TOWARD DISTRIBUTIONALLY ROBUST DECISION-FOCUSED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decision-focused learning often relies on point predictions or decides under a single learned distribution, leaving downstream decisions vulnerable to model error. We propose a distributionally robust framework for decision-focused learning that learns a conditional distribution and conformally calibrates a Wasserstein ambiguity set around it. We use empirical Sinkhorn divergences as nonconformity scores, with corrections for entropic regularization and finite-sample estimation error. Under exchangeability and suitable concentration assumptions, the resulting Wasserstein-1 and squared Wasserstein-2 ambiguity sets contain the true conditional distribution with finite-sample marginal coverage. On the coverage event, the robust objective upper-bounds the true conditional expected task loss. We derive gradients through calibration and downstream optimization under regularity conditions, enabling end-to-end task-aware training. Experiments on portfolio allocation demonstrate improved decision quality over two-stage and stochastic decision-focused baselines.

1 INTRODUCTION

In learning-based decision pipelines such as portfolio allocation, energy scheduling, and resource allocation, the final objective is an optimal decision induced by an intermediate prediction. Standard prediction losses may penalize errors that are irrelevant to the downstream decision, while possibly underweighting errors that change the optimal decision. Decision-focused learning (DFL) addresses this mismatch by training a predictive model to align with a downstream optimization task, using either differentiable optimization layers (Donti et al., 2017; Amos & Kolter, 2017; Agrawal et al., 2019) or decision-aware losses and its convex surrogate such as SPO+ (Elmachtoub & Grigas, 2022).

Despite the integration of prediction and decision-making, much of the DFL literature omits accounting for the uncertainty of the prediction and its influence in downstream decision-making (Amos & Kolter, 2017; Agrawal et al., 2019; Elmachtoub & Grigas, 2022). This omission limits the value of DFL in stochastic environments. In such cases, a point predictor can be accurate under a standard statistical loss while discarding distributional information that is decisive for the optimizer. This limitation has recently motivated stochastic DFL methods that learn richer conditional distributions rather than single nominal parameters (Zhao et al., 2026). Yet a learned distribution $P_\theta(\cdot \mid \boldsymbol{x})$ remains an estimate due to finite data and model misspecification. Representing uncertainty therefore does not eliminate the risk of optimizing under an inaccurate distribution.

Distributionally robust optimization (DRO) addresses this model risk by optimizing over an ambiguity set of plausible distributions (Blanchet & Murthy, 2019; Kuhn et al., 2019). An ambiguity set that is too small can miss the true distribution, whereas an overly large set can lead to unnecessarily conservative decisions. Its size must therefore be statistically calibrated. For distributional predictors, this requires controlling the error of distributional discrepancies estimated from finite samples while allowing gradients to propagate through calibration and optimization.

We address these challenges through distributionally robust decision-focused learning (DRDFL). Our method learns a conditional distribution for the uncertain parameter and constructs a Wasserstein-p ambiguity set $\{Q : W_p^p(Q, P_\theta(\cdot \mid \boldsymbol{x})) \leq \rho_\theta\}$ for an ambiguity radius ρ_θ around the predicted distribution followed by solving a downstream distributionally robust optimization problem. The key design choice is the ambiguity radius ρ_θ. We use empirical Sinkhorn divergences as nonconformity scores and calibrate ρ_θ using a conformal quantile corrected for entropic regularization and finite-sample estimation error. Final calibration is performed on held-out data after the predictor is frozen.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We summarize our main contributions below:

(i) **End-to-end learning.** We develop a distributionally robust DFL framework which trains distribution predictors by minimizing the task loss of the robust decisions induced by their predictions. The framework integrates distribution prediction, calibration of uncertainty, and robust optimization within a single training procedure (Section 4).

(ii) **Calibration with Sinkhorn scores.** We use Sinkhorn divergences as tractable surrogates for Wasserstein distances when calibrating the ambiguity of the predicted distributions, which robustifies the task-aware training of the predictive models. To our best knowledge, this is the first work integrating conformal calibration for distributions in DFL by using Sinkhorn divergences as the nonconformity scores (Section 3).

(iii) **Finite-sample guarantees.** We derive corrections for entropic regularization and sampling error that yield finite-sample marginal coverage guarantees for Wasserstein-1 and squared Wasserstein-2 ambiguity sets. On the coverage event, the DRO objective also upper-bounds the true conditional expected task loss (Section 3.1 and Section 3.2).

(iv) **Empirical evaluation.** We evaluate DRDFL on portfolio allocation, demonstrating improved decision quality over two-stage predict-then-optimize and stochastic DFL baselines (Section 5).

1.1 Related Work

Decision-focused and generative learning. DFL, also known as task-based learning, trains predictive models informed by the quality of decisions induced by their predictions rather than the prediction error alone. Key approaches include task-based learning for stochastic optimization (Donti et al., 2017), differentiable optimization layers (Amos & Kolter, 2017; Agrawal et al., 2019), and surrogate decision losses such as SPO and SPO+ (Elmachtoub & Grigas, 2022). These approaches typically predict the parameters of a downstream optimization problem but do not explicitly protect the resulting decisions against distributional misspecification. Recent extensions instead learn full conditional distributions. Diffusion-DFL (Zhao et al., 2026) uses a conditional diffusion model, and optimizes decisions using samples from the learned distribution, while Gen-DFL (Wang et al., 2026) trains a conditional generator using CVaR-based decision regret. Although these approaches capture stochasticity and, in the latter case, tail risk, they optimize under the learned distribution itself. Our method instead constructs a conformally calibrated ambiguity set around the learned conditional law $P_\theta(\cdot \mid \boldsymbol{x})$, and thereby accounts for misspecification of that law.

Conformal and distributionally robust decision-making. DRO minimizes the worst-case expected loss over an ambiguity set of distributions. Wasserstein ambiguity sets are particularly useful because they allow perturbations of the distributional support and often be tractable (Kuhn et al., 2019). Conformal methods provide a data-driven mechanism for calibrating uncertainty sets. For example, Yeh et al. (2025) calibrate sets of possible parameter realizations for conditional robust optimization, whereas we calibrate a set of distributions centered at $P_\theta(\cdot \mid \boldsymbol{x})$. More closely related, Patel et al. (2024) perform conformal prediction over distributions and use the resulting region for downstream stochastic optimization. In our framework, however, the calibrated ambiguity set is embedded in an end-to-end DFL objective, allowing predictive distributions to receive direct feedback from the downstream robust decision loss. Gen-WDRO (Xu & Yeh, 2026) is closest to our architecture, as it combines a conditional normalizing flow with a differentiable Wasserstein DRO layer. Its ambiguity radius is learned from the task objective, whereas ours is determined by split conformal calibration of the predicted distribution. This gives the radius an explicit statistical interpretation and yields a prescribed finite-sample marginal coverage guarantee together with a downstream decision guarantee.

Generative ambiguity modeling. There are other approaches parametrizing the distributional ambiguity using generative models. WEN & Yang (2026) constrain adversarial generators through their reconstruction loss relative to a fixed nominal distribution, while Zhang et al. (2026) robustify decisions from a frozen conditional sampler using a generator-family adversary constrained by Sinkhorn divergence. Our framework instead jointly learns the nominal conditional distribution, conformally quantifies distributional uncertainty around it, and propagates that uncertainty through an end-to-end DRO layer. It also supports risk-aware mean–CVaR objectives.

2 Problem Statement

We consider a stochastic decision problem in which a decision $\boldsymbol{z} \in \mathbb{R}^{n_z}$ must be made based on observed features $\boldsymbol{x} \in \mathbb{R}^{n_x}$, while the task loss $f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{c})$ depends on an uncertain parameter $\boldsymbol{c} \in \mathbb{R}^{n_c}$.

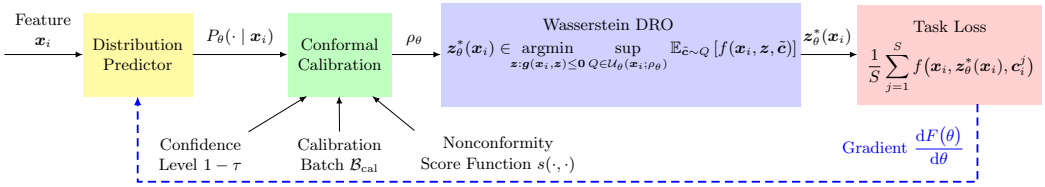


Figure 1: An illustration of the DRDFL framework, processing one data point $(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S)$.

We denote the true conditional distribution of $\boldsymbol{c}$ given $\boldsymbol{x}$ by $P_\star(\cdot|\boldsymbol{x})$. If this distribution were known, the optimal decision for a given $\boldsymbol{x}$ could be obtained by solving

$$\min_{\boldsymbol{z}:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\leq \boldsymbol{0}} \mathbb{E}_{\boldsymbol{c}\sim P_\star(\cdot|\boldsymbol{x})}[f(\boldsymbol{x},\boldsymbol{z},\boldsymbol{c})],$$

where $\boldsymbol{g} = (g_1, \ldots, g_{n_g}) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_g}$ constrain the feasible decisions.

Our goal is to learn a decision rule that accounts for uncertainty in the learned conditional distribution when $P_\star(\cdot|\boldsymbol{x})$ is unknown. To this end, assume access to data generated as follows. Let P be the joint distribution of $(\boldsymbol{x}, \boldsymbol{c})$, with the marginal $P_{\boldsymbol{x}}$ over features. The features $\boldsymbol{x}_i$ are sampled i.i.d. from $P_{\boldsymbol{x}}$, and, conditional on each $\boldsymbol{x}_i$, we observe S independent samples $\boldsymbol{c}_i^1, \ldots, \boldsymbol{c}_i^S$ from $P_\star(\cdot|\boldsymbol{x}_i)$. Hence, each data point is the tuple $(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S)$. The training, calibration, and test datasets, denoted by $\mathcal{D}_{\text{tra}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{test}}$, consist of N_{tra}, N_{cal}, and N_{tes} such data points, respectively.

Proposition 2.1 (Exchangeability)**.** *The calibration and test data points are exchangeable.*

See Appendix C for the proof.

Since the true conditional distribution is unknown, DFL relies on a parameterized predictive distribution $P_\theta(\cdot|\boldsymbol{x}) \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c})$, where $\mathcal{P}_{1,2}(\mathbb{R}^{n_c})$ is the set of Borel probability measures with finite first and second moments. Given $P_\theta(\cdot|\boldsymbol{x})$, the downstream optimization problem defines a decision. The predictive model is then trained end-to-end by minimizing the empirical task loss for this decision.

However, since the downstream decision is optimized for $P_\theta(\cdot|\boldsymbol{x})$, discrepancies between $P_\theta(\cdot|\boldsymbol{x})$ and $P_\star(\cdot|\boldsymbol{x})$ can be amplified by the optimization problem, leading to poor out-of-sample decisions. To account for this distributional uncertainty, we replace the nominal downstream problem with a DR counterpart. For a given feature vector $\boldsymbol{x}$, we consider the Wasserstein ambiguity set

$$\mathcal{U}_\theta(\boldsymbol{x};\rho_\theta) := \left\{Q \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c}) : W_p^p(Q, P_\theta(\cdot \mid \boldsymbol{x})) \leq \rho_\theta\right\},$$

where $p \in \{1, 2\}$ indicate Wasserstein-1 and the squared Wasserstein-2 distances, while $\rho_\theta > 0$ controls the size of the ambiguity set. We then formulate the DRDFL problem as

$$\min_\theta \frac{1}{N_{\text{tra}} S} \sum_{i=1}^{N_{\text{tra}}} \sum_{j=1}^{S} f\left(\boldsymbol{x}_i, \boldsymbol{z}_\theta^*(\boldsymbol{x}_i), \boldsymbol{c}_i^j\right) \quad \text{s.t.} \quad \boldsymbol{z}_\theta^*(\boldsymbol{x}) \in \underset{\boldsymbol{z}:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\leq \boldsymbol{0}}{\operatorname{argmin}} \sup_{Q\in\mathcal{U}_\theta(\boldsymbol{x};\rho_\theta)} \mathbb{E}_{\tilde{\boldsymbol{c}}\sim Q}\left[f(\boldsymbol{x},\boldsymbol{z},\tilde{\boldsymbol{c}})\right]. \quad (1)$$

In the sequel, $F(\theta)$ denotes the empirical task loss in the objective of (1); see Appendix A for a justification of this empirical approximation. Figure 1 depicts the DRDFL framework.

Our analysis relies on the following assumptions.

