Robust Generalization with Adaptive Optimal Transport Priors for Decision-Focused Learning

Haixiang Sun **Andrew L. Liu**
Purdue University

Abstract

Few-shot learning requires models to generalize under limited supervision while remaining robust to distribution shifts. Existing Sinkhorn Distributionally Robust Optimization (DRO) methods provide theoretical guarantees but rely on a fixed reference distribution, which limits their adaptability. We propose a Prototype-Guided Distributionally Robust Optimization (PG-DRO) framework that learns class-adaptive priors from abundant base data via hierarchical optimal transport and embeds them into the Sinkhorn DRO formulation. This design enables few-shot information to be organically integrated into producing class-specific robust decisions that are both theoretically grounded and efficient, and further aligns the uncertainty set with transferable structural knowledge. Experiments show that PG-DRO achieves stronger robust generalization in few-shot scenarios, outperforming both standard learners and DRO baselines.

1 INTRODUCTION

Few-shot learning aims to enable models to generalize to new tasks with only a handful of labeled examples. In many real-world scenarios, one may have access to abundant regular data from common conditions, but only a very limited number of examples under rare or extreme circumstances. To prepare models for such situations, it is crucial to design approaches that can learn effectively even when supervision is scarce. While metric-based Snell et al. (2017a); Vinyals et al. (2016); Wang et al. (2018) and meta-learning approaches Chen et al. (2020b); Finn et al. (2017) have shown strong results in clean benchmarks, they struggle in practice when data are imbalanced, shifted, or adversarially perturbed. This fragility directly impacts the reliability of the decision rules derived from few-shot learners.

In practice, the data distributions encountered by few-shot learners are rarely stable. Even small perturbations in input space, regardless of sensor noise, domain shifts, or adversarial manipulation, can push examples across tight decision margins and yield unstable predictions Hendrycks and Dietterich (2019); Koh et al. (2020); Madry et al. (2018). Robust decision making therefore requires models not only to interpolate among a few examples, but also to anticipate the worst-case deviations that could occur around them Esfahani and Kuhn (2015); Goldblum et al. (2019). Preparing for such perturbations reduces overfitting to idiosyncrasies of the limited support set and encourages invariances that transfer across environments. In this sense, robustness is not a luxury but a necessity: it determines whether few-shot learning methods can survive the realities of open-world deployment.

A principled approach to ensure robustness is distributionally robust optimization, which minimizes the worst-case risk over an ambiguity set centered at the empirical support distribution Kuhn et al. (2024); Rahimian and Mehrotra (2019). The ambiguity set can be defined through different notions of distance between the historical data and a surrogate distribution. A widely used choice is the Wasserstein distance, and its entropic regularization leads to the Sinkhorn distance, which offers favorable computational properties Blanchet et al. (2019); Cuturi (2013); Esfahani and Kuhn (2015). Existing studies, such as Azizian et al. (2023); Wang et al. (2025), reformulate the resulting dual problems into low-dimensional parameterizations, making optimization tractable. Nevertheless, these approaches remain tied to a single predefined reference distribution, which restricts their ability to adapt when the data distribution shifts. This fixed-reference design represents a fundamental limitation in

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

real-world few-shot learning scenarios, where the support distribution is sparse and rarely representative of future test conditions.

In this work, we address two central challenges of robust decision making under limited data: achieving reliable generalization from scarce supervision and maintaining stability under distributional shift. We propose Prototype-Guided Distributionally Robust Optimization (PG-DRO), a class-adaptive framework that constructs class-specific priors via optimal transport from base prototypes and incorporates them directly into the robust decision rule. By aligning novel classes with structural knowledge distilled from a rich base dataset, PG-DRO improves generalization when only a handful of labeled examples are available. At the same time, it explicitly accounts for worst-case perturbations within each class, enhancing robustness and generating informative outliers that highlight the decision boundaries most vulnerable to shift. These properties make PG-DRO valuable for domains where models must extrapolate from scarce and noisy observations while remaining stable under unexpected change across many real-world applications.

Contribution

- We introduce Prototype-Guided Distributionally Robust Optimization (PG-DRO), a novel decision-making framework for low-data regimes that leverages hierarchical optimal transport to construct class-adaptive priors within entropic DRO, enabling scalable and robust learning under limited supervision.
- We demonstrate that this framework yields strong robust generalization: it maintains accuracy under perturbations and distribution shifts, particularly benefiting few-shot minority classes.
- We also provide theoretical analysis establishing conditions for robust optimization with adaptive priors, and validate our framework through extensive experiments showing consistent gains in both accuracy and robustness under distributional shifts.

2 RELATED WORKS

Few-shot Learning Few-shot learning (FSL) aims to rapidly adapt a model to novel classes from only a handful of labeled examples by leveraging transferable structure learned from base tasks Finn et al. (2017); Snell et al. (2017b). Beyond the canonical setting, a growing body of work investigates domain-adaptive FSL, where base and novel classes come from different or even mismatched domains, requiring methods to bridge distribution gaps and improve cross-domain transfer Guo et al. (2023); Pal et al. (2023); Zhao et al. (2020). More recently, researchers have also begun to study robust FSL, seeking resilience to adversarial perturbations and distributional shifts Goldblum et al. (2020); Sagawa et al. (2019); Wang et al. (2021). Such robustness-oriented approaches have been explored across downstream tasks, including natural language understanding Nookala et al. (2023), image classification Dong et al. (2022), and reinforcement learning Greenberg et al. (2023), yet they often remain limited by task-agnostic assumptions about uncertainty sets, leaving open the question of how to construct class-adaptive robust priors under few-shot regimes.

Optimal Transport Optimal Transport (OT) provides a geometric framework for comparing probability measures by modeling the minimal cost of transporting mass between them Cuturi (2013); Peyré and Cuturi (2018). A widely used extension is the entropically regularized formulation, which introduces a smoothing term and leads to the so-called Sinkhorn distance Carlier et al. (2015); Cuturi and Doucet (2014). Formally, the entropic OT problem seeks a coupling $T \in \mathbb{R}^{B\times N}$ solving

$$\min_{T\geq 0}; \langle C, T\rangle + \varepsilon H(T), \quad \text{s.t. } T\mathbf{1} = \alpha, \;\; T^\top \mathbf{1} = \beta, \quad (1)$$

where $H(T) = -\sum_{bn} T_{bn} \log T_{bn}$ is the entropy and α, β are the prescribed marginals.

OT has since become a versatile tool across several domains. In domain adaptation, OT-based methods align source and target distributions either at the feature level or jointly with labels Chang et al. (2022); Courty et al. (2014); Damodaran et al. (2018), with extensions to multi-source, target-shift settings Redko et al. (2018) and deep architectures Kerdoncuff et al. (2020); Li et al. (2020). OT also works in representation learning, which underpins cross-domain embedding alignment Chen et al. (2020a), self-supervised objectives Caron et al. (2020); Shi et al. (2023), and structure-preserving mappings Alvarez-Melis and Jaakkola (2018). Beyond these, OT has been applied to downstream tasks ranging from generative modeling with Wasserstein losses Deshpande et al. (2018); Kolouri et al. (2019); Simsekli et al. (2018) to structured prediction and sequence modeling. Beyond classical OT applications, several works adapt OT for few-shot learning. Hierarchical OT adapts base-class statistics to novel classes for better priors Guo et al. (2022b), while Guo et al. (2022a); Yurochkin et al. (2019); Zhang et al. (2020) use OT for structured representation and prototype aggregation. Our work resonates with these methods in using hierarchical OT to capture cross-class relations, but differs by embedding

it into a Sinkhorn DRO framework to obtain class-adaptive robust logits rather than solely performing distribution calibration or metric alignment.

Robust Optimization It is formulated as a min-max problem mathematically,

$$\min_{\theta\in\Theta} \sup_{P\in\mathcal{U}(\mu,\nu)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P}\big[\ell_\theta(\mathbf{x},\mathbf{y})\big], \tag{2}$$

where the ambiguity set $\mathcal{U}$ collects all candidate joint distributions P whose distance to a reference distribution does not exceed a prescribed radius.

For current DRO, common choices for the ambiguity set Θ include ϕ-divergences Namkoong and Duchi (2016a,b), Wasserstein balls Esfahani and Kuhn (2015); Gao and Kleywegt (2016), and kernel metrics such as MMD Feydy et al. (2018); Staib and Jegelka (2019); Zhu et al. (2021). Wasserstein DRO leverages the geometry of the sample space and provides strong guarantees, but its dual entails a hard supremum, with worst-case distributions degenerating to finite discrete measures, problematic under small support Gao and Kleywegt (2016). Entropic regularization smooths the Wasserstein dual's supremum into a log-sum-exp, yielding the Sinkhorn distance solvable via fast matrix scaling Cuturi (2013); Peyré and Cuturi (2018). Recent analyses further show that this regularization leads to a single scalar dual variable while preserving convexity and tractability Azizian et al. (2023); Wang et al. (2025). Decision-focused DRO trains prediction and decisions jointly Chenreddy and Delage (2024); Costa and Iyengar (2023); Donti et al. (2017); Ma et al. (2024), while adversarial training links to Wasserstein-DRO, unifying attacks with distributional models Bai et al. (2023); Dong et al. (2020); Schmidt et al. (2018) and providing certificates Sinha et al. (2018).

