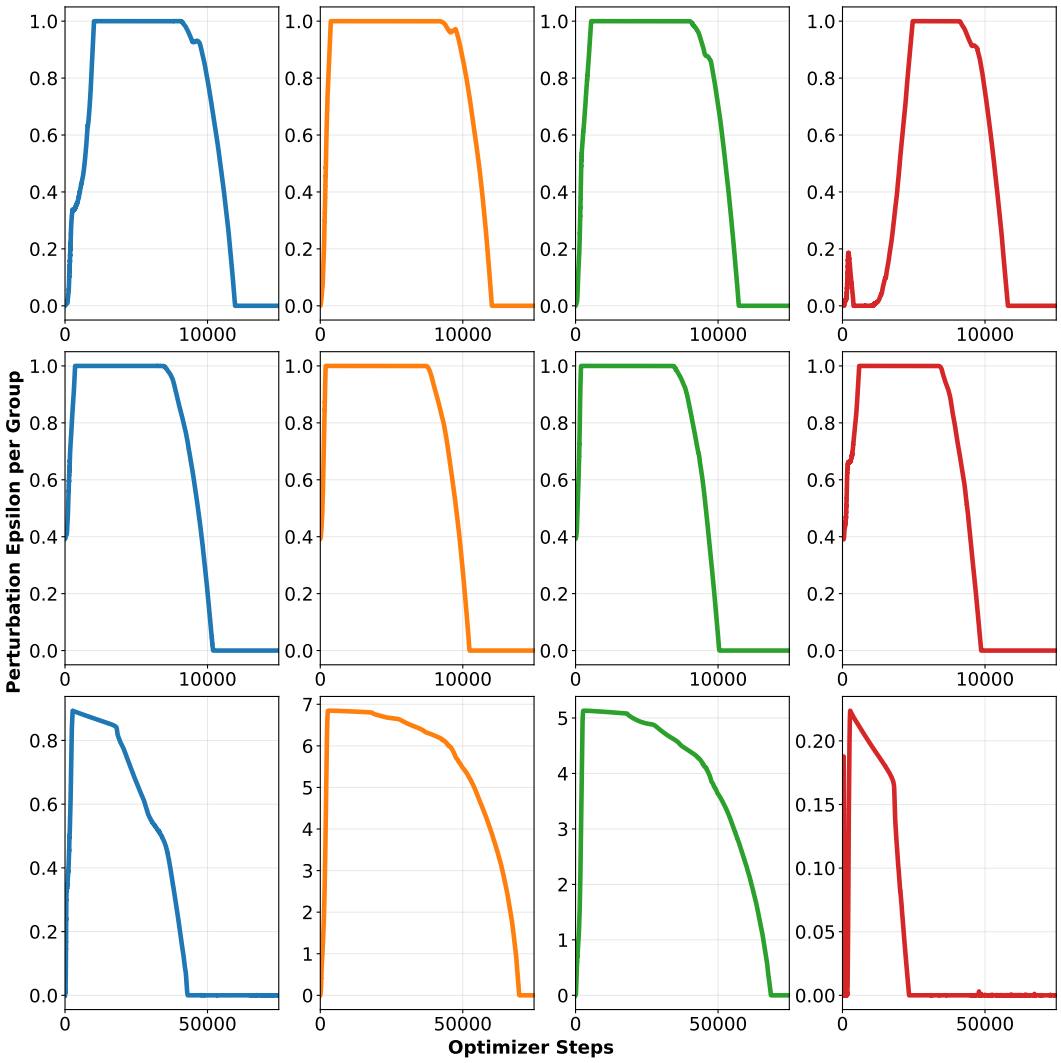
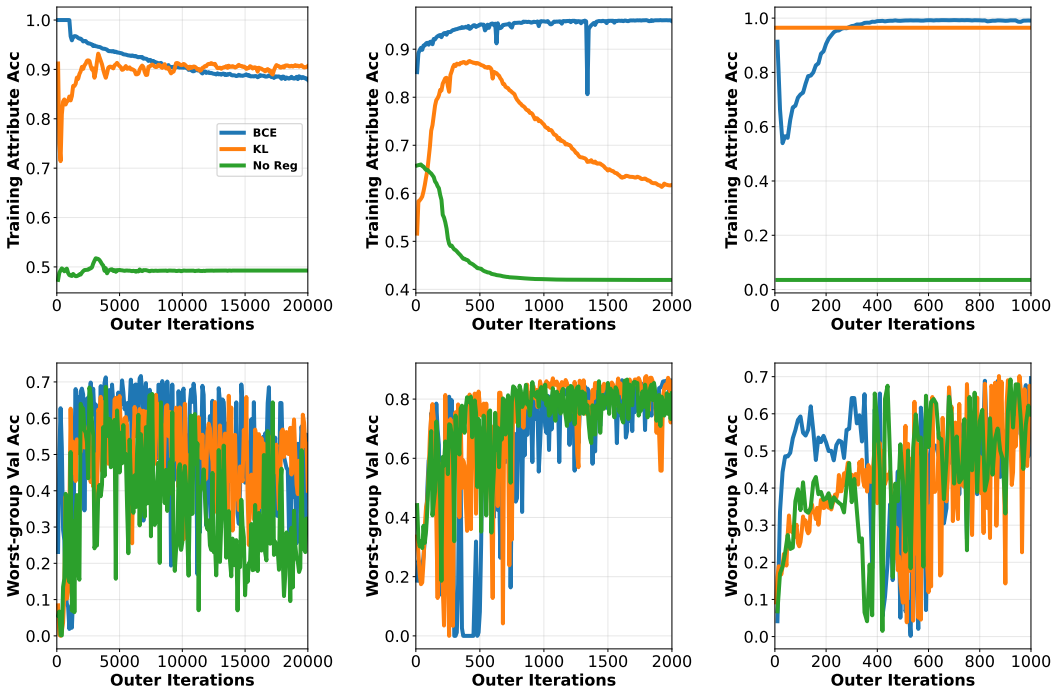


Figure 2: ϵ_g converges to 0 for Bi-PG-DRO. First row: $\epsilon_{\text{init}} = 0$. Second row: $\epsilon_{\text{init}} = 96/255$. Third row: Unclipped $\epsilon_{\text{init}} = 0$. Last column is the shifted minority group.

E.3 Should Membership Inference be Precise for Worst-Group Performance?

As long as worst-group validation loss is focused on during bilevel optimization, predicted group labels need not be precise. In Figure 3, we show that post-hoc attribute accuracy over group-unlabeled training set without upper level regularization is poor, but the validation performance does not degrade as much. Moreover, we show that KL regularization instead of explicit BCE training provides closer attribute prediction accuracy results and gives a small boost in validation worst-group accuracy.



(a) CMNIST (b) CelebA (c) CivilComments

Figure 3: Attribute prediction accuracy on group-unlabeled training data (top) and worst-group validation accuracy (bottom) across three datasets organized by column.

F Generalization Theory for Continuous Bilevel Hyperparameter Tuning

In this section, we analyze the generalization guarantees of continuous bilevel hyperparameter tuning. Our primary theoretical goal is to establish a *Continuous Oracle Inequality* for Group Distributionally Robust Optimization (DRO). This inequality characterizes how optimizing hyperparameters on a validation set allows the algorithm to seamlessly navigate the fundamental trade-off between structural complexity and worst-case group robustness.

Roadmap. Because establishing continuous generalization bounds requires multiple statistical learning tools, our analysis proceeds in four stages.

1. We formalize the bilevel setup and explicitly prove that the lower-level optimization algorithm induces a bounded, Lipschitz-continuous hypothesis space with respect to the hyperparameters (Appendix F.1).

2. We establish the mathematical machinery required to bound the upper-level validation error across the entire continuous hyperparameter path using Rademacher complexity (Appendix F.3).

3. As a theoretical warm-up, we apply this machinery to standard Regularized Empirical Risk Minimization, which isolates how the upper-level continuous tuning error cleanly decouples from the lower-level uniform stability (Appendix F.4).

4. We derive our main result: the Group DRO Oracle Inequality (Appendix F.5).

5. Finally, we extend this continuous generalization theory to Hierarchical DRO (Bi-HDRO), demonstrating that tuning multi-dimensional perturbation radii preserves the algorithmic Lipschitz continuity and resulting Oracle Inequalities (Appendix F.6).

F.1 Setup and Assumptions

Throughout this section, we make the following formal assumptions about the learning problem and the bilevel setup:

1. **Loss Function:** The instantaneous loss function $\ell(W, z)$ is convex and L-Lipschitz with respect to W. Furthermore, its absolute value is bounded by M (i.e., $|\ell(W, z)| \leq M$) over a bounded optimization domain of radius B (i.e., $\|W\| \leq B$).
2. **Continuous Hyperparameter Space:** The generic hyperparameter $\boldsymbol{\phi}$ is tuned over a continuous k-dimensional bounded domain $\Phi \subset \mathbb{R}^k$. We assume the maximum ℓ_2 distance between any two hyperparameters in Φ is bounded by a diameter R.
3. **Strong Convexity of Regularization:** The lower-level objective includes a regularization term $\Omega_\lambda(W)$ that is λ-strongly convex with respect to W.
4. **Algorithmic Mapping and Lipschitz Continuity:** The lower-level optimization acts as a mapping algorithm $\mathcal{A} : \Phi \to \mathbb{R}^d$ on a training set S_{train} of size n^{tr}, outputting a parameter $\widehat{W}_{\boldsymbol{\phi}} = \mathcal{A}(\boldsymbol{\phi})$. We require this mapping $\mathcal{A}$ to be $\rho_{\mathcal{A}}$-Lipschitz continuous with respect to $\boldsymbol{\phi}$ in the ℓ_2 norm: $\|\mathcal{A}(\boldsymbol{\phi}_1) - \mathcal{A}(\boldsymbol{\phi}_2)\| \leq \rho_{\mathcal{A}}\|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|$. While formalized here as a high-level requirement for our general theorems, we explicitly demonstrate later that this continuity is a consequence of strong convexity (Assumption 3) for our specific learning objectives.

Based on this mapping, the upper-level problem evaluates the induced predictors on an independent validation set S_{val} of size n^{val}. We define the effective **algorithmic hypothesis class** explored by the validation process as the k-dimensional continuous manifold induced by $\mathcal{A}$:

$$\mathcal{H}_\Phi = \{\widehat{W}_{\boldsymbol{\phi}} : \boldsymbol{\phi} \in \Phi\} \tag{20}$$

To establish our generalization guarantees, we rely on the classical framework of *uniform stability*. For completeness, we defer the standard formal definitions and stability derivations to Appendix F.7.

F.2 Algorithmic Lipschitz Continuity Examples

To provide concrete intuition for Assumption 4 (Algorithmic Lipschitz Continuity), we walk through two representative cases: a warm-up for standard Regularized Loss Minimization, and a formal proof for our primary application of Group DRO.

Example 1: (Warm-up) Standard Regularized Loss Minimization. Consider a standard bilevel formulation for Regularized Empirical Risk Minimization (ERM), where the goal is to tune the continuous regularization penalty. Here, the one-dimensional continuous hyperparameter is the regularization weight $\lambda \in \Lambda = [\lambda_{\min}, \lambda_{\max}]$ where $\lambda_{\min} > 0$. The lower-level objective minimizes the regularized empirical risk over the training set:

$$F_\lambda(W) = L^{\text{tr}}(W) + \Omega_\lambda(W) \tag{21}$$

yielding the optimal predictor $\widehat{W}_\lambda = \arg\min_W F_\lambda(W)$. The upper-level problem evaluates these predictors to minimize the unregularized risk on a validation set: $\min_{\lambda\in\Lambda} L^{\text{val}}(\widehat{W}_\lambda)$.