Assumption 2.1. We assume $f(\boldsymbol{x},\boldsymbol{z},\boldsymbol{c}) = \max_{k\in[N_{\text{K}}]}\{\boldsymbol{a}_k(\boldsymbol{x},\boldsymbol{z})^\top \boldsymbol{c} + b_k(\boldsymbol{x},\boldsymbol{z})\}$, where $N_{\text{K}} \leq 2$, each $\boldsymbol{a}_k(\boldsymbol{x},\cdot)$ is affine and each $b_k(\boldsymbol{x},\cdot)$ is finite, convex, and twice continuously differentiable.

Assumption 2.2. Each function $g_i(\boldsymbol{x},\boldsymbol{z})|_{i=1}^{n_g}$ is convex in $\boldsymbol{z}$ and Slater's condition holds for $\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})$.

A key question is how to choose the radius ρ_θ of the ambiguity set. In the next section, we address this question using split conformal calibration with a prescribed coverage level $1-\tau$, where $\tau \in (0,1)$.

3 Conformal Calibration for a Distribution

Conformal calibration constructs uncertainty sets with a prescribed coverage level based on a predictor's nonconformity scores (Shafer & Vovk, 2008). We use split conformal calibration; given

a trained predictor and a target coverage level $1-\tau$, we evaluate the nonconformity scores on a separate calibration set and use their empirical quantile to calibrate the radius of the ambiguity set.

We use split conformal calibration for two reasons: (i) it provides finite-sample marginal coverage guarantees on unseen data; (ii) the calibration step is differentiable almost everywhere under suitable conditions on the nonconformity scores (Yeh et al., 2025). This allows gradients from the downstream loss to propagate through the calibrated radius, which is crucial for training end-to-end.

The finite-sample coverage guarantee relies on exchangeability of the calibration and test nonconformity scores (Vovk et al., 2005). By Proposition 2.1, the calibration and test data points are exchangeable. Thus, for a fixed predictor, the associated nonconformity scores are also exchangeable.

For brevity, we write $P_{\star,i} := P_\star(\cdot|\boldsymbol{x}_i)$ and $P_{\theta,i} := P_\theta(\cdot|\boldsymbol{x}_i)$. The radius ρ_θ accounts for the discrepancy between the predicted distribution $P_{\theta,i}$ and true conditional distribution $P_{\star,i}$. A natural nonconformity score is the Wasserstein-1 or squared Wasserstein-2 distance between these distributions. However, these distances are generally nonsmooth with respect to the model parameters, which complicates end-to-end training. We instead use an entropically regularized surrogate and establish conditions under which its calibration provides coverage guarantees for the corresponding ambiguity set.

We construct the nonconformity scores using the Sinkhorn divergence (Genevay et al., 2018). For probability measures P and Q, define the entropically regularized optimal transport cost

$$\mathrm{OT}_\epsilon^p(P,Q) := \inf_{\pi\in\Pi(P,Q)} \left\{ \int \|\boldsymbol{p}-\boldsymbol{q}\|_2^p \,\mathrm{d}\pi(\boldsymbol{p},\boldsymbol{q}) + \epsilon\, D_{\mathrm{KL}}(\pi\|P\otimes Q) \right\}, \qquad p\in\{1,2\}, \tag{2}$$

where $D_{\mathrm{KL}}(\cdot|\cdot)$ is the Kullback–Leibler divergence, $\epsilon>0$ is the regularization parameter, $\otimes$ denotes the product measure, and $\Pi(P,Q)$ denotes the set of couplings of P and Q. The corresponding debiased Sinkhorn divergence is

$$\mathrm{SD}_\epsilon^p(P,Q) := \mathrm{OT}_\epsilon^p(P,Q) - \frac{1}{2}\mathrm{OT}_\epsilon^p(P,P) - \frac{1}{2}\mathrm{OT}_\epsilon^p(Q,Q). \tag{3}$$

Thus, SD_ϵ^1 and SD_ϵ^2 provide regularized surrogates for W_1 and W_2^2, respectively

3.1 Coverage guarantee with W_1

For the Wasserstein-1 ambiguity set, we use SD_ϵ^1 as the nonconformity score. Strict convexity of OT_ϵ^1 makes the optimal π unique, dense, and analytic in θ, so SD_ϵ^1 has a smooth, full-support gradient with respect to θ. In practice, $P_{\star,i}$ and $P_{\theta,i}$ are accessed through finite samples. We therefore consider the empirical measures

$$\widehat{P}_{\star,i} := \frac{1}{S}\sum_{j=1}^{S} \delta_{\boldsymbol{c}_i^j}, \qquad \widehat{P}_{\theta,i} := \frac{1}{M}\sum_{j=1}^{M} \delta_{\tilde{\boldsymbol{c}}_i^j},$$

where $\{\tilde{\boldsymbol{c}}_i^j\}_{j=1}^M \overset{\text{i.i.d.}}{\sim} P_{\theta,i}$ and $\delta_{\boldsymbol{c}}$ denotes the Dirac measure at $\boldsymbol{c}$. Computing SD_ϵ^1 from these empirical measures introduces an estimation error, characterized by the following lemma.

Lemma 3.1 (Estimation error for W_1)**.** *For each $\boldsymbol{x}_i$,*

$$\left|\mathrm{SD}_\epsilon^1(\widehat{P}_{\star,i},\widehat{P}_{\theta,i}) - W_1(P_{\star,i},P_{\theta,i})\right| \le \frac{\epsilon}{2}\log(SM) + W_1(\widehat{P}_{\star,i},P_{\star,i}) + W_1(\widehat{P}_{\theta,i},P_{\theta,i}). \tag{4}$$

See Appendix D for the proof. Building on Lemma 3.1, we next establish a probabilistic error bound for SD_ϵ^1, which will be used to derive the conformal coverage guarantee.

Proposition 3.1 (Probabilistic error bound for W_1)**.** *Assume that $P_{\star,i}$ is supported on a set $\mathcal{C}_i \subseteq \mathbb{R}^{n_c}$ of diameter at most D, and $P_{\theta,i}$ is supported on a set of diameter at most $\widehat{D}$, for every i. Then, for every $\eta\in(0,1)$,*

$$\mathbb{P}\Big(W_1(P_{\star,i},P_{\theta,i}) \le \mathrm{SD}_\epsilon^1(\widehat{P}_{\star,i},\widehat{P}_{\theta,i}) + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(1)}(\eta)\Big) \ge 1-\eta, \tag{5}$$

with $r_{S,M}^{(1)}(\eta) := \sqrt{\frac{\log(1/\eta)}{2}\left(\frac{D^2}{S}+\frac{\widehat{D}^2}{M}\right)}$.

The probabilistic correction $r^{(1)}_{S,M}(\eta)$ decays as $O(S^{-1/2} + M^{-1/2})$. Since the samples $\tilde{\boldsymbol{c}}_i^j$ are generated from $P_{\theta,i}$, the sample size M can typically be chosen much larger than S, making the correction due to finite M negligible in practice. Combining Proposition 3.1 with the split conformal sorting argument yields the following finite-sample marginal coverage guarantee.

Theorem 3.1 (Finite-sample marginal coverage for W_1)**.** *Fix the distribution predictor P_θ and assume the conditions of Proposition 3.1. For each calibration point $i = 1, \dots, N_{\mathrm{cal}}$, define the nonconformity score $s_i := \mathrm{SD}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$. Let $s_{(1)} \le \cdots \le s_{(N_{\mathrm{cal}})}$ denote the corresponding order statistics, with $s_{(N_{\mathrm{cal}}+1)} := +\infty$. For $\tau \in (0,1)$ and $\eta \in (0, \tau)$, define*

$$k := \lceil (N_{\mathrm{cal}} + 1)(1 - \tau + \eta) \rceil, \quad \rho_\theta := s_{(k)} + \frac{\epsilon}{2} \log(SM) + r^{(1)}_{S,M}(\eta).$$

Then, for a test point $\boldsymbol{x}^{\mathrm{test}}$, we have

$$\mathbb{P}\left(W_1\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta \right) \ge 1 - \tau. \tag{6}$$

Thus, the ambiguity set in (1) contains the true conditional distribution with probability at least $1 - \tau$.

See Appendix E.3 for the proof. The probability in Theorem 3.1 is taken jointly over the calibration sample and the test point. In this setting, distribution-free conditional coverage is generally unattainable (Foygel Barber et al., 2021).

Remark 3.1 (Extension beyond bounded support). The bounded-support assumption in Proposition 3.1 is used only to invoke the concentration bound. When $P_{\star,i}$ and $P_{\theta,i}$ satisfy quadratic transport-entropy inequalities uniformly in $\boldsymbol{x}_i$ (see Appendix E.2 for more details), analogous results hold without the bounded-support assumptions requirement.

3.2 Coverage guarantee with W_2^2

For the squared Wasserstein-2 ambiguity set, we use SD^2_ϵ as the nonconformity score. Following the same argument as for W_1, we obtain the corresponding finite-sample marginal coverage guarantee.

Theorem 3.2 (Finite-sample marginal coverage for W_2^2)**.** *Fix the distribution predictor P_θ. Assume that $P_{\star,i}$ is supported on a set $\mathcal{C}_i \subseteq \mathbb{R}^{n_c}$ of diameter at most D, and $P_{\theta,i}$ is supported on a set of diameter at most $\widehat{D}$, for every i. We further assume that $\mathrm{diam}(\mathcal{C}_i \cup \mathrm{supp}\, P_{\theta,i}) \le D_{\mathrm{uni}}$ for every i. For each calibration point $i = 1, \dots, N_{\mathrm{cal}}$, define the nonconformity score $s_i := \mathrm{SD}^2_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$. Let $s_{(1)} \le \cdots \le s_{(N_{\mathrm{cal}})}$ denote the corresponding order statistics, with $s_{(N_{\mathrm{cal}}+1)} := +\infty$. For $\tau \in (0,1)$ and $\eta \in (0, \tau)$, define*

$$k := \lceil (N_{\mathrm{cal}}+1)(1-\tau+\eta) \rceil, \quad \rho_\theta := s_{(k)} + \frac{\epsilon}{2} \log(SM) + r^{(2)}_{S,M}(\eta), \quad r^{(2)}_{S,M}(\eta) := 2D_{\mathrm{uni}}\, r^{(1)}_{S,M}(\eta).$$

Then, for a test point $\boldsymbol{x}^{\mathrm{test}}$, we have

$$\mathbb{P}\left(W_2^2\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta \right) \ge 1 - \tau. \tag{7}$$

Thus, the ambiguity set in (1) contains the true conditional distribution with probability at least $1 - \tau$.

See Appendix G for the proof.

Remark 3.2 (Extension beyond bounded support). The bounded-support assumptions in Theorem 3.2 can similarly be replaced by suitable quadratic transport-entropy conditions; see Appendix G.2.

3.3 Extension to temporal dependence

The probabilistic error bounds in Proposition 3.1 and Proposition G.1 do not rely on exchangeability. This is important for temporally dependent data, where the exchangeability assumption generally does not hold. Building on (Barber & Pananjady, 2026), we extend the marginal coverage guarantees in Theorems 3.1 and 3.2 to temporally dependent data under suitable β-mixing conditions[1].

[1] $\beta_{\mathrm{mix}}(\ell)$ quantifies how far the process is from independent across a gap of ℓ periods.

Corollary 3.1 (Marginal coverage under temporal dependence)**.** *Assume that the data-generating process is stationary with β-mixing coefficients $\beta_{\mathrm{mix}}(\cdot)$ and that the nonconformity score has memory L, i.e., the score at time t depends only on observations from times $t-L,\ldots,t$. Then, the coverage guarantees in Theorems 3.1 and 3.2 hold with $1-\tau$ replaced by $1-\tau-\Delta_{\mathrm{mix}}$, where*

$$\Delta_{\mathrm{mix}} := \min_{\substack{\ell,\ell_*\geq 0 \\ \ell+\ell_*\leq m-L}} \left\{ \frac{\ell + L + (\tau-\eta)\ell_*}{m-\ell_*+1} + 2\beta_{\mathrm{mix}}(\ell) + 2\beta_{\mathrm{mix}}(\ell_*) \right\}. \tag{8}$$

See Appendix H for the proof.

4 Distributionally Robust Decision-Focused Learning

In this section, we describe how a predictive model is trained end-to-end using the DRDFL approach. Our DRDFL is compatible with conditional normalizing flow (CNF), conditional diffusion model (CDM) and Gaussian Mixture Density Network (GMDN), see Appendix B for details. The training of the model is robustified through a conformally calibrated Wasserstein ambiguity set.

