OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Per-Group Distributionally Robust Optimization (Per-GDRO) with Learnable Ambiguity Set Sizes via Bilevel Optimization

Seobeom Jung[1] NASNAGA1541@SKKU.EDU
Woojae Lee[1] DLDNWO@SKKU.EDU
Jihun Hamm[3] JHAMM3@TULANE.EDU
Jangho Park[1,2*] JANGHOPARK@SKKU.EDU

[1]*Department of Systems Management Engineering, Sungkyunkwan University, Suwon, Republic of Korea*
[2]*Department of Industrial Engineering, Sungkyunkwan University, Suwon, Republic of Korea*
[3]*Department of Computer Science, Tulane University, New Orleans, LA 70118, United States*
[*]*Corresponding author*

Abstract

Group structures frequently influence model behavior; however, group membership is often unobserved during inference, limiting explicit control over group-specific performance. This can result in models performing well on certain groups but underperforming in others, leading to concerns about fairness. Moreover, each group may follow a different distribution, and a subset of groups may be more susceptible to distributional shifts due to external factors such as policy changes or environmental variation. To address these challenges, we propose a Per-Group Distributionally Robust Optimization (Per-GDRO) framework. It ensures fairness across groups and robustness to group-specific distributional shifts. In this framework, a ϕ-divergence ambiguity set governs adversarial group reweighting, and Wasserstein ambiguity sets capture local uncertainty within each group. We then develop an iterative algorithm that alternates between model updates and adversarial distributions across and within groups. We employ a derivative-free surrogate optimization method to determine the size of these ambiguity sets in an adaptive manner.

1. Introduction

Modern machine learning models exhibit outstanding performance across diverse domains yet often underperform in real-world deployments because of distributional differences between the training data and the deployment environment [8, 21]. Such distribution shifts—frequently driven by policy changes or environmental factors—motivate the use of Distributionally Robust Optimization (DRO) [23, 31]. From a group robustness perspective, DRO is particularly effective, since many datasets exhibit underlying group structures that significantly influence model behavior. Errors often concentrate in minority subpopulations—e.g., lower accuracy for certain regional accents or for particular racial/gender groups [5, 9, 20]. Such disparities largely arise from group imbalance and spurious correlations and persist despite abundant data.

To address these issues, GroupDRO explicitly protects worst-group performance, thereby reducing inter-group disparities [32]. Extensions such as FairDRO [19] and PGDRO [13] further improve inter-group fairness but largely overlook within-group distributional differences.

Motivated by these considerations, we propose Per-Group Distributionally Robust Optimization (Per-GDRO). By addressing both intra-group and inter-group distributions, it learns group-specific

levels of robustness. This enables us to avoid unnecessary conservatism for well-represented groups while effectively protecting vulnerable groups. Such an adaptive approach preserves the advantages of DRO for small groups while achieving both scalability and generalization performance in large-scale machine learning environments.

In summary, our contributions are as follows. (i) We propose Per-GDRO, a framework that assigns group-specific robustness levels, enabling the model to adaptively allocate robustness based on each group's exposure to distributional shifts and to offer stronger protection to more vulnerable groups without uniformly inflating conservatism. (ii) Our theoretical analysis shows that the Per-GDRO robust objective admits an explicit upper bound comprising a variance-like penalty that captures dispersion across group losses and a Lipschitz-weighted sum of Wasserstein radii; the first term encourages group fairness by implicitly penalizing uneven performance across groups, while the second controls within-group robustness. (iii) To avoid the limitations of manually tuning ambiguity set sizes, we treat the inter-group and intra-group robustness levels as learnable hyperparameters, enabling data-driven calibration that avoids overly conservative or insufficiently protective fixed sets and yields a more effective trade-off between worst-group robustness and overall performance. (iv) Finally, the Per-GDRO framework unifies and generalizes existing approaches; it reduces to standard DRO when all groups share a single ambiguity radius, and it recovers GroupDRO when intra-group robustness is disabled and only worst-group reweighting is applied.

A summary of related work is provided in Appendix A.

2. Proposed Method

Per-GDRO addresses both intra-group and inter-group distributional shifts by learning group-specific robustness levels: intra-group uncertainty is captured via the Wasserstein distance, and inter-group uncertainty is modeled with a ϕ-divergence. We next introduce the notation and definitions.

Notation. Let $\mathcal{G} = \{1, \dots, G\}$ denote the set of groups, and let $\mathcal{X}$ and $\mathcal{Y}$ be the input and label spaces, respectively; $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the set of probability measures on $\mathcal{X} \times \mathcal{Y}$. For each $g \in \mathcal{G}$, let $\mathcal{D}^g_{\text{train}}$, $\mathcal{D}^g_{\text{val}} \subseteq \mathcal{X} \times \mathcal{Y}$ denote the training and validation datasets with sizes $n_g := |\mathcal{D}^g_{\text{train}}|$ and $m_g := |\mathcal{D}^g_{\text{val}}|$, and set $n := \sum_{g=1}^G n_g$ and $m := \sum_{g=1}^G m_g$. Let $P_g \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denote the (unknown) data-generating distribution for group g. Assume that, for each g, $\mathcal{D}^g_{\text{train}}$ and $\mathcal{D}^g_{\text{val}}$ are disjoint and independent, with elements drawn i.i.d. from P_g and group labels observed; the validation sets are used only to form the upper-level objective (see Eq. (1) below).

Model parameters are $\theta \in \Theta$, and the sample-wise loss at a sample (x, y) is $\ell(\theta; (x, y))$. In the upper-level objective, expectations over a finite sample set $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y}$ are taken as empirical averages $\mathbb{E}\big[\ell(\theta; \mathcal{S})\big] := \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \ell(\theta; (x, y))$.