However, despite their differences, all these approaches posit an ambiguity set centered around a single, fixed reference distribution, which in few-shot regimes can only be estimated from a handful of supports and thus is prone to misalignment with rare or shifted classes. Our method instead adapts the reference per class via adaptive OT, thereby retaining Sinkhorn DRO's tractability while aligning the uncertainty set with transferable base-class structure.

3 PRELIMINARY

We begin by recalling the formulation of distributionally robust optimization under entropic optimal transport. Let $\mathcal{X}$ denote the input space, let $\mu \in \mathcal{P}(\mathcal{X})$ be the empirical distribution of observed inputs, and let $\ell : \mathcal{Y} \to \mathbb{R}$ be the prediction loss function defined on the output space $\mathcal{Y}$. Consider a reference measure ν on $\mathcal{Y}$ and a ground cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ that encodes admissible perturbations between input $\mathbf{x} \in \mathcal{X}$ and output $\mathbf{y} \in \mathcal{Y}$. The entropic optimal transport distance between two distributions μ and ν is defined as

$$\begin{aligned} W_\varepsilon(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\big[c(\mathbf{x},\mathbf{y})\big] \\ + \varepsilon \mathrm{KL}\left(\gamma \| \mu\otimes\nu\right), \end{aligned} \tag{3}$$

where $\Gamma(\mu,\nu)$ denotes the set of couplings between μ and ν, $\varepsilon > 0$ is the entropic regularization parameter, and $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence.

Definition 3.1 (Sinkhorn DRO)**.** *The worst-case expected function f over the Sinkhorn ball $\{\mu \in \mathcal{P}(\mathcal{X}) : W_\varepsilon(\mu,\nu) \le \rho\}$ admits the form*

$$V = \sup_{W_\varepsilon(\mu,\nu)\le\rho} \mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{y}\sim\gamma_\mathbf{x}}\big[f(\mathbf{y})\big], \tag{4}$$

where $\rho > 0$ specifies the radius of the uncertainty set, ν is a fixed reference measure on $\mathcal{Y}$, and $\gamma_\mathbf{x}$ is the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ under some coupling $\Pi \in \Gamma(\mu,\nu)$.

By strong duality, this problem can be equivalently expressed in terms of a convex dual formulation

$$V_D(\lambda) = \lambda\rho + \lambda\varepsilon\mathbb{E}_\mathbf{x}\left[\log \mathbb{E}_{\mathbf{y}\sim Q^\nu_{\varepsilon,\mathbf{x}}} \exp\left(\frac{f(\mathbf{y})}{\lambda\varepsilon}\right)\right], \tag{5}$$

where $\lambda > 0$ is a scalar dual variable, and $Q^\nu_{\varepsilon,\mathbf{x}}$ is a Gibbs kernel defined as

$$\mathrm{d}Q^\nu_{\varepsilon,\mathbf{x}}(\mathbf{y}) = \frac{\exp\left(-c(\mathbf{x},\mathbf{y})/\varepsilon\right)}{Z(\mathbf{x})}\mathrm{d}\nu(\mathbf{y}), \tag{6}$$

with the normalizing constant $Z(\mathbf{x}) = \int \exp\left(-c(\mathbf{x},\mathbf{y})/\varepsilon\right)\mathrm{d}\nu(\mathbf{y})$. This dual formulation highlights that Sinkhorn DRO reduces the infinite-dimensional adversarial optimization to the minimization of a one-dimensional convex function in λ, where robustness is achieved through the log-moment generating function of ℓ under the kernelized distribution $Q^\nu_{\varepsilon,\mathbf{x}}$. Intuitively, the adversary inflates the expected loss by softmax-smoothing over plausible perturbations of each sample, controlled jointly by the ground cost c and the reference distribution ν.

Lemma 3.2 (Convexity and uniqueness of the dual minimizer; cf. Wang et al. (2025), Thm. 1(III) and Lemma 3)**.** *$V_D(\lambda)$ is convex in λ on $[0,\infty)$. Moreover, unless ℓ is ν-a.s. constant, $V_D(\lambda)$ is* strictly *convex on $(0,\infty)$ and thus admits a unique minimizer $\lambda^\star > 0$. In the degenerate case where ℓ is ν-a.s. constant, the unique minimizer is attained at $\lambda^\star = 0$.*

Since $V_D(\lambda)$ is strictly convex and satisfies $V_D(\lambda) \ge \lambda\rho$, it is coercive for $\rho > 0$ and hence $V_D(\lambda) \to \infty$ as

$\lambda \to \infty$. Therefore the minimizer not only exists but is also unique, with $\lambda^\star = 0$ exactly in the degenerate constant-loss case. This ensures a well-posed one-dimensional optimization for any fixed prior ν. However, the statistics still hinge on how well ν reflects the target task. In few-shot settings, generally, such a rigid prior cannot encode any task-specific structure. To remedy this, we introduce hierarchical OT priors ν_c for different classes that transfer geometry and distributional knowledge from base classes, thereby enabling a class-specific robust decision guidance.

4 METHODOLOGY

Adaptive Optimal Transport Priors. While Sinkhorn DRO offers a principled route to distributional robustness, a fixed reference distribution ν is agnostic to the downstream few-shot task. It ignores class-specific geometry, which leads to a mismatch of ambiguity set under scarce supervision. To make robustness class-aware by leveraging the few-shot support information, we transfer knowledge from abundant base classes and construct class-adaptive priors via a hierarchical optimal transport procedure Guo et al. (2022b).

Let the source domain contain B classes, indexed by $b = 1, \dots, B$, with few-shot prototypes $\{\mathbf{x}_{b,i}\}_{i=1}^{m_b}$ for class b, class mean $\mu_b = \frac{1}{m_b}\sum_{i=1}^{m_b} \mathbf{x}_{b,i}$, and covariance $\Sigma_b = \frac{1}{m_b - 1}\sum_{i=1}^{m_b}(\mathbf{x}_{b,i} - \mu_b)(\mathbf{x}_{b,i} - \mu_b)^\top$. For a new few-shot task with labeled support $S_k = \{(\tilde{\mathbf{x}}_n, \mathbf{y}_n)\}_{n=1}^{N}$, we compute for each support example a soft-min Sinkhorn cost to every base class

$$C_b(\tilde{\mathbf{x}}_n) = -\varepsilon_{\mathrm{s}} \log \sum_{i=1}^{m_b} \exp\Big(-\frac{1}{\varepsilon_{\mathrm{s}}}\, c(\tilde{\mathbf{x}}_n, \mathbf{x}_{bi})\Big), \tag{7}$$

where $c(\cdot,\cdot)$ is the ground cost and ε_{s} is the entropic temperature. Stacking $\{C_b(\tilde{\mathbf{x}}_n)\}$ gives a cost matrix $\mathbf{C} \in \mathbb{R}^{B \times N}$.

At the class level, we solve the entropic OT problem (1) on $\mathbf{C}$ to obtain a transport plan $\mathbf{T} \in \mathbb{R}^{B \times N}$ that aligns support examples with base classes. Aggregating mass over the support of class c yields mixture weights

$$\tilde{w}_{bc} = \frac{\sum_{n:\, \ell_n = c} T_{bn}}{\sum_{b'=1}^{B} \sum_{n:\, \ell_n = c} T_{b'n}}, \tag{8}$$

and the resulting class-adaptive prior is the Gaussian mixture

$$\nu_c \;=\; \sum_{b=1}^{B} \tilde{w}_{bc}\, \mathcal{N}(\mu_b, \Sigma_b). \tag{9}$$

In summary, the hierarchical OT procedure yields class-adaptive prototype-guided priors $\{\nu_c\}$. Each prior is explicitly anchored on transferable prototypes from base classes and refined by the few-shot support set. Rather than serving as a generic preprocessing step, these priors act as structured guides that can be injected into the Sinkhorn DRO dual.

Prototype-Guided DRO (PG-DRO) Plugging ν_c into the Sinkhorn DRO kernel $Q^{\nu_c}_{\varepsilon,\mathbf{x}}$ yields the class-specific robust logit

$$V_c(\mathbf{x};\theta) = \min_{\lambda \geq 0} \Big\{ \lambda\rho + \lambda\varepsilon \log \mathbb{E}_{\mathbf{y}\sim Q^{\nu_c}_{\varepsilon,\mathbf{x}}} \exp\big(\tfrac{f_c(\mathbf{y};\theta)}{\lambda\varepsilon}\big)\Big\}. \tag{10}$$

This yields class-specific robust logits $V_c(\mathbf{x};\theta)$, which capture adversarial uncertainty in a manner consistent with the structural prior of class c, where θ denotes the parameter. Since the objective in λ is smooth and strictly convex, (10) can be solved efficiently via a one-dimensional Newton step per class. Compared with a single fixed prior ν, prototype-guided priors sharpen the ambiguity set of ν_c by some prototypes around class-consistent regions, align worst-case perturbations $\mathbf{y} \sim Q^{\nu_c}_{\varepsilon,\mathbf{x}}$ with transferable geometry, and deliver decision-focused gains through the robust logit $V_c(\mathbf{x};\theta)$. This turns robustness from a global assumption into a class-level inductive bias.