In each training iteration, we sample a batch $\mathcal{B}_{\mathrm{tra}}$ from the training dataset $\mathcal{D}_{\mathrm{tra}}$ and split it into a prediction batch $\mathcal{B}_{\mathrm{pre}}$ and a calibration batch $\mathcal{B}_{\mathrm{cal}}$. The calibration batch is used to compute the nonconformity scores and the training radius ρ_θ. For each data point $(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S) \in \mathcal{B}_{\mathrm{pre}}$, the feature $\boldsymbol{x}_i$ is fed to the current predictive model to obtain $P_\theta(\cdot \mid \boldsymbol{x}_i)$. The resulting robust decision $\boldsymbol{z}_\theta^*(\boldsymbol{x}_i)$ is evaluated using the corresponding target samples $\{\boldsymbol{c}_i^j\}_{j=1}^S$. The parameter θ is then updated using the gradient of the resulting task loss. Here, we give the computation of this gradient

$$\frac{\mathrm{d}F(\theta)}{\mathrm{d}\theta} = \frac{\partial F(\theta)}{\partial \boldsymbol{z}_\theta^*}\frac{\mathrm{d}\boldsymbol{z}_\theta^*}{\mathrm{d}\theta} = \frac{1}{N_{\mathrm{tra}}}\sum_{i=1}^{N_{\mathrm{tra}}}\frac{1}{S}\sum_{j=1}^{S}\left[\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{z}_\theta^*, \boldsymbol{c}_i^j)}{\partial \boldsymbol{z}_\theta^*}\frac{\mathrm{d}\boldsymbol{z}_\theta^*}{\mathrm{d}\theta}\right]. \tag{9}$$

To compute $\frac{\mathrm{d}\boldsymbol{z}_\theta^*}{\mathrm{d}\theta}$, we first reformulate the lower-level DRO problem in (1) as a stochastic optimization problem. For Wasserstein-1, it would be

$$\boldsymbol{z}_\theta^*(\boldsymbol{x}_i) \in \operatorname*{argmin}_{\boldsymbol{z}:\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z})\leq \boldsymbol{0}} \Phi_{1,i}(\boldsymbol{z},\theta) := \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\left[f(\boldsymbol{x}_i,\boldsymbol{z},\tilde{\boldsymbol{c}})\right] + \rho_\theta \max_{k\in[N_{\mathrm{K}}]} \|\boldsymbol{a}_k(\boldsymbol{x}_i,\boldsymbol{z})\|_*, \tag{10}$$

which is convex in $\boldsymbol{z}$. Under Assumption 2.2 the KKT conditions are necessary and sufficient for its optimality whenever $\Phi_{1,i}(\cdot,\theta)$ is differentiable at the solution. Details of the derivations can be found in Appendix I. Here we denote the KKT conditions as $\boldsymbol{K}\big(\boldsymbol{y}_i^*(\theta),\theta\big) = \boldsymbol{0}$, with $\boldsymbol{y}^*(\theta) := (\boldsymbol{z}_\theta^*(\boldsymbol{x}_i), \boldsymbol{\nu}_{\mathcal{A}}^*)$, where $\boldsymbol{\nu}$ is the dual variable and $\mathcal{A} := \big\{j : g_j\big(\boldsymbol{x}_i, \boldsymbol{z}_\theta^*(\boldsymbol{x}_i)\big) = 0\big\}$. We further differentiate the KKT conditions and obtain $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{y}}\frac{\mathrm{d}\boldsymbol{y}^*}{\mathrm{d}\theta} + \frac{\partial \boldsymbol{K}}{\partial \theta} = \boldsymbol{0}$, as below:

$$\frac{\mathrm{d}\boldsymbol{z}_\theta^*(\boldsymbol{x}_i)}{\mathrm{d}\theta} = -\begin{bmatrix}\boldsymbol{I}_{n_z} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{H}_{1,i} & \nabla_{\boldsymbol{z}}\boldsymbol{g}_{\mathcal{A}_i}\big(\boldsymbol{x}_i, \boldsymbol{z}_\theta^*(\boldsymbol{x}_i)\big)^\top \\ \nabla_{\boldsymbol{z}}\boldsymbol{g}_{\mathcal{A}_i}\big(\boldsymbol{x}_i, \boldsymbol{z}_\theta^*(\boldsymbol{x}_i)\big) & \boldsymbol{0}\end{bmatrix}^{-1}\frac{\partial \boldsymbol{K}}{\partial \theta}, \tag{11}$$

where

$$\boldsymbol{H}_1 := \nabla_{\boldsymbol{z}\boldsymbol{z}}^2\,\mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\big[f(\boldsymbol{x}_i,\boldsymbol{z}^*,\tilde{\boldsymbol{c}})\big] + \rho_\theta\,\nabla_{\boldsymbol{z}\boldsymbol{z}}^2 \max_{k\in[N_{\mathrm{K}}]}\big\|\boldsymbol{a}_k(\boldsymbol{x}_i,\boldsymbol{z}^*)\big\|_* + \sum_{j\in\mathcal{A}}\nu_j^*\,\nabla_{\boldsymbol{z}\boldsymbol{z}}^2 g_j(\boldsymbol{x}_i,\boldsymbol{z}^*),$$

$$\frac{\partial \boldsymbol{K}}{\partial \theta} = \begin{bmatrix}\underbrace{\nabla_\theta \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\Big[\nabla_{\boldsymbol{z}} f\big(\boldsymbol{x}_i,\boldsymbol{z}^*,\tilde{\boldsymbol{c}}\big)\Big]}_{:=\boldsymbol{T}_1 \text{ predicted distribution}} + \underbrace{\nabla_{\boldsymbol{z}}\max_{k\in[N_{\mathrm{K}}]}\big\|\boldsymbol{a}_k(\boldsymbol{x}_i,\boldsymbol{z}^*)\big\|_*\big(\nabla_\theta\rho_\theta\big)^\top}_{:=\boldsymbol{T}_2 \text{ calibrated radius}} \\ \boldsymbol{0}\end{bmatrix}. \tag{12}$$

As shown in (12), the gradient is composed of 2 parts, namely with respect to the change of predicted distribution and with respect to the change of the calibrated radius, which matches design of DRDFL. The computation of $\boldsymbol{T}_1$ depends on the choice of the distribution predictor and is covered in Appendix J.1. $\boldsymbol{T}_2$ involves $\nabla_\theta\rho_\theta$, which requires differentiation of the conformal calibration

algorithm and the nonconformity score $s(\cdot,\cdot)$. According to Yeh et al. (2025, Theorem D.1), the conformal calibration can be written as a differentiable function. Assume the nonconformity score is unique for each calibration data point, we have

$$\frac{\mathrm{d}\rho_\theta}{\mathrm{d}\theta} = \begin{cases} \frac{\mathrm{d}s_{(r)}}{\mathrm{d}\theta} = \frac{\mathrm{d}}{\mathrm{d}\theta}\, s_\theta\big(\boldsymbol{x}_{\sigma(p)}, \{\boldsymbol{c}^j_{\sigma(p)}\}_{j=1}^S\big), & \text{if } \tau \geq \frac{1}{N_{\mathrm{cal}}+1} + \eta, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where $\sigma := \{1, ..., N_{\mathrm{cal}}+1\} \to \{1, ..., N_{\mathrm{cal}}+1\}$ denotes a permutation function, which sorts the nonconformity scores of the calibration data points $\{s_i\}_{i=1}^{N_{\mathrm{cal}}+1}$ in ascending order, so that $s_{(i)} \leq s_{(j)}$ for $i < j$, with $s_{(i)} := s_{\sigma(i)}$ and $r = \lceil (N_{\mathrm{cal}}+1)(1-\tau+\eta) \rceil$.

The computation of $\mathrm{d}s_\theta/\mathrm{d}\theta$, namely the derivative of the debiased Sinkhorn divergence with respect to the predictive model's parameter θ is given in Appendix I for W_1 and in Appendix J.2 for W_2. After the DRDFL training ends (θ is fixed), we use $\mathcal{D}_{\mathrm{cal}}$ for the final calibration and obtain the radius of the ambiguity set ρ_θ, the trained model gives a marginal coverage guarantee on the decision quality.

Theorem 4.1 (Decision marginal coverage guarantee)**.** *Fix the distribution predictor P_θ trained by DRDFL and let ρ_θ be obtained from the independent held-out calibration dataset $\mathcal{D}_{\mathrm{cal}}$. For data from $\mathcal{D}_{\mathrm{test}}$ and under the conditions of Theorem 3.1 for $p = 1$, or Theorem 3.2 for $p = 2$, we have*

$$\mathbb{P}\left(\mathbb{E}_{\boldsymbol{c}\sim P_\star(\cdot|\boldsymbol{x})}\left[f\big(\boldsymbol{x}, \boldsymbol{z}^*_\theta(\boldsymbol{x}), \boldsymbol{c}\big)\right] \leq \min_{\boldsymbol{z}:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})\leq \boldsymbol{0}} \sup_{Q\in\mathcal{U}_\theta(\boldsymbol{x};\rho_\theta)} \mathbb{E}_{\tilde{\boldsymbol{c}}\sim Q}\left[f\big(\boldsymbol{x}, \boldsymbol{z}, \tilde{\boldsymbol{c}}\big)\right] \right) \geq 1-\tau. \tag{14}$$

Remark 4.1. The calibration batch $\mathcal{B}_{\mathrm{cal}}$ is sampled from $\mathcal{D}_{\mathrm{tra}}$ and is used only during DRDFL training to construct the training radius ρ_θ. The held-out dataset $\mathcal{D}_{\mathrm{cal}}$ is not used during training. It is used only after θ has been frozen to obtain the final deployment radius ρ_θ.

5 Experiments

Here, we validate the conformal calibration's coverages in Theorems 3.1 and 3.2 with experiments. In addition, we evaluate the performance of our DRDFL framework on a synthetic example and a real-world portfolio allocation problem against the following baselines:

1. **Distributionally robust predict-then-optimize baselines (DRPtO):** GMDN or CNF is trained to minimize the prediction error, conformally calibrate the ambiguity set and the decision is made by solving a DRO problem with respect to the predicted distribution and the ambiguity;
2. **Stochastic predict-then-optimize baselines (StoPtO):** GMDN or CNF is trained to minimize the prediction error and the decision is made by solving a stochastic optimization problem with respect to the predicted distribution;
3. **Stochastic DFL (SDFL):** GMDN or CNF trained to minimize the task regret with the lower-level problem as stochastic optimization task;
4. **Deterministic DFL (DetDFL):** A multi-layer perceptron (MLP) is trained to minimize the task regret with the lower-level problem as deterministic optimization task.

Details of the baselines can be found in Appendix L.1. We denote the task regret as the difference between the objective value induced by the decision parametrized by the learned model parameter and the optimal objective value, $R_i := \frac{1}{S}\sum_{j=1}^S f(\boldsymbol{x}_i, \boldsymbol{z}^*_\theta, \boldsymbol{c}^j_i) - \min_{\boldsymbol{z}} \frac{1}{S}\sum_{j=1}^S f(\boldsymbol{x}_i, \boldsymbol{z}, \boldsymbol{c}^j_i)$. The comparison of all methods' results can be found in Table 1.

5.1 Coverage Validation

We train a GMDN using data $(\boldsymbol{x}_i, \{\boldsymbol{c}^j_i\}_{j=1}^S)$ is generated by Warped K-Sinusoidal with $n_{\boldsymbol{x}} = 10$ and $n_{\boldsymbol{c}} = 20$. The details can be found in Appendix L.2. Given this trained model and a user-defined confidence level $1-\tau$, we conformally calibrate the radius ρ_θ of the Wasserstein ambiguity set and compute the empirical coverage of the true distribution $P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}})$ staying within the ambiguity set centered at the predicted $P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})$ over 100 test data points. As shown in Figure 2, the empirical coverage is close to $1-\tau$, which validates our coverage theorems.

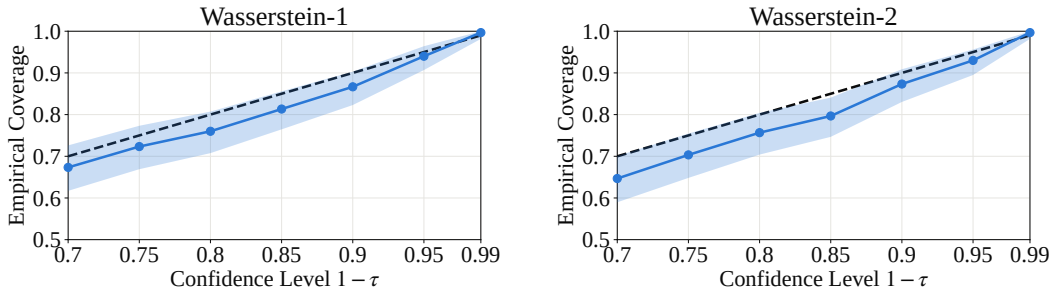


Figure 2: Empirical coverages for W_1 and W_2 are averaged over 300 test data points. The radius ρ_θ is calibrated using 280 data points. The number of samples from the target distribution $P_\star$ is 50 while the number of samples from the predicted distribution P_θ is 100.