The nominal inter-group proportions are $\widehat{\mathbf{p}}^{\text{train}} = \big(\frac{n_1}{n}, \dots, \frac{n_G}{n}\big)$, $\widehat{\mathbf{p}}^{\text{val}} = \big(\frac{m_1}{m}, \dots, \frac{m_G}{m}\big) \in \Delta_G$; the empirical within-group distributions are, for each $g \in \mathcal{G}$, $\widehat{q}_g^{\text{train}} = \frac{1}{n_g} \sum_{(x,y)\in\mathcal{D}^g_{\text{train}}} \delta_{(x,y)}$, $\widehat{q}_g^{\text{val}} = \frac{1}{m_g} \sum_{(x,y)\in\mathcal{D}^g_{\text{val}}} \delta_{(x,y)}$ in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where $\Delta_G := \{\mathbf{p} \in \mathbb{R}^G : p_g \geq 0\ , \forall g \in \mathcal{G},\ \sum_{g\in\mathcal{G}} p_g = 1\}$ and $\delta_{(x,y)}$ denotes the Dirac measure at (x, y).

Definition 1 (ϕ-divergence) *Let $\phi : [0, \infty) \to \mathbb{R}$ be convex on $[0, \infty)$ with $\phi(1) = 0$, and adopt the conventions $0\,\phi\big(\frac{a}{0}\big) := a \lim_{t\to\infty} \frac{\phi(t)}{t}$ for $a > 0$ and $0\,\phi\big(\frac{0}{0}\big) := 0$ (see [4, Sec. 3.1]). For probability vectors $\mathbf{p}, \mathbf{q} \in \Delta_G$, define $I_\phi(\mathbf{p} \,\|\, \mathbf{q}) := \sum_{g=1}^G q_g\, \phi\Big(\frac{p_g}{q_g}\Big)$. We call ϕ the ϕ-divergence function and*

I_ϕ the ϕ-divergence; representative divergences (e.g., KL divergence, χ^2–distance, Cressie–Read power divergence [6, 18]) are listed in Appendix D.

Definition 2 (Wasserstein distance) *Fix $p \in [1, \infty)$ and a ground norm $\|\cdot\|$ on the feature space $\mathcal{X}$. For probability measures $Q, Q' \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with finite p-th moments, define $W_p(Q, Q') := \left(\inf_{\pi \in \Pi(Q,Q')} \int \|x - x'\|^p \pi\big(d(x,y),\, d(x',y')\big) \right)^{1/p}$, i.e., the transport cost depends only on the feature displacement $\|x - x'\|$. Here, $\|\cdot\|$ is a norm on $\mathcal{X}$, and $\Pi(Q, Q')$ denotes the set of all joint probability distributions on $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$ with marginals Q and Q', respectively. (See [22, Def. 1] for the general definition.)*

We define the ambiguity sets used throughout as follows: inter-group (ϕ-divergence) balls $P_\phi^{\text{train}}(\rho) := \{\mathbf{p} \in \Delta_G : I_\phi(\mathbf{p} \,\|\, \widehat{\mathbf{p}}^{\text{train}}) \le \rho\}$ and $P_\phi^{\text{val}}(\rho) := \{\mathbf{p} \in \Delta_G : I_\phi(\mathbf{p} \,\|\, \widehat{\mathbf{p}}^{\text{val}}) \le \rho\}$; intra-group Wasserstein balls $Q_g^{\text{train}}(\epsilon_g) := \{\, q_g \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : W_p(q_g, \widehat{q}_g^{\text{train}}) \le \epsilon_g \,\}$ and $Q_g^{\text{val}}(\epsilon_g) := \{\, q_g \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : W_p(q_g, \widehat{q}_g^{\text{val}}) \le \epsilon_g \,\}$ for all $g \in \mathcal{G}$.

Bilevel formulation of Per-GDRO. We treat the inter-group radius ρ and the per-group radii $\{\epsilon_g\}_{g \in \mathcal{G}}$ as upper-level decision variables, and train θ at the lower level under worst-case inter- and within-group perturbations. Per-GDRO solves

$$\min_{\rho, \{\epsilon_g\}} \max_{\mathbf{p} \in P_\phi^{\text{val}}(\rho)} \sum_{g \in \mathcal{G}} p_g \max_{q_g \in Q_g^{\text{val}}(\epsilon_g)} \mathbb{E}_{q_g}\left[\ell\big(\theta^*(\rho, \{\epsilon_g\}); \mathcal{D}_{\text{val}}^g\big)\right], \tag{1}$$

$$\text{where,} \quad \theta^*(\rho, \{\epsilon_g\}) = \arg\min_\theta \max_{\mathbf{p} \in P_\phi^{\text{train}}(\rho)} \sum_{g \in \mathcal{G}} p_g \max_{q_g \in Q_g^{\text{train}}(\epsilon_g)} \mathbb{E}_{q_g}\left[\ell\big(\theta; \mathcal{D}_{\text{train}}^g\big)\right]. \tag{2}$$

Here, $P_\phi^{\text{train}}(\rho)$ and $P_\phi^{\text{val}}(\rho)$ are the inter-group ϕ-divergence balls on the train and validation splits, and $Q_g^{\text{train}}(\epsilon_g)$ and $Q_g^{\text{val}}(\epsilon_g)$ are the corresponding per-group Wasserstein balls; superscripts denote the split, and the radii $(\rho, \{\epsilon_g\})$ are shared.