Because the lower-level objective $F_\lambda(W)$ is λ-strongly convex (Assumption 3), the Lipschitz continuity of the mapping $\lambda \mapsto \widehat{W}_\lambda$ is naturally guaranteed without requiring differentiability of the loss. By the property of strong convexity at the optimum $\widehat{W}_\lambda$, for any W we have $F_\lambda(W) \geq F_\lambda(\widehat{W}_\lambda) + \frac{\lambda}{2}\|W - \widehat{W}_\lambda\|^2$. Evaluating this for λ_1 at $W = \widehat{W}_{\lambda_2}$ and for λ_2 at $W = \widehat{W}_{\lambda_1}$ gives two inequalities:

$$F_{\lambda_1}(\widehat{W}_{\lambda_2}) \geq F_{\lambda_1}(\widehat{W}_{\lambda_1}) + \frac{\lambda_1}{2}\|\widehat{W}_{\lambda_2} - \widehat{W}_{\lambda_1}\|^2$$
$$F_{\lambda_2}(\widehat{W}_{\lambda_1}) \geq F_{\lambda_2}(\widehat{W}_{\lambda_2}) + \frac{\lambda_2}{2}\|\widehat{W}_{\lambda_1} - \widehat{W}_{\lambda_2}\|^2$$

Expanding $F_\lambda(W) = L^{\text{tr}}(W) + \Omega_\lambda(W)$ and summing them, the empirical loss terms $L^{\text{tr}}(\widehat{W}_{\lambda_1})$ and $L^{\text{tr}}(\widehat{W}_{\lambda_2})$ exactly cancel out on both sides. Rearranging the remaining regularization terms leaves:

$$\frac{\lambda_1 + \lambda_2}{2}\|\widehat{W}_{\lambda_1} - \widehat{W}_{\lambda_2}\|^2 \leq \Big(\Omega_{\lambda_1}(\widehat{W}_{\lambda_2}) - \Omega_{\lambda_1}(\widehat{W}_{\lambda_1})\Big) + \Big(\Omega_{\lambda_2}(\widehat{W}_{\lambda_1}) - \Omega_{\lambda_2}(\widehat{W}_{\lambda_2})\Big) \tag{22}$$

Consider the general proximal regularization case where $\Omega_\lambda(W) = \frac{\lambda}{2}\|W - \mathbf{a}\|^2$ for some reference vector $\mathbf{a}$ (standard Ridge is recovered when $\mathbf{a} = \mathbf{0}$). The right side simplifies to $\frac{\lambda_1-\lambda_2}{2}(\|\widehat{W}_{\lambda_2} - \mathbf{a}\|^2 - \|\widehat{W}_{\lambda_1} - \mathbf{a}\|^2)$. Using the algebraic identity $\|x\|^2 - \|y\|^2 = \langle x - y, x + y\rangle$, we can rewrite this difference as:

$$\frac{\lambda_1 - \lambda_2}{2}\langle \widehat{W}_{\lambda_2} - \widehat{W}_{\lambda_1}, \widehat{W}_{\lambda_2} + \widehat{W}_{\lambda_1} - 2\mathbf{a}\rangle$$

Applying the Cauchy-Schwarz inequality, and noting that the average $\frac{\lambda_1+\lambda_2}{2}$ is strictly lower-bounded by $\lambda_{\min}$, yields:

$$\lambda_{\min}\|\widehat{W}_{\lambda_1} - \widehat{W}_{\lambda_2}\|^2 \leq \frac{|\lambda_1 - \lambda_2|}{2}\|\widehat{W}_{\lambda_1} - \widehat{W}_{\lambda_2}\|\|\widehat{W}_{\lambda_1} + \widehat{W}_{\lambda_2} - 2\mathbf{a}\| \tag{23}$$

By the triangle inequality, and assuming the optimization domain is bounded by a constant radius B (Assumption 1), the second term is bounded by $\|\widehat{W}_{\lambda_1}\| + \|\widehat{W}_{\lambda_2}\| + 2\|\mathbf{a}\| \leq 2(B + \|\mathbf{a}\|)$. Dividing by $\|\widehat{W}_{\lambda_1} - \widehat{W}_{\lambda_2}\|$ proves that the mapping $\mathcal{A}$ is strictly bounded by a constant $\rho_{\mathcal{A}} = (B + \|\mathbf{a}\|)/\lambda_{\min}$ and is therefore $\rho_{\mathcal{A}}$-Lipschitz:

$$\|\widehat{W}_{\lambda_1} - \widehat{W}_{\lambda_2}\| \leq \rho_{\mathcal{A}}|\lambda_1 - \lambda_2| \tag{24}$$

Alternatively, if the loss and regularizer are strictly twice continuously differentiable, one can recover this same Lipschitz constant $\rho_{\mathcal{A}} \leq B/\lambda_{\min}$ by directly bounding the spectral norm of the Implicit Function Theorem Jacobian: $\frac{d\widehat{W}_\lambda}{d\lambda} = -[\nabla^2_W F_\lambda(\widehat{W}_\lambda)]^{-1}\nabla^2_{\lambda W} F_\lambda(\widehat{W}_\lambda)$.

Example 2: Group Distributionally Robust Optimization (DRO). As a concrete application, consider a linearized Group DRO setting where we optimize a linear classifier W directly on fixed inputs. Given training data partitioned into G groups with empirical group losses $L^{\text{tr}}(W) \in \mathbb{R}^G$, the lower level optimizes the classifier against a worst-case group distribution $q \in \Delta_G$ (the probability simplex weighting the groups). The deviation of this worst-case distribution from the uniform distribution is penalized by a robustness hyperparameter $\eta \in H = [\eta_{\min}, \eta_{\max}]$ where $\eta_{\min} > 0$. The lower-level problem is:

$$\widehat{W}_\eta = \arg\min_W \max_{q\in\Delta_G} \left[q^T L^{\text{tr}}(W) - \frac{\eta}{2}\left\| q - \frac{1}{G}\mathbf{1}\right\|^2 + \frac{\lambda}{2}\|W\|^2 \right] \tag{25}$$

where λ is a fixed L_2 regularization coefficient. In the upper level, we tune the robustness hyperparameter $\eta \in H$ using an independent validation set D_{val} (partitioned into G groups) to minimize the worst-group validation loss. Here, η naturally assumes the role of the generic continuous tuning hyperparameter $\boldsymbol{\phi}$ (with $k = 1$) from Section F.1, while λ serves strictly as the fixed strong convexity constant required for stability:

$$\min_{\eta\in H}\max_{g\in[G]} L_g^{\text{val}}(\widehat{W}_\eta) \tag{26}$$

By continuously tuning η, the bilevel formulation dynamically discovers the optimal trade-off between average-case empirical risk and worst-group robustness on unseen data.

Remarkably, this formulation strictly satisfies the algorithmic Lipschitz condition (Assumption 4). We formally state this property below.

Lemma F.1 (Algorithmic Lipschitz Continuity of Group DRO)**.** *Suppose the instantaneous group losses $L_g^{\text{tr}}(W)$ are L-Lipschitz and bounded by M over a bounded optimization domain of radius B (Assumption 1). For any fixed strong convexity regularizer $\lambda > 0$ (Assumption 3), the max-marginalized lower-level objective $J_\eta(W)$ induces an algorithmic mapping $\widehat{W}_\eta$ that is $\rho_{\mathcal{A}}$-Lipschitz continuous with respect to the robustness hyperparameter $\eta \geq \eta_{\min} > 0$, with explicit constant:*

$$\rho_{\mathcal{A}} \leq \frac{GLM}{\lambda\eta_{\min}^2} \tag{27}$$

Proof. Let $J_\eta(W) = \max_{q\in\Delta_G} F(W, q, \eta)$ be the max-marginalized lower-level objective, where $F(W, q, \eta) = q^T L^{\text{tr}}(W) - \frac{\eta}{2}\|q - \frac{1}{G}\mathbf{1}\|^2 + \frac{\lambda}{2}\|W\|^2$. By Danskin's theorem (Danskin, 1967), the gradient of the max-marginalized function is simply the gradient of the objective evaluated at the optimal inner variable. Thus, its gradient is $\nabla J_\eta(W) = \nabla_W F(W, q^*_\eta(W), \eta) =$

$J(W)q_\eta^*(W) + \lambda W$, where $J(W)$ is the $d \times G$ Jacobian matrix of group losses. By Assumption 1, each group loss is L-Lipschitz, so the operator norm of $J(W)$ is bounded by its Frobenius norm $\sqrt{\sum_{g=1}^{G} \|\nabla L_g^{\mathrm{tr}}(W)\|^2} \leq \sqrt{G}L$.

Next, we bound how much this gradient shifts with respect to η. The optimal inner distribution is exactly the simplex projection $q_\eta^*(W) = \Pi_{\Delta_G}(\mathbf{u}_\eta(W))$, where $\mathbf{u}_\eta(W) = \frac{1}{\eta}L^{\mathrm{tr}}(W) + \frac{1}{G}\mathbf{1}$. Because the projection Π_{Δ_G} is non-expansive (1-Lipschitz), the shift in the optimal inner distribution is bounded by the shift in the unprojected vector: $\|q_{\eta_1}^*(W) - q_{\eta_2}^*(W)\| \leq \|\mathbf{u}_{\eta_1}(W) - \mathbf{u}_{\eta_2}(W)\| = \left\|(\frac{1}{\eta_1} - \frac{1}{\eta_2})L^{\mathrm{tr}}(W)\right\|$. By Assumption 1, each group loss is bounded by M, so $\|L^{\mathrm{tr}}(W)\| \leq \sqrt{G}M$. Requiring $\eta \geq \eta_{\min} > 0$, we have $|1/\eta_1 - 1/\eta_2| \leq |\eta_1 - \eta_2|/\eta_{\min}^2$. This directly provides the shift bound: $\|q_{\eta_1}^*(W) - q_{\eta_2}^*(W)\| \leq \frac{\sqrt{G}M}{\eta_{\min}^2}|\eta_1 - \eta_2|$, which in turn limits the overall gradient shift to $\|\nabla J_{\eta_1}(W) - \nabla J_{\eta_2}(W)\| \leq \frac{GLM}{\eta_{\min}^2}|\eta_1 - \eta_2|$.

Finally, let $\widehat{W}_{\eta_1}$ and $\widehat{W}_{\eta_2}$ be the optimal lower-level classifiers for hyperparameters η_1 and η_2. Because J_η is λ-strongly convex, its gradient is λ-strongly monotone:

$$\langle \nabla J_{\eta_1}(\widehat{W}_{\eta_1}) - \nabla J_{\eta_1}(\widehat{W}_{\eta_2}), \widehat{W}_{\eta_1} - \widehat{W}_{\eta_2} \rangle \geq \lambda \|\widehat{W}_{\eta_1} - \widehat{W}_{\eta_2}\|^2$$

For any closed convex domain (in particular $\|W\| \leq B$), the first-order optimality condition dictates:

$$\langle -\nabla J_{\eta_1}(\widehat{W}_{\eta_1}), \widehat{W}_{\eta_1} - \widehat{W}_{\eta_2} \rangle \geq 0$$
$$\langle \nabla J_{\eta_2}(\widehat{W}_{\eta_2}), \widehat{W}_{\eta_1} - \widehat{W}_{\eta_2} \rangle \geq 0$$

Summing the above three inequalities and applying the Cauchy-Schwarz inequality yields:

$$\begin{aligned} \lambda \|\widehat{W}_{\eta_1} - \widehat{W}_{\eta_2}\|^2 &\leq \langle \nabla J_{\eta_2}(\widehat{W}_{\eta_2}) - \nabla J_{\eta_1}(\widehat{W}_{\eta_2}), \widehat{W}_{\eta_1} - \widehat{W}_{\eta_2} \rangle \\ &\leq \|\nabla J_{\eta_2}(\widehat{W}_{\eta_2}) - \nabla J_{\eta_1}(\widehat{W}_{\eta_2})\| \|\widehat{W}_{\eta_1} - \widehat{W}_{\eta_2}\| \end{aligned} \tag{28}$$

Dividing by $\lambda \|\widehat{W}_{\eta_1} - \widehat{W}_{\eta_2}\|$ and plugging in our bound for the gradient shift establishes the explicit Lipschitz constant $\rho_{\mathcal{A}}$. □

Thus, tuning the DRO robustness parameter over a continuous space is theoretically well-behaved and Lipschitz-stable as long as the search space is bounded away from zero ($\eta \geq \eta_{\min} > 0$).

F.3 Uniform Convergence over Multi-dimensional Continuous Hypothesis Spaces

Before specializing to specific learning algorithms like ERM or Group DRO, we first establish a general uniform convergence bound for the k-dimensional continuous hypothesis class $\mathcal{H}_\Phi$. By leveraging the Lipschitz continuity of the algorithmic mapping, we can bound the Rademacher complexity of this manifold as a function of the hyperparameter space dimensionality k.

Lemma F.2 (Rademacher Complexity of k-dimensional Algorithmic Hypothesis Class)**.** *Suppose the loss function ℓ is L-Lipschitz and bounded by M (Assumption 1). Let $\Phi \subset \mathbb{R}^k$ be a bounded k-dimensional hyperparameter space with ℓ_2 diameter R (Assumption 2). If the lower-level mapping $\mathcal{A}(\boldsymbol{\phi}) = \widehat{W}_{\boldsymbol{\phi}}$ is $\rho_{\mathcal{A}}$-Lipschitz (Assumption 4), the empirical Rademacher complexity of the validation loss class over $\mathcal{H}_\Phi$ on a set S_{val} of size n^{val} is bounded by:*

$$\hat{\mathcal{R}}_{S_{val}}(\mathcal{H}_\Phi) \leq \frac{6M}{\sqrt{n^{\mathrm{val}}}}\sqrt{k}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}}RL}{M}\right)} + 2\sqrt{\log 2}\right) \tag{29}$$

Proof. The proof relies on bounding the continuous covering number and applying discrete chaining. For a k-dimensional space Φ with ℓ_2 diameter R, its r-covering number in ℓ_2 norm is bounded by $N(r, \Phi) \leq (3R/r)^k$ for $r \leq R$. This follows from a standard volumetric argument: a maximal r-separated set of size N in Φ induces N disjoint balls of radius $r/2$. Since Φ has diameter R, these disjoint balls are entirely contained within a larger ball of radius $R + r/2$. Comparing their volumes yields $N(r/2)^k \leq (R + r/2)^k$, which implies $N \leq (1 + 2R/r)^k \leq (3R/r)^k$ when $r \leq R$. Due

to the $\rho_{\mathcal{A}}$-Lipschitz property of the mapping, an r-cover of Φ projects to an $(\rho_{\mathcal{A}} \cdot r)$-cover of the hypothesis class $\mathcal{H}_\Phi$. Thus, the covering number of the effective hypothesis class is bounded by $N(r, \mathcal{H}_\Phi) \leq (3\rho_{\mathcal{A}} R/r)^k$.