5.2 Portfolio Allocation

Synthetic Data. We use synthetic data generated by Warped K-Sniusoidal as in Section 5.1 together with a distribution shift between the training data points and the test data points (a Gaussian noise, $\mathcal{N}(0, 0.4)$). with such data we study a portfolio allocation problem with the objective function as $\mathbb{E}_{\boldsymbol{c}\sim P(\cdot|\boldsymbol{x})}[-\boldsymbol{c}^\top \boldsymbol{z}] + \beta\,\mathrm{CVaR}_\alpha^{P(\cdot|\boldsymbol{x})}(-\boldsymbol{c}^\top \boldsymbol{z})$ and the constraints are $\mathbf{1}^\top \boldsymbol{z} = 1 \;\;\&\;\; \boldsymbol{z} \geq \mathbf{0}$. In Appendix K we show that this objective function can be reformulated as f satisfying Assumption 2.1. In Figure 3, we compare the task regrets on test data of DRDFL and the baselines with respect to different confidence levels.

Real-World Data. We compare the performance of DRDFL and the baselines on real-world portfolio allocation. The objective function and the constraints are the same as Section 5.2. We use the US stock market data offered Alpaca (2026). We construct the feature $\boldsymbol{x}_i$ using data up to day $i-1$ and $\{\boldsymbol{c}_i^j\}_{j=1}^S$ are samples from the return on day i with $n_x = 50$, $n_c = 30$ and $S = 27$. The task regret comparison on test data with respect to different confidence levels is in Figure 4.

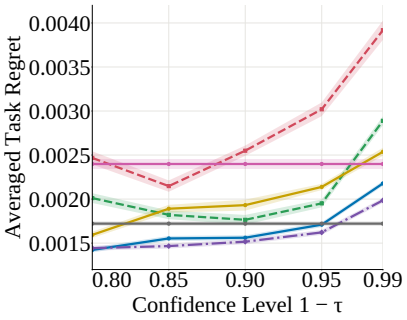


Figure 3: Synthetic example.

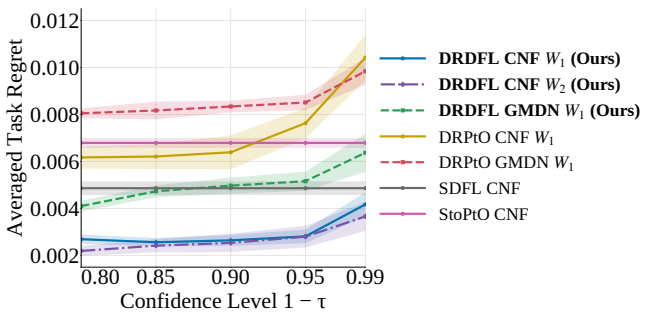


Figure 4: Real-world portfolio allocation.

5.3 Discussion of Results

Across the considered settings, the DFL methods outperform their predict-then-optimize (PtO) counterparts. Whereas PtO trains the predictive model solely to minimize prediction error, without accounting for downstream decision quality, DFL directly incorporates the task objective into end-to-end training. Moreover, methods based on distributional predictors substantially outperform DetDFL. Because DetDFL relies on point predictions, it cannot represent the uncertainty or complex structure of the underlying conditional distribution, which can lead to inferior downstream task performance.

Confidence Level's Influence. The results in Figures 3 and 4 reveal a trade-off in selecting the confidence level. As the confidence level increases, the calibrated ambiguity radius generally gets larger. As discussed in Section 4, this strengthens the regularization induced by the DRO formulation and produces more conservative decisions, which can increase task regret. However, reducing the

Table 1: Results of all methods on 2 tasks. We perform each experiment with 7 random seeds and give the mean $\pm$ standard error of the task regret averaged on test data of each method. CNF denotes using a conditional normalizing flow as the predictive model while GMDN denotes Gaussian Mixture Density Network. W_1 or W_2 denotes the a W_1 or W_2 ambiguity set in the DRO respectively. For each task, the bolded value are the top two. The unevaluated SDFL GMDN on the synthetic example can be estimated by SDFL CNF.

Methods	**Synthetic Example**	**Real-World Portfolio Allocation**
Predict-then-Optimize		
StoPtO CNF	0.0023986 ± 0.000056	0.00679 ± 0.00021
DRPtO GMDN W_1	0.0021474 ± 0.000064	0.00805 ± 0.00020
DRPtO CNF W_1	0.0015936 ± 0.000035	0.00617 ± 0.00047
Decision-Focused Learning		
DetDFL	0.0086479 ± 0.000017	0.02619 ± 0.00012
SDFL GMDN	—	0.00614 ± 0.00033
SDFL CNF	0.0017213 ± 0.000025	0.00486 ± 0.00029
DRDFL GMDN W_1 (Ours)	0.0017646 ± 0.000056	0.00410 ± 0.00023
DRDFL CNF W_1 (Ours)	**0.0014227 ± 0.000029**	**0.00256 ± 0.00018**
DRDFL CNF W_2 (Ours)	**0.0014429 ± 0.000028**	**0.00219 ± 0.00021**

confidence level does not necessarily improve performance since if it is too small, the resulting ambiguity set may fail to contain the test distribution, leaving the decision insufficiently protected against distributional error. Consistent with this trade-off, the lowest task regret for DRDFL CNF with W_1 in the real-world portfolio experiment is attained at a confidence level of approximately 0.85.

DRDFL vs. SDFL. In both tasks with distribution shift, DRDFL outperforms SDFL, consistent with the protection offered by its conformally calibrated ambiguity set. The shift is introduced explicitly in the synthetic experiment, while the chronological split in the portfolio experiment exposes the methods to temporal changes in returns. In contrast, SDFL performs better in the no-shift synthetic setting, as expected, since robustness may introduce unnecessary conservatism when the learned and test distributions agree.

DRDFL CNF vs. DRDFL GMDN. Across both tasks, DRDFL CNF outperforms DRDFL GMDN, while DRPtO CNF similarly outperforms DRPtO GMDN. This pattern is consistent with the superior predictive performance of CNF in our experiments. Compared to GMDN, CNF gives more flexible distribution shapes and has a better tail behavior, which captures more complicated distributions. This reveals that integrating a powerful distribution predictor improves the task performance of DRDFL.

6 Conclusion

We introduced DRDFL, an end-to-end framework that optimizes robust decisions through learning a conditional distribution which lies in a calibrates a Wasserstein ambiguity set using differentiable Sinkhorn scores. We proposed finite-sample marginal coverage guarantees for Wasserstein-1 and squared Wasserstein-2 ambiguity sets followed by an upper bound on the true conditional expected task loss. The portfolio allocation experiment demonstrated the value of coupling prediction, calibration, and robust optimization rather than treating them as separate stages.

The present results establish this pipeline under a deliberately structured setting, which also identifies several substantive directions for future work. The exact coverage guarantee relies on exchangeability; although our extension to stationary β-mixing processes quantifies the degradation under temporal dependence, financial returns may exhibit nonstationarity or dependencies beyond this model. Moreover, the tractable DRO reformulations and gradient analysis rely on functional assumptions. Extending the framework beyond this setting is therefore an important direction. These extensions can be built directly on the main technical infrastructure developed in this work.

AI USE STATEMENT

In this work, we used generative AI tools to search related references, polish Figure 1, check the correctness of the proofs of our theorems and implement experiments. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or assist with translation.

All AI-assisted mathematical arguments were independently checked by the authors. Suggested references and literature claims were verified against the original sources, and all AI-assisted manuscript text was manually reviewed and revised. We take responsibility for the final content of this work, including text, claims, proofs, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a methodological framework for distributionally robust decision-focused learning and studies portfolio allocation as an application. The work does not involve human subjects or personally identifiable information. Potential risks arise from downstream use of learned decision systems in high-stakes settings such as finance, where model misspecification, distribution shift, or inappropriate interpretation of uncertainty guarantees may lead to undesirable decisions. The proposed guarantees should therefore be interpreted under the assumptions stated in the paper and should not be viewed as eliminating all forms of model or deployment risk. The portfolio experiments are intended for research evaluation and not as financial advice.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs of the theoretical results in the appendix, including the exchangeability, the finite-sample Wasserstein approximation bounds, and the conformal coverage guarantees. The manuscript specifies the construction of the training and final calibration sets, the computation of the conformal ambiguity radius, the Wasserstein DRO reformulations, and the end-to-end training procedure. We also document the predictive models, dataset construction and preprocessing, data splits, evaluation metrics, optimization settings, and hyperparameters used in the experiments. Code and experiment configurations for reproducing the reported results are provided in the supplementary material.

REFERENCES

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. Advances in neural information processing systems, 32, 2019.

Alpaca. Historical stock data. https://docs.alpaca.markets/docs/historical-stock-data, 2026. Alpaca Market Data API. Accessed: 2026-09-24.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In International conference on machine learning, pp. 136–145. PMLR, 2017.

Rina Foygel Barber and Ashwin Pananjady. Predictive inference for time series: why is split conformal effective despite temporal dependence? In 37th International Conference on Algorithmic Learning Theory, 2026.

Christopher M Bishop. Mixture density networks. 1994.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. Mathematics of Operations Research, 44(2):565–600, 2019.

SR Dalal and WJ Hall. Approximating priors by mixtures of natural conjugate priors. Journal of the Royal Statistical Society: Series B (Methodological), 45(2):278–286, 1983.

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.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". Management science, 68(1): 9–26, 2022.

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 The 22nd international conference on artificial intelligence and statistics, pp. 2681–2690. PMLR, 2019.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA, 10(2):455–482, 2021.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In International conference on artificial intelligence and statistics, pp. 1608–1617. PMLR, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

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, pp. 130–166. Informs, 2019.

Colin McDiarmid. On the method of bounded differences. In Surveys in Combinatorics, 1989, volume 141 of London Mathematical Society Lecture Note Series, pp. 148–188. Cambridge University Press, 1989.

Yash Patel, Guyang Cao, and Ambuj Tewari. Non-parameteric conformal distributionally robust optimization. In ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling, 2024.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In International conference on machine learning, pp. 1530–1538. PMLR, 2015.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. The Journal of Risk, 2(3):21–41, 2000.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. Journal of machine learning research, 9(3), 2008.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pp. 2256–2265. pmlr, 2015.

Cédric Villani. Optimal Transport: Old and New, volume 338 of Grundlehren der mathematischen Wissenschaften. Springer, 2009.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer, 2005.

Prince Zizhuang Wang, Shuyi Chen, Jinhao Liang, Ferdinando Fioretto, and Shixiang Zhu. Gen-DFL: Decision-focused generative learning for robust decision making. In The Fourteenth International Conference on Learning Representations, 2026.

JIAQI WEN and Jianyi Yang. Distributionally robust optimization via generative ambiguity modeling. In The Fourteenth International Conference on Learning Representations, 2026.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. arXiv preprint arXiv:1912.00042, 2019.

Han Xu and Christopher Yeh. Robust decisions via generative wasserstein distributionally robust optimization. In NeurIPS 2025 Workshop: Generative AI in Finance, 2026.

Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. Transactions on Machine Learning Research, 2025. ISSN 2835-8856.

Ziwei Zhang, Jonathan Yu-Meng Li, and Zhihao Jin. Generative distributionally robust optimization. arXiv preprint arXiv:2607.24983, 2026.

Zihao Zhao, Christopher Yeh, Lingkai Kong, and Kai Wang. Diffusion-DFL: Decision-focused diffusion models for stochastic optimization. In The Fourteenth International Conference on Learning Representations, 2026.