The lower-level objective Eq. (2) explicitly protects the worst group via adversarial reweighting across groups ($\mathbf{p} \in P_\phi^{\text{train}}(\rho)$, in the spirit of GroupDRO [32]) and attains robustness to group-specific distributional shifts via per-group Wasserstein sets ($q_g \in Q_g^{\text{train}}(\epsilon_g)$). Using group-specific radii $\{\epsilon_g\}$ from the upper level allows the robustness level to adapt by group, avoiding uniform over-conservatism on well-represented groups while guarding vulnerable ones.

In the upper-level objective Eq. (1), Per-GDRO optimizes the ambiguity set sizes ρ and $\{\epsilon_g\}$ by minimizing the worst-case validation risk: an inter-group adversary reweights groups within $P_\phi^{\text{val}}(\rho)$ and, for each group, an intra-group adversary perturbs the validation distribution within $Q_g^{\text{val}}(\epsilon_g)$. This criterion mirrors the adversarial structure used in training and directly aligns model selection with the lower-level robust objective. Accordingly, the lower-level problem enforces worst-group protection and robustness on the training split, while the upper-level problem tunes $(\rho, \{\epsilon_g\})$ under the same structure on the validation split.

Solving the Lower-Level Problem. Algorithm 1 solves the lower-level optimization problem (Eq. 2) in our bilevel framework for given robustness parameters ρ and $\{\epsilon_g\}_{g \in \mathcal{G}}$.

Algorithm 1 Per-Group DRO (Per-GDRO) Optimization Algorithm

Input: Ambiguity set sizes ρ, $\{\epsilon_g\}_{g\in\mathcal{G}}$; learning rate schedules η_θ^t, η_p^t; initial θ^0, $\mathbf{p}^0$, $\{q_g^0\}$
for $t = 0, 1, \ldots, T$ **do**
  // Update model parameters θ
  $\theta^{t+1} \leftarrow \theta^t - \eta_\theta^t \nabla_\theta \Big(\sum_{g\in\mathcal{G}} p_g^t \, \mathbb{E}_{q_g^t}\big[\ell(\theta^t;\, \mathcal{D}_{\text{train}}^g)\big] \Big)$
  // Update intra-group adversarial distributions q_g
  for each group $g \in \mathcal{G}$ **do**
    $q_g^{*,t} \leftarrow \arg\max_{q_g \in Q_g^{\text{train}}(\epsilon_g)} \mathbb{E}_{q_g}\big[\ell(\theta^{t+1};\, \mathcal{D}_{\text{train}}^g)\big]$
    $q_g^{t+1} \leftarrow q_g^{*,t}$
  end for
  // Update group distribution $\mathbf{p}$
  $\mathbf{p}^{*,t} \leftarrow \arg\max_{\mathbf{p}\in P_\phi^{\text{train}}(\rho)} \mathbb{E}_{\mathbf{p}}\Big[\sum_{g\in\mathcal{G}} \mathbb{E}_{q_g^{t+1}}\big[\ell(\theta^{t+1};\, \mathcal{D}_{\text{train}}^g)\big]\Big]$
  $\mathbf{p}^{t+1} \leftarrow (1-\eta_p^t)\,\mathbf{p}^t + \eta_p^t\,\mathbf{p}^{*,t}$
end for
Output: θ^T

We use Iterative Best Response (IBR) [42] (see also [19]); at each epoch, it computes the inner adversary's best responses $\big(q_g^*, \mathbf{p}^*\big)$ and plugs them into the loss; then it updates θ via stochastic gradient descent. In implementation, q_g^* is approximated by projected gradient descent (PGD) over the group-wise Wasserstein balls [24, 33]. By contrast, $\mathbf{p}^*$ is obtained exactly by solving the convex subproblem under the ϕ-divergence constraint. The updates of θ, $\{q_g\}$, and $\mathbf{p}$ are alternated until convergence.

Solving the Upper-Level Problem via Surrogate Optimization. We consider several strategies for optimizing bilevel objectives and adopt a gradient-free, surrogate-model approach because the lower-level problem is solved by external solvers that do not expose gradients, making hypergradients impractical. Accordingly, we optimize the ambiguity set sizes $(\rho, \{\epsilon_g\})$ via surrogate optimization inspired by [28]—Gaussian process models and expected improvement.

Algorithm 2 outlines this process; see Appendix C for details. At each iteration, we solve the lower-level problem (Algorithm 1) to obtain $\theta^*(\rho, \{\epsilon_g\})$, which is then evaluated on validation data using the robust-validation objective aligned with Eq. (1).

2.1. Theoretical Analysis

We analyze an explicit upper bound on the Per-GDRO robust risk because it separates the two sources of worst-case degradation—across-group reweighting and within-group perturbations—while retaining the nominal risk. This decomposition clarifies the dependence on $(\rho, \{\epsilon_g\})$ and performs basic consistency checks (e.g., reduced to the nominal risk when $\rho = 0$).

Proposition 1 *Let $(\rho, \{\epsilon_g\})$ be fixed, $\widehat{\mathbf{p}} \in \Delta_G$ the nominal group proportions, and $\mathbf{p} \in P_\phi(\rho)$ the inter-group adversarial weights. For each group $g \in \mathcal{G}$, let $\mathcal{L}(\theta; \mathcal{D}_g) := \mathbb{E}\big[\ell(\theta; \mathcal{D}_g)\big]$, $\widetilde{\mathcal{L}}(\theta; \mathcal{D}_g) := \max_{q_g \in Q_g(\epsilon_g)} \mathbb{E}_{q_g}\big[\ell(\theta; \mathcal{D}_g)\big]$ for some $p \geq 1$. Assume, for any fixed θ, the map $(x, y) \mapsto \ell(\theta; (x, y))$*

is $\mathrm{Lip}(\ell)$*-Lipschitz. Then*

$$\sum_{i=1}^{G} p_i \widetilde{\mathcal{L}}(\theta; \mathcal{D}_i) \leq \sum_{i=1}^{G} \widehat{p}_i \mathcal{L}(\theta; \mathcal{D}_i) + \|\mathbf{p}\|_2 \sqrt{\sum_{i=1}^{G} \Big(\mathcal{L}(\theta; \mathcal{D}_i) - \sum_{j=1}^{G} \widehat{p}_j \mathcal{L}(\theta; \mathcal{D}_j) \Big)^2} + \mathrm{Lip}(\ell) \sum_{i=1}^{G} p_i \epsilon_i. \tag{3}$$