Corollary 4.1 (Multi-class convexity)**.** *For each class c, the dual objective $V_c(\mathbf{x};\theta)$ is (strictly) convex in λ_c on $(0,\infty)$ in the sense of Lemma 3.2. Consequently, the joint objective*

$$\begin{aligned}&\mathcal{V}(\lambda_1, \dots, \lambda_C)\\ &:= \sum_{c=1}^{C} \Big\{ \lambda_c\rho + \lambda_c\varepsilon \log \mathbb{E}_{\mathbf{y}\sim Q^{\nu_c}_{\varepsilon,\mathbf{x}}} \exp\big(f_c(\mathbf{y};\theta)/(\lambda_c\varepsilon)\big)\Big\}\end{aligned} \tag{11}$$

is convex on $(0,\infty)^C$. Moreover, unless $f_c(\cdot;\theta)$ is ν_c-a.s. constant for every c, $\mathcal{V}$ is strictly convex in each non-degenerate coordinate and the corresponding minimizer components $\lambda_c^\star$ are unique; if $f_c(\cdot;\theta)$ is a.s. constant for some c, then the unique minimizer in that coordinate is $\lambda_c^\star = 0$.

Decision rule and training objective. With class-specific robust logits in hand, our prediction is $\hat{c} = \arg\max_c V_c(\mathbf{x};\theta)$. Training minimizes the softmax cross-entropy applied to these robust logits:

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{x}, c^\star; \theta) = -\log \frac{\exp\big(V_{c^\star}(\mathbf{x};\theta)\big)}{\sum_{c=1}^{C} \exp\big(V_c(\mathbf{x};\theta)\big)}. \tag{12}$$

Note that this coincides with the standard cross-entropy when V_c are the raw logits from the classifier head. Our goal is to make the decision reliable under shift. Since each $V_c(\mathbf{x};\theta)$ in (10) is the worst-case class score within a Sinkhorn ball shaped by the class-adaptive prior ν_c, the likelihood $V_c(\mathbf{x};\theta)$

is guarded against per-class perturbations. Leveraging few-shot supports to form ν_c aligns the ambiguity set with transferable base-class structure, keeping novel-class predictions reliable under domain noise and scarcity. Besides, PG-DRO increases it increases the robust margin $V_{c^\star}(\mathbf{x};\theta) - \max_{c\neq c^\star} V_c(\mathbf{x};\theta)$ and thereby lowers worst-case misclassification.

To study how adaptive OT priors affect the robust objective, we track the Sinkhorn DRO value $V_\nu(\theta)$ as the class prior ν is updated. Here for brevity, we fix a target class c and omit the subscript c. Denote $w^{(t)}$ as the mixture weights over base prototypes in iteration t and $\nu^{(t)} = \sum_b w_b^{(t)}\mu_b$, the update is given by a relaxed fixed-point iteration with the stochastic OT map $\widehat{T}$. Based on some wild and common assumptions, we could obtain the following contraction property:

Theorem 4.2 (Contraction of V under adaptive OT)**.** *Let $w^{(t+1)} = (1-\eta_t)w^{(t)} + \eta_t\widehat{T}\big(w^{(t)}\big)$ with $\eta_t \in (0,1)$, and define $\nu^{(t)} = \sum_b w_b^{(t)}\mu_b$ and $\Delta_t(\theta) := |V_{\nu^{(t)}}(\theta) - V_{\nu^\star}(\theta)|$. Suppose the OT map T is locally contractive near $w^\star$ with contraction rate $\kappa \in (0,1)$. Then, for all t in the contraction neighborhood,*

$$\Delta_{t+1}(\theta) \;\le\; (1-\eta_t\kappa)\Delta_t(\theta) \;+\; \mathcal{O}\big(N^{-1/2}\big), \qquad (13)$$

where κ is the contraction constant of the population map T around $w^\star$, and N is the number of few-shot samples.

The proof is provided in Appendix E. Theorem 4.2 shows that the adaptive prior update drives the robust value $V_{\nu^{(t)}}(\theta)$ toward the oracle $V_{\nu^\star}(\theta)$ at a linear rate $1-\eta_t\kappa$, modulo a stochastic term $\mathcal{O}(N^{-1/2})$ arising from few-shot estimation. So whenever the number of prototypes $N \to \infty$ and $\sum_t \eta_t = \infty$, the gap $\Delta_t(\theta)$ vanishes. In particular, the adaptive prior family concentrates around $\nu^\star$ faster than any fixed-reference alternative, which under the Lipschitz continuity of $V_\nu(\theta)$ in $W_1(\nu,\cdot)$ directly supports the worst-case excess-risk comparison.

Theorem 4.3 (Consistency)**.** *Assume the ground cost $c(x,y)$ is continuous and bounded below, $f_\theta(\mathbf{x},\mathbf{y})$ is continuous in $\mathbf{x}$, $\mathbf{y}$ and locally Lipschitz in θ, the output space $\mathcal{Y}$ is compact, then for each $\mathbf{x}$ and class c,*

$$V_c^{(N)}(\mathbf{x};\theta) \to V_c^\star(\mathbf{x};\theta) \;\; \text{and} \;\; \lambda_c^{(N)}(\mathbf{x}) \to \lambda_c^\star(\mathbf{x}), \;\; (14)$$

where $V_c^{(N)}$ (resp. $V_c^\star$) is the Sinkhorn DRO dual with kernel induced by $\nu_c^{(N)}$ (resp. $\nu_c^\star$), and $\lambda_c^{(N)}$ (resp. $\lambda_c^\star$) is any minimizer of the corresponding dual variable.

This establishes robust risk consistency: solutions trained with finitely many supports converge to those of the population-level prototype-guided objective, thereby guarantee the validity of robust decisions and heir potential for generalization under distributional shift.

In summary, by equipping each class with an adaptive prior ν_c via adaptive OT, PG-DRO embeds class-specific reference distributions into Sinkhorn DRO. This unifies hierarchical priors with distributionally robust optimization, yielding robustness that adapts to the variability of novel classes, improves generalization under limited supervision, and strengthens the reliability of decision making in few-shot scenarios.

5 EXPERIMENTS

We aim to answer the following Research Questions (RQs) through our experiments:

- **RQ1:** Does PG-DRO improve generalization under distribution shifts compared to standard methods?
- **RQ2:** Can PG-DRO provide stronger robustness and decision quality in worst-case scenarios, particularly for weakly-supervised and severely shifted classes?
- **RQ3:** Does the adaptive OT prior in PG-DRO effectively generalize the cross-class knowledge to enhance the robustness of few-shot learners under limited supervision?

We next investigate these questions on both synthetic datasets and real benchmark datasets.

5.1 Simulation

To answer **RQ1**, we construct synthetic datasets that simulate distributional shifts between a source domain and a target domain. In the source domain, each of the C latent classes is modeled as a Gaussian

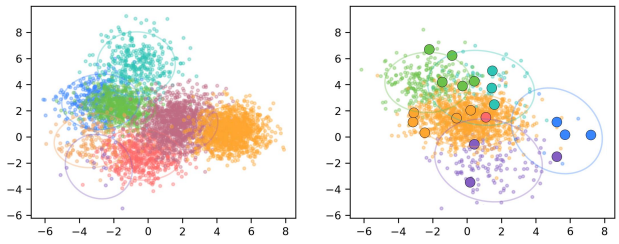
$$\mathbf{x} \sim \mathcal{N}(\mu_c, \Sigma_c), \quad c \in \{1,\dots,C\},$$

yielding supervised pairs $(\mathbf{x},\mathbf{z})$.

The target domain contains the same C classes, but their distributions are perturbed versions of the source ones. Specifically, we generate target Gaussians by applying

$$\mu_c' = \mu_c + \lambda_{\text{mean}}\Delta\mu_c, \qquad \Sigma_c' = \mathbf{R}\big(\lambda_{\text{cov}}\Sigma_c\big)\mathbf{R}^\top,$$

where $\Delta\mu_c$ denotes a mean displacement vector, $\mathbf{R}$ a random rotation matrix, and both are controlled by the hyperparameters λ_{mean} and λ_{cov}. To emulate the few-shot adaptation scenario, we retain only a limited number of labeled support samples per class in the target domain.



(a) Source domain per class $\mathcal{N}(\mu_c, \Sigma_c)$ (b) Target domain with K-shot supports

Figure 1: Comparison of source and target domain with 2D projection

Support P_s and P_t denote source and target domain respectively, we compare PG-DRO against two baselines.

- **Empirical Risk Minimization (ERM)**. ERM trains predictors only on labeled source-domain samples:

$$\mathcal{L}_{\text{ERM}} = \mathbb{E}_{(\mathbf{x}_s,\mathbf{z}_s)\sim P_s}\big[\ell\big(\mathbf{z}_s, f_\theta(\mathbf{x}_s)\big)\big], \tag{15}$$

  where ℓ evaluates decision quality, instantiated as cross-entropy for classification and squared error for regression.

- **Classical OT adaptation**. OT learns a transport map $\mathcal{T}$ that maps source distributions to target ones, minimizing

$$\mathcal{L}_{\text{OT}} = \mathbb{E}_{(\mathbf{x}_s,\mathbf{z}_s)\sim P_s,(\mathbf{x}_t,\mathbf{z}_t)\sim P_t}\big[\ell\big(\mathbf{z}_t, f_\theta(\mathcal{T}(\mathbf{x}_s))\big)\big], \tag{16}$$

  where $(\mathbf{x}_s, \mathbf{z}_s)$ with $\mathbf{z}_s = \mathbf{z}_t = c$ and $(\mathbf{x}_t, \mathbf{z}_t)$ are paired from the same class in source and target.