Let $A \subset \mathbb{R}^{n^{\text{val}}}$ be the set of loss evaluations on $S_{\text{val}} = \{z_1, \ldots, z_{n^{\text{val}}}\}$ for all predictors in $\mathcal{H}_\Phi$. The maximum ℓ_2 norm of any vector in A is $C = M\sqrt{n^{\text{val}}}$. Since the loss is L-Lipschitz, the ℓ_2 distance between two loss vectors $\mathbf{a}_{\boldsymbol{\phi}_1}$ and $\mathbf{a}_{\boldsymbol{\phi}_2}$ generated by predictors $\widehat{W}_{\boldsymbol{\phi}_1}$ and $\widehat{W}_{\boldsymbol{\phi}_2}$ is bounded by $\|\mathbf{a}_{\boldsymbol{\phi}_1} - \mathbf{a}_{\boldsymbol{\phi}_2}\|_2 \leq L\sqrt{n^{\text{val}}}\|\widehat{W}_{\boldsymbol{\phi}_1} - \widehat{W}_{\boldsymbol{\phi}_2}\|_2$. Therefore, guaranteeing an ϵ-cover of A requires at most an r-cover of $\mathcal{H}_\Phi$ with $r = \frac{\epsilon}{L\sqrt{n^{\text{val}}}}$. Thus, the covering number of A is bounded by $N(\epsilon, A) \leq N(r, \mathcal{H}_\Phi) \leq \left(\frac{3\rho_{\mathcal{A}} RL\sqrt{n^{\text{val}}}}{\epsilon}\right)^k$.

By the discrete chaining lemma (Shalev-Shwartz & Ben-David, 2014, Lemma 27.4), we evaluate the complexity over discrete scales $\epsilon_i = C2^{-i}$. For any $i \geq 1$:

$$\sqrt{\log N(C2^{-i}, A)} \leq \sqrt{k \log\left(\frac{3\rho_{\mathcal{A}} RL\sqrt{n^{\text{val}}}}{M\sqrt{n^{\text{val}}}} 2^i\right)} \leq \sqrt{k}(\alpha + \beta i) \tag{30}$$

where $\alpha = \sqrt{\log\left(\frac{3\rho_{\mathcal{A}} RL}{M}\right)}$ and $\beta = \sqrt{\log 2}$. Evaluating the full infinite chaining sum gives the explicit Rademacher complexity (Shalev-Shwartz & Ben-David, 2014, Lemma 27.5):

$$\hat{\mathcal{R}}_{S_{\text{val}}}(\mathcal{H}_\Phi) \leq \frac{6C}{n^{\text{val}}}\sqrt{k}(\alpha + 2\beta) = \frac{6M}{\sqrt{n^{\text{val}}}}\sqrt{k}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}} RL}{M}\right)} + 2\sqrt{\log 2}\right) \tag{31}$$

□

This lemma provides a deterministic uniform convergence bound. By applying standard Rademacher concentration bounds (Shalev-Shwartz & Ben-David, 2014, Theorem 26.5) combined with McDiarmid's inequality, the uniform deviation $\sup_{\boldsymbol{\phi}\in\Phi} |L_{\mathcal{D}}(\widehat{W}_{\boldsymbol{\phi}}) - L^{\text{val}}(\widehat{W}_{\boldsymbol{\phi}})| \leq \epsilon_{\text{val}}$ holds with probability $1 - \delta$, where:

$$\epsilon_{\text{val}} = \frac{12M}{\sqrt{n^{\text{val}}}}\sqrt{k}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}} RL}{M}\right)} + 2\sqrt{\log 2}\right) + M\sqrt{\frac{2\log(2/\delta)}{n^{\text{val}}}} \tag{32}$$

F.4 Warm-up: Continuous Oracle Inequality for Regularized ERM

In this subsection, we formalize the generalization guarantees for Regularized Empirical Risk Minimization (ERM), where the generic continuous tuning parameter is instantiated as the L_2 regularization coefficient (i.e., we set $\boldsymbol{\phi} = \lambda$ with $k = 1$). In this setting, the lower level trains a regularized model $\widehat{W}_\lambda$ on a training set $S_{\text{train}} = \{z_1^{\text{train}}, \ldots, z_{n^{\text{tr}}}^{\text{train}}\}$ drawn i.i.d. from a single data distribution $\mathcal{D}$:

$$\widehat{W}_\lambda = \arg\min_W \left(\frac{1}{n^{\text{tr}}}\sum_{i=1}^{n^{\text{tr}}} \ell(W, z_i^{\text{train}}) + \Omega_\lambda(W)\right) \tag{33}$$

The upper level evaluates this continuous hypothesis class $\mathcal{H}_\Lambda = \{\widehat{W}_\lambda : \lambda \in \Lambda\}$ on an independent validation set $S_{\text{val}} = \{z_1^{\text{val}}, \ldots, z_{n^{\text{val}}}^{\text{val}}\}$ also drawn from $\mathcal{D}$:

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \frac{1}{n^{\text{val}}}\sum_{j=1}^{n^{\text{val}}} \ell(\widehat{W}_\lambda, z_j^{\text{val}}) \tag{34}$$

This serves as the foundational continuous oracle inequality, clearly distinct from the worst-case Group DRO formulation analyzed subsequently.

The following theorem combines the lower-level high-probability generalization bound for a fixed λ with the uniform convergence bound over the 1-dimensional continuous hypothesis class $\mathcal{H}_\Lambda$.

Theorem F.3 (Continuous Oracle Inequality for ERM)**.** *Suppose Assumptions 1–4 hold: the loss ℓ is L-Lipschitz and bounded by M, the regularization Ω_λ is λ-strongly convex, the continuous tuning interval Λ has length R, and the lower-level mapping $\mathcal{A}(\lambda) = \widehat{W}_\lambda$ is $\rho_{\mathcal{A}}$-Lipschitz. Let W^* be an arbitrary reference predictor (e.g., the population risk minimizer). Let $\hat{\lambda} \in \Lambda$ be the hyperparameter chosen by minimizing the validation risk over $\mathcal{H}_\Lambda$. With probability at least $1-\delta$ over the random draw of both S_{train} and S_{val}, the true risk $L_{\mathcal{D}}(\widehat{W}_{\hat{\lambda}}) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(\widehat{W}_{\hat{\lambda}}, z)]$ satisfies:*

$$L_{\mathcal{D}}(\widehat{W}_{\hat{\lambda}}) \leq L_{\mathcal{D}}(W^*) + \min_{\lambda\in\Lambda}\left(\Omega_\lambda(W^*) + \frac{2L^2}{\lambda n^{\mathrm{tr}}} + \left(\frac{4L^2}{\lambda n^{\mathrm{tr}}} + \frac{4M}{n^{\mathrm{tr}}}\right)\sqrt{\frac{n^{\mathrm{tr}}\log(4/\delta)}{2}}\right) + \frac{24M}{\sqrt{n^{\mathrm{val}}}}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}}RL}{M}\right)} + 2\sqrt{\log 2}\right) + 2M\sqrt{\frac{2\log(4/\delta)}{n^{\mathrm{val}}}}$$

Interpretation of the Continuous Oracle Inequality. In standard Regularized Loss Minimization (as noted in Shalev-Shwartz & Ben-David (2014, Corollary 13.8)), finding the optimal hyperparameter λ requires prior knowledge of the optimal predictor's norm $\|W^*\|$. If $\|W^*\|$ is known, one can analytically set λ to perfectly balance the bias (the $\Omega_\lambda(W^*)$ term) and the variance (the stability gap, which scales as $O(1/\lambda m)$), thereby achieving an optimal generalization rate of $O(1/\sqrt{m})$.

However, in practice, the true optimal predictor W^* and its norm are unknown. The standard approach to circumvent this is Structural Risk Minimization (SRM), where one trains models on a finite, discrete grid of λ values and selects the best one using a validation set. While SRM guarantees learning, it restricts the solution to the predefined grid, introducing discretization error. Furthermore, as established by Shalev-Shwartz & Ben-David (2014, Theorem 11.2), to theoretically guarantee that the validation set does not overfit to any model in the finite grid, one applies the union bound. Plugging the union bound over all $|Grid|$ discrete options into Hoeffding's inequality incurs a uniform convergence penalty scaling with $O\left(\sqrt{\log(|Grid|)/n^{\mathrm{val}}}\right)$. Consequently, attempting to reduce discretization error by making the grid denser degrades the theoretical generalization guarantee.

Our Continuous Oracle Inequality demonstrates that continuous bilevel tuning effectively acts as an "oracle," automatically discovering the theoretically optimal bias-variance tradeoff $\min_{\lambda\in\Lambda}(\Omega_\lambda(W^*) + \epsilon_{\text{train}}(\lambda))$ for the *unknown* reference predictor W^*. Crucially, it achieves this without discretization error, paying only a logarithmic uniform convergence penalty $O\big(\sqrt{\log(\rho_{\mathcal{A}}RL/M)/n^{\mathrm{val}}}\big)$. Conceptually, the inner term $\rho_{\mathcal{A}}RL/M$ acts as an "effective grid size"—representing the finite number of distinguishable models within the continuous interval—allowing us to bypass discretization error while paying a statistical penalty no worse than a dense discrete grid. Furthermore, this continuous formulation allows us to directly traverse the hyperparameter space using efficient continuous optimization techniques, such avoiding the prohibitive computational cost of repeatedly training independent models associated with standard grid search.

Proof. Given the Lipschitz continuity of the algorithmic mapping established in the assumptions, the proof decomposes the true risk using the uniform convergence of the validation loss (via our general Rademacher bound) and the stability of the lower-level algorithm.

Step 1: Upper-Level Uniform Convergence. By substituting $k = 1$ into the general uniform deviation bound derived in Equation (32), the two-sided uniform deviation $\sup_{\lambda\in\Lambda}|L_{\mathcal{D}}(\widehat{W}_\lambda) - L^{\mathrm{val}}(\widehat{W}_\lambda)| \leq \epsilon_{\text{val}}$ holds with probability $1-\delta/2$, where:

$$\epsilon_{\text{val}} = \frac{12M}{\sqrt{n^{\mathrm{val}}}}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}}RL}{M}\right)} + 2\sqrt{\log 2}\right) + M\sqrt{\frac{2\log(4/\delta)}{n^{\mathrm{val}}}} \tag{35}$$

Step 2: Final Continuous Oracle Inequality. Let W^* be any fixed reference predictor (such as the population risk minimizer). We define the optimal regularization parameter for this predictor as $\lambda^* = \arg\min_{\lambda\in\Lambda}(\Omega_\lambda(W^*) + \epsilon_{\text{train}}(\lambda))$, where $\epsilon_{\text{train}}(\lambda)$ will be defined shortly. Since $\hat{\lambda}$ minimizes the empirical validation risk, we have $L^{\mathrm{val}}(\widehat{W}_{\hat{\lambda}}) \leq L^{\mathrm{val}}(\widehat{W}_{\lambda^*})$. Applying the uniform deviation

bound (which holds over all $\mathcal{H}_\Lambda$ with probability $1-\delta/2$) to both $\hat{\lambda}$ and λ^* yields:

$$\begin{aligned} L_{\mathcal{D}}(\widehat{W}_{\hat{\lambda}}) &\leq L^{\text{val}}(\widehat{W}_{\hat{\lambda}}) + \epsilon_{\text{val}} \\ &\leq L^{\text{val}}(\widehat{W}_{\lambda^*}) + \epsilon_{\text{val}} \\ &\leq L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) + 2\epsilon_{\text{val}} \end{aligned}$$

To bound $L_{\mathcal{D}}(\widehat{W}_{\lambda^*})$, we compare it against the fixed W^*. By the empirical optimality of $\widehat{W}_{\lambda^*}$, we have $L^{\text{tr}}(\widehat{W}_{\lambda^*}) + \Omega_{\lambda^*}(\widehat{W}_{\lambda^*}) \leq L^{\text{tr}}(W^*) + \Omega_{\lambda^*}(W^*)$. Because $\Omega_\lambda \geq 0$, we can decompose the true risk as:

$$\begin{aligned} L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) &= L^{\text{tr}}(\widehat{W}_{\lambda^*}) + \Big(L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) - L^{\text{tr}}(\widehat{W}_{\lambda^*})\Big) \\ &\leq L^{\text{tr}}(W^*) + \Omega_{\lambda^*}(W^*) + \Big(L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) - L^{\text{tr}}(\widehat{W}_{\lambda^*})\Big) \\ &= L_{\mathcal{D}}(W^*) + \Omega_{\lambda^*}(W^*) + \underbrace{\Big(L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) - L^{\text{tr}}(\widehat{W}_{\lambda^*})\Big)}_{\text{Stability gap}} + \underbrace{\big(L^{\text{tr}}(W^*) - L_{\mathcal{D}}(W^*)\big)}_{\text{Hoeffding gap}} \end{aligned}$$