The proof of Proposition 1 is provided in Appendix B, and a numerical test of the proposition is presented in Appendix E.3. A specific choice of ϕ-divergence yields a ρ-dependent inter-group term; see, e.g., [37, Proposition 1]. For instance, if the ambiguity set is constructed with the χ^2–distance, the second term corresponds to a variance-like regularization term [19, 37]. The within-group component follows from the standard Lipschitz–Wasserstein bound applied per group and then averaged with weights p_i [22, Thm. 5].

The upper bound Eq. (3) decomposes the robust risk into three components: (i) the nominal risk under the reference group proportions, (ii) an inter-group dispersion term that measures variance of group losses, and (iii) a within-group perturbation cost determined by the given Wasserstein radii. The bound holds under a general ϕ-divergence ambiguity set over the inter-group distribution and any Wasserstein balls. It allows us to analyze the sources of the worst-case objective independently.

The within-group term grows linearly with the first-order transport budget permitted by the Wasserstein radius. This inequality follows solely from the assumption that the loss is $\mathrm{Lip}(\ell)$-Lipschitz with respect to the ground norm that induces W_p, providing a dimension-free, linearly scaled sensitivity control. The result is ϕ-agnostic on the inter-group side and thus can be coupled with any ϕ-divergence ambiguity set; once a particular ϕ is fixed, $P_\phi(\rho)$ yields ρ-dependent control of the inter-group term. The same Lipschitz condition makes the bound apply verbatim to both classification and regression, and it admits immediate corollaries such as pure inter-group robustness ($\epsilon_g = 0$) and pure within-group robustness ($\rho = 0$).

3. Conclusion

We introduced Per-GDRO, a bilevel framework that unifies inter–group and intra–group robustness by learning the inter–group radius ρ and group-specific radii $\{\epsilon_g\}$. We established a robust risk bound (Eq. (3)) that serves as an interpretable certificate, decomposing the risk into nominal, inter–group, and intra–group components. This formulation clarifies how robustness tightens or relaxes, narrowing group-wise performance gaps while maintaining robustness to distributional shifts.

In practice, the lower–level optimization shows robust within feasible parameter regions but can destabilize under extreme misalignments of ambiguity set sizes, as detailed in our preliminary synthetic experiments (Appendix E). These results demonstrate that Per–GDRO effectively addresses both inter–group and intra–group distributional shifts, successfully recovering minority performance where baselines fail. However, the existence of a better solution suggests that the current objective operates conservatively. Developing a refined upper–level formulation capable of identifying such better solutions represents a valuable direction for future research.

Finally, the bilevel min–max–max structure can incur substantial computational cost, as the inner maximization is repeated and an outer-level search is required. Mitigating this complexity while preserving theoretical guarantees remains an important direction for future work. We intend to investigate acceleration strategies that maintain stability in subsequent studies.

References

[1] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International conference on machine learning*, pages 120–129. PMLR, 2019.

[2] Sushant Agarwal, Amit Deshpande, Rajmohan Rajaraman, and Ravi Sundaram. Optimal fair learning robust to adversarial distribution shift. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=TGcXwWdQQj.

[3] Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pages 1–19. Informs, 2015.

[4] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[6] Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3):440–464, 1984.

[7] Siran Dai, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Drauc: an instance-wise distributionally robust auc optimization framework. *Advances in Neural Information Processing Systems*, 36:44658–44670, 2023.

[8] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

[9] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[10] Ahmad Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. Designing ambiguity sets for distributionally robust optimization using structural causal optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16462–16470, 2025.

[11] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

[12] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

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

[14] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[15] Jiaming Hu, Debarghya Mukherjee, and Ioannis Ch Paschalidis. Dro-augment framework: Robustness by synergizing wasserstein distributionally robust optimization and data augmentation. *arXiv preprint arXiv:2506.17874*, 2025.

[16] Shu Hu and George H Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, pages 62–87. PMLR, 2022.

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

[18] Leah Jager and Jon A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5), October 2007. ISSN 0090-5364. doi: 10.1214/0009053607000000244. URL http://dx.doi.org/10.1214/000905360700000244.

[19] Sangwon Jung, Taeeon Park, Sanghyuk Chun, and Taesup Moon. Re-weighting based group fairness regularization via classwise robust optimization. *arXiv preprint arXiv:2303.00442*, 2023.

[20] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689, 2020.

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

[22] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[23] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12): 2346–2363, 2018.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Naren Sarayu Manoj and Kumar Kshitij Patel. Distributionally robust linear regression with block lewis weights. In *OPT 2024: Optimization for Machine Learning*, 2024.

[26] Akshay Mehra and Jihun Hamm. Penalty method for inversion-free deep bilevel optimization. In *Asian conference on machine learning*, pages 347–362. PMLR, 2021.

[27] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[28] Juliane Müller, Jangho Park, Reetik Sahu, Charuleka Varadharajan, Bhavna Arora, Boris Faybishenko, and Deborah Agarwal. Surrogate optimization of deep neural networks for groundwater predictions. *Journal of Global Optimization*, 81:203–231, 2021.