CONTENTS

1 Introduction **1**

1.1 Related Work 2

2 Problem Statement **2**

3 Conformal Calibration for a Distribution **3**

3.1 Coverage guarantee with W_1 4

3.2 Coverage guarantee with W_2^2 5

3.3 Extension to temporal dependence 5

4 Distributionally Robust Decision-Focused Learning **6**

5 Experiments **7**

5.1 Coverage Validation 7

5.2 Portfolio Allocation 8

5.3 Discussion of Results 8

6 Conclusion **9**

A Empirical Task Loss Formulation under the Multiple Realizations Setting **15**

B Details of Various Predictive Models **16**

B.1 Conditional Diffusion Model 16

B.2 Conditional Normalizing Flow 16

B.3 Gaussian Mixture Density Network 17

C Proof of Exchangeability of the Data Points **18**

D Proof of the Estimation Error of SD_ϵ^1 for Wasserstein-1 Distance **19**

D.1 Proof of Lemma 3.1 19

E Proof of the Marginal Coverage Guarantee of using SD_ϵ^1 for W_1 **20**

E.1 Proof of Proposition 3.1 20

E.2 Explanation of Remark 3.1 21

E.3 Proof of Theorem 3.1 23

F Proof of the Estimation Error of SD_ϵ^2 for Squared Wasserstein-2 Distance **24**

G Proof of the Marginal Coverage Guarantee of using SD_ϵ^2 for W_2^2 **26**

G.1 Proof ofTheorem 3.2 27

G.2 Explanation of Remark 3.2 28

H Marginal Coverage under Temporal Dependence **30**

I Reformulations of the Wasserstein DRO Problems as Stochastic Optimization Problems **32**

I.1 Wasserstein-1 DRO 32

I.2 Wasserstein-2 DRO 32

J Computation of the Gradient in DRDFL **34**

J.1 Computation of the Gradient of the Optimal W_1 DRO Decision with Respect to θ . 34

J.2 Computation of the Gradient of the Optimal W_2 DRO Decision with Respect to θ . 35

J.3 Computation of the Gradient of the Nonconformity Score $\frac{\mathrm{d}s\left(\widehat{P}_{\star,i},\widehat{P}_{\theta,i}\right)}{\mathrm{d}\theta}$ 36

K Reformulation of Mean-Risk Optimization as Distributionally Robust Optimization **39**

L Details of Experiments **40**

L.1 Details of the Baselines 40

L.2 Details of the Synthetic Data 41

L.3 Details of the Real-World Portfolio Allocation Data 42

L.4 Further Discussion of the Experiment Results 42

A Empirical Task Loss Formulation under the Multiple Realizations Setting

Let P denote the joint distribution of $(\boldsymbol{x}, \boldsymbol{c})$, with marginal $P_{\boldsymbol{x}}$ and conditional distribution $P(\cdot \mid \boldsymbol{x})$. We assume that the data are generated according to the hierarchical sampling scheme

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{x}}, \qquad i = 1, \dots, N_{\text{tra}}, \tag{15}$$

$$\boldsymbol{c}_i^1, \dots, \boldsymbol{c}_i^S \mid \boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} P(\cdot \mid \boldsymbol{x}_i), \qquad j = 1, \dots, S, \tag{16}$$

where the pairs $\left\{ \left(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S \right) \right\}_{i=1}^{N_{\text{tra}}}$ are independent. This sampling structure is a standard two-stage Monte Carlo construction.

For an integrable function f, our target quantity is

$$\mu_f := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{c})\sim P} \left[f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{c}) \right].$$

By the law of iterated expectation,

$$\mu_f = \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}} \left[\mathbb{E}_{\boldsymbol{c}\sim P(\cdot|\boldsymbol{x})} [f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{c}) \mid \boldsymbol{x}] \right]. \tag{17}$$

Define the conditional expectation

$$m_f(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{c}\sim P(\cdot|\boldsymbol{x})} [f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{c})].$$

For each observed $\boldsymbol{x}_i$, we approximate $m_f(\boldsymbol{x}_i)$ using the S conditionally i.i.d. samples,

$$\widehat{m}_f(\boldsymbol{x}_i) := \frac{1}{S} \sum_{j=1}^{S} f(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{c}_i^j).$$

Since the inner samples are drawn from $P(\cdot \mid \boldsymbol{x}_i)$, this estimator is conditionally unbiased:

$$\mathbb{E} \left[\widehat{m}_f(\boldsymbol{x}_i) \mid \boldsymbol{x}_i \right] = m_f(\boldsymbol{x}_i). \tag{18}$$

We therefore approximate equation (17) by the two-stage empirical average

$$\widehat{\mu}_f := \frac{1}{N_{\text{tra}}} \sum_{i=1}^{N_{\text{tra}}} \widehat{m}_f(\boldsymbol{x}_i) = \frac{1}{N_{\text{tra}} S} \sum_{i=1}^{N_{\text{tra}}} \sum_{j=1}^{S} f(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{c}_i^j). \tag{19}$$

Applying equation (18) and the tower property gives

$$\mathbb{E}[\widehat{\mu}_f] = \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}} [m_f(\boldsymbol{x})] = \mu_f.$$

Therefore equation (19) is unbiased for any finite N_{tra} and S. Moreover, if $\mathbb{E}_P[|f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{c})|] < \infty$, then for any fixed $S \geq 1$, the group averages $\{\widehat{m}_f(\boldsymbol{x}_i)\}_{i=1}^{N_{\text{tra}}}$ are i.i.d. with mean μ_f. Hence, by the strong law of large numbers,

$$\widehat{\mu}_f \xrightarrow[N_{\text{tra}}\to\infty]{\text{almost surely}} \mu_f.$$

B Details of Various Predictive Models

B.1 Conditional Diffusion Model

Diffusion models generate samples by learning to reverse a fixed noising process, (Sohl-Dickstein et al., 2015). Fix a horizon T and a variance schedule $\beta_1, \dots, \beta_T \in (0,1)$, and write $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Starting from $\boldsymbol{c}_0 := \boldsymbol{c}$, the forward process has the closed form (Ho et al., 2020)

$$\boldsymbol{c}_t = \sqrt{\bar{\alpha}_t}\, \boldsymbol{c}_0 + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{h}, \qquad \boldsymbol{h} \sim p_{\boldsymbol{h}} = \mathcal{N}(0, \boldsymbol{I}),$$

so that $\boldsymbol{c}_T$ is approximately distributed as $p_{\boldsymbol{h}}$ when $\bar{\alpha}_T \approx 0$. Given a feature $\boldsymbol{x}$, a CDM reverses this process by Gaussian transitions $p_\theta(\boldsymbol{c}_{t-1} \mid \boldsymbol{c}_t, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{c}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{c}_t, t; \boldsymbol{x}), \sigma_t^2 \boldsymbol{I})$ with a fixed variance schedule σ_t^2, typically $\sigma_t^2 = \beta_t$. The mean is parameterized through a noise-prediction network $\boldsymbol{h}_\theta(\boldsymbol{c}_t, t; \boldsymbol{x})$, where θ denotes the learnable parameters. Sampling from the model is performed by drawing $\boldsymbol{h}_1, \dots, \boldsymbol{h}_T \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{h}}$, setting $\boldsymbol{c}_T = \boldsymbol{h}_T$ and $\boldsymbol{h}_0 := \boldsymbol{0}$, and iterating for $t = T, \dots, 1$

$$\boldsymbol{c}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{c}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\, \boldsymbol{h}_\theta(\boldsymbol{c}_t, t; \boldsymbol{x}) \right) + \sigma_t\, \boldsymbol{h}_{t-1}. \tag{20}$$

The recursion (20) composes into a deterministic map $\boldsymbol{c} = \boldsymbol{c}_0 = F_\theta(\boldsymbol{h}_{1:T}; \boldsymbol{x})$ of the latent variables. We denote the resulting conditional distribution by $P_\theta(\cdot \mid \boldsymbol{x})$. For any measurable set A,

$$P_\theta(A \mid \boldsymbol{x}) = \int \mathbf{1}\{F_\theta(\boldsymbol{h}_{1:T}; \boldsymbol{x}) \in A\} \prod_{t=1}^{T} p_{\boldsymbol{h}}(\boldsymbol{h}_t)\, d\boldsymbol{h}_{1:T}.$$

To use CDM in the end-to-end training, we use the score function estimator proposed by Zhao et al. (2026) to avoid backpropagation through the sampling.

B.2 Conditional Normalizing Flow

Normalizing flows model complex distributions by transforming a simple base distribution through invertible mappings, (Rezende & Mohamed, 2015). Let a latent variable $\boldsymbol{h} \in \mathbb{R}^{d_{lat}}$ follow a fixed base distribution $p_{\boldsymbol{h}}$, typically $\mathcal{N}(0, \boldsymbol{I})$. Given a feature $\boldsymbol{x}$, a conditional normalizing flow defines an invertible and differentiable transformation $\boldsymbol{z} = f_\theta(\boldsymbol{c}; \boldsymbol{x})$ (Winkler et al., 2019), where θ denotes the learnable parameters of the transformation. According to Winkler et al. (2019) by the change-of-variables formula, the induced conditional density is

$$p_\theta(\boldsymbol{c} \mid \boldsymbol{x}) = p_{\boldsymbol{h}}(f_\theta(\boldsymbol{c}; \boldsymbol{x})) \left| \det \frac{\partial f_\theta(\boldsymbol{c}; \boldsymbol{x})}{\partial \boldsymbol{c}} \right|.$$

Equivalently sampling from the model is performed by drawing $\boldsymbol{h} \sim p_{\boldsymbol{h}}$ and computing $\boldsymbol{c} = f_\theta^{-1}(\boldsymbol{h}; \boldsymbol{x})$. We denote the resulting conditional distribution by $P_\theta(\cdot \mid \boldsymbol{x})$. For any measurable set A,

$$P_\theta(A \mid \boldsymbol{x}) = \int \mathbf{1}\{ f_\theta^{-1}(\boldsymbol{h}; \boldsymbol{x}) \in A \}\, p_{\boldsymbol{h}}(\boldsymbol{h})\, d\boldsymbol{h}.$$

The parameters θ, typically neural-network weights inside the flow transformations, are learned by maximizing the conditional log-likelihood $\sum_i \log p_\theta(\boldsymbol{c}_i \mid \boldsymbol{x}_i)$. CNF can capture complex, high-dimensional, and potentially multimodal conditional distributions while retaining tractable likelihood evaluation and efficient sampling. Besides, the invertible and differentiable transformation of CNF fits decision-focused training very well.

B.3 Gaussian Mixture Density Network

A Gaussian mixture density network models a conditional distribution as a finite mixture of Gaussian components whose parameters are produced by a neural network (Bishop, 1994). Given a feature $\boldsymbol{x}$, a GMDN with K components defines the conditional density

$$p_\theta(\boldsymbol{c} \mid \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}; \theta)\, \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu}_k(\boldsymbol{x}; \theta), \Sigma_k(\boldsymbol{x}; \theta))\,, \tag{21}$$

where θ denotes the learnable parameters of the network. For each $\boldsymbol{x}$ the network outputs the mixture weights $\pi_k(\boldsymbol{x}; \theta) \geq 0$ with $\sum_{k=1}^{K} \pi_k(\boldsymbol{x}; \theta) = 1$, the component means $\boldsymbol{\mu}_k(\boldsymbol{x}; \theta) \in \mathbb{R}^{n_c}$ and the component covariances $\Sigma_k(\boldsymbol{x}; \theta) \succ 0$, the last of these through a Cholesky factor $\boldsymbol{L}_k(\boldsymbol{x}; \theta)$ so that $\Sigma_k(\boldsymbol{x}; \theta) = \boldsymbol{L}_k(\boldsymbol{x}; \theta)\boldsymbol{L}_k(\boldsymbol{x}; \theta)^\top$.

According to Dalal & Hall (1983), a linear combination of Gaussian distributions is a flexible universal approximator of broad classes of densities under suitable conditions. This ensures the representation power of a GMDN. Another popular choice for the probabilistic model is a conditional diffusion model (Sohl-Dickstein et al., 2015).

C Proof of Exchangeability of the Data Points

Assume the same sampling schemes described in equation (15) and (16) in Appendix A.