By Lemma F.15 with probability $1-\delta/4$, the stability gap is bounded by:

$$\frac{2L^2}{\lambda^* n^{\text{tr}}} + \left(\frac{4L^2}{\lambda^* n^{\text{tr}}} + \frac{2M}{n^{\text{tr}}}\right)\sqrt{\frac{n^{\text{tr}}\log(4/\delta)}{2}}$$

Simultaneously, since W^* is fixed independent of S_{train}, we apply Hoeffding's inequality (Shalev-Shwartz & Ben-David, 2014, Lemma B.6). Because the absolute loss is bounded by M, the loss variables fall in a range of $2M$. For a one-sided bound with local confidence $\delta_{\text{local}} = \delta/4$, Hoeffding's inequality exactly bounds the second gap by $M\sqrt{\frac{2\log(4/\delta)}{n^{\text{tr}}}} = \frac{2M}{n^{\text{tr}}}\sqrt{\frac{n^{\text{tr}}\log(4/\delta)}{2}}$ with probability $1-\delta/4$. Summing these bounds gives:

$$L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) \leq L_{\mathcal{D}}(W^*) + \Omega_{\lambda^*}(W^*) + \underbrace{\frac{2L^2}{\lambda^* n^{\text{tr}}} + \left(\frac{4L^2}{\lambda^* n^{\text{tr}}} + \frac{4M}{n^{\text{tr}}}\right)\sqrt{\frac{n^{\text{tr}}\log(4/\delta)}{2}}}_{:=\epsilon_{\text{train}}(\lambda^*)} \tag{36}$$

By the definition of λ^*, this is exactly $L_{\mathcal{D}}(W^*) + \min_{\lambda\in\Lambda}\left(\Omega_\lambda(W^*) + \epsilon_{\text{train}}(\lambda)\right)$. Substituting this bound directly into $L_{\mathcal{D}}(\widehat{W}_{\hat{\lambda}}) \leq L_{\mathcal{D}}(\widehat{W}_{\lambda^*}) + 2\epsilon_{\text{val}}$, and applying the union bound over all three events (total probability $1-\delta$), we obtain the Continuous Oracle Inequality. Expanding $\epsilon_{\text{train}}(\lambda)$ and $2\epsilon_{\text{val}}$ precisely matches the theorem statement, completing the proof. □

F.5 Continuous Oracle Inequality for Group DRO

While Theorem F.3 outlines the oracle inequality for standard ERM over a unified dataset, applying this continuous generalization bound to the Group DRO formulation requires substituting the sample complexities. In this setting, the generic continuous tuning parameter is instantiated as the robustness penalty (i.e., we set $\phi = \eta$ with $k = 1$), while the L_2 regularization coefficient λ is held fixed strictly to satisfy the required strong convexity for algorithmic stability. Because DRO is evaluated on a worst-case basis at both levels, uniform stability in the lower level and uniform convergence in the upper level are strictly bottlenecked by the most scarcely represented groups.

Lemma F.4 (Modified Uniform Stability of Group DRO)**.** *Let* $J_\eta(W;S) = \max_{q\in\Delta_G}[\sum_{g=1}^G q_g L_g^{\text{tr}}(W) - \frac{\eta}{2}\|q - \frac{1}{G}\mathbf{1}\|^2 + \frac{\lambda}{2}\|W\|^2]$ *be the lower-level Group DRO objective. Under the assumptions of bounded and Lipschitz continuous group losses (Assumption 1), the lower-level Group DRO algorithm* $\mathcal{A}(S) = \arg\min_W J_\eta(W;S)$ *is uniformly stable with modified constant:*

$$\beta_{DRO} = \frac{2L^2}{\lambda \min_g n_g^{\text{tr}}}\left(1 + \frac{M\sqrt{G}}{\eta_{\min}}\right) \tag{37}$$

where $\min_g n_g^{\text{tr}}$ *is the size of the smallest group in the training set* S*.*

Proof. This proof follows the same standard gradient-based strong convexity argument as the standard stability result (Lemma F.14), but must account for the coupled min-max optimization over all groups. When the training set S is perturbed by changing exactly one example z into z', this perturbation occurs in exactly one group, say group k. Because each instantaneous loss is bounded by M, the empirical loss of group k changes by at most $\frac{1}{n_k^{\mathrm{tr}}}|\ell(W,z)-\ell(W,z')| \le \frac{2M}{\min_g n_g^{\mathrm{tr}}}$.

To bound the uniform stability, we evaluate how much the gradient of J_η shifts due to this perturbation. By Danskin's theorem, the gradient is exactly $\nabla J_\eta(W;S) = \sum_{g=1}^G q_g \nabla L_g^{\mathrm{tr}}(W) + \lambda W$, where $q = \Pi_{\Delta_G}\left(\frac{1}{\eta}L^{\mathrm{tr}}(W) + \frac{1}{G}\mathbf{1}\right)$ are the optimal simplex weights. Let S' be the perturbed dataset with corresponding optimal weights q'. The shift in the gradient of the loss term is bounded by:

$$\begin{aligned}
&\left\|\sum_{g=1}^G q_g \nabla L_g^{\mathrm{tr}}(S) - \sum_{g=1}^G q'_g \nabla L_g^{\mathrm{tr}}(S')\right\| \\
&\le \underbrace{\left\|\sum_{g=1}^G q_g \left(\nabla L_g^{\mathrm{tr}}(S) - \nabla L_g^{\mathrm{tr}}(S')\right)\right\|}_{\text{Direct gradient shift}} + \underbrace{\left\|\sum_{g=1}^G (q_g - q'_g)\nabla L_g^{\mathrm{tr}}(S')\right\|}_{\text{Weight perturbation shift}}
\end{aligned}$$

For the first term, only group k's empirical gradient changes (by at most $\frac{2L}{\min_g n_g^{\mathrm{tr}}}$). Since $q_k \le 1$, this direct shift is bounded by $\frac{2L}{\min_g n_g^{\mathrm{tr}}}$. For the second term, the simplex projection is 1-Lipschitz, so the shift in optimal weights is bounded by the scaled shift in the group losses: $\|q - q'\| \le \frac{1}{\eta}\|L^{\mathrm{tr}}(S) - L^{\mathrm{tr}}(S')\| \le \frac{1}{\eta_{\min}}\frac{2M}{\min_g n_g^{\mathrm{tr}}}$. Multiplying by the Jacobian norm of the group losses ($\sqrt{G}L$), the second term is bounded by $\frac{2ML\sqrt{G}}{\eta_{\min}\min_g n_g^{\mathrm{tr}}}$.

Summing these, the entire gradient shifts by at most $\frac{2L}{\min_g n_g^{\mathrm{tr}}}\left(1 + \frac{M\sqrt{G}}{\eta_{\min}}\right) := \Delta$. Because J_η is λ-strongly convex, applying the exact same strong monotonicity displacement argument from Equation (43) in Lemma F.14 guarantees $\|\widehat{W}_\eta(S) - \widehat{W}_\eta(S')\| \le \frac{\Delta}{\lambda}$. Multiplying by the L-Lipschitz constant of the worst-group loss yields the final modified uniform stability constant β_{DRO}. □

Theorem F.5 (Group DRO Continuous Oracle Inequality)**.** *Let $\min_g n_g^{\mathrm{tr}}$ and $\min_g n_g^{\mathrm{val}}$ denote the sizes of the smallest groups in the training and validation sets, respectively, where both sets consist of G groups. For any reference predictor W^*, let $\Omega_\eta(W^*) = \frac{\lambda}{2}\|W^*\|^2 + \frac{\eta}{2}$ be the approximation bias penalty. Let $\hat{\eta} = \arg\min_{\eta\in H} f_{val}(\widehat{W}_\eta)$ be the hyperparameter chosen by minimizing the empirical validation worst-group risk f_{val} over the continuous hypothesis class $\mathcal{H}_H$, where $H = [\eta_{\min}, \eta_{\max}]$. Let $\beta_{DRO} = \frac{2L^2}{\lambda \min_g n_g^{\mathrm{tr}}}\left(1 + \frac{M\sqrt{G}}{\eta_{\min}}\right)$ be the modified uniform stability. With probability at least $1-\delta$, the true worst-group risk $L_{\mathcal{D}}^{worst}(\widehat{W}_{\hat{\eta}}) = \max_g L_{\mathcal{D},g}(\widehat{W}_{\hat{\eta}})$ satisfies:*

$$\begin{aligned}
L_{\mathcal{D}}^{worst}(\widehat{W}_{\hat{\eta}}) \le & L_{\mathcal{D}}^{worst}(W^*) + M\sqrt{\frac{2\log(4G/\delta)}{\min_g n_g^{\mathrm{tr}}}} + 2M\sqrt{\frac{2\log(4G/\delta)}{\min_g n_g^{\mathrm{val}}}} \\
&+ \min_{\eta\in H}\left(\Omega_\eta(W^*) + \beta_{DRO} + \Sigma\sqrt{\frac{\log(4G/\delta)}{2}}\right) \\
&+ \frac{24M}{\sqrt{\min_g n_g^{\mathrm{val}}}}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}}RL}{M}\right)} + 2\sqrt{\log 2}\right)
\end{aligned}$$

where $\rho_{\mathcal{A}} \le \frac{GLM}{\lambda\eta_{\min}^2}$ is the Lipschitz constant of the DRO algorithmic mapping, and $\Sigma^2 = 4n^{\mathrm{tr}}\beta_{DRO}^2 + 8M\beta_{DRO} + \frac{4M^2}{\min_g n_g^{\mathrm{tr}}}$ bounds the stability variance.

Proof. The proof mirrors the three-step structure of Theorem F.3, isolating the exact points where worst-case group bounds modify the complexities.

Step 1: Lower-Level Uniform Stability (via Union Bound). By Lemma F.4, the lower-level Group DRO objective is uniformly stable with the modified constant $\beta_{\text{DRO}} = \frac{2L^2}{\lambda \min_g n_g^{\text{tr}}}\left(1 + \frac{M\sqrt{G}}{\eta_{\min}}\right)$. Because the empirical worst-group risk is a maximum over empirical averages, its expected generalization gap is not bounded directly by β_{DRO} due to Jensen's inequality ($\max_g \mathbb{E}[\cdot] \leq \mathbb{E}[\max_g \cdot]$). Instead, we decouple the maximum operator. By the subadditivity of the maximum, the worst-group generalization gap is bounded by the maximum of the individual group generalization gaps:

$$L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_\eta) - L_{\text{worst}}^{\text{tr}}(\widehat{W}_\eta) = \max_g L_{\mathcal{D},g}(\widehat{W}_\eta) - \max_g L_g^{\text{tr}}(\widehat{W}_\eta) \leq \max_{g\in[G]} \underbrace{\left(L_{\mathcal{D},g}(\widehat{W}_\eta) - L_g^{\text{tr}}(\widehat{W}_\eta)\right)}_{:=Z_g(S)}$$

For any specific group g, the standard expected-loss stability result applies perfectly: $\mathbb{E}_S[Z_g(S)] \leq \beta_{\text{DRO}}$. To bound the maximum over all groups with high probability, we rigorously evaluate the sensitivity of $Z_g(S)$ to a single point perturbation. Suppose we perturb exactly one training example $z_i \to z'$ to form $S^{(i)}$. Let $\widehat{W}$ and $\widehat{W}^{(i)}$ be the optimal predictors for S and $S^{(i)}$, respectively. By the triangle inequality, the sensitivity of the group generalization gap is bounded by:

$$|Z_g(S) - Z_g(S^{(i)})| \leq \underbrace{\left|L_{\mathcal{D},g}(\widehat{W}) - L_{\mathcal{D},g}(\widehat{W}^{(i)})\right|}_{\text{True risk sensitivity}} + \underbrace{\left|L_g^{\text{tr}}(\widehat{W}) - L_g^{\text{tr},(i)}(\widehat{W}^{(i)})\right|}_{\text{Empirical risk sensitivity}}$$

For the true risk sensitivity, the uniform stability property guarantees that the absolute loss difference on any arbitrary point z is deterministically bounded by β_{DRO}. Therefore, its expectation over the target group distribution $z \sim \mathcal{D}_g$ is identically bounded: $\left|L_{\mathcal{D},g}(\widehat{W}) - L_{\mathcal{D},g}(\widehat{W}^{(i)})\right| \leq \mathbb{E}_{z\sim\mathcal{D}_g}[|\ell(\widehat{W}, z) - \ell(\widehat{W}^{(i)}, z)|] \leq \beta_{\text{DRO}}$. For the empirical risk sensitivity, the bound depends on whether the perturbed index i belongs to group g ($i \in G_g$):