[29] Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2018.

[30] Qi Qian, Juhua Hu, and Hao Li. Hierarchically robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7336–7344, 2020.

[31] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[32] 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. *arXiv preprint arXiv:1911.08731*, 2019.

[33] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[34] Irina Wang, Cole Becker, Bart Van Parys, and Bartolomeo Stellato. Learning decision-focused uncertainty sets in robust optimization. *arXiv preprint arXiv:2305.19225*, 2023.

[35] Zifan Wang, Yi Shen, Michael M Zavlanos, and Karl H Johansson. Outlier-robust distributionally robust optimization via unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 37:52189–52214, 2024.

[36] Mingyang Wu, Li Lin, Wenbin Zhang, Xin Wang, Zhenhuan Yang, and Shu Hu. Preserving auc fairness in learning with noisy protected groups. *arXiv preprint arXiv:2505.18532*, 2025.

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

[38] Yufeng Yang, Yi Zhou, and Zhaosong Lu. A stochastic algorithm for sinkhorn distance-regularized distributionally robust optimization. In *OPT 2024: Optimization for Machine Learning*, 2024.

[39] Qi Zhang, Yi Zhou, Ashley Prater-Bennette, Lixin Shen, and Shaofeng Zou. Large-scale non-convex stochastic constrained distributionally robust optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8217–8225, 2024.

[40] Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24746–24755, 2024.

[41] Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.

[42] Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. Distributionally robust multilingual machine translation. *arXiv preprint arXiv:2109.04020*, 2021.

Appendix A. Related Works

Distributionally Robust Optimization (DRO). DRO pursues distributional robustness by defining an ambiguity set of distributions around a baseline distribution estimated from data and optimizing under the worst-case distribution within that set [31]. The ambiguity set can be defined in various ways; the most common approach is to use a discrepancy between distributions, most notably ϕ-divergences [3, 4] and the Wasserstein distance [12, 22, 27]. ϕ-divergence–based ambiguity sets are convex, which makes the optimization tractable; under standard regularity conditions, strong duality holds, and the DRO objective admits simple reformulations. However, ϕ-divergence has been criticized for failing to adequately capture nearby distributions compared to the Wasserstein distance or for potentially inducing overfitting rather than distributional robustness [12]. Accordingly, we model inter-group variation using ϕ-divergence ambiguity sets, while modeling within-group shifts with Wasserstein balls around each group's distribution.

In general, the size of the ambiguity set is determined statistically [11, 29], but in machine learning there is a view that such DRO can be overly conservative [17]. To alleviate this, there is ongoing work on defining ambiguity sets in alternative ways [10, 35, 38] or on learning the size of the ambiguity set [34]. There is also continued research on applying DRO effectively to large-scale datasets [39] and on data augmentation using DRO [15].

Group fairness. GroupDRO was proposed to reduce inter-group performance gaps caused by spurious correlations, with extensions such as PGDRO, which models probabilistic group membership [13], and FairDRO, which applies DRO on a per-class basis [19]. In classification, common strategies include minimizing worst-group risk and equalizing group-wise losses via reweighting or resampling; from a ranking-metrics perspective, there are also approaches that directly optimize AUC using instance-level DRO [7]. In addition, hierarchical representation learning has been proposed to minimize a distributionally robust upper bound over the full dataset rather than employing group-wise ambiguity sets [30]. Classification-centric work is active, including methods that control worst-subpopulation loss without explicit protected-group labels [14], approaches that preserve AUC fairness under noisy protected-group labels [36], fairness-aware adversarial training [40], and theoretically optimal fair learning under adversarial distribution shift [2]. In contrast, for regression and continuous prediction, two representative directions are particularly relevant; (i) reductions that encode fairness criteria as constraints or weights on group-wise errors [1], and (ii) a least-squares formulation that directly controls worst-group mean-squared error and remains efficient with many groups via Block Lewis weights [25]; additionally, survival-analysis losses have been proposed to control risk for potentially vulnerable subpopulations via DRO even without protected-group labels [16].

Bilevel Optimization. Bilevel optimization jointly defines an upper-level and a lower-level optimization problem and solves them simultaneously; in machine learning, it is primarily used for hyperparameter optimization, data denoising, few-shot learning, and training-data poisoning [26, 41]. In the DRO context, regarding the size of the ambiguity set as a hyperparameter can mitigate excessive pessimism. Methods to optimize it include using hypergradients [26] and surrogate models such as Gaussian Process (GP) or Radial Basis Function (RBF) [28]. We adopt the latter—a derivative-free, surrogate-based bilevel approach (GP)—since the lower-level DRO subproblems are solved by gradient-inaccessible external solvers; see Section 2 for details.

Appendix B. Proofs

Proof [Proof of Proposition 1] For brevity write $\mathcal{L}_i := \mathcal{L}(\theta; \mathcal{D}_i)$ and $\widetilde{\mathcal{L}}_i := \widetilde{\mathcal{L}}(\theta; \mathcal{D}_i)$. Let $\mathbf{q} \in \Delta_G$ denote a reference weight vector (e.g., $\mathbf{q} = \widehat{\mathbf{p}}$).

Inter-group step (Cauchy–Schwarz). We decompose and bound the p-weighted risk:

$$\begin{aligned}
\sum_{i=1}^{G} p_i \mathcal{L}_i &= \sum_{i=1}^{G} q_i \mathcal{L}_i + \sum_{i=1}^{G} (p_i - q_i) \mathcal{L}_i && (4)\\
&= \sum_{i=1}^{G} q_i \mathcal{L}_i + \sum_{i=1}^{G} p_i \left(\mathcal{L}_i - \sum_{j=1}^{G} q_j \mathcal{L}_j \right) && (5)\\
&\leq \sum_{i=1}^{G} q_i \mathcal{L}_i + \sqrt{\sum_{i=1}^{G} p_i^2} \sqrt{\sum_{i=1}^{G} \left(\mathcal{L}_i - \sum_{j=1}^{G} q_j \mathcal{L}_j \right)^2}, && (6)
\end{aligned}$$

where (6) follows from Cauchy–Schwarz.