Proposition C.1 (Exchangeability of conditional distributions)**.** *Let $(\boldsymbol{x}, \boldsymbol{c}) \sim P$ on standard Borel spaces $\mathcal{X} \times \mathcal{C}$. $P_{\boldsymbol{x}}$ denotes the $\boldsymbol{x}$-marginal and $\boldsymbol{x}_i, \ldots, \boldsymbol{x}_N \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{x}}$. Let $K(\boldsymbol{x}, B) := P(\boldsymbol{c} \in B | \boldsymbol{x})$, where $B \in \mathcal{C}$ be a regular conditional distribution. Then the sequence $\{(\boldsymbol{x}_i, K(\boldsymbol{x}_i, \cdot))\}_{i=1}^N$ is exchangeable.*

Proof. Define $T(\boldsymbol{x}) := (\boldsymbol{x}, K(\boldsymbol{x}, \cdot))$. Since K is a probability kernel, T is measurable and $\big(\boldsymbol{x}_i, K(\boldsymbol{x}_i, \cdot)\big) = T(\boldsymbol{x}_i)$. Since $\boldsymbol{x}_i$ are i.i.d. and the same measurable map T is applied to every $\boldsymbol{x}_i$, the random elements $T(\boldsymbol{x}_i)$ are i.i.d. Therefore, if Q denotes their common distribution, then for any N and any permutation π of $\{1, \ldots, N\}$,

$$\mathcal{L}(T(\boldsymbol{x}_1), \ldots, T(\boldsymbol{x}_N)) = Q^{\otimes N} = \mathcal{L}(T(\boldsymbol{x}_{\pi(1)}), \ldots, T(\boldsymbol{x}_{\pi(N)})).$$

Hence $\{(\boldsymbol{x}_i, K(\boldsymbol{x}_i, \cdot))\}_{i=1}^N$ is exchangeable. □

Proposition 2.1 (Exchangeability)**.** *The calibration and test data points are exchangeable.*

Proof. Use the same notations and abbreviations as in Proposition C.1. Conditional on $\{\boldsymbol{x}_i\}_{i=1}^N$, $\{\{\boldsymbol{c}_i^j\}_{j=1}^S\}_{i=1}^N$ are mutually independent. Let $W_i := \left(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S\right)$. For measurable $A_i \subseteq \mathcal{X}$ and $B_i^1, \ldots, B_i^S \subseteq \mathcal{C}$, conditional independence implies

$$\Pr\left(\boldsymbol{x}_i \in A_i,\, C_i^j \in B_i^j \text{ for all } i, j\right) \quad = \mathbb{E}\left[\prod_{i=1}^N \mathbf{1}_{\{\boldsymbol{x}_i \in A_i\}} \prod_{j=1}^S K(\boldsymbol{x}_i, B_i^j)\right]. \tag{22}$$

Since $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are independent, equation (22) factorizes as $\prod_{i=1}^N \int_{A_i} \prod_{j=1}^S K(\boldsymbol{x}, B_i^j)\, P_{\boldsymbol{x}}(d\boldsymbol{x})$. Thus $W_1, \ldots, W_N$ are independent. Moreover, each W_i has the same distribution

$$Q_S(d\boldsymbol{x}, d\boldsymbol{c}^1, \ldots, d\boldsymbol{c}^S) = P_{\boldsymbol{x}}(d\boldsymbol{x}) \prod_{j=1}^S K(\boldsymbol{x}, d\boldsymbol{c}^j).$$

Hence $W_1, W_2, \ldots$ are i.i.d. and for every permutation π of $\{1, \ldots, N\}$, $(W_1, \ldots, W_N) \overset{d}{=} (W_{\pi(1)}, \ldots, W_{\pi(N)})$. Thus the sequence is exchangeable. □

D Proof of the Estimation Error of SD^1_ϵ for Wasserstein-1 Distance

D.1 Proof of Lemma 3.1

Lemma 3.1 (Estimation error for W_1)**.** *For each $\boldsymbol{x}_i$,*

$$\left|\mathrm{SD}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) - W_1(P_{\star,i}, P_{\theta,i})\right| \le \frac{\epsilon}{2}\log(SM) + W_1(\widehat{P}_{\star,i}, P_{\star,i}) + W_1(\widehat{P}_{\theta,i}, P_{\theta,i}). \tag{4}$$

Proof. Fix $\boldsymbol{x}_i$. Since $P_{\star,i}, P_{\theta,i} \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c})$, all Wasserstein-1 distances below are finite.

We first bound the difference between the Sinkhorn divergence and the Wasserstein distance between the two empirical measures. The diagonal coupling of $\widehat{P}_{\star,i}$ has zero transport cost, and its Kullback–Leibler divergence with respect to $\widehat{P}_{\star,i} \otimes \widehat{P}_{\star,i}$ is the Shannon entropy of the atomic weights of $\widehat{P}_{\star,i}$. Since $\widehat{P}_{\star,i}$ has at most S distinct atoms, this entropy is at most $\log S$. The same argument applies to $\widehat{P}_{\theta,i}$, which has at most M distinct atoms. Consequently,

$$\begin{aligned} 0 \le \mathrm{OT}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\star,i}\big) \le \epsilon \log S, \\ 0 \le \mathrm{OT}^1_\epsilon\big(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i}\big) \le \epsilon \log M. \end{aligned} \tag{23}$$

Moreover, nonnegativity of the Kullback–Leibler divergence gives

$$W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \le \mathrm{OT}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big). \tag{24}$$

For the reverse bound, let π_0 be an optimal coupling for $W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$. After merging repeated samples, every positive atomic weight of $\widehat{P}_{\star,i}$ is at least $1/S$, and every positive atomic weight of $\widehat{P}_{\theta,i}$ is at least $1/M$. The marginal constraints therefore imply, for every pair of atoms $(\boldsymbol{u}, \boldsymbol{v})$,

$$\frac{\pi_0(\{(\boldsymbol{u}, \boldsymbol{v})\})}{\widehat{P}_{\star,i}(\{\boldsymbol{u}\})\widehat{P}_{\theta,i}(\{\boldsymbol{v}\})} \le \frac{1}{\max\Big\{\widehat{P}_{\star,i}(\{\boldsymbol{u}\}), \widehat{P}_{\theta,i}(\{\boldsymbol{v}\})\Big\}} \le \min\{S, M\}.$$

It follows that

$$D_{\mathrm{KL}}\Big(\pi_0 \,\Big\|\, \widehat{P}_{\star,i} \otimes \widehat{P}_{\theta,i}\Big) \le \log \min\{S, M\}.$$

Evaluating the entropically regularized transport objective at π_0 yields

$$\mathrm{OT}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \le W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \epsilon \log \min\{S, M\}. \tag{25}$$

Using (23) and (24) in the definition of SD^1_ϵ gives

$$\begin{aligned} \mathrm{SD}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) &\ge W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - \frac{\epsilon}{2}\log S - \frac{\epsilon}{2}\log M \\ &= W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - \frac{\epsilon}{2}\log(SM). \end{aligned}$$

Conversely, the two self-transport terms are nonnegative, so (25) gives

$$\begin{aligned} \mathrm{SD}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) &\le W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \epsilon \log \min\{S, M\} \\ &\le W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM), \end{aligned}$$

where the last inequality follows from $\min\{S, M\} \le \sqrt{SM}$. Hence,

$$\Big|\mathrm{SD}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\Big| \le \frac{\epsilon}{2}\log(SM). \tag{26}$$

Since W_1 is a metric, the triangle inequality yields

$$\Big|W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_1\big(P_{\star,i}, P_{\theta,i}\big)\Big| \le W_1\big(\widehat{P}_{\star,i}, P_{\star,i}\big) + W_1\big(\widehat{P}_{\theta,i}, P_{\theta,i}\big). \tag{27}$$

Combining (26) and (27) proves

$$\Big|\mathrm{SD}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_1\big(P_{\star,i}, P_{\theta,i}\big)\Big| \le \frac{\epsilon}{2}\log(SM) + W_1\big(\widehat{P}_{\star,i}, P_{\star,i}\big) + W_1\big(\widehat{P}_{\theta,i}, P_{\theta,i}\big).$$

□

E Proof of the Marginal Coverage Guarantee of using SD^1_ϵ for W_1

E.1 Proof of Proposition 3.1

Proposition 3.1 (Probabilistic error bound for W_1)**.** *Assume that $P_{\star,i}$ is supported on a set $\mathcal{C}_i \subseteq \mathbb{R}^{n_c}$ of diameter at most D, and $P_{\theta,i}$ is supported on a set of diameter at most $\widehat{D}$, for every i. Then, for every $\eta \in (0,1)$,*

$$\mathbb{P}\Big(W_1(P_{\star,i}, P_{\theta,i}) \le \mathrm{SD}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) + \frac{\epsilon}{2}\log(SM) + r^{(1)}_{S,M}(\eta)\Big) \ge 1-\eta, \tag{5}$$

with $r^{(1)}_{S,M}(\eta) := \sqrt{\frac{\log(1/\eta)}{2}\left(\frac{D^2}{S} + \frac{\widehat{D}^2}{M}\right)}$.

Proof. Fix $\boldsymbol{x}_i$. The proof has three steps: a deterministic step that removes the entropic regularization, an averaging step that relates the empirical Wasserstein distance to the target distance, and a concentration step that relates the empirical Wasserstein distance to its expectation.

Step 1: from the nonconformity score to the plug-in distance. By (26),

$$W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \le \mathrm{SD}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM). \tag{28}$$

This inequality is deterministic. It holds for every realization of the samples from the conditional distributions and for arbitrary support. As a result, it consumes no failure probability.

Step 2: the plug-in distance dominates the target in expectation. By Kantorovich–Rubinstein duality (Villani, 2009, Theorem 5.10),

$$W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) = \sup_{\substack{f:\,\mathrm{Lip}(f)\le 1\\ f(\mathbf{0})=0}} \left\{\int f\,\mathrm{d}\widehat{P}_{\star,i} - \int f\,\mathrm{d}\widehat{P}_{\theta,i}\right\}.$$

For every such f,

$$\mathbb{E}\left[\int f\,\mathrm{d}\widehat{P}_{\star,i}\right] = \int f\,\mathrm{d}P_{\star,i}, \qquad \mathbb{E}\left[\int f\,\mathrm{d}\widehat{P}_{\theta,i}\right] = \int f\,\mathrm{d}P_{\theta,i}.$$

Therefore, using the fact that the expectation of a supremum is at least the supremum of the expectations gives

$$\begin{aligned}\mathbb{E}\left[W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right] &\ge \sup_{\substack{f:\,\mathrm{Lip}(f)\le 1\\ f(\mathbf{0})=0}} \left\{\int f\,\mathrm{d}P_{\star,i} - \int f\,\mathrm{d}P_{\theta,i}\right\}\\ &= W_1\big(P_{\star,i}, P_{\theta,i}\big).\end{aligned} \tag{29}$$

This is again a sure inequality with no additive slack.

Step 3: concentration of the plug-in distance. View $W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ as a function of the $S+M$ independent random vectors

$$\big(\boldsymbol{c}_i^1, \dots, \boldsymbol{c}_i^S, \tilde{\boldsymbol{c}}_i^1, \dots, \tilde{\boldsymbol{c}}_i^M\big).$$

Replacing a single sample $\boldsymbol{c}_i^j$ by $\boldsymbol{c}_i' \in \mathcal{C}_i$ produces an empirical measure $\widehat{P}'_{\star,i}$ satisfying

$$W_1\big(\widehat{P}_{\star,i}, \widehat{P}'_{\star,i}\big) \le \frac{1}{S}\|\boldsymbol{c}_i^j - \boldsymbol{c}_i'\|_2 \le \frac{D}{S}.$$

The triangle inequality for W_1 therefore gives

$$\Big|W_1\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_1\big(\widehat{P}'_{\star,i}, \widehat{P}_{\theta,i}\big)\Big| \le \frac{D}{S}.$$

The same argument shows that replacing one generated sample $\tilde{\boldsymbol{c}}_i^j$ changes $W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ by at most $\widehat{D}/M$.

McDiarmid's inequality (McDiarmid, 1989) consequently gives, for every $t > 0$,

$$\mathbb{P}\left(\mathbb{E}\left[W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})\right] - W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) \geq t\right) \leq \exp\left(-\frac{2t^2}{S(D/S)^2 + M(\widehat{D}/M)^2}\right) = \exp\left(-\frac{2t^2}{D^2/S + \widehat{D}^2/M}\right).$$

Setting the right-hand side equal to η shows that, with probability at least $1 - \eta$,

$$\mathbb{E}\left[W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})\right] \leq W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) + r^{(1)}_{S,M}(\eta). \tag{30}$$

Conclusion. Combining (29), (30), and (28) shows that, with probability at least $1 - \eta$,

$$\begin{aligned} W_1(P_{\star,i}, P_{\theta,i}) &\leq \mathbb{E}\left[W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})\right] \\ &\leq W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) + r^{(1)}_{S,M}(\eta) \\ &\leq \mathrm{SD}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) + \frac{\epsilon}{2}\log(SM) + r^{(1)}_{S,M}(\eta), \end{aligned}$$

which proves (5). □

E.2 Explanation of Remark 3.1

Step 3 in the proof of Proposition 3.1 is the only place where boundedness is used. Suppose that, uniformly in i, $P_{\star,i}$ and $P_{\theta,i}$ satisfy the quadratic transport-entropy inequalities $T_2(\nu_P^2)$ and $T_2(\nu_\theta^2)$, respectively, for some $\nu_P, \nu_\theta > 0$.

A probability measure μ satisfies $T_2(\nu^2)$ if

$$W_2(\gamma, \mu) \leq \nu\sqrt{2D_{\mathrm{KL}}(\gamma\|\mu)} \qquad \text{for every } \gamma \ll \mu.$$

Transport-entropy inequalities are stable under Lipschitz pushforwards. Hence, after scaling the observed samples by $1/\nu_P$ and the generated samples by $1/\nu_\theta$, every factor satisfies $T_2(1)$. Tensorization then implies that the joint law of the scaled sample configuration satisfies $T_2(1)$ with respect to the product Euclidean norm (Villani, 2009).