1. **Case 1 ($i \notin G_g$):** The subset of points belonging to group g is identical between S and $S^{(i)}$. The empirical risk shifts solely due to the change in the algorithmic output $\widehat{W}$. Averaging the uniform stability bound over these n_g^{tr} unperturbed points yields an empirical shift of exactly β_{DRO}.
2. **Case 2 ($i \in G_g$):** Group g shares $n_g^{\text{tr}} - 1$ identical points between the two sets, but one point differs ($z_i \to z'$). For the identical points, the loss difference is bounded by β_{DRO}. For the single swapped point, the loss difference is naively bounded by $2M$ (Assumption 1). Averaging these yields:

$$\begin{aligned}
&\left|L_g^{\text{tr}}(\widehat{W}) - L_g^{\text{tr},(i)}(\widehat{W}^{(i)})\right| \\
&\leq \frac{1}{n_g^{\text{tr}}} \sum_{j\in G_g, j\neq i} \underbrace{|\ell(\widehat{W}, z_j) - \ell(\widehat{W}^{(i)}, z_j)|}_{\leq \beta_{\text{DRO}}} + \frac{1}{n_g^{\text{tr}}} \underbrace{|\ell(\widehat{W}, z_i) - \ell(\widehat{W}^{(i)}, z')|}_{\leq 2M} \\
&\leq \frac{n_g^{\text{tr}} - 1}{n_g^{\text{tr}}}\beta_{\text{DRO}} + \frac{2M}{n_g^{\text{tr}}} \leq \beta_{\text{DRO}} + \frac{2M}{n_g^{\text{tr}}}
\end{aligned}$$

Summing the true and empirical sensitivities, the bounded difference constants c_i for $Z_g(S)$ satisfy $c_i \leq 2\beta_{\text{DRO}} + \frac{2M}{n_g^{\text{tr}}}$ for $i \in G_g$, and $c_i \leq 2\beta_{\text{DRO}}$ for $i \notin G_g$. The sum of squared differences over all n^{tr} independent examples is bounded by:

$$\begin{aligned}
\sum_{i=1}^{n^{\text{tr}}} c_i^2 &= \sum_{i\notin G_g} (2\beta_{\text{DRO}})^2 + \sum_{i\in G_g} \left(2\beta_{\text{DRO}} + \frac{2M}{n_g^{\text{tr}}}\right)^2 \\
&= n^{\text{tr}}(2\beta_{\text{DRO}})^2 + n_g^{\text{tr}} \cdot 2(2\beta_{\text{DRO}})\frac{2M}{n_g^{\text{tr}}} + n_g^{\text{tr}}\frac{4M^2}{(n_g^{\text{tr}})^2} \leq 4n^{\text{tr}}\beta_{\text{DRO}}^2 + 8M\beta_{\text{DRO}} + \frac{4M^2}{\min_g n_g^{\text{tr}}} \\
&:= \Sigma^2
\end{aligned}$$

Applying the one-sided McDiarmid's inequality (Lemma F.16) to bound the deviation of $Z_g(S)$ above its expectation, and noting that the expected generalization gap is bounded by uniform stability $\mathbb{E}_S[Z_g(S)] \leq \beta_{\text{DRO}}$, we obtain that with probability at least $1 - \frac{\delta}{4G}$:

$$Z_g(S) \leq \beta_{\text{DRO}} + \Sigma\sqrt{\frac{\log(4G/\delta)}{2}}$$

Applying a union bound over all G groups ensures that this bound holds simultaneously for the maximum $\max_g Z_g(S)$ with total failure probability $\delta/4$:

$$L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_\eta) \leq L_{\text{worst}}^{\text{tr}}(\widehat{W}_\eta) + \beta_{\text{DRO}} + \Sigma\sqrt{\frac{\log(4G/\delta)}{2}} \tag{38}$$

Step 2: Upper-Level Uniform Convergence Decomposition. The upper-level empirical objective is the worst-group validation loss $f_{\text{val}}(W) = \max_{g\in[G]} L_g^{\text{val}}(W)$, and the target population risk is $L_{\mathcal{D}}^{\text{worst}}(W) = \max_{g\in[G]} L_{\mathcal{D},g}(W)$. By the non-expansive property of the maximum operator (Lemma F.19), the uniform deviation over the continuous path $\mathcal{H}_H$ cleanly decomposes. Furthermore, because the supremum and finite maximum operators commute, we can isolate the supremum to each individual group:

$$\begin{aligned}
\sup_{\eta\in H}\left|f_{\text{val}}(\widehat{W}_\eta) - L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_\eta)\right| &= \sup_{\eta\in H}\left|\max_{g\in[G]} L_g^{\text{val}}(\widehat{W}_\eta) - \max_{g\in[G]} L_{\mathcal{D},g}(\widehat{W}_\eta)\right| \\
&\leq \sup_{\eta\in H}\max_{g\in[G]}\left|L_g^{\text{val}}(\widehat{W}_\eta) - L_{\mathcal{D},g}(\widehat{W}_\eta)\right| \\
&= \max_{g\in[G]}\sup_{\eta\in H}\left|L_g^{\text{val}}(\widehat{W}_\eta) - L_{\mathcal{D},g}(\widehat{W}_\eta)\right|
\end{aligned}$$

This isolates the continuous uniform convergence problem entirely into G independent group-wise continuous uniform convergence bounds.

Step 3: Rademacher Complexity and Union Bound. For any individual group g, bounding the continuous 1-dimensional path $\mathcal{H}_H$ relies on the Rademacher complexity of the independent subset $S_{\text{val},g}$ of size n_g^{val} . Following exactly the uniform deviation derivation in Lemma F.2 (Equation (32) with $k = 1$), we evaluate the continuous uniform deviation bound for group g. To ensure this bound holds simultaneously across all G groups with a total failure probability of $\delta/2$, we apply a discrete union bound allocating confidence $\delta/(2G)$ to each group. As a result, for all $g \in [G]$ with probability $1 - \delta/2$:

$$\sup_{\eta\in H}\left|L_g^{\text{val}}(\widehat{W}_\eta) - L_{\mathcal{D},g}(\widehat{W}_\eta)\right| \leq \underbrace{\frac{12M}{\sqrt{n_g^{\text{val}}}}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}}RL}{M}\right)} + 2\sqrt{\log 2}\right) + M\sqrt{\frac{2\log(4G/\delta)}{n_g^{\text{val}}}}}_{:=\epsilon_{\text{val},g}}$$

Taking the maximum over all groups conservatively bounds the overall uniform deviation by the worst-case group size $\min_g n_g^{\text{val}}$:

$$\max_{g\in[G]}\sup_{\eta\in H}\left|L_g^{\text{val}}(\widehat{W}_\eta) - L_{\mathcal{D},g}(\widehat{W}_\eta)\right| \leq \max_{g\in[G]}\epsilon_{\text{val},g} := \epsilon_{\text{val}}^{\text{worst}}$$

Step 4: Final Decomposition. Let $\epsilon_{\text{train}}(\eta) = \beta_{\text{DRO}} + \Sigma\sqrt{\frac{\log(4/\delta)}{2}}$ denote the lower-level stability gap bound from Step 1. We define $\eta^* = \arg\min_{\eta\in H}(\Omega_\eta(W^*) + \epsilon_{\text{train}}(\eta))$ as the optimal hyperparameter for the reference predictor W^*. Because $\hat{\eta} = \arg\min_{\eta\in H} f_{\text{val}}(\widehat{W}_\eta)$ minimizes the empirical worst-group validation loss f_{val}, we have $f_{\text{val}}(\widehat{W}_{\hat{\eta}}) \leq f_{\text{val}}(\widehat{W}_{\eta^*})$. Recall that $f_{\text{val}}(W) = \max_g L_g^{\text{val}}(W)$. Applying the uniform deviation bound (which holds over all $\eta \in H$ with probability $1 - \delta/2$) to both $\hat{\eta}$ and η^* yields:

$$\begin{aligned}
L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_{\hat{\eta}}) &\leq f_{\text{val}}(\widehat{W}_{\hat{\eta}}) + \epsilon_{\text{val}}^{\text{worst}} \\
&\leq f_{\text{val}}(\widehat{W}_{\eta^*}) + \epsilon_{\text{val}}^{\text{worst}} \\
&\leq L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_{\eta^*}) + 2\epsilon_{\text{val}}^{\text{worst}}
\end{aligned}$$

To bound the target $L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_{\eta^*})$, we decompose the true risk via the empirical training risks. Crucially, while the algorithm optimizes the 3-term Group DRO objective $J_\eta(W)$, defined as:

$$J_\eta(W) = \max_{q \in \Delta_G} \left[\sum_{g=1}^{G} q_g L_g^{\text{tr}}(W) - \frac{\eta}{2} \left\| q - \frac{1}{G}\mathbf{1} \right\|^2 \right] + \frac{\lambda}{2}\|W\|^2$$

we can rigorously relate its minimizer back to the pure unregularized worst-group loss by bounding the $-\frac{\eta}{2}\|q - \frac{1}{G}\mathbf{1}\|^2$ penalty term. Let $\Omega_{\eta^*}(W^*) = \frac{\lambda}{2}\|W^*\|^2 + \frac{\eta^*}{2}$ encompass the deterministic penalties. By the optimality of $\widehat{W}_{\eta^*}$ on J_{η^*}, we have $J_{\eta^*}(\widehat{W}_{\eta^*}) \leq J_{\eta^*}(W^*)$.

For the left side, the maximum over $q \in \Delta_G$ is lower bounded by evaluating it at the specific one-hot vector $q = \mathbf{e}_k$ corresponding to the worst group $k = \arg\max_g L_g^{\text{tr}}(\widehat{W}_{\eta^*})$. The η-penalty for this one-hot vector evaluates to exactly $\frac{\eta^*}{2}\|\mathbf{e}_k - \frac{1}{G}\mathbf{1}\|^2 = \frac{\eta^*}{2}(1 - \frac{1}{G}) \leq \frac{\eta^*}{2}$. Thus, retaining the non-negative regularizer $\frac{\lambda}{2}\|\widehat{W}_{\eta^*}\|^2 \geq 0$, we have:

$$J_{\eta^*}(\widehat{W}_{\eta^*}) \geq L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) - \frac{\eta^*}{2} + \frac{\lambda}{2}\|\widehat{W}_{\eta^*}\|^2 \geq L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) - \frac{\eta^*}{2}$$

For the right side, because the η-penalty term $-\frac{\eta^*}{2}\|q - \frac{1}{G}\mathbf{1}\|^2$ is strictly non-positive for any q, we can trivially upper bound the inner maximum by dropping the penalty entirely:

$$J_{\eta^*}(W^*) \leq \max_{q \in \Delta_G} \left[\sum_g q_g L_g^{\text{tr}}(W^*) \right] + \frac{\lambda}{2}\|W^*\|^2 = L_{\text{worst}}^{\text{tr}}(W^*) + \frac{\lambda}{2}\|W^*\|^2$$

Chaining these two inequalities ($L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) - \frac{\eta^*}{2} \leq J_{\eta^*}(\widehat{W}_{\eta^*}) \leq J_{\eta^*}(W^*) \leq L_{\text{worst}}^{\text{tr}}(W^*) + \frac{\lambda}{2}\|W^*\|^2$) securely isolates the empirical risks:

$$L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) \leq L_{\text{worst}}^{\text{tr}}(W^*) + \Omega_{\eta^*}(W^*)$$

We can now algebraically inject this upper bound into the true risk:

$$\begin{aligned}
L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_{\eta^*}) &= L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) + \left(L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_{\eta^*}) - L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) \right) \\
&\leq L_{\text{worst}}^{\text{tr}}(W^*) + \Omega_{\eta^*}(W^*) + \underbrace{\left(L_{\mathcal{D}}^{\text{worst}}(\widehat{W}_{\eta^*}) - L_{\text{worst}}^{\text{tr}}(\widehat{W}_{\eta^*}) \right)}_{\text{Stability gap}} \\
&= L_{\mathcal{D}}^{\text{worst}}(W^*) + \Omega_{\eta^*}(W^*) + \text{Stability gap} + \underbrace{\left(L_{\text{worst}}^{\text{tr}}(W^*) - L_{\mathcal{D}}^{\text{worst}}(W^*) \right)}_{\text{Hoeffding gap}}
\end{aligned}$$

Simultaneously, since W^* is fixed independent of S_{train}, we bound the one-sided deviation of its pure unregularized worst-group training loss. For any group $g \in [G]$, the one-sided Hoeffding's inequality bounds the deviation $L_g^{\text{tr}}(W^*) - L_{\mathcal{D},g}(W^*)$. Applying a union bound over all G training groups with total confidence $\delta/4$ yields the Hoeffding gap:

$$L_{\text{worst}}^{\text{tr}}(W^*) - L_{\mathcal{D}}^{\text{worst}}(W^*) \leq \max_{g \in [G]} \left(L_g^{\text{tr}}(W^*) - L_{\mathcal{D},g}(W^*) \right) \leq M\sqrt{\frac{2\log(4G/\delta)}{\min_g n_g^{\text{tr}}}}$$