To answer **RQ2**, we assess model performance using two complementary metrics. Average accuracy over all target-domain test samples reflects overall generalization under distribution shift. We also report the worst-10% accuracy, defined as the accuracy restricted to the lowest-performing 10% of classes. This criterion highlights whether a method can sustain reliable predictions even for severely shifted or minority classes. For **RQ3**, we draw a heatmap to show the concentration degree to see if the adaptive OT could capture the mapping through the learned transport weights.

Few-shot simulation To emulate realistic prior shift, we sample domain-specific class proportions from Dirichlet distributions. The source domain uses a nearly uniform prior drawn from Dir(1), while the target domain adopts a long-tailed prior drawn from $\text{Dir}(\alpha_{\text{test}})$ with $\alpha_{\text{test}} \ll 1$. Because such priors naturally yield highly imbalanced class frequencies, the resulting support sets contain very few prototypes for minority classes, producing a challenging head–tail few-shot scenario.

Table 1: Results under different levels of disturbance

disturb λ_{cov}	model	avg acc.	worst 10% acc.
0.0	Pure ERM	0.07 ± 0.00	0.67 ± 0.02
	Classical OT	74.06 ± 1.20	95.68 ± 0.15
	PG-DRO	**78.10 ± 0.06**	**98.93 ± 0.75**
1.0	Pure ERM	10.03 ± 0.29	29.07 ± 1.78
	Classical OT	62.47 ± 0.11	46.33 ± 0.63
	PG-DRO	**65.31 ± 0.39**	**48.07 ± 0.38**
2.0	Pure ERM	17.14 ± 0.30	30.20 ± 0.61
	Classical OT	52.42 ± 0.26	32.40 ± 0.38
	PG-DRO	**54.47 ± 0.64**	**32.73 ± 0.32**
3.0	Pure ERM	20.01 ± 1.73	25.67 ± 1.53
	Classical OT	47.15 ± 0.27	26.40 ± 3.27
	PG-DRO	**51.39 ± 1.84**	**32.53 ± 1.45**
4.0	Pure ERM	21.03 ± 1.38	22.93 ± 0.49
	Classical OT	40.39 ± 0.16	28.00 ± 0.21
	PG-DRO	**49.73 ± 0.21**	**39.93 ± 0.13**
5.0	Pure ERM	21.68 ± 0.26	23.27 ± 0.95
	Classical OT	17.43 ± 0.13	22.00 ± 0.21
	PG-DRO	**32.53 ± 0.04**	**26.60 ± 0.49**

Based on this framework, we evaluate PG-DRO against other baselines on both classification and regression tasks to assess its decision-making ability. It is worth noting that all evaluations are performed on the target domain test set, ensuring that our reported metrics directly reflect generalization and robustness under domain shift. And all results are presented as mean ± standard deviation over repeated runs. All experiments are conducted on a machine equipped with 13th Gen Intel(R) Core(TM) i9-13900HX CPU (24 cores) and NVIDIA A100 GPU (40GB). The code is available at

5.1.1 Classification

For classification, each training instance is a pair $(\mathbf{x}, \mathbf{z})$, where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector generated from the Gaussian mixture construction, and $\mathbf{z} \in \{1, \ldots, C\}$ is its class label. We employ the standard cross-entropy loss (12), with logits obtained from a linear classification head applied to the encoder output. For the ERM and classic OT baselines, logits are computed directly from the encoder–head network. In contrast, for our PG-DRO, logits are given by the robust dual $V_c(\mathbf{x}; \theta)$, which integrates Sinkhorn-based class priors $\{\nu_c\}$ and solves λ_c via Newton updates. To emulate the few-shot setting in the target domain, we randomly sample between 3 and 8 labeled support instances per class which are used both for supervi-

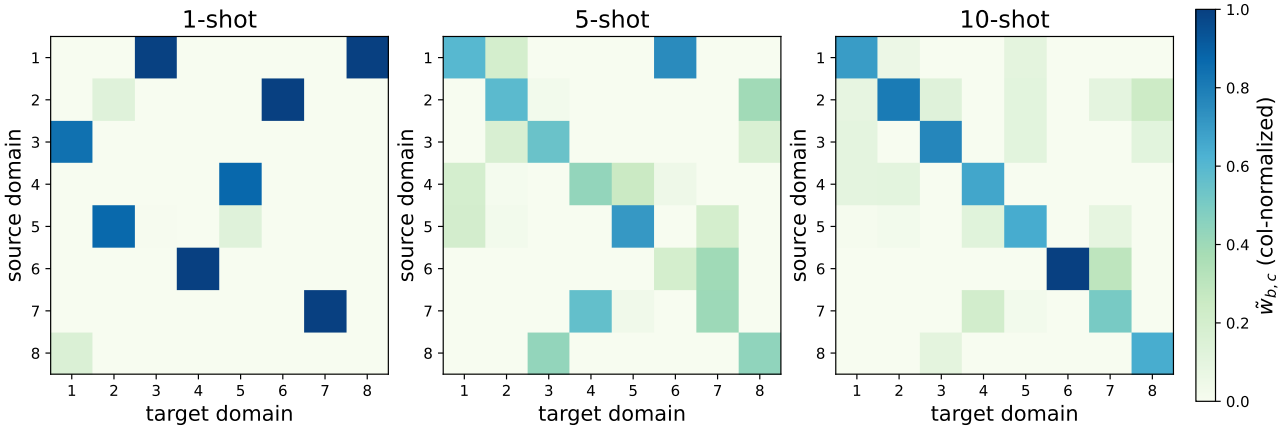


Figure 2: Normalized weights $\tilde{w}_{bc}$ for adaptive OT from source to target domain

sion and for refining the class-specific priors in our method. As shown in Table 1, PG-DRO consistently improves both average accuracy and worst-case performance compared to ERM and classical OT. And the test is carried out on the target domain. This highlights the benefit of integrating adaptive OT priors with Sinkhorn DRO, yielding robustness not only in overall generalization but also in the most challenging shifted classes.

To evaluate whether adaptive OT effectively captures the correct relationship between the source and target domains, we present the results of correlation mapping in Figure 2. The heatmaps clearly show that as the number of support samples increases, the alignments progressively concentrate along the diagonal. This indicates that with sufficient support data, target classes can be more accurately anchored to their corresponding source classes. The trend becomes increasingly clear, demonstrating sharper and more reliable alignments between source and target domains with more support.

5.1.2 Regression

We follow the same setting to generate source and target samples $(\mathbf{x}, \mathbf{z})$, where $\mathbf{x}$ is drawn from the synthetic geometry and the response is defined by $\mathbf{z} = \beta^\top \mathbf{x} + \varepsilon$. Thus we obtain paired data $(\mathbf{x}, \mathbf{z})$ for regression. Similarly, three regressors are trained with a shared encoder and a linear head. The ERM model and classical OT are optimized on training dataset using the (15) and (16) respectively. We also include our PG-DRO model minimizing a robust dual objective induced by the hierarchical OT prior. For adaptation, we sample a small number of labeled supports per target class as few-shot supervision and fine-tune accordingly. Evaluation is also carried out exclusively on the target domain. We report Mean Squared Error (MSE), Mean Absolute Error (MAE) set as well as worst-10% MSE, computed over the 10% largest absolute errors in the target distribution.

We train all models with Adam Optimizer Kingma and Ba (2014) using a short warm-up schedule. Figure 3 reports the regression performance of PG-DRO across varying numbers of classes and noise types. The results show that PG-DRO consistently achieves competitive results in all settings. As expected, the overall regression error increases as the noise level becomes larger, but the relative advantage of PG-DRO remains clear, demonstrating its robustness to distributional perturbations.

5.2 Image Classification

Beyond the synthetic experiments, we further evaluate the decision quality of PG-DRO on real-world vision benchmarks, including CIFAR-10 and CIFAR-100 Krizhevsky et al. (2009), as well as mini-ImageNet and tiered-ImageNet Ren et al. (2018). Due to space constraints, we present the CIFAR results in the main text and defer the ImageNet-family results to Appendix E.

We use CIFAR-100 and as the source domain and CIFAR-10 as the target domain, and evaluate adaptation performance under extremely low-label conditions to answer **RQ1**. Specifically, we first pretrain a ResNet-18 encoder E_ϕ on CIFAR-100, and then transfer it to CIFAR-10. For each target class, we randomly sample $S \in \{1, 5, 10\}$ labeled supports in CIFAR-10 dataset. For each query $\mathbf{x}$, we obtain its representation $m = E_\phi(\mathbf{x})$, and then compute the logits $\sigma_c(\mathbf{x}) = w_c^\top m + \epsilon_c$ according to different strategies.

To benchmark PG-DRO, we compare against two standard robust adaptation strategies.