Consider two sample configurations and the corresponding empirical measures $\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}$ and $\widehat{P}'_{\star,i}, \widehat{P}'_{\theta,i}$. The triangle inequality and the coupling that matches samples with the same indices give

$$\begin{aligned} &\left|W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) - W_1(\widehat{P}'_{\star,i}, \widehat{P}'_{\theta,i})\right| \\ &\quad\leq W_1(\widehat{P}_{\star,i}, \widehat{P}'_{\star,i}) + W_1(\widehat{P}_{\theta,i}, \widehat{P}'_{\theta,i}) \\ &\quad\leq \frac{1}{S}\sum_{j=1}^{S}\|\boldsymbol{c}_i^j - \boldsymbol{c}_i'^j\|_2 + \frac{1}{M}\sum_{j=1}^{M}\|\tilde{\boldsymbol{c}}_i^j - \tilde{\boldsymbol{c}}_i'^j\|_2 \\ &\quad\leq \sqrt{\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}}\left(\sum_{j=1}^{S}\left\|\frac{\boldsymbol{c}_i^j - \boldsymbol{c}_i'^j}{\nu_P}\right\|_2^2 + \sum_{j=1}^{M}\left\|\frac{\tilde{\boldsymbol{c}}_i^j - \tilde{\boldsymbol{c}}_i'^j}{\nu_\theta}\right\|_2^2\right)^{1/2}, \end{aligned}$$

where the last inequality follows by applying the Cauchy–Schwarz inequality within each sample block and then across the two blocks. Thus, as a function of the scaled sample configuration, $W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ is Lipschitz with constant

$$\sqrt{\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}}.$$

The concentration inequality implied by $T_2(1)$ therefore gives

$$\mathbb{P}\left(\mathbb{E}\left[W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})\right] - W_1(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) \geq t\right) \leq \exp\left(-\frac{t^2}{2\left(\nu_P^2/S + \nu_\theta^2/M\right)}\right).$$

Setting the right-hand side equal to η gives

$$t = \sqrt{2\left(\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}\right)\log(1/\eta)}.$$

Replacing $r_{S,M}^{(1)}(\eta)$ by this quantity in Proposition 3.1 and Theorem 3.1 gives the unbounded-support extension stated in Remark 3.1.

Algebraically, substituting $\nu_P = D/2$ and $\nu_\theta = \widehat{D}/2$ into this expression recovers $r_{S,M}^{(1)}(\eta)$.

E.3 Proof of Theorem 3.1

Theorem 3.1 (Finite-sample marginal coverage for W_1)**.** *Fix the distribution predictor P_θ and assume the conditions of Proposition 3.1. For each calibration point $i = 1, \dots, N_{\mathrm{cal}}$, define the nonconformity score $s_i := \mathrm{SD}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$. Let $s_{(1)} \le \cdots \le s_{(N_{\mathrm{cal}})}$ denote the corresponding order statistics, with $s_{(N_{\mathrm{cal}}+1)} := +\infty$. For $\tau \in (0,1)$ and $\eta \in (0, \tau)$, define*

$$k := \lceil (N_{\mathrm{cal}} + 1)(1 - \tau + \eta) \rceil, \quad \rho_\theta := s_{(k)} + \frac{\epsilon}{2} \log(SM) + r^{(1)}_{S,M}(\eta).$$

Then, for a test point $\boldsymbol{x}^{\mathrm{test}}$, we have

$$\mathbb{P}\left(W_1\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta \right) \ge 1 - \tau. \tag{6}$$

Thus, the ambiguity set in (1) contains the true conditional distribution with probability at least $1 - \tau$.

Proof. Since P_θ is fixed independently of the calibration data, the mapping from each augmented data point

$$\left(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S, \{\tilde{\boldsymbol{c}}_i^j\}_{j=1}^M \right)$$

to its Sinkhorn nonconformity score is fixed and measurable. By Proposition 2.1, the data points $(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S)$ are exchangeable. Moreover, the samples $\{\tilde{\boldsymbol{c}}_i^j\}_{j=1}^M$ are generated independently across data points from the same fixed predictor P_θ, with the same number M of samples for every point. Hence, the $N_{\mathrm{cal}} + 1$ augmented data points, consisting of the N_{cal} calibration points and one test point, are exchangeable.

Let s_{test} denote the Sinkhorn nonconformity score of the test point, constructed in the same way as the calibration scores $s_1, \dots, s_{N_{\mathrm{cal}}}$. The empirical measures used to define s_{test} are introduced only for the proof and are not required at deployment. The scores

$$s_1, \dots, s_{N_{\mathrm{cal}}}, s_{\mathrm{test}}$$

are therefore exchangeable.

The rank event. With ties broken uniformly at random, exchangeability makes the rank of s_{test} among the $N_{\mathrm{cal}} + 1$ scores uniform on $\{1, \dots, N_{\mathrm{cal}} + 1\}$. Therefore, s_{test} is at most the k-th smallest of all $N_{\mathrm{cal}} + 1$ scores with probability at least

$$\frac{k}{N_{\mathrm{cal}} + 1} \ge 1 - (\tau - \eta).$$

Removing the test score cannot decrease the k-th order statistic, where $s_{(N_{\mathrm{cal}}+1)} = +\infty$ by convention. Consequently,

$$\mathbb{P}\left(s_{\mathrm{test}} \le s_{(k)} \right) \ge 1 - (\tau - \eta). \tag{31}$$

No continuity of the score distribution is required.

The sampling event. Applying Proposition 3.1 at the test point gives

$$\mathbb{P}\left(W_1\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le s_{\mathrm{test}} + \frac{\epsilon}{2} \log(SM) + r^{(1)}_{S,M}(\eta) \right) \ge 1 - \eta. \tag{32}$$

Composition. On the intersection of the events in (31) and (32),

$$\begin{aligned} W_1\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) &\le s_{\mathrm{test}} + \frac{\epsilon}{2} \log(SM) + r^{(1)}_{S,M}(\eta) \\ &\le s_{(k)} + \frac{\epsilon}{2} \log(SM) + r^{(1)}_{S,M}(\eta) \\ &= \rho_\theta. \end{aligned}$$

The union bound therefore yields

$$\begin{aligned} &\mathbb{P}\left(W_1\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta \right) \\ &\quad \ge 1 - (\tau - \eta) - \eta = 1 - \tau, \end{aligned}$$

which proves (6). No independence between the rank event and the sampling event is required.

The event in (6) states precisely that

$$P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}) \in \left\{ Q : W_1\big(Q, P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta \right\},$$

which is the Wasserstein ambiguity set in (1) with radius ρ_θ. □

F Proof of the Estimation Error of SD_ϵ^2 for Squared Wasserstein-2 Distance

Lemma F.1 (Estimation error for W_2^2)**.** *For each $\boldsymbol{x}_i$, define $\delta_i := W_2\big(\widehat{P}_{\star,i}, P_{\star,i}\big) + W_2\big(\widehat{P}_{\theta,i}, P_{\theta,i}\big)$. Then,*

$$\Big|\mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_2^2\big(P_{\star,i}, P_{\theta,i}\big)\Big| \le \frac{\epsilon}{2}\log(SM) + 2\delta_i W_2\big(P_{\star,i}, P_{\theta,i}\big) + \delta_i^2. \tag{33}$$

Proof. Fix $\boldsymbol{x}_i$. Since $P_{\star,i}, P_{\theta,i} \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c})$, all Wasserstein-2 distances below are finite.

The arguments used in the proof of Lemma 3.1 depend only on the nonnegativity of the transport cost, its vanishing on the diagonal, and the marginal constraints on the transport plans. They therefore apply to the quadratic cost and give

$$\begin{aligned} 0 \le \mathrm{OT}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\star,i}\big) &\le \epsilon \log S, \\ 0 \le \mathrm{OT}_\epsilon^2\big(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i}\big) &\le \epsilon \log M, \end{aligned} \tag{34}$$

and

$$\begin{aligned} W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) &\le \mathrm{OT}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \\ &\le W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \epsilon \log \min\{S, M\}. \end{aligned} \tag{35}$$

Using (34) and the lower bound in (35) in the definition of the Sinkhorn divergence gives

$$\begin{aligned} \mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) &\ge W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - \frac{\epsilon}{2}\log S - \frac{\epsilon}{2}\log M \\ &= W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - \frac{\epsilon}{2}\log(SM). \end{aligned}$$

Conversely, the two self-transport terms are nonnegative, so the upper bound in (35) gives

$$\begin{aligned} \mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) &\le W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \epsilon \log \min\{S, M\} \\ &\le W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM). \end{aligned}$$

Thus,

$$\Big|\mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\Big| \le \frac{\epsilon}{2}\log(SM). \tag{36}$$

Since W_2 is a metric, the quadrilateral inequality yields

$$\begin{aligned} &\Big|W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_2\big(P_{\star,i}, P_{\theta,i}\big)\Big| \\ &\quad\le W_2\big(\widehat{P}_{\star,i}, P_{\star,i}\big) + W_2\big(\widehat{P}_{\theta,i}, P_{\theta,i}\big) = \delta_i. \end{aligned} \tag{37}$$

It follows that

$$W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + W_2\big(P_{\star,i}, P_{\theta,i}\big) \le 2W_2\big(P_{\star,i}, P_{\theta,i}\big) + \delta_i.$$

Consequently,

$$\Big|W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_2^2\big(P_{\star,i}, P_{\theta,i}\big)\Big| \le \delta_i\left(2W_2\big(P_{\star,i}, P_{\theta,i}\big) + \delta_i\right) = 2\delta_i W_2\big(P_{\star,i}, P_{\theta,i}\big) + \delta_i^2.$$

Combining this inequality with (36) proves (33). □

The entropic correction in (33) is the same as for Wasserstein-1 and vanishes as $\epsilon \to 0$, whereas the sampling error is amplified by the size of the quantity being estimated.

Remark F.1 (Unsquared form). Since $e^{-\|\boldsymbol{p}-\boldsymbol{q}\|_2^2/\epsilon}$ is a Gaussian positive-definite kernel, $\mathrm{SD}_\epsilon^2 \ge 0$ (Feydy et al., 2019). Therefore,

$$\left|\sqrt{\mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)} - W_2\big(P_{\star,i}, P_{\theta,i}\big)\right| \le \sqrt{\frac{\epsilon}{2}\log(SM) + 2\delta_i W_2\big(P_{\star,i}, P_{\theta,i}\big) + \delta_i^2}.$$

If $W_2(P_{\star,i}, P_{\theta,i}) > 0$, we also have

$$\left| \sqrt{\mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)} - W_2\big(P_{\star,i}, P_{\theta,i}\big) \right| \leq \frac{\frac{\epsilon}{2}\log(SM) + 2\delta_i W_2\big(P_{\star,i}, P_{\theta,i}\big) + \delta_i^2}{W_2\big(P_{\star,i}, P_{\theta,i}\big)}.$$

These bounds follow from $|\sqrt{u} - \sqrt{v}| \leq \sqrt{|u - v|}$ and

$$|\sqrt{u} - \sqrt{v}| = \frac{|u - v|}{\sqrt{u} + \sqrt{v}} \leq \frac{|u - v|}{\sqrt{v}}, \qquad u, v \geq 0.$$

G Proof of the Marginal Coverage Guarantee of using SD_ϵ^2 for W_2^2

We condition on the training data used to obtain the fixed predictor P_θ. We first establish the probabilistic error bound for SD_ϵ^2 and then use it to prove the marginal coverage guarantee.

Proposition G.1 (Probabilistic error bound for W_2^2)**.** *Assume that, for every i, $P_{\star,i}$ is supported on a set $\mathcal{C}_i \subseteq \mathbb{R}^{n_c}$ of diameter at most D, $P_{\theta,i}$ is supported on a set of diameter at most $\widehat{D}$, and $\operatorname{diam}\big(\mathcal{C}_i \cup \operatorname{supp} P_{\theta,i}\big) \le D_{\mathrm{uni}}$. Then, for every $\eta \in (0,1)$,*

$$\mathbb{P}\Big(W_2^2\big(P_{\star,i}, P_{\theta,i}\big) \le \mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(2)}(\eta)\Big) \ge 1-\eta, \tag{38}$$

where $r_{S,M}^{(2)}(\eta) := 2D_{\mathrm{uni}}\, r_{S,M}^{(1)}(\eta)$. The deterministic correction $\frac{\epsilon}{2}\log(SM)$ may be replaced by the sharper data-dependent correction

$$\frac{1}{2}\mathrm{OT}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\star,i}\big) + \frac{1}{2}\mathrm{OT}_\epsilon^2\big(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i}\big).$$

Proof. Fix $\boldsymbol{x}_i$. The proof has three steps.

Step 1: from the nonconformity score to the plug-in distance. Nonnegativity of the relative entropy implies

$$\begin{aligned} W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) &\le \mathrm{OT}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \\ &= \mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{1}{2}\mathrm{OT}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\star,i}\big) + \frac{1}{2}\mathrm{OT}_\epsilon^2\big(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i}\big) \\ &\le \mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM), \end{aligned} \tag{39}$$

where the last inequality follows from (34). This bound is deterministic and therefore consumes no failure probability.