Combining the uniform convergence over $\mathcal{H}_H$ (probability $1-\delta/2$), the stability gap for η^* evaluated in Step 1 (probability $1-\delta/4$), and the Hoeffding bound for the fixed reference W^* (probability $1-\delta/4$), the total failure probability sums exactly to δ, yielding the final continuous oracle inequality. □

Corollary F.6 (Joint Hyperparameter Tuning of λ and η)**.** *Suppose the assumptions of Theorem F.5 hold. Let the joint hyperparameter vector be $\boldsymbol{\psi} = (\lambda, \eta) \in \Psi = \Lambda \times H \subset \mathbb{R}^2$, where $\Lambda = [\lambda_{\min}, \lambda_{\max}]$, $H = [\eta_{\min}, \eta_{\max}]$, and $R = diam(\Psi)$. Let $\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi} \in \Psi} f_{val}(\widehat{W}_{\boldsymbol{\psi}})$. The lower-level algorithmic mapping is jointly Lipschitz with respect to $\boldsymbol{\psi}$ with constant $\rho_{\mathcal{A}} \leq \frac{L}{\lambda_{\min}^2} + \frac{GLM}{\lambda_{\min}\eta_{\min}^2}$.*

With probability at least $1-\delta$*, the true worst-group risk of the jointly tuned predictor satisfies:*

$$
\begin{aligned}
&L_{\mathcal{D}}^{\textit{worst}}(\widehat{W}_{\hat{\boldsymbol{\psi}}}) \\
&\leq L_{\mathcal{D}}^{\textit{worst}}(W^*) + M\sqrt{\frac{2\log(4G/\delta)}{\min_g n_g^{\mathrm{tr}}}} + 2M\sqrt{\frac{2\log(4G/\delta)}{\min_g n_g^{\mathrm{val}}}} \\
&+ \min_{\lambda\in\Lambda,\eta\in H}\left(\Omega_{\lambda,\eta}(W^*) + \beta_{\textit{DRO}}(\lambda,\eta) + \Sigma(\lambda,\eta)\sqrt{\frac{\log(4G/\delta)}{2}}\right) \\
&+ \frac{24M}{\sqrt{\min_g n_g^{\mathrm{val}}}}\sqrt{2}\left(\sqrt{\log\left(\frac{3\rho_{\mathcal{A}}RL}{M}\right)} + 2\sqrt{\log 2}\right)
\end{aligned}
$$

where $\Omega_{\lambda,\eta}(W^*) = \frac{\lambda}{2}\|W^*\|^2 + \frac{\eta}{2}$*, and* $\beta_{\textit{DRO}}(\lambda,\eta) = \frac{2L^2}{\lambda \min_g n_g^{\mathrm{tr}}}\left(1 + \frac{M\sqrt{G}}{\eta}\right)$.

Proof of Corollary F.6. To establish the joint continuous uniform convergence bound, we must prove the mapping $\boldsymbol{\psi} \mapsto \widehat{W}_{\boldsymbol{\psi}}$ is Lipschitz over Ψ. We directly apply the exact strong monotonicity argument used in Lemma F.14. Let $J(W;\lambda,\eta)$ denote the lower-level Group DRO objective. For two hyperparameter configurations $\boldsymbol{\psi} = (\lambda,\eta)$ and $\boldsymbol{\psi}' = (\lambda',\eta')$, let $\widehat{W}$ and $\widehat{W}'$ be their respective minimizers.

Because $J(W;\lambda,\eta)$ is λ-strongly convex with respect to W, its gradient is λ-strongly monotone. Evaluating at the two optima and exploiting the first-order optimality condition $\nabla J(\widehat{W};\lambda,\eta) = \nabla J(\widehat{W}';\lambda',\eta') = 0$, we bound the shift in the predictors by the shift in the gradients evaluated at the fixed point $\widehat{W}'$:

$$
\begin{aligned}
\lambda\|\widehat{W} - \widehat{W}'\|^2 &\leq \langle \nabla J(\widehat{W}';\lambda,\eta) - \nabla J(\widehat{W};\lambda,\eta), \widehat{W}' - \widehat{W}\rangle \\
&\leq \langle \nabla J(\widehat{W}';\lambda,\eta) - \nabla J(\widehat{W}';\lambda',\eta'), \widehat{W}' - \widehat{W}\rangle \\
&\leq \|\nabla J(\widehat{W}';\lambda,\eta) - \nabla J(\widehat{W}';\lambda',\eta')\|\|\widehat{W} - \widehat{W}'\|
\end{aligned}
$$

Dividing by $\lambda\|\widehat{W} - \widehat{W}'\|$ isolates the deviation. The exact gradient of the objective is $\nabla J(W;\lambda,\eta) = \sum q_g \nabla L_g^{\mathrm{tr}}(W) + \lambda W$. Thus, the gradient shift decomposes into a regularization shift and a robust weight shift:

$$
\|\nabla J(\widehat{W}';\lambda,\eta) - \nabla J(\widehat{W}';\lambda',\eta')\| \leq |\lambda-\lambda'|\|\widehat{W}'\| + \left\|\sum_{g=1}^{G}(q_g - q_g')\nabla L_g^{\mathrm{tr}}(\widehat{W}')\right\|
$$

For the first term, the optimality condition for $\widehat{W}'$ implies $\lambda'\widehat{W}' = -\sum q_g' \nabla L_g^{\mathrm{tr}}(\widehat{W}')$. Since $\|\nabla L_g^{\mathrm{tr}}\| \leq L$, we have $\|\widehat{W}'\| \leq \frac{L}{\lambda'} \leq \frac{L}{\lambda_{\min}}$. For the second term, following the weight perturbation derivation in Lemma F.4, the optimal simplex weights shift by at most $\|q - q'\|_2 \leq \left\|\left(\frac{1}{\eta} - \frac{1}{\eta'}\right) L^{\mathrm{tr}}(\widehat{W}')\right\|_2$. Since $|L_g^{\mathrm{tr}}| \leq M$, this is bounded by $\frac{|\eta-\eta'|}{\eta_{\min}^2}\sqrt{G}M$. Multiplying by the Jacobian norm ($\sqrt{G}L$) bounds the robust weight shift by $\frac{GML}{\eta_{\min}^2}|\eta - \eta'|$.

Combining these and dividing by $\lambda \geq \lambda_{\min}$ yields the final perturbation bound:

$$
\|\widehat{W} - \widehat{W}'\| \leq \frac{L}{\lambda_{\min}^2}|\lambda-\lambda'| + \frac{GLM}{\lambda_{\min}\eta_{\min}^2}|\eta-\eta'| \leq \left(\frac{L}{\lambda_{\min}^2} + \frac{GLM}{\lambda_{\min}\eta_{\min}^2}\right)\|\boldsymbol{\psi} - \boldsymbol{\psi}'\|_2
$$

This establishes the joint Lipschitz constant $\rho_{\mathcal{A}}$ over Ψ. Injecting this $\rho_{\mathcal{A}}$ into a 2-dimensional variant of Lemma F.2 yields the $\sqrt{2}$ dimension scaling on the validation uniform deviation. The minimization trades off this against the local stability gap $\beta_{\mathrm{DRO}}(\lambda,\eta)$, completing the proof. □

Corollary F.7 (Joint Tuning with Deep Encoder Parameters)**.** *Suppose the assumptions of Corollary F.6 hold. Further assume that the input features are produced by a deep encoder* $z = f_\theta(x)$ *parameterized by* $\theta \in \Theta$*, where* $\Theta \subset \mathbb{R}^{d_\theta}$*. Assume (1) the encoder output* $f_\theta(x)$ *is* L_f*-Lipschitz*

with respect to θ, and (2) the classification loss gradient $\nabla_W \ell(W, z)$ is L_{grad}-Lipschitz with respect to z. Let the extended joint hyperparameter vector be $\boldsymbol{\psi} = (\lambda, \eta, \theta) \in \Psi \times \Theta$.

The lower-level algorithmic mapping is jointly Lipschitz with respect to $\boldsymbol{\psi}$ with constant:

$$\rho_{\mathcal{A}} \leq \frac{L}{\lambda_{\min}^2} + \frac{GLM}{\lambda_{\min}\eta_{\min}^2} + \frac{L_{grad} L_f}{\lambda_{\min}} \tag{39}$$

Crucially, the lower-level uniform stability constant $\beta_{DRO}(\lambda, \eta)$ remains unchanged, as θ is fixed during the lower-level optimization. Consequently, the true worst-group risk bound takes the identical structural form as Corollary F.6, with the dimension factor k increasing to $2 + d_\theta$ and the covering radius scaling to diam$(\Psi \times \Theta)$.

Proof. To derive the joint algorithmic Lipschitz constant, we follow the exact strong monotonicity argument used in Corollary F.6. Let $J(W; \lambda, \eta, \theta) = \sum q_g(\eta) L_g^{\text{tr}}(W, \theta) + \frac{\lambda}{2}\|W\|^2$ denote the lower-level objective. For any two joint configurations $\boldsymbol{\psi} = (\lambda, \eta, \theta)$ and $\boldsymbol{\psi}' = (\lambda', \eta', \theta')$, the optimal predictors shift according to the total gradient shift at the fixed point $\widehat{W}'$:

$$\lambda \|\widehat{W} - \widehat{W}'\| \leq \|\nabla J(\widehat{W}'; \lambda, \eta, \theta) - \nabla J(\widehat{W}'; \lambda', \eta', \theta')\|$$

By the triangle inequality, this gradient shift decomposes into three distinct perturbations corresponding to the regularization, the group weights, and the encoder representations:

$$\|\nabla J(\widehat{W}'; \lambda, \eta, \theta) - \nabla J(\widehat{W}'; \lambda', \eta', \theta')\| \leq \underbrace{|\lambda - \lambda'| \|\widehat{W}'\|}_{\leq \frac{L}{\lambda_{\min}}|\lambda - \lambda'|} + \underbrace{\left\| \sum_{g=1}^{G} (q_g - q_g') \nabla_W L_g^{\text{tr}}(\widehat{W}', \theta) \right\|}_{\leq \frac{GLM}{\eta_{\min}^2}|\eta - \eta'|}$$
$$+ \underbrace{\left\| \sum_{g=1}^{G} q_g' \left(\nabla_W L_g^{\text{tr}}(\widehat{W}', \theta) - \nabla_W L_g^{\text{tr}}(\widehat{W}', \theta') \right) \right\|}_{\text{Encoder shift}}$$

The first two bounds follow identically from Corollary F.6. For the third term, because the simplex weights satisfy $\sum q_g' = 1$, the encoder shift is bounded by the maximum gradient deviation across groups. Applying the smoothness of the loss and the Lipschitz property of the encoder, this shift evaluates to:

$$\max_g \|\nabla_W L_g^{\text{tr}}(\widehat{W}', \theta) - \nabla_W L_g^{\text{tr}}(\widehat{W}', \theta')\| \leq L_{\text{grad}} L_f \|\theta - \theta'\|$$

Combining these three bounds and dividing by $\lambda \geq \lambda_{\min}$ yields the final mapping deviation:

$$\|\widehat{W} - \widehat{W}'\| \leq \left(\frac{L}{\lambda_{\min}^2} + \frac{GLM}{\lambda_{\min}\eta_{\min}^2} + \frac{L_{\text{grad}} L_f}{\lambda_{\min}} \right) \|\boldsymbol{\psi} - \boldsymbol{\psi}'\|_2$$

For the stability term, uniform stability measures the sensitivity of the learning algorithm to a single training point perturbation for a *fixed* hyperparameter configuration. Because θ is an upper-level variable, it acts as a constant mapping $x \mapsto z$ during the lower-level optimization. Provided the loss $\ell(W, z)$ is L-Lipschitz over the bounded representation space, the stability constant β_{DRO} relies solely on the loss properties and the fixed regularization λ, remaining identical to the linear case. □

Remark F.8 (Neural Tangent Kernel). While this generic continuous treatment seamlessly maintains the logical flow of our bilevel framework, the sample complexity bounds could be further tightened by incorporating specialized neural network generalization theories, such as the Neural Tangent Kernel (NTK) (Jacot et al., 2018).

F.6 Extension to Hierarchical DRO (Bi-HDRO)

The continuous generalization theory seamlessly extends to Hierarchical DRO (HDRO) where the continuous hyperparameters being tuned include the multi-dimensional inner perturbation radii $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_G)$. In this formulation, the joint hyperparameter is $\boldsymbol{\psi} = (\lambda, \eta, \boldsymbol{\epsilon})$. As long as the smoothed robust loss $\ell_{\text{rob},g}(W, \epsilon_g)$ is used, the lower-level mapping remains Lipschitz continuous.