Within-group step (Wasserstein–Lipschitz bound). By assumption, for any fixed θ the map $(x, y) \mapsto \ell(\theta; (x, y))$ is $\mathrm{Lip}(\ell)$-Lipschitz with respect to the norm that induces W_p ($p \geq 1$). Hence, for each group i with $Q_i := \{\, q_i : W_p(q_i, \mathcal{D}_i) \leq \epsilon_i \,\}$,

$$\widetilde{\mathcal{L}}_i = \max_{q_i \in Q_i} \mathbb{E}_{q_i}[\ell(\theta; \mathcal{D}_i)] \;\leq\; \mathcal{L}_i + \mathrm{Lip}(\ell)\, \epsilon_i. \tag{7}$$

Combine. Summing (7) with weights p_i and applying (6) yields

$$\sum_{i=1}^{G} p_i \widetilde{\mathcal{L}}_i \;\leq\; \sum_{i=1}^{G} q_i \mathcal{L}_i + \sqrt{\sum_{i=1}^{G} p_i^2} \sqrt{\sum_{i=1}^{G} \left(\mathcal{L}_i - \sum_{j=1}^{G} q_j \mathcal{L}_j \right)^2} + \mathrm{Lip}(\ell) \sum_{i=1}^{G} p_i \epsilon_i, \tag{8}$$

which is exactly (3). ■

Appendix C. Algorithm for surrogate optimization

$$\mathcal{L}_{\text{val}}(\theta;\, \rho, \{\epsilon_g\}) \;:=\; \max_{\mathbf{p}\in P_\phi^{\text{val}}(\rho)} \sum_{g\in\mathcal{G}} p_g \max_{q_g \in Q_g^{\text{val}}(\epsilon_g)} \mathbb{E}_{q_g}\left[\ell\big(\theta;\, \mathcal{D}_{\text{val}}^g\big)\right]. \tag{9}$$

Algorithm 2 Surrogate Optimization for Upper-Level Parameters $(\rho, \{\epsilon_g\})$

Input: Initial samples $\{(\rho_i, \{\epsilon_{g,i}\})\}_{i=1}^n$; validation loss $\mathcal{L}_{\text{val}}$ (Eq. (9)); number of iterations T
for $i = 1, \ldots, n$ **do**
  Solve lower-level optimization using Algorithm 1 with fixed $\rho_i, \{\epsilon_{g,i}\}$ to get θ_i^*
  Evaluate validation loss: $y_i \leftarrow \mathcal{L}_{\text{val}}(\theta_i^*;\, \mathcal{D}_{\text{val}})$
end for
for $t = n+1, \ldots, T$ **do**
  Fit a surrogate model (e.g., Gaussian Process) to $\{(\rho_i, \{\epsilon_{g,i}\}), y_i\}_{i=1}^{t-1}$
  Select new candidate $(\rho_t, \{\epsilon_{g,t}\})$ via an acquisition function
  Solve lower-level optimization using Algorithm 1 with $(\rho_t, \{\epsilon_{g,t}\})$ to get θ_t^*
  Evaluate validation loss: $y_t \leftarrow \mathcal{L}_{\text{val}}(\theta_t^*;\, \mathcal{D}_{\text{val}})$
end for
Output: $(\rho^*, \{\epsilon_g^*\}) \leftarrow \arg\min_{(\rho_i, \{\epsilon_{g,i}\})} y_i$

Appendix D. Common ϕ-divergences

Table 1: Common ϕ-divergences.

Divergence	$\boldsymbol{\phi(t),\ t \geq 0}$	$\boldsymbol{I_\phi(p, q)}$
Kullback–Leibler Divergence	$t\log t - t + 1$	$\sum_\omega p_\omega \log\left(\frac{p_\omega}{q_\omega}\right)$
Burg Entropy	$-\log t + t - 1$	$\sum_\omega q_\omega \log\left(\frac{q_\omega}{p_\omega}\right)$
χ^2-Distance	$\frac{(t-1)^2}{t}$	$\sum_\omega \frac{(p_\omega - q_\omega)^2}{p_\omega}$
Modified χ^2-Distance	$(t-1)^2$	$\sum_\omega \frac{(p_\omega - q_\omega)^2}{q_\omega}$
Total Variation Distance	$\frac{1}{2}\lvert t-1\rvert$	$\frac{1}{2}\sum_\omega \lvert p_\omega - q_\omega\rvert$
Hellinger Distance	$\left(\sqrt{t}-1\right)^2$	$\sum_\omega \left(\sqrt{p_\omega} - \sqrt{q_\omega}\right)^2$
Cressie–Read Power Divergence	$\frac{1-\theta+\theta t - t^\theta}{\theta(1-\theta)},\ \theta \neq 0, 1$	$\frac{1 - \sum_\omega p_\omega^\theta\, q_\omega^{1-\theta}}{\theta(1-\theta)},\ \theta \neq 0, 1$

Appendix E. Synthetic Experiments

E.1. Experimental Setup

We construct a synthetic binary classification task in $\mathbb{R}^2$ designed to probe model behavior under heterogeneous distributional shifts. Features are generated from Gaussian distributions $x|g \sim \mathcal{N}(\mu_g, I_2)$ with means symmetric across the decision boundary: $\mu_0 = (5/3, 5/3)$, $\mu_1 = (-5/3, -5/3)$, $\mu_2 = (5/3, -5/3)$, and $\mu_3 = (-5/3, 5/3)$. Class labels are assigned as $y = 0$ for $g \in \{0, 2\}$ and $y = 1$ for $g \in \{1, 3\}$. Given this symmetric geometry, the ideal decision boundary for the nominal distribution is a vertical line ($x_1 \approx 0$) parallel to the y-axis.