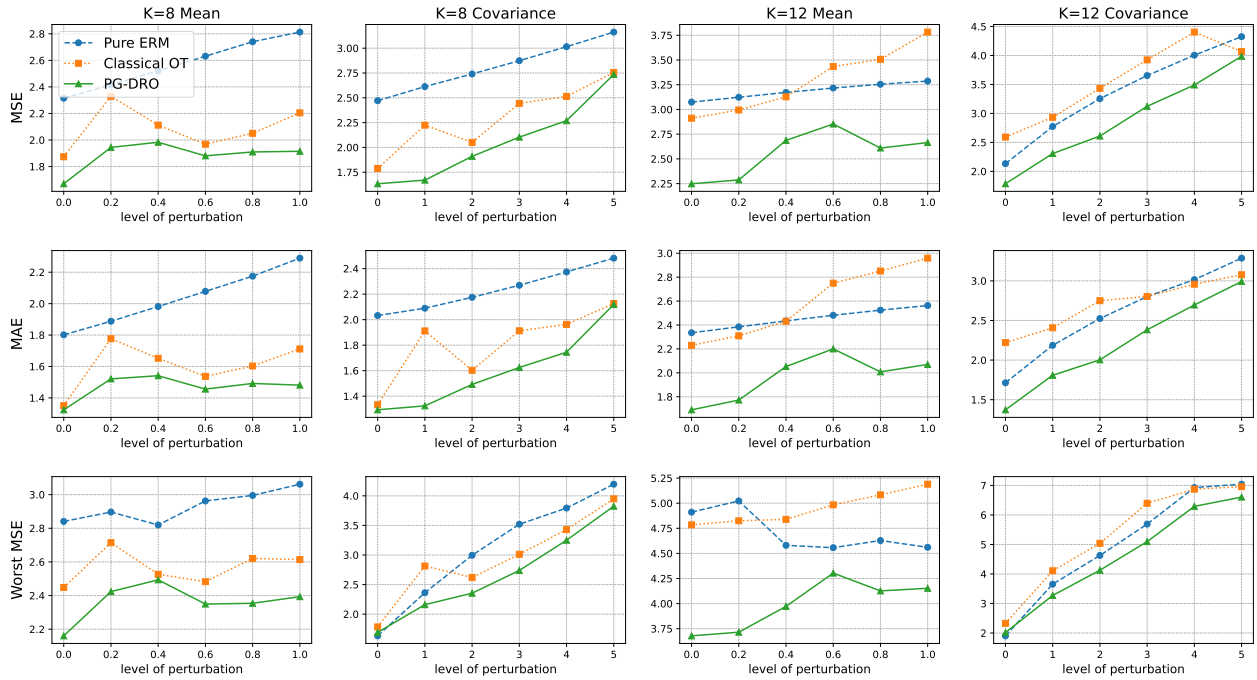


Figure 3: Comparison of robustness under prior shift: PG-DRO consistently improves both average accuracy and worst-case performance over ERM and OT baselines.

Table 2: Accuracy comparison for CIFAR-10 with Laplace Noise

	Laplace1			Laplace2			Laplace5		
Method	k=1	k=5	k=10	k=1	k=5	k=10	k=1	k=5	k=10
Few-shot	10.85 ± 0.44	27.25 ± 0.47	59.16 ± 0.50	10.03 ± 0.13	41.38 ± 0.41	52.75 ± 0.48	10.05 ± 0.20	25.58 ± 0.40	32.24 ± 0.33
SAA	18.81 ± 1.48	50.33 ± 0.61	59.56 ± 0.40	21.91 ± 0.77	48.03 ± 0.91	58.93 ± 0.67	15.25 ± 0.42	39.42 ± 0.83	49.68 ± 0.55
W-DRO	20.23 ± 0.18	49.34 ± 0.73	60.41 ± 0.50	22.07 ± 0.95	49.35 ± 0.73	58.15 ± 0.66	15.96 ± 0.81	40.49 ± 0.69	50.21 ± 0.28
PG-DRO	**20.88 ± 0.57**	**50.45 ± 0.54**	**61.54 ± 0.34**	**22.28 ± 0.71**	**50.21 ± 0.86**	**59.57 ± 0.45**	**17.71 ± 0.24**	**41.53 ± 0.39**	**50.58 ± 0.41**

- **Sample Average Approximation (SAA).** Augments each support $(\mathbf{x}_i, \mathbf{z}_i)$ with perturbations drawn from noise distribution $\mathcal{D}_{\text{noise}}$, and minimizes empirical risk $\mathbb{E}_{\eta\sim\mathcal{D}_{\text{noise}}}\big[\ell(\mathbf{z}_i, f_\theta(\mathbf{x}_i + \eta))\big]$ via Monte Carlo approximation.
- **Wasserstein DRO (W-DRO).** Builds on the score $z_c(\mathbf{x})$ but replaces the empirical distribution of supports with the worst-case distribution within a Wasserstein ball, thereby minimizing the worst-case expected loss.

For PG-DRO, we regard $\sigma_c(\mathbf{x})$ as the objective ℓ in (4) and strictly follow the steps of PG-DRO. We first build class-adaptive priors by performing hierarchical OT between base classes in CIFAR-100 and few-shot supports in CIFAR-10. Sinkhorn couplings are aggregated into weights $w_{b,c}$, yielding Gaussian mixture priors $\nu_c = \sum_b w_{b,c}\mathcal{N}(\mu_b, \Sigma_b)$ for each target class.

For a query $\mathbf{x}$, we kernelize its embedding $m = E_\phi(\mathbf{x})$ against ν_c, obtaining a posterior mixture. Rather than using the plain logit $w_c^\top m + \epsilon_c$, we compute the robust logit $V_c(\mathbf{x};\theta)$ by evaluating the robust objective with respect to the posterior mixture, thereby shaping the decision rule under worst-case perturbations. Training then proceeds with standard cross-entropy (12) over $V_c(\mathbf{x})$. Unlike W-DRO, our priors vary across classes and are induced by the adaptive OT, enabling improved robust generalization under low-label conditions, which could exactly answer **RQ3**.

In addition, we also compare with the simplest baseline, where we directly fine-tune a pretrained ResNet-18 on the few-shot supports from CIFAR-10. To answer **RQ2**, evaluation is conducted on the standard CIFAR-10 test set on perturbed features where the encoder output is corrupted as $m^\dagger = m + \eta$, with different levels of Gaussian or Laplace noise for testing both generalization and robustness.

To conclude, all experiments are decision-focused and end-to-end. We train by minimizing cross-entropy on the robust logits $V_c(\mathbf{x};\theta)$ with few-shot prototypes and

Table 3: Accuracy comparison for CIFAR-10 with Gaussian Noise

	Gaussian1			Gaussian2			Gaussian5		
Method	k=1	k=5	k=10	k=1	k=5	k=10	k=1	k=5	k=10
Few-shot	10.07 ± 0.15	26.11 ± 0.46	59.38 ± 0.35	9.95 ± 0.22	41.95 ± 0.33	52.87 ± 0.28	9.69 ± 0.34	25.48 ± 0.70	33.25 ± 0.46
SAA	16.37 ± 0.91	48.94 ± 0.67	60.15 ± 0.42	17.80 ± 0.75	48.68 ± 0.45	58.37 ± 0.76	17.11 ± 0.75	39.48 ± 0.47	50.17 ± 0.70
W-DRO	19.11 ± 1.75	50.70 ± 0.56	59.12 ± 1.25	17.48 ± 1.02	49.77 ± 0.53	58.93 ± 0.70	17.45 ± 1.11	39.45 ± 0.47	49.73 ± 0.60
PG-DRO	**20.53 ± 1.06**	**53.62 ± 1.08**	**60.78 ± 0.44**	**22.74 ± 1.44**	**51.00 ± 1.15**	**60.34 ± 0.48**	**18.80 ± 0.34**	**41.86 ± 0.37**	**50.47 ± 0.48**

therefore compare only to end-to-end baselines with the same training and evaluation interface under identical backbones and compute. Results in Table 2 and 3 show that PG-DRO consistently outperforms the baselines across different disturbance strengths, various shot settings, and under diverse noise conditions.

6 CONCLUSION

In this work, we introduced Prototype-Guided Distributionally Robust Optimization (PG-DRO), a framework that unifies hierarchical optimal transport priors with entropic DRO. By transferring structure from abundant base classes, PG-DRO constructs adaptive priors that embed robustness directly into the decision. This design enables the model not only to generalize from limited supervision but also to withstand distributional shifts and worst-case perturbations.

Theoretical analysis established that adaptive priors preserve consistency and reduce excess risk compared to the traditional DRO. Empirical results on synthetic and real datasets further validated our framework, showing consistent improvements in both accuracy and robustness under challenging few-shot settings. Future work will extend PG-DRO to online regimes, updating class-adaptive priors and dual variables on the fly as new supports arrive, with targets of no-regret guarantees under nonstationary shifts.

References

David Alvarez-Melis and T. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:33, 2023.

Xingjian Bai, Guangyi He, Yifan Jiang, and Jan Obloj. Wasserstein distributional robustness of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

José H. Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in wasserstein distributionally robust estimation. *Biometrika*, 2019.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49:1385–1418, 2015.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020.

Wanxing Chang, Ye Shi, Hoang Duong Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. In *Advances in Neural Information Processing Systems*, 2022.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020a.

Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9042–9051, 2020b.

Abhilash Reddy Chenreddy and Erick Delage. End-to-end conditional robust optimization. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Giorgio Costa and Garud N Iyengar. Distributionally robust end-to-end portfolio construction. *Quantitative Finance*, 23(10):1465–1482, 2023.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1853–1865, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas

Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *European Conference on Computer Vision*, 2018.

Ishani Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced wasserstein distance. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9015–9024, 2022.

Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. *Advances in neural information processing systems*, 33:8270–8283, 2020.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30, 2017.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115 – 166, 2015.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Chelsea Finn, P. Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Math. Oper. Res.*, 48:603–655, 2016.

Micah Goldblum, Liam H. Fowl, and Tom Goldstein. Robust few-shot learning with adversarially queried meta-learners. *ArXiv*, abs/1910.00982, 2019.

Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33:17886–17895, 2020.