Lemma F.9 (Algorithmic Lipschitz Continuity of Bi-HDRO)**.** *Assume the smoothed robust loss $\ell_{rob,g}(W, \epsilon_g) = \mathbb{E}[\sup_{\|\delta_g\| \le \epsilon_g} \ell(W, z+\delta_g, y)]$ has bounded gradient shifts with respect to ϵ_g, satisfying $\|\nabla_W \ell_{rob,g}(W, \epsilon_g^{(1)}) - \nabla_W \ell_{rob,g}(W, \epsilon_g^{(2)})\|_2 \le C_\epsilon |\epsilon_g^{(1)} - \epsilon_g^{(2)}|$, and is L_ϵ-Lipschitz with respect to ϵ_g. Let the lower-level objective maintain λ-strong convexity via L_2 regularization. Then, the HDRO algorithmic mapping $\widehat{W}_{\boldsymbol{\psi}}$ is jointly Lipschitz with resp ect to the combined hyperparameter $\boldsymbol{\psi} = (\lambda, \eta, \boldsymbol{\epsilon})$ with the joint Lipschitz constant bounded by:*

$$\rho_{\mathcal{A},\boldsymbol{\psi}} \le \underbrace{\frac{L}{\lambda_{\min}^2} + \frac{GLM}{\lambda_{\min}\eta_{\min}^2}}_{\textit{Shift from } \lambda,\eta} + \underbrace{\frac{C_\epsilon}{\lambda_{\min}} + \frac{\sqrt{G}LL_\epsilon}{\lambda_{\min}\eta_{\min}}}_{\textit{Shift from } \boldsymbol{\epsilon}} \tag{40}$$

Remark F.10 (Validity of the Lipschitz Assumption for Classification). The assumption that the robust loss has bounded gradient shifts with respect to ϵ_g naturally holds for the exact closed-form perturbations used in binary classification (derived in Appendix C). Analytically solving the inner adversarial maximization $\min_{\|\delta\| \le \epsilon} y(W^\top(z+\delta)+b)$ yields the robust margin $y(W^\top z+b) - \epsilon\|W\|_*$, where $\|W\|_*$ is the dual norm of the perturbation constraint.

Consider L_2 norm perturbations where the dual norm is simply $\|W\|_* = \|W\|_2$. For the globally smooth robust logistic/BCE loss $\ell_{\text{rob}}(W, \epsilon) = \log(1+\exp(A))$, where $A = -y(W^\top z+b)+\epsilon\|W\|_2$, the gradient with respect to W everywhere is $\nabla_W \ell_{\text{rob}}(W, \epsilon) = \sigma(A)\left(-yz + \epsilon\frac{W}{\|W\|_2}\right)$. Using the $1/4$-Lipschitz continuity of the sigmoid function $\sigma(\cdot)$, the gradient shift between two perturbation radii $\epsilon^{(1)}$ and $\epsilon^{(2)}$ is explicitly bounded by the triangle inequality:

$$\begin{aligned}
&\|(\sigma(A^{(1)}) - \sigma(A^{(2)}))(-yz) + (\sigma(A^{(1)})\epsilon^{(1)} - \sigma(A^{(2)})\epsilon^{(2)})\frac{W}{\|W\|_2}\|_2 \\
&\le \frac{1}{4}|A^{(1)} - A^{(2)}|\|z\|_2 + 1 \cdot |\epsilon^{(1)} - \epsilon^{(2)}| + \epsilon_{\max}\frac{1}{4}|A^{(1)} - A^{(2)}| \\
&= \left(1 + \frac{1}{4}\|W\|_2\|z\|_2 + \frac{\epsilon_{\max}}{4}\|W\|_2\right)|\epsilon^{(1)} - \epsilon^{(2)}|.
\end{aligned}$$

Because the lower-level objective enforces λ-strong convexity via $\frac{\lambda}{2}\|W\|_2^2$, the optimal weights $\|W\|_2$ are bounded by a constant M_w. Assuming bounded $\|z\|_2$, this ensures the gradient shift is strictly bounded by $C_\epsilon|\epsilon^{(1)} - \epsilon^{(2)}|$ where C_ϵ is a finite constant. Furthermore, this justifies the L_ϵ-Lipschitzness of the loss value itself: the derivative $\frac{\partial \ell_{\text{rob}}}{\partial \epsilon} = \sigma(A)\|W\|_2$ is strictly bounded by $\|W\|_2$, meaning the robust loss is L_ϵ-Lipschitz with $L_\epsilon \le M_w$. Thus, all Lipschitz continuity assumptions are rigorously satisfied globally in our implementation.

Proof. Let $F(W, q, \boldsymbol{\epsilon}) = \sum_{g=1}^G q_g \ell_{\text{rob},g}(W, \epsilon_g) - \frac{\eta}{2}\|q - \frac{1}{G}\mathbf{1}\|^2 + \frac{\lambda}{2}\|W\|^2$ be the lower-level HDRO objective. By Danskin's theorem, the gradient of the max-marginalized objective $J_{\boldsymbol{\epsilon}}(W)$ with respect to W is $\nabla_W J_{\boldsymbol{\epsilon}}(W) = \sum_{g=1}^G q_g^* \nabla_W \ell_{\text{rob},g}(W, \epsilon_g) + \lambda W$, where $q^* = \Pi_{\Delta_G}(\frac{1}{\eta}\boldsymbol{\ell}_{\text{rob}}(W, \boldsymbol{\epsilon}) + \frac{1}{G}\mathbf{1})$, with $\boldsymbol{\ell}_{\text{rob}}(W, \boldsymbol{\epsilon}) \in \mathbb{R}^G$ denoting the vector of robust losses across all groups.

If the perturbation hyperparameter shifts from $\boldsymbol{\epsilon}^{(1)}$ to $\boldsymbol{\epsilon}^{(2)}$, the gradient shift is bounded by the triangle inequality:

$$\begin{aligned}
&\|\nabla_W J_{\boldsymbol{\epsilon}^{(1)}}(W) - \nabla_W J_{\boldsymbol{\epsilon}^{(2)}}(W)\|_2 \\
&\le \left\|\sum_{g=1}^G q_g^{(1)}\left(\nabla_W \ell_{\text{rob},g}^{(1)} - \nabla_W \ell_{\text{rob},g}^{(2)}\right)\right\|_2 + \left\|\sum_{g=1}^G \left(q_g^{(1)} - q_g^{(2)}\right)\nabla_W \ell_{\text{rob},g}^{(2)}\right\|_2 \\
&\le \sum_{g=1}^G q_g^{(1)} C_\epsilon |\epsilon_g^{(1)} - \epsilon_g^{(2)}| + \sum_{g=1}^G |q_g^{(1)} - q_g^{(2)}| \underbrace{\|\nabla_W \ell_{\text{rob},g}^{(2)}\|_2}_{\le L} \\
&\le C_\epsilon \|\boldsymbol{\epsilon}^{(1)} - \boldsymbol{\epsilon}^{(2)}\|_2 + L\|q^{(1)} - q^{(2)}\|_1 \\
&\le C_\epsilon \|\boldsymbol{\epsilon}^{(1)} - \boldsymbol{\epsilon}^{(2)}\|_2 + L\sqrt{G}\|q^{(1)} - q^{(2)}\|_2
\end{aligned}$$

Because the simplex projection Π_{Δ_G} is 1-Lipschitz, the shift in the adversarial weights is strictly bounded by the shift in the robust loss terms scaled by the fixed penalty $\eta \geq \eta_{\min}$:

$$\|q^{(1)} - q^{(2)}\|_2 \leq \frac{1}{\eta_{\min}} \|\boldsymbol{\ell}_{\text{rob}}(W, \boldsymbol{\epsilon}^{(1)}) - \boldsymbol{\ell}_{\text{rob}}(W, \boldsymbol{\epsilon}^{(2)})\|_2 \leq \frac{L_\epsilon}{\eta_{\min}} \|\boldsymbol{\epsilon}^{(1)} - \boldsymbol{\epsilon}^{(2)}\|_2$$

Substituting this bound into the gradient shift yields a total shift bounded by $\left(C_\epsilon + \frac{\sqrt{G} L L_\epsilon}{\eta_{\min}}\right) \|\boldsymbol{\epsilon}^{(1)} - \boldsymbol{\epsilon}^{(2)}\|_2$. Because the objective is λ-strongly convex, applying the exact same strong monotonicity argument from Lemma F.1 divides this gradient shift by λ, proving the mapping is Lipschitz continuous with respect to $\boldsymbol{\epsilon}$. Summing this $\boldsymbol{\epsilon}$-specific constant with the Lipschitz bounds for λ and η derived in Corollary 3.8 establishes the combined joint Lipschitz constant $\rho_{\mathcal{A},\boldsymbol{\psi}}$ over the entire hyperparameter space. □

Remark F.11 (Uniform Stability of Bi-HDRO). While tuning $\boldsymbol{\epsilon}$ expands the algorithmic Lipschitz constant, the uniform stability of the lower-level algorithm remains unchanged. For a fixed configuration, Bi-HDRO optimizes the robust loss $\ell_{\text{rob},g}(W) = \sup_{\|\delta\| \leq \epsilon} \ell(W, z + \delta)$. Because the global constants L and M bound the base loss across all possible inputs, they naturally bound any perturbed input $z + \delta$. Thus, Bi-HDRO inherits the exact same stability constant $\beta_{\text{DRO}}(\lambda, \eta) = \frac{2L^2}{\lambda \min_g n_g^{\text{tr}}} \left(1 + \frac{M\sqrt{G}}{\eta}\right)$ as standard group DRO.

Corollary F.12 (Joint Tuning of Bi-HDRO with Deep Encoder Parameters)**.** *Suppose the assumptions of Lemma F.9 hold. Further assume the input features are generated by a deep encoder $z = f_\theta(x)$ parameterized by $\theta \in \Theta$, such that the encoder output is L_f-Lipschitz with respect to θ, and the gradient of the robust loss $\nabla_W \ell_{rob,g}(W, \theta, \epsilon_g)$ is L_{grad}^{rob}-Lipschitz with respect to the representation z. Let the fully extended joint hyperparameter vector be $\boldsymbol{\psi} = (\lambda, \eta, \boldsymbol{\epsilon}, \theta) \in \Psi \times \Theta$.*

The Bi-HDRO algorithmic mapping is jointly Lipschitz with respect to $\boldsymbol{\psi}$ with constant:

$$\rho_{\mathcal{A},\boldsymbol{\psi}} \leq \underbrace{\frac{L}{\lambda_{\min}^2} + \frac{GLM}{\lambda_{\min}\eta_{\min}^2}}_{\textit{Shift from } \lambda,\eta} + \underbrace{\frac{C_\epsilon}{\lambda_{\min}} + \frac{\sqrt{G} L L_\epsilon}{\lambda_{\min}\eta_{\min}}}_{\textit{Shift from } \boldsymbol{\epsilon}} + \underbrace{\frac{L_{grad}^{rob} L_f}{\lambda_{\min}}}_{\textit{Shift from } \theta} \tag{41}$$

Moreover, as established in the preceding remark, the uniform stability constant $\beta_{HDRO}(\lambda, \eta)$ remains identical.

Proof. The proof follows immediately by combining the derivation of Lemma F.9 with the triangle inequality decomposition established in Corollary F.7. The gradient shift now contains a fourth additive term arising from the variation in θ, which evaluates to $\max_g \|\nabla_W \ell_{\text{rob},g}(\widehat{W}', \theta, \epsilon_g) - \nabla_W \ell_{\text{rob},g}(\widehat{W}', \theta', \epsilon_g)\| \leq L_{\text{grad}}^{\text{rob}} L_f \|\theta - \theta'\|$. Dividing by the strong convexity constant $\lambda_{\min}$ yields the additive θ-shift term. □

F.7 Background: Uniform Stability

To establish generalization guarantees, we rely on the framework of uniform stability. We first explicitly recall the uniform stability of the lower-level algorithm, adapting the standard analysis for Tikhonov regularization (Shalev-Shwartz & Ben-David, 2014, Section 13.3) to our general λ-strongly convex regularizer Ω_λ.