However, we introduce specific challenges to disrupt this ideal boundary. First, we simulate severe group imbalance with sample counts of $(500, 420, 140, 140)$ for groups 0 to 3. Under standard training (ERM), this imbalance induces spurious correlations, causing the decision boundary to rotate counter-clockwise away from the vertical axis to fit the majority variance. Second, to evaluate robustness against feature drift, we inject an additive shift noise $\epsilon_3 \sim \mathcal{N}((-2, -0.5), 0.3I_2)$ exclusively to the training/validation sets of the minority group G_3.

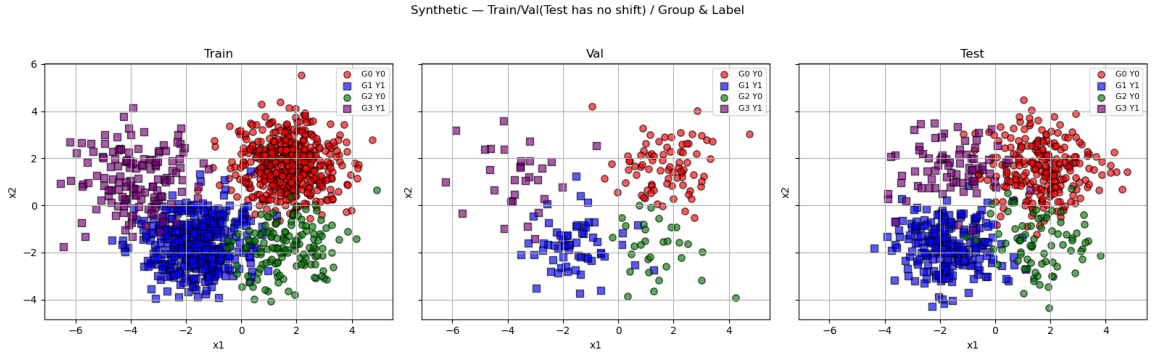


Figure 1: Visualization of the synthetic dataset generation. The training and validation samples for the minority group G_3 (purple squares) are subjected to an additive shift ϵ_3, while the test samples follow the nominal distribution.

For optimization, we use a fixed 4-layer MLP trained for 100 epochs. We employ our bilevel strategy to find the optimal ambiguity set sizes. Leveraging our knowledge of the group structure, we tailor the search space to allow greater flexibility for minority groups: the radius range is set to $\epsilon \in [0, 3]$ for majority groups (G_0, G_1) and $\epsilon \in [0, 5]$ for minority groups (G_2, G_3). The global radius ρ is searched within $[0, 3]$. The optimization is initialized with a grid of 243 points and proceeds with Bayesian optimization for 300 evaluations.

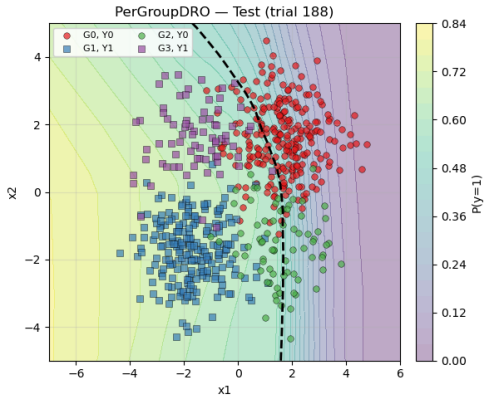
E.2. Results and Analysis

Figure 2 compares decision boundaries across four representative scenarios. The collapse case (Trial 188, $\rho = 3.0, \epsilon = (0.0, 3.0, 5.0, 2.5)$) in Figure 2(a) highlights the instability of misalignment. Excessive robustness widens uncertainty ($P(y = 1) \approx 0.5$) and forces the boundary to collapse into the Class 0 region. This results in a trivial model predicting mostly Class 1; while G_1 and G_3 achieve 100% accuracy, performance on Class 0 suffers significantly, with the minority G_2 falling to 44.0%.

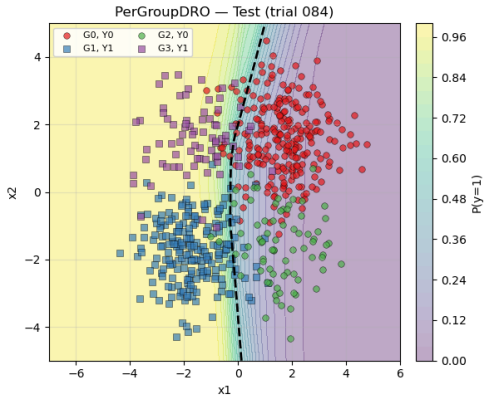
This trend suggests that such pathological behavior could become even more severe with further misalignment, potentially rendering the model completely uninformative.

In contrast, the targeted case (Trial 84, $\rho = 1.5, \epsilon = (0.0, 0.0, 0.0, 5.0)$) in Figure 2(b) demonstrates effective parameterization. Targeting a high radius strictly to the shifted group G_3 induces a clockwise curvature that counteracts the spurious tilt, ensuring full coverage of the shifted distribution.