Ido Greenberg, Shie Mannor, Gal Chechik, and Eli Meirom. Train hard, fight easy: Robust meta reinforcement learning. *Advances in Neural Information Processing Systems*, 36:68276–68299, 2023.

Dandan Guo, Long Tian, Minghe Zhang, Mingyuan Zhou, and Hongyuan Zha. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*, 2022a.

Dandan Guo, Long Tian, He Zhao, Mingyuan Zhou, and Hongyuan Zha. Adaptive distribution calibration for few-shot learning with hierarchical optimal transport. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b.

Yurong Guo, Ruoyi Du, Yuan Dong, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Task-aware adaptive learning for cross-domain few-shot learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1590–1599, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. Metric learning in optimal transport for domain adaptation. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2162–2168. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

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, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Meghan Beery, Jure Leskovec, Anshul B Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2020.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Kunde Rohde. Generalized sliced wasserstein distances. In *Neural Information Processing Systems*, 2019.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical Report, CIFAR-10 and CIFAR-100 datasets. `https://www.cs.toronto.edu/~kriz/cifar.html`.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiese-

mann. Distributionally robust optimization. *Acta Numerica*, 34:579 – 804, 2024.

Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13933–13941, 2020.

Xutao Ma, Chao Ning, and Wenli Du. Differentiable distributionally robust optimization layers. *ArXiv*, abs/2406.16571, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2216–2224, 2016a. ISBN 9781510838819.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Neural Information Processing Systems*, 2016b.

Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee, and Srijan Kumar. Adversarial robustness of prompt-based few-shot learning for natural language understanding. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Debabrata Pal, Deeptej More, Sai Bhargav, Dipesh Tamboli, Vaneet Aggarwal, and Biplab Banerjee. Domain adaptive few-shot open-set learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18785–18794, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11: 355–607, 2018.

Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open J. Math. Optim.*, 3:1–85, 2019.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially robust generalization requires more data. *ArXiv*, abs/1804.11285, 2018.

Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31408–31421. PMLR, 23–29 Jul 2023.

Umut Simsekli, Antoine Liutkus, Szymon Majewski, and Alain Oliviero Durmus. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, 2018.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, 2017b.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Neural Information Processing Systems*, 2016.

H. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Operations Research*, 2025.

Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. In *International Conference on Learning Representations*, 2021.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. *Advances in neural information processing systems*, 32, 2019.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12200–12210, 2020.

An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, Ji rong Wen, and Ping Luo. Domain-adaptive few-shot learning. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1389–1398, 2020.

Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes. We provide full mathematical definitions (Sinkhorn DRO, OT formulation, adaptive priors) and pseudocode in Section 3,4 and appendix.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes. The paper analyzes convexity of the dual objective, error bound, and consistency analysis in Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes. The code is available at `https://github.com/hSun08/PG-DRO`.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes. We explicitly lists assumptions in our proof in Appendix.
   (b) Complete proofs of all theoretical results. Yes. Proofs of Theorems 4.2 and 4.3 are included in Appendix.
   (c) Clear explanations of any assumptions. Yes. The meaning of smoothness, positivity, and few-shot sampling assumptions are explained in the method and Appendix.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes. Dataset sources are given and we have uploaded our code to supplementary materials.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes. Details on support size, optimizer, noise levels, and hyperparameter selection are described in Section 5 and Appendix.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes. All results are reported as mean $\pm$ standard deviation, with worst-10% accuracy clearly defined.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes. Intel i9-13900HX CPU and NVIDIA A100 GPU (40GB) are specified.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes. CIFAR-10/100 datasets are cited.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable. No human subjects involved.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

A Pseudo-code for PG-DRO

Algorithm 1 PG-DRO: Prototype-Guided DRO

Require: Base data $\mathcal{D}_{\text{base}}$, support set $\mathcal{S}$, model f_θ, cost c, radius ρ, temperature ε

Phase I: Prototype-Guided Priors
Compute statistics (μ_b, Σ_b) for each base class b
Solve hierarchical OT between $\mathcal{S}$ and $\mathcal{D}_{\text{base}} \to$ transport plan T
For each novel class c: set prior $\nu_c = \sum_b \tilde{w}_{bc} \mathcal{N}(\mu_b, \Sigma_b)$
Phase II: Robust Training
for each minibatch $\{(x, c^\star)\}$ **do**
  for each sample x and class c **do**
    Define Gibbs kernel $Q^{\nu_c}_{\varepsilon,x}$ from cost c and prior ν_c
    Solve 1D convex problem $\min_{\lambda \geq 0} \phi_c(\lambda)$
    Robust logit $V_c(x) = \phi_c(\lambda_c^\star)$
  end for
  Loss $L(x, c^\star) = -\log \frac{e^{V_{c^\star}}}{\sum_c e^{V_c}}$
  Update $\theta \leftarrow \theta - \eta \nabla_\theta L$
end for

B Proof of Theorem 4.2

Based on the following assumption, we proved Algorithm 4.2.

Assumption B.1 (Smooth lower level)**.** *$C(w)$ is C^1 and Lipschitz in a neighborhood of $w^\star$.*

Assumption B.2 (Interior/positivity)**.** *$\tau > 0$ and the optimal Sinkhorn coupling at $w^\star$ has strictly positive entries bounded away from zero; also $\min_b w_b^\star \geq \gamma > 0$.*

Assumption B.3 (Few-shot sampling)**.** *$\mathbb{E}\|\hat{b} - b\|_1 = \mathcal{O}(N_t^{-1/2})$ for a support size N_t.*

Assumption B.4 (Regularity for stability of V)**.** *Fix $\varepsilon > 0$ and define $k_x(y) := \exp(-c(x,y)/\varepsilon)$. There exist constants $0 < m \leq M < \infty$, $L_k < \infty$, $D < \infty$, $B_f < \infty$, $L_f < \infty$ such that for every x:*

1. *$m \leq k_x(y) \leq M$ and $\text{Lip}_y(k_x) \leq L_k$ on Y;*
2. *either Y is compact with $\text{diam}(Y) \leq D$, or all measures considered have uniformly bounded first moment so that 1-Lipschitz test functions are uniformly bounded by D;*
3. *$f_\theta(x, \cdot)$ is bounded and Lipschitz on Y: $|f_\theta(x,y)| \leq B_f$ and $\text{Lip}_y(f_\theta(x,\cdot)) \leq L_f$.*

Theorem 4.2 (Core contraction of DRO objective for V under adaptive OT) *Fix a target class c and omit the subscript c for brevity. Let $w^{(t+1)} = (1-\eta_t)w^{(t)} + \eta_t \widehat{T}(w^{(t)})$ with $\eta_t \in (0,1]$, and define $\nu^{(t)} = \sum_b w_b^{(t)} \mu_b$ and $\Delta_t(\theta) := |V_{\nu^{(t)}}(\theta) - V_{\nu^\star}(\theta)|$. Suppose the OT map T is locally contractive near $w^\star$ with contraction rate $\kappa \in (0,1)$. Then, for all t in the contraction neighborhood,*

$$\Delta_{t+1}(\theta) \;\leq\; (1 - \eta_t \kappa)\Delta_t(\theta) \;+\; \mathcal{O}(N^{-1/2}), \tag{17}$$

where κ is the contraction constant of the population map T around $w^\star$, and N is the number of few-shot samples.

Proof. We first prove the following lemma:

Lemma B.5. *Under Assumption B.4, there exists a constant $C < \infty$ such that for all mixtures μ, ν and all parameters θ,*

$$\left|V_\mu(\theta) - V_\nu(\theta)\right| \;\leq\; C W_1(\mu, \nu). \tag{18}$$

Proof. Write the tilted normalization $\Xi^x(\nu)(dy) := \dfrac{k_x(y)\nu(dy)}{\int k_x d\nu}$.

For any 1-Lipschitz $h : Y \to \mathbb{R}$,

$$\int h d(\Xi^x(\nu) - \Xi^x(\mu)) = \frac{\langle hk_x, \nu\rangle}{\langle k_x, \nu\rangle} - \frac{\langle hk_x, \mu\rangle}{\langle k_x, \mu\rangle} =: A + B.$$

Using $m \le \langle k_x, \cdot\rangle \le M$, $\|h\|_\infty \le D$, and $\mathrm{Lip}(hk_x) \le M \cdot 1 + L_k \cdot D$, we obtain

$$|A| \le \tfrac{M+L_kD}{m} W_1(\nu, \mu), \qquad |B| \le \tfrac{DML_k}{m^2} W_1(\nu, \mu).$$

Taking the supremum over h and invoking Kantorovich–Rubinstein duality yields

$$W_1\big(\Xi^x(\nu), \Xi^x(\mu)\big) \;\le\; K_x W_1(\nu, \mu), \quad K_x := \frac{M + L_kD}{m} + \frac{DML_k}{m^2}. \tag{19}$$

Fix any $\lambda_0 > 0$ and set $g(y) := f_\theta(x, y)/(\lambda_0\varepsilon)$. Then $\mathrm{Lip}(g) = L_f/(\lambda_0\varepsilon)$ and $|g| \le B_f/(\lambda_0\varepsilon)$, so $\mathrm{Lip}(e^g) \le e^{B_f/(\lambda_0\varepsilon)} \cdot L_f/(\lambda_0\varepsilon)$ and $\min\{\mathbb{E}[e^g]\} \ge e^{-B_f/(\lambda_0\varepsilon)}$. Hence

$$\Big|\lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\nu)}[e^g] - \lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\mu)}[e^g]\Big| \le L_f e^{\frac{2B_f}{\lambda_0\varepsilon}}\; W_1\big(\Xi^x(\nu), \Xi^x(\mu)\big).$$