Definition F.13 (Uniform Stability (Bousquet & Elisseeff, 2002))**.** A learning algorithm $\mathcal{A}$ is β-uniformly stable with respect to a loss function ℓ if, for any two training sets $S, S^{(i)}$ of size m that differ by exactly one example, and for any arbitrary test point z, the following holds:

$$\sup_z |\ell(\mathcal{A}(S), z) - \ell(\mathcal{A}(S^{(i)}), z)| \leq \beta \tag{42}$$

Lemma F.14 (Uniform Stability of λ-Strongly Convex RLM)**.** *Assume the loss function $\ell(W, z)$ is convex and L-Lipschitz with respect to W. Let $\Omega_\lambda(W)$ be a λ-strongly convex regularization function. Then the Regularized Loss Minimization rule $\mathcal{A}(S) = \arg\min_W (L_S(W) + \Omega_\lambda(W))$ is β-uniformly stable with $\beta = \frac{2L^2}{\lambda m}$.*

Proof. Let $S = (z_1, \ldots, z_m)$ be a training set, z' an additional example, and $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$. Denote $f_S(W) = L_S(W) + \Omega_\lambda(W)$. Because f_S is λ-strongly convex, its gradient is λ-strongly monotone. Evaluating this for the optimal predictors $W = \mathcal{A}(S)$ and $W^{(i)} = \mathcal{A}(S^{(i)})$ yields:

$$\lambda \|W^{(i)} - W\|^2 \leq \langle \nabla f_S(W^{(i)}) - \nabla f_S(W), W^{(i)} - W \rangle$$

In view of the first-order optimality conditions of W and $W^{(i)}$, we have:

$$\langle -\nabla f_S(W), W^{(i)} - W \rangle \leq 0$$
$$\langle \nabla f_{S^{(i)}}(W^{(i)}), W^{(i)} - W \rangle \leq 0$$

Summing these three inequalities yields:

$$\lambda \|W^{(i)} - W\|^2 \leq \langle \nabla f_S(W^{(i)}) - \nabla f_{S^{(i)}}(W^{(i)}), W^{(i)} - W \rangle$$

Applying the Cauchy-Schwarz inequality, we can bound the distance strictly by the shift in the gradients:

$$\lambda \|W^{(i)} - W\| \leq \|\nabla f_S(W^{(i)}) - \nabla f_{S^{(i)}}(W^{(i)})\| \tag{43}$$

Expanding the empirical risk gradients, the shift is exactly:

$$\|\nabla f_S(W^{(i)}) - \nabla f_{S^{(i)}}(W^{(i)})\| = \left\| \frac{\nabla \ell(W^{(i)}, z_i) - \nabla \ell(W^{(i)}, z')}{m} \right\| \leq \frac{2L}{m}$$

Dividing by λ gives the optimal parameter displacement $\|W^{(i)} - W\| \leq \frac{2L}{\lambda m}$. Finally, the L-Lipschitzness of ℓ implies that for any test point z, the difference in loss is bounded by:

$$|\ell(\mathcal{A}(S^{(i)}), z) - \ell(\mathcal{A}(S), z)| \leq L \|\mathcal{A}(S^{(i)}) - \mathcal{A}(S)\| \leq \frac{2L^2}{\lambda m} \tag{44}$$

Thus, the learning rule is $\frac{2L^2}{\lambda m}$-uniformly stable. □

Lemma F.15 (High Probability Generalization via Stability)**.** *Let the learning algorithm $\mathcal{A}$ be β-uniformly stable, and assume the loss function ℓ is bounded by M. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of a training set S_{train} of size n^{tr}, the true risk of the output hypothesis is bounded by:*

$$L_{\mathcal{D}}(\mathcal{A}(S_{train})) \leq L^{\mathrm{tr}}(\mathcal{A}(S_{train})) + \beta + \left(2\beta + \frac{2M}{n^{\mathrm{tr}}}\right) \sqrt{\frac{n^{\mathrm{tr}} \log(1/\delta)}{2}} \tag{45}$$

For the λ-strongly convex Regularized Loss Minimization rule defined in Lemma F.14, we substitute $\beta = \frac{2L^2}{\lambda n^{\mathrm{tr}}}$.

Proof. Let $f(S_{\text{train}}) = L_{\mathcal{D}}(\mathcal{A}(S_{\text{train}})) - L^{\mathrm{tr}}(\mathcal{A}(S_{\text{train}}))$ denote the generalization gap. A fundamental result in stability theory (Shalev-Shwartz & Ben-David, 2014, Section 13.2) guarantees that the expected generalization gap is bounded by the uniform stability: $\mathbb{E}_{S_{\text{train}}}[f(S_{\text{train}})] \leq \beta$.

To obtain a high-probability bound, we analyze the sensitivity of $f(S_{\text{train}})$ to the replacement of a single training example. Let S_{train} and $S^{(i)}_{\text{train}}$ be two training sets differing by exactly one example $z_i \to z'$. By definition of β-uniform stability, the loss on any arbitrary point z changes by at most β:

$$\sup_z |\ell(\mathcal{A}(S_{\text{train}}), z) - \ell(\mathcal{A}(S^{(i)}_{\text{train}}), z)| \leq \beta \tag{46}$$

Taking the expectation over $z \sim \mathcal{D}$, the difference in true risk is bounded by this uniform difference:

$$|L_{\mathcal{D}}(\mathcal{A}(S_{\text{train}})) - L_{\mathcal{D}}(\mathcal{A}(S^{(i)}_{\text{train}}))| \leq \mathbb{E}_{z \sim \mathcal{D}} \left[\sup_{z'} |\ell(\mathcal{A}(S_{\text{train}}), z') - \ell(\mathcal{A}(S^{(i)}_{\text{train}}), z')| \right] \leq \beta \tag{47}$$

Furthermore, we can bound the change in the empirical risk between the two sets. Noting that S_{train} and $S^{(i)}_{\text{train}}$ share $n^{\text{tr}} - 1$ identical points and only differ at the i-th point (z_i vs z'), we have:

$$\begin{aligned}
|L^{\text{tr}}(\mathcal{A}(S_{\text{train}})) - L^{\text{tr},(i)}(\mathcal{A}(S^{(i)}_{\text{train}}))| &= \left| \frac{1}{n^{\text{tr}}} \sum_{j=1}^{n^{\text{tr}}} \left(\ell(\mathcal{A}(S_{\text{train}}), z_j) - \ell(\mathcal{A}(S^{(i)}_{\text{train}}), z^{(i)}_j) \right) \right| \\
&\leq \frac{1}{n^{\text{tr}}} \sum_{j \neq i} \underbrace{|\ell(\mathcal{A}(S_{\text{train}}), z_j) - \ell(\mathcal{A}(S^{(i)}_{\text{train}}), z_j)|}_{\leq \beta \text{ (uniform stability)}} \\
&\quad + \frac{1}{n^{\text{tr}}} \underbrace{|\ell(\mathcal{A}(S_{\text{train}}), z_i) - \ell(\mathcal{A}(S^{(i)}_{\text{train}}), z')|}_{\leq 2M \text{ (bounded loss)}} \\
&\leq \frac{n^{\text{tr}} - 1}{n^{\text{tr}}} \beta + \frac{2M}{n^{\text{tr}}} \leq \beta + \frac{2M}{n^{\text{tr}}}
\end{aligned}$$

Consequently, the change in the function f when one point is perturbed is bounded by:

$$\begin{aligned}
|f(S_{\text{train}}) - f(S^{(i)}_{\text{train}})| &\leq |L_{\mathcal{D}}(\mathcal{A}(S_{\text{train}})) - L_{\mathcal{D}}(\mathcal{A}(S^{(i)}_{\text{train}}))| + |L^{\text{tr}}(\mathcal{A}(S_{\text{train}})) - L^{\text{tr},(i)}(\mathcal{A}(S^{(i)}_{\text{train}}))| \\
&\leq \beta + \left(\beta + \frac{2M}{n^{\text{tr}}} \right) = 2\beta + \frac{2M}{n^{\text{tr}}}
\end{aligned}$$

Thus, $f(S_{\text{train}})$ satisfies the bounded differences property with constant $c = 2\beta + \frac{2M}{n^{\text{tr}}}$. Applying the one-sided McDiarmid's inequality (Lemma F.16), we have that with probability at least $1 - \delta$:

$$f(S_{\text{train}}) \leq \mathbb{E}[f(S_{\text{train}})] + c\sqrt{\frac{n^{\text{tr}} \log(1/\delta)}{2}} \leq \beta + \left(2\beta + \frac{2M}{n^{\text{tr}}} \right) \sqrt{\frac{n^{\text{tr}} \log(1/\delta)}{2}} \tag{48}$$

which yields the final result. □

F.8 Helpful Lemmas

For completeness, we include the explicit derivations for properties utilized in the main theorems.

Lemma F.16 (McDiarmid's Inequality)**.** *Let $X_1, \ldots, X_m$ be independent random variables, and let $f(X_1, \ldots, X_m)$ be a function that satisfies the bounded differences property with constants $c_1, \ldots, c_m$:*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x'_i, \ldots, x_m)| \leq c_i \tag{49}$$

Then for any $\epsilon > 0$, the one-sided deviation is bounded by:

$$P(f(X_1, \ldots, X_m) - \mathbb{E}[f] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2} \right) \tag{50}$$

By symmetry, the two-sided absolute deviation is bounded by:

$$P(|f(X_1, \ldots, X_m) - \mathbb{E}[f]| \geq \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2} \right) \tag{51}$$

Lemma F.17 (Sub-Gaussian Variance Proxy)**.** *Let $X_1, \ldots, X_m$ be independent random variables, and let $f(X_1, \ldots, X_m)$ be a function that satisfies the bounded differences property with constants $c_1, \ldots, c_m$:*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x'_i, \ldots, x_m)| \leq c_i \tag{52}$$

Then the random variable $Z = f(X_1, \ldots, X_m)$ is a sub-Gaussian random variable with variance proxy $\sigma^2 = \frac{1}{4} \sum_{i=1}^m c_i^2$.

Proof. By the one-sided McDiarmid's inequality (Lemma F.16), for any $t \geq 0$, the probability of deviation from the expected value is bounded by:

$$P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2} \right) \tag{53}$$

A random variable Z is formally defined as sub-Gaussian with variance proxy σ^2 if its tail distribution satisfies $P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$. By equating the exponents of the bounds, we have:

$$\frac{t^2}{2\sigma^2} = \frac{2t^2}{\sum_{i=1}^m c_i^2} \implies 2\sigma^2 = \frac{1}{2}\sum_{i=1}^m c_i^2 \implies \sigma^2 = \frac{1}{4}\sum_{i=1}^m c_i^2$$

□

Lemma F.18 (Maximal Inequality for Sub-Gaussian Random Variables)**.** *Let $Z_1, \ldots, Z_n$ be a finite collection of sub-Gaussian random variables, where each Z_i has variance proxy σ^2 and expectation $\mu_i = \mathbb{E}[Z_i]$. The expected maximum of these random variables is bounded by:*

$$\mathbb{E}\left[\max_{1\leq i\leq n} Z_i\right] \leq \max_{1\leq i\leq n} \mu_i + \sigma\sqrt{2\log n} \tag{54}$$

Proof. Let $Y_i = Z_i - \mu_i$. By definition, each Y_i is a zero-mean sub-Gaussian random variable with variance proxy σ^2, satisfying the moment generating function bound $\mathbb{E}[\exp(sY_i)] \leq \exp\left(\frac{s^2\sigma^2}{2}\right)$ for any $s > 0$. We wish to bound $\mathbb{E}[\max_i Y_i]$.

By Jensen's inequality, since the exponential function is strictly convex for $s > 0$:

$$\begin{aligned}
\exp\left(s\mathbb{E}\left[\max_{1\leq i\leq n} Y_i\right]\right) &\leq \mathbb{E}\left[\exp\left(s\max_{1\leq i\leq n} Y_i\right)\right] \\
&= \mathbb{E}\left[\max_{1\leq i\leq n} \exp(sY_i)\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^n \exp(sY_i)\right] \\
&= \sum_{i=1}^n \mathbb{E}[\exp(sY_i)] \leq n\exp\left(\frac{s^2\sigma^2}{2}\right)
\end{aligned}$$

Taking the natural logarithm of both sides and dividing by s yields:

$$\mathbb{E}\left[\max_{1\leq i\leq n} Y_i\right] \leq \frac{\log n}{s} + \frac{s\sigma^2}{2} \tag{55}$$

To minimize this upper bound, we select the optimal parameter $s = \sqrt{2\log n/\sigma^2}$. Substituting this into the inequality gives:

$$\mathbb{E}\left[\max_{1\leq i\leq n} Y_i\right] \leq \frac{\log n}{\sqrt{2\log n}/\sigma} + \frac{\sigma^2\sqrt{2\log n/\sigma^2}}{2} = \sigma\sqrt{\frac{\log n}{2}} + \sigma\sqrt{\frac{\log n}{2}} = \sigma\sqrt{2\log n} \tag{56}$$

Finally, because $\max_i Z_i \leq \max_i \mu_i + \max_i Y_i$, applying the expectation yields $\mathbb{E}[\max_i Z_i] \leq \max_i \mu_i + \mathbb{E}[\max_i Y_i] \leq \max_i \mu_i + \sigma\sqrt{2\log n}$. □

Lemma F.19 (Non-Expansive Property of the Maximum Operator)**.** *For any two finite sets of real numbers $A = \{A_1, \ldots, A_G\}$ and $B = \{B_1, \ldots, B_G\}$, the absolute difference between their maximums is bounded by the maximum of their element-wise absolute differences:*

$$\left|\max_{g\in[G]} A_g - \max_{g\in[G]} B_g\right| \leq \max_{g\in[G]} |A_g - B_g| \tag{57}$$

Proof. Without loss of generality, assume that $\max_g A_g \geq \max_g B_g$. Let $k = \arg\max_g A_g$ be the index that achieves the maximum for A. We can then write the difference as:

$$\max_g A_g - \max_g B_g = A_k - \max_g B_g$$

Since the maximum of the set B must be at least as large as any specific element in B, we know that $\max_g B_g \geq B_k$. Substituting this lower bound can only increase the difference:

$$A_k - \max_g B_g \leq A_k - B_k$$

Since a quantity is always bounded by its absolute value, and the k-th element's difference is bounded by the maximum absolute difference across all elements, we have:

$$A_k - B_k \leq |A_k - B_k| \leq \max_g |A_g - B_g|$$

This establishes the bound, completing the proof. □