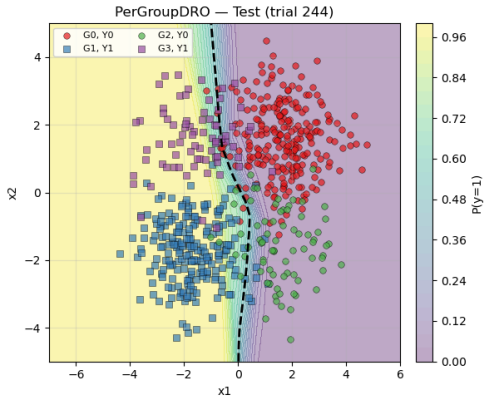
Crucially, our bilevel optimization automatically discovers this strategy (Trial 244, $\rho = 0.0, \epsilon = (0.0, 0.0, 0.0, 1.8)$) in Figure 2(c). By selectively applying robustness only where needed (G_3), it establishes a stable boundary that effectively separates the classes. We also visualize an ideal candidate (Trial 253, $\rho = 2.2, \epsilon = (0.6, 1.2, 0.4, 4.9)$) in Figure 2(d). Although Trial 253 yields higher average and worst-group accuracy, our optimizer favored Trial 244 due to its lower upper-level objective Eq. (1). This observation suggests that the current upper-level objective acts conservatively, indicating that further refinement could better align robust guarantees with empirical optimality.



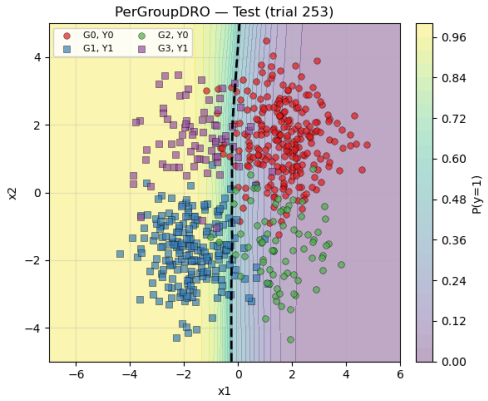
(a) Collapse (Trial 188)
$\rho = 3.0, \epsilon = (0.0, 3.0, 5.0, 2.5)$



(b) Targeted (Trial 84)
$\rho = 1.5, \epsilon = (0.0, 0.0, 0.0, 5.0)$



(c) Ours (Trial 244)
$\rho = 0.0, \epsilon = (0.0, 0.0, 0.0, 1.8)$



(d) Ideal (Trial 253)
$\rho = 2.2, \epsilon = (0.6, 1.2, 0.4, 4.9)$

Figure 2: Comparison of decision boundaries. (a) Collapse due to misalignment. (b) Tilt correction by targeting G_3. (c) Stable boundary via bilevel optimization. (d) Better solution in search space.

Table 2 confirms the effectiveness of our approach. While ERM and GDRO suffer on the shifted group G_3 (accuracies of 68.0% and 72.0%), Per-GDRO significantly recovers performance to 81.3%, achieving the best worst-group accuracy among automated methods. Notably, the ideal case found within the search space achieves an even higher worst-group accuracy of 90.7%, indicating the potential for further improvements.

Table 2: Test Performance Comparison

Method	Avg	Worst	G0	G1	G2	G3
ERM	0.93	0.68	**0.99**	0.96	0.89	0.68
GDRO	0.93	0.72	0.98	0.95	**0.95**	0.72
Ours	0.94	**0.81**	0.97	**0.97**	0.89	**0.81**
Ideal	*0.95*	*0.91*	*0.96*	*0.93*	*0.97*	*0.91*

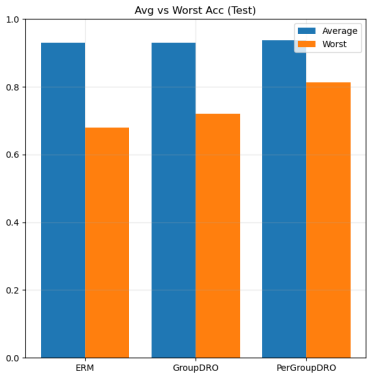


Figure 3: Comparison of Average and Worst-group accuracy.

E.3. Numerical Test of Theoretical Bounds

Finally, we numerically verify the theoretical bound presented in Proposition 1. Figure 4 tracks the optimization dynamics of the optimal configuration (Trial 244). The plot confirms that the robust loss (LHS) remains strictly bounded by the theoretical upper bound (RHS) throughout training. The stable convergence of both curves indicates that the optimization successfully minimizes the robust risk within the learned ambiguity set.

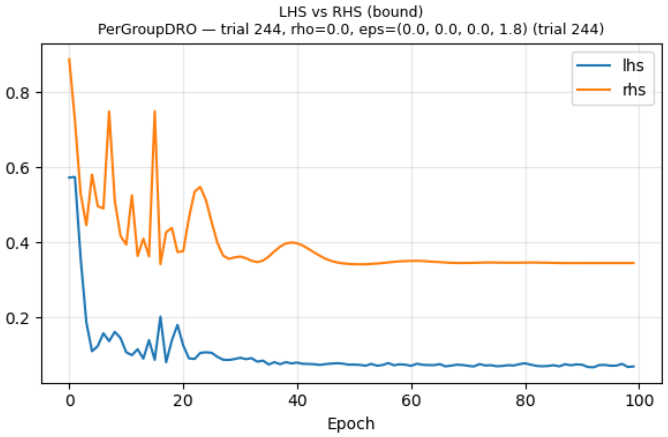


Figure 4: Numerical test of Proposition 1 (Trial 244). The robust loss (LHS) is strictly bounded by the theoretical upper bound (RHS), validating the stability of our proposed relaxation.