Combine with (19) to get

$$\Big|\lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\nu)}[e^g] - \lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\mu)}[e^g]\Big| \le L_f e^{\frac{2B_f}{\lambda_0\varepsilon}}\; K_x\; W_1(\nu, \mu). \tag{20}$$

By definition of V (the inner minimization over $\lambda > 0$), for any fixed $\lambda_0 > 0$,

$$V_\nu(x) \;\le\; \lambda_0\rho + \lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\nu)}[e^g], \qquad V_\mu(x) \;\le\; \lambda_0\rho + \lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\mu)}[e^g],$$

so

$$|V_\nu(x) - V_\mu(x)| \;\le\; \Big|\lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\nu)}[e^g] - \lambda_0\varepsilon \log \mathbb{E}_{\Xi^x(\mu)}[e^g]\Big|.$$

Apply (20) and take $\sup_x K_x$ (finite by Assumption B.4) to conclude

$$|V_\nu(\theta) - V_\mu(\theta)| \;\le\; CW_1(\nu, \mu),$$

with C as stated. □

Then we prove the theorem:

Fix a dictionary of measures $\{\mu_b\}_{b=1}^B$. Any prior is a mixture $\nu = \sum_{b=1}^B w_b\mu_b$ with $w \in \Delta_B := \{w \ge 0,\ \sum_b w_b = 1\}$. The upper level OT uses entropic OT with cost matrix $C(w) \in \mathbb{R}^{B\times m}$ and kernel

$$K(w) := \exp\big(-C(w)/\tau\big) \in \mathbb{R}^{B\times m}_{++}, \qquad \tau > 0.$$

Its Sinkhorn coupling has the form

$$\Pi(w) = \mathrm{diag}(u)K(w)\mathrm{diag}(v), \quad u \odot (K(w)v) = w, \quad v \odot (K(w)^\top u) = b,$$

where $b \in \Delta_m$ is the population target marginal. The population map from w to new mixture weights is

$$T(w) = \mathrm{Norm}\big(R\Pi(w)\mathbf{1}_m\big) \in \Delta_B,$$

with a fixed linear aggregator R (often $R = I$). The empirical map $\widehat{T}(w)$ replaces b by an empirical $\hat{b}$. The algorithm uses the damped update

$$w^{(t+1)} = (1 - \eta_t)w^{(t)} + \eta_t\widehat{T}(w^{(t)}), \qquad \eta_t \in (0, 1]. \tag{21}$$

Distances between mixtures are measured by W_1.

Write the upper-level Lagrangian

$$\mathcal{L}(\Pi, \alpha, \beta; w) = \langle C(w), \Pi \rangle + \tau \sum_{ij} \Pi_{ij}(\log \Pi_{ij} - 1) + \alpha^\top(\Pi \mathbf{1} - w) + \beta^\top(\Pi^\top \mathbf{1} - b).$$

Because $\tau > 0$, the Hessian in Π on the constraint tangent space obeys $\nabla^2_{\Pi\Pi} \succeq \tau \mathrm{Diag}(1/\Pi^\star) \succeq \tau/m_\Pi \cdot I$ with $m_\Pi = \min_{ij} \Pi^\star_{ij} > 0$ by Assumption B.2. The KKT system $F(\Pi, \alpha, \beta; w) = 0$ has a locally unique solution $(\Pi^\star(w), \alpha^\star(w), \beta^\star(w))$ by the implicit function theorem. Differentiating in w yields

$$D\Pi^\star(w^\star) = -\big(\nabla^2_{\Pi\Pi}\mathcal{L}\big)^{-1} \nabla^2_{\Pi w}\mathcal{L},$$

so $\|D\Pi^\star(w^\star)\| \le L_C/(\tau/m_\Pi) =: L_\Pi$ by Assumption B.1. The map $w \mapsto T(w) = \mathrm{Norm}(R\Pi^\star(w)\mathbf{1})$ is a composition of linear maps and a smooth normalization in an interior neighborhood, hence $\|DT(w^\star)\| \le c_{\mathrm{agg}} L_\Pi$ for some finite constant c_{agg}. Choosing (or verifying) parameters so that $c_{\mathrm{agg}} L_\Pi < 1$ and shrinking the neighborhood if necessary gives $\sup_{w \in \mathcal{U}} \|DT(w)\| \le 1 - \kappa$, which implies

$$\|T(w) - T(w^\star)\|_1 \;\le\; (1 - \kappa)\|w - w^\star\|_1. \tag{22}$$

by the mean-value theorem.

Applying the implicit function argument w.r.t. the marginal b gives $\|\widehat{\Pi}(w) - \Pi(w)\| \le L_b \|\hat{b} - b\|_1$. Aggregation and normalization are Lipschitz in an interior neighborhood, hence $\|\widehat{T}(w) - T(w)\|_1 \le c_{\mathrm{agg}} L_b \|\hat{b} - b\|_1$. Taking expectations and using Assumption B.3 yields

$$\mathbb{E}\big[\|\widehat{T}(w) - T(w)\|_1 \big| N_t\big] \;\le\; c_T \mathbb{E}\|\hat{b} - b\|_1 \;=\; \mathcal{O}(N_t^{-1/2}). \tag{23}$$

With a fixed dictionary supported on a set of diameter D, we have $W_1(\sum_b (w_b - w'_b)\mu_b) \le (D/2)\|w - w'\|_1$. Set $L_{\mathrm{mix}} := D/2$ to obtain

$$W_1\Big(\sum_b w_b \mu_b, \sum_b w'_b \mu_b\Big) \;\le\; L_{\mathrm{mix}} \|w - w'\|_1, \tag{24}$$

where all $w, w' \in \Delta_B$.

By convexity of W_1 and the update (21),

$$W_1(\nu^{(t+1)}, \nu^\star) \le (1 - \eta_t) W_1(\nu^{(t)}, \nu^\star) + \eta_t W_1(\widehat{T}(w^{(t)}), \nu^\star).$$

Using the triangle inequality with Steps 1–3 gives $W_1(\widehat{T}(w^{(t)}), \nu^\star) \le (1 - \kappa) W_1(\nu^{(t)}, \nu^\star) + B + \xi_t$, which yields

$$E_{t+1} \;\le\; (1 - \eta_t \kappa) E_t + \eta_t\big(B + \xi_t\big), \qquad \mathbb{E}[\xi_t | N_t] = \mathcal{O}(N_t^{-1/2}). \tag{25}$$

In particular, if $B = 0$ and $N_t \to \infty$ (or $\sum_t \eta_t \mathbb{E}\xi_t < \infty$), then $E_t \to 0$; with constant $\eta_t \equiv \eta$,

$$\limsup_{t \to \infty} E_t \;\le\; \frac{B + \overline{\xi}}{\kappa}, \qquad \overline{\xi} := \limsup_t \mathbb{E}\xi_t.$$

Apply (18) to transfer bounds from E_t to $\Delta_t(\theta)$; standard epi-convergence arguments then yield convergence of L_t and minimizers.

$$\Delta_t(\theta) := |V_{\nu^{(t)}}(\theta) - V_{\nu^\star}(\theta)| \;\le\; C E_t,$$

so $\Delta_t(\theta)$ satisfies the same contraction conclusions as E_t. Under standard integrability/strict convexity already assumed in the DRO dual, $L_t(\theta) \to L^\star(\theta)$ and (when unique) minimizers $\theta_t \to \theta^\star$. □

Multi-class extension. Apply the theorem per class c and sum over c; all constants can be chosen uniformly on a neighborhood where each class remains interior.

C Proof of Theorem 4.3

Theorem 4.3 (Consistency) *Assume the ground cost $c(x, y)$ is continuous and bounded below, $f_\theta(\mathbf{x}, \mathbf{y})$ is continuous in $\mathbf{x}$, $\mathbf{y}$ and locally Lipschitz in θ, the output space $\mathcal{Y}$ is compact, then for each $\mathbf{x}$ and class c,*

$$V_c^{(N)}(\mathbf{x};\theta) \to V_c^\star(\mathbf{x};\theta) \quad \text{and} \quad \lambda_c^{(N)}(\mathbf{x}) \to \lambda_c^\star(\mathbf{x}), \tag{26}$$

where $V_c^{(N)}$ (resp. $V_c^\star$) is the Sinkhorn DRO dual with kernel induced by $\nu_c^{(N)}$ (resp. $\nu_c^\star$), and $\lambda_c^{(N)}$ (resp. $\lambda_c^\star$) is any minimizer of the corresponding dual variable.

Proof. By the definition of Gibbs kernel

$$dQ_{\varepsilon,x}^{\nu}(y) = \frac{e^{-c(x,y)/\varepsilon}}{\int e^{-c(x,u)/\varepsilon} d\nu(u)} d\nu(y).$$

If $\nu^{(N)} \Rightarrow \nu^\star$, then $Q_{\varepsilon,x}^{\nu^{(N)}} \Rightarrow Q_{\varepsilon,x}^{\nu^\star}$. Indeed, since c is continuous and bounded below, $k_x(y) = e^{-c(x,y)/\varepsilon}$ is bounded and continuous, hence $\int k_x d\nu^{(N)} \to \int k_x d\nu^\star$. For any bounded continuous g,

$$\int g dQ_{\varepsilon,x}^{\nu^{(N)}} = \frac{\int g k_x d\nu^{(N)}}{\int k_x d\nu^{(N)}} \to \frac{\int g k_x d\nu^\star}{\int k_x d\nu^\star} = \int g dQ_{\varepsilon,x}^{\nu^\star}.$$

For simplicity, we denote

$$\Phi_c^{(N)}(\lambda; x) = \lambda\rho + \lambda\varepsilon \log \mathbb{E}_{y\sim Q_{\varepsilon,x}^{\nu_c^{(N)}}} \exp\{f(x,y)/(\lambda\varepsilon)\},$$

and similarly $\Phi_c^\star$ with $\nu_c^\star$. Since for any compact $\Lambda \subset (0, \infty)$,

$$\sup_{\lambda\in\Lambda} |\Phi_c^{(N)}(\lambda; x) - \Phi_c^\star(\lambda; x)| \to 0.$$

f bounded and continuous $\Rightarrow g_\lambda(y) = \exp(f(x, y)/(\lambda\varepsilon))$ bounded continuous and equicontinuous in λ, so expectations converge uniformly.

From (Wang et al., 2025, Thm. 1(IV), Lem. EC.5), we have

$$\lim_{\lambda\to\infty} \Phi_\nu(\lambda; x) = \lim_{\lambda\to\infty} \left(\lambda\rho + \lambda\varepsilon \mathbb{E}_x \log \mathbb{E}_{Q_{x,\varepsilon}} e^{f/(\lambda\varepsilon)}\right) = +\infty \quad (\rho > 0),$$

so minimizers cannot escape to infinity. Similarly, the condition for $\lambda^\star = 0$ is characterized therein. Thus all minimizers lie in some compact interval $\Lambda \subset (0, \infty)$ independent of N. Moreover, by lemma 3.1, $\Phi_\nu(\cdot; x)$ is convex in λ, strictly convex unless f is a.s. constant, hence the minimizer is unique (except trivial degeneracy). Therefore $\Phi_c^{(N)} \to \Phi_c^\star$ uniformly on Λ. By standard convex optimization stability (epi-convergence plus uniqueness),

$$\lambda_c^{(N)}(x) \to \lambda_c^\star(x).$$

Since we already proved uniform convergence and argmin convergence, we have

$$V_c^{(N)}(x) = \Phi_c^{(N)}(\lambda_c^{(N)}(x); x) \to \Phi_c^\star(\lambda_c^\star(x); x) = V_c^\star(x).$$

Besides, in the multi-class setting we use the robust logits $V_c(x)$ inside the softmax cross-entropy. Under $f_\theta(x, \cdot)$ is continuous (hence bounded), the terms $\exp(V_c^{(N)})$ are uniformly integrable. Hence by dominated convergence,

$$\log \sum_c e^{V_c^{(N)}(x;\theta)} \to \log \sum_c e^{V_c^\star(x;\theta)}.$$

Thus $\mathcal{L}_N(\theta) \to \mathcal{L}_\star(\theta)$ pointwise and locally uniformly in θ. By epi-convergence of convex functions, any limit point of minimizers θ_N lies in $\arg\min \mathcal{L}_\star$, and uniqueness implies $\theta_N \to \theta^\star$. □

D Hyperparameters

To produce our experimental results, the hyperparameter are list as follows:

Classification: We use $K = 8$ classes with input dimensions $d = 10$. We sample $n_{\text{train}} = 6000$ and $n_{\text{test}} = 3000$ points, with the target domain mixture ratio set to $\alpha_{\text{test}} = 0.15$. For distributional perturbations, we apply a mean shift of 0.6, covariance scaling of 1.0, and a 15-degree rotation. Optimal transport is regularized with entropic weight 0.2. For DRO, we set the $\varepsilon_{\text{sample}} = 1.0$, $\varepsilon_{\text{class}} = 0.8$, and balance parameter ρ as 1.0. The model is trained using a hidden dimension of 256, batch size 256, learning rate 10^{-3}, and 200 epochs, with Newton iterations set to 8. We also report performance on the worst 10% quantile of test samples to assess robustness.

Regression: Similarly as the classification task, we also use $K = 8$ classes with input dimension $d = 10$. We sample $n_{\text{train}} = 6000$ and $n_{\text{test}} = 3000$ with target mixture $\alpha_{\text{test}} = 0.20$. Shifts are applied with mean shift 0.6, covariance scaling 1.15, rotation 15°. Optimal transport uses entropic weight 0.2; hierarchical OT priors use $\varepsilon_{\text{sample}} = 1.0$, $\varepsilon_{\text{class}} = 0.8$ (200 iterations), and covariance inflation 3.0. The regression targets use $w \sim \mathcal{N}(0, I_d)$ and additive noise $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. Training uses a single hidden layer of size 256, batch size 256, learning rate 10^{-3}, Huber loss with beta = 1.0, and 200 epochs. The DRO prior penalty employs temperature 0.1 and loss weight $\lambda = 1.0$. Evaluation reports RMSE, MAE and worst-10% RMSE.

Image Classification: We use a ResNet-18 encoder with feature dimension 512 and a linear head. Few-shot supports are sampled with $S \in \{1, 5, 10\}$ per class. Training uses batch size 64, learning rate 10^{-4}, and 50 epochs. In PG-DRO, we set the sample-level softmin temperature $\varepsilon_{\text{sample}} = 0.1$, the class-level entropic weight $\varepsilon_{\text{class}} = 0.1$ with 100 Sinkhorn iterations, and a covariance inflation factor of 1.5. The ambiguity radius is fixed to $\rho = 5.0$, and the robust logit solver runs 5 Newton iterations. Evaluation is performed on both clean features and perturbed ones, with Gaussian or Laplace noise of relative radius $\epsilon \in \{1, 2, 5\}$.

E Additional experimental results

In the main text we reported results on CIFAR-based transfer tasks. This appendix provides additional few-shot cross-domain image classification results on ImageNet-family benchmarks under feature-space distribution shifts. Following Section 5.2, we evaluate robustness by corrupting the encoder representation with additive Laplace or Gaussian noise. We report mean accuracy (in %) ± standard deviation over repeated runs for $k \in \{1, 5, 10\}$ labeled supports per class.

Overall, PG-DRO consistently achieves the best robust accuracy across all three transfer settings and both noise types. The advantage over robust baselines typically grows as more supports are available, which aligns with the motivation that adaptive OT priors can better leverage limited support information to form class-specific reference distributions for Sinkhorn DRO. Besides, we can observe when using more complex and stronger datasets, the performance gains of PG-DRO become more pronounced over merely simple CIFAR-10 based baselines, which further demonstate the advantage of PG-DRO.

Table 4: Accuracy comparison for miniImageNet→CIFAR-100 with noise

	Laplace2			Gaussian2		
Method	k=1	k=5	k=10	k=1	k=5	k=10
Few-shot	1.05 ± 0.13	7.95 ± 0.27	11.55 ± 0.30	1.12 ± 0.10	8.02 ± 0.21	12.10 ± 0.26
SAA	4.95 ± 0.20	12.14 ± 0.57	17.32 ± 0.47	4.84 ± 0.14	12.49 ± 0.57	18.13 ± 0.49
W-DRO	6.93 ± 0.31	14.53 ± 0.71	22.97 ± 0.71	6.10 ± 0.21	13.61 ± 1.17	21.37 ± 1.08
PG-DRO	**7.45 ± 0.36**	**18.78 ± 0.64**	**28.11 ± 0.62**	**7.84 ± 0.42**	**17.86 ± 0.52**	**26.90 ± 0.48**

Table 5: Accuracy comparison for tieredImageNet→CIFAR-100 with Noise

	Laplace2			Gaussian2		
Method	k=1	k=5	k=10	k=1	k=5	k=10
Few-shot	1.01 ± 0.18	8.41 ± 0.25	11.61 ± 0.35	1.15 ± 0.14	7.93 ± 0.29	11.96 ± 0.21
SAA	6.20 ± 0.16	23.94 ± 0.35	31.15 ± 0.25	5.40 ± 0.45	24.65 ± 0.37	32.58 ± 0.32
W-DRO	7.06 ± 0.29	25.66 ± 0.58	34.85 ± 0.30	6.92 ± 0.33	24.31 ± 0.58	34.35 ± 0.51
PG-DRO	**9.21 ± 0.47**	**29.95 ± 0.44**	**40.53 ± 0.36**	**9.26 ± 0.24**	**30.25 ± 0.41**	**40.05 ± 0.04**

Table 6: Accuracy comparison for tieredImageNet→miniImageNet with Noise

	Laplace2			Gaussian2		
Method	k=1	k=5	k=10	k=1	k=5	k=10
Few-shot	1.12 ± 0.09	9.43 ± 0.14	13.99 ± 0.34	1.03 ± 0.08	9.10 ± 0.13	14.03 ± 0.36
SAA	20.57 ± 0.41	30.89 ± 0.76	39.60 ± 1.28	15.37 ± 0.27	27.47 ± 1.55	38.15 ± 0.32
W-DRO	19.79 ± 0.43	33.88 ± 0.37	41.87 ± 0.59	16.48 ± 0.93	32.30 ± 0.35	43.77 ± 0.12
PG-DRO	**22.10 ± 0.75**	**40.04 ± 0.39**	**49.66 ± 0.70**	**19.95 ± 0.44**	**39.73 ± 1.15**	**50.84 ± 0.85**