Step 2: the plug-in distance dominates the target in expectation. The squared Wasserstein distance is jointly convex in its two arguments. Since

$$\mathbb{E}[\widehat{P}_{\star,i}] = P_{\star,i}, \qquad \mathbb{E}[\widehat{P}_{\theta,i}] = P_{\theta,i},$$

joint convexity gives

$$W_2^2\big(P_{\star,i}, P_{\theta,i}\big) \le \mathbb{E}\left[W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right]. \tag{40}$$

Step 3: concentration of the plug-in distance. Replacing one sample $\boldsymbol{c}_i^j$ by $\boldsymbol{c}_i' \in \mathcal{C}_i$ produces an empirical measure $\widehat{P}'_{\star,i}$. Rerouting the mass $1/S$ transported from $\boldsymbol{c}_i^j$ so that it instead departs from $\boldsymbol{c}_i'$ and then exchanging the two sample configurations gives

$$\begin{aligned} \left|W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_2^2\big(\widehat{P}'_{\star,i}, \widehat{P}_{\theta,i}\big)\right| &\le \frac{1}{S}\sup_{\boldsymbol{v}\in\operatorname{supp} P_{\theta,i}} \left|\|\boldsymbol{c}_i' - \boldsymbol{v}\|_2^2 - \|\boldsymbol{c}_i^j - \boldsymbol{v}\|_2^2\right| \\ &\le \frac{2D_{\mathrm{uni}}D}{S}. \end{aligned}$$

The same argument shows that replacing one generated sample changes $W_2^2(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ by at most $2D_{\mathrm{uni}}\widehat{D}/M$.

McDiarmid's inequality therefore gives, for every $t > 0$,

$$\mathbb{P}\left(\mathbb{E}\left[W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right] - W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \ge t\right) \le \exp\left(-\frac{2t^2}{4D_{\mathrm{uni}}^2\left(D^2/S + \widehat{D}^2/M\right)}\right).$$

Setting the right-hand side equal to η shows that, with probability at least $1-\eta$,

$$\mathbb{E}\left[W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right] \le W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + r_{S,M}^{(2)}(\eta). \tag{41}$$

Conclusion. Combining (40), (41), and (39) shows that, with probability at least $1-\eta$,

$$\begin{aligned} W_2^2\big(P_{\star,i}, P_{\theta,i}\big) &\le \mathbb{E}\left[W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right] \\ &\le W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + r_{S,M}^{(2)}(\eta) \\ &\le \mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(2)}(\eta), \end{aligned}$$

which proves (38). □

G.1 Proof of Theorem 3.2

Theorem 3.2 (Finite-sample marginal coverage for W_2^2)**.** *Fix the distribution predictor P_θ. Assume that $P_{\star,i}$ is supported on a set $\mathcal{C}_i \subseteq \mathbb{R}^{n_c}$ of diameter at most D, and $P_{\theta,i}$ is supported on a set of diameter at most $\widehat{D}$, for every i. We further assume that* $\mathrm{diam}(\mathcal{C}_i \cup \mathrm{supp}\, P_{\theta,i}) \le D_{\mathrm{uni}}$ *for every i. For each calibration point $i = 1, \dots, N_{\mathrm{cal}}$, define the nonconformity score $s_i := \mathrm{SD}_\epsilon^2(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$. Let $s_{(1)} \le \cdots \le s_{(N_{\mathrm{cal}})}$ denote the corresponding order statistics, with $s_{(N_{\mathrm{cal}}+1)} := +\infty$. For $\tau \in (0,1)$ and $\eta \in (0,\tau)$, define*

$$k := \lceil (N_{\mathrm{cal}}+1)(1-\tau+\eta) \rceil, \quad \rho_\theta := s_{(k)} + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(2)}(\eta), \quad r_{S,M}^{(2)}(\eta) := 2D_{\mathrm{uni}}\, r_{S,M}^{(1)}(\eta).$$

Then, for a test point $\boldsymbol{x}^{\mathrm{test}}$, we have

$$\mathbb{P}\left(W_2^2\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta\right) \ge 1-\tau. \tag{7}$$

Thus, the ambiguity set in (1) contains the true conditional distribution with probability at least $1-\tau$.

Proof. Since P_θ is fixed independently of the calibration data, the mapping from each augmented data point

$$\left(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S, \{\tilde{\boldsymbol{c}}_i^j\}_{j=1}^M\right)$$

to its Sinkhorn nonconformity score is fixed and measurable. By Proposition 2.1, the data points $(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S)$ are exchangeable. Moreover, the particles $\{\tilde{\boldsymbol{c}}_i^j\}_{j=1}^M$ are generated independently across data points using the same fixed predictor P_θ and the same number M of particles. Hence, the $N_{\mathrm{cal}}+1$ augmented data points are exchangeable.

Let s_{test} denote the Sinkhorn nonconformity score of the test point, constructed in the same way as the calibration scores. The empirical measures used to construct s_{test} are introduced only for the proof and are not required at deployment. Thus, $s_1, \dots, s_{N_{\mathrm{cal}}}, s_{\mathrm{test}}$ are exchangeable.

The rank event. With ties broken uniformly at random, the rank of s_{test} among the $N_{\mathrm{cal}}+1$ scores is uniform on $\{1, \dots, N_{\mathrm{cal}}+1\}$. Therefore,

$$\mathbb{P}\left(s_{\mathrm{test}} \le s_{(k)}\right) \ge \frac{k}{N_{\mathrm{cal}}+1} \ge 1-(\tau-\eta). \tag{42}$$

Here $s_{(N_{\mathrm{cal}}+1)} = +\infty$ by convention. No continuity of the score distribution is required.

The sampling event. Applying Proposition G.1 at the test point gives

$$\mathbb{P}\left(W_2^2\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le s_{\mathrm{test}} + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(2)}(\eta)\right) \ge 1-\eta. \tag{43}$$

The bound holds conditionally on each test feature and hence also after averaging over $\boldsymbol{x}^{\mathrm{test}}$.

Composition. On the intersection of the events in (42) and (43),

$$\begin{aligned} W_2^2\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) &\le s_{\mathrm{test}} + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(2)}(\eta) \\ &\le s_{(k)} + \frac{\epsilon}{2}\log(SM) + r_{S,M}^{(2)}(\eta) \\ &= \rho_\theta. \end{aligned}$$

The union bound therefore gives

$$\mathbb{P}\left(W_2^2\big(P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}), P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta\right) \ge 1 - (\tau - \eta) - \eta = 1 - \tau,$$

which proves (7). No independence between the rank event and the sampling event is required.

The event in (7) states precisely that

$$P_\star(\cdot \mid \boldsymbol{x}^{\mathrm{test}}) \in \left\{Q : W_2^2\big(Q, P_\theta(\cdot \mid \boldsymbol{x}^{\mathrm{test}})\big) \le \rho_\theta\right\},$$

which is the Wasserstein ambiguity set in (1) with radius ρ_θ. □

G.2 Explanation of Remark 3.2

Suppose that, for $P_{\boldsymbol{x}}$-almost every $\boldsymbol{x}$, $P_\star(\cdot \mid \boldsymbol{x})$ and $P_\theta(\cdot \mid \boldsymbol{x})$ satisfy the quadratic transport-entropy inequalities $T_2(\nu_P^2)$ and $T_2(\nu_\theta^2)$, respectively, with finite positive constants ν_P and ν_θ that are uniform in $\boldsymbol{x}$.

After scaling the observed samples by $1/\nu_P$ and the generated samples by $1/\nu_\theta$, the joint law of the sample configuration satisfies $T_2(1)$ with respect to the product Euclidean norm. For two sample configurations, the triangle inequality and the couplings that match samples with the same indices give

$$\begin{aligned}
\left|W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) - W_2\big(\widehat{P}'_{\star,i}, \widehat{P}'_{\theta,i}\big)\right| &\le W_2\big(\widehat{P}_{\star,i}, \widehat{P}'_{\star,i}\big) + W_2\big(\widehat{P}_{\theta,i}, \widehat{P}'_{\theta,i}\big) \\
&\le \left(\frac{1}{S}\sum_{j=1}^{S} \|\boldsymbol{c}_i^j - \boldsymbol{c}_i'^j\|_2^2\right)^{1/2} + \left(\frac{1}{M}\sum_{j=1}^{M} \|\tilde{\boldsymbol{c}}_i^j - \tilde{\boldsymbol{c}}_i'^j\|_2^2\right)^{1/2} \\
&\le \sqrt{\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}} \left(\sum_{j=1}^{S} \left\|\frac{\boldsymbol{c}_i^j - \boldsymbol{c}_i'^j}{\nu_P}\right\|_2^2 + \sum_{j=1}^{M} \left\|\frac{\tilde{\boldsymbol{c}}_i^j - \tilde{\boldsymbol{c}}_i'^j}{\nu_\theta}\right\|_2^2\right)^{1/2}.
\end{aligned}$$

Thus, $W_2(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ is Lipschitz as a function of the scaled sample configuration with constant

$$\sqrt{\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}}.$$

The concentration inequality implied by $T_2(1)$ gives

$$\mathbb{P}\left(\mathbb{E}\left[W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right] - W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \ge \sqrt{2\left(\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}\right)\log(1/\eta)}\right) \le \eta.$$

The same sub-Gaussian concentration inequality implies

$$\mathrm{Var}\left(W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right) \le \frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}.$$

By joint convexity of W_2^2,

$$\begin{aligned}
W_2^2\big(P_{\star,i}, P_{\theta,i}\big) &\le \mathbb{E}\left[W_2^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right] \\
&= \mathrm{Var}\left(W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right) + \left(\mathbb{E}\left[W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)\right]\right)^2.
\end{aligned}$$

Consequently, with probability at least $1 - \eta$,

$$\begin{aligned}
W_2^2\big(P_{\star,i}, P_{\theta,i}\big) \le \frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M} \\
+ \left(W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \sqrt{2\left(\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}\right)\log(1/\eta)}\right)^2.
\end{aligned}$$

Moreover, (39) implies

$$W_2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) \le \sqrt{\mathrm{SD}_\epsilon^2\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) + \frac{\epsilon}{2}\log(SM)}.$$

Combining the last two displays and then applying the conformal rank argument gives the radius

$$\rho_\theta := \left(\sqrt{s_{(k)} + \frac{\epsilon}{2}\log(SM)} + \sqrt{2\left(\frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}\right)\log(1/\eta)}\right)^2 + \frac{\nu_P^2}{S} + \frac{\nu_\theta^2}{M}.$$

With this radius, the conclusion of Theorem 3.2 continues to hold without the bounded-support assumptions.

H Marginal Coverage under Temporal Dependence

Let θ be the trained distribution-predictor parameter, which is frozen throughout calibration and testing. Let $(\boldsymbol{Z}_t)_{t\geq 1}$ be the raw process. For example, in portfolio allocation, $\boldsymbol{Z}_t$ collects the return path of trading day t. A data point is assembled from a window of this process: the feature is a fixed measurable function of the previous L periods and the S targets are functions of the current one,

$$\boldsymbol{x}_t = \phi\big(\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-L}\big), \qquad \{\boldsymbol{c}_t^j\}_{j=1}^S = \psi(\boldsymbol{Z}_t). \tag{44}$$

The M particles $\{\tilde{\boldsymbol{c}}_t^j\}_{j=1}^M$ are drawn from $P_\theta(\cdot \mid \boldsymbol{x}_t)$. The nonconformity score of period t is

$$s_t = s\big(\widehat{P}_{\star,t}, \widehat{P}_{\theta,t}\big) = \bar{s}_\theta\big(\boldsymbol{Z}_t; \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-L}\big). \tag{45}$$

Here, $\widehat{P}_{\star,t}$ and $\widehat{P}_{\theta,t}$ are the empirical measures associated with $P_{\star,t}$ and $P_{\theta,t}$, respectively. The map $\bar{s}_\theta$ composes the construction in (44) of the data point, the draw of the particles from $P_\theta(\cdot \mid \boldsymbol{x}_t)$, and the score $s(\cdot,\cdot)$. It depends on the data only through θ, so once θ is frozen, the score is a fixed measurable function of the last $L+1$ entries of $\boldsymbol{Z}$. In the terminology of Barber & Pananjady (2026), the score is a predictor with memory L.

Let

$$\boldsymbol{s} := (s_1, \ldots, s_m, s^{\text{test}})$$

be the vector of usable calibration scores followed by the test score. Its length is $m+1$. Following Barber & Pananjady (2026), for a gap $\ell \geq 0$ and a position k, compare the two subvectors of $\boldsymbol{s}$ obtained by deleting a block of ℓ entries:

$$\boldsymbol{s}_{k,\ell}^{[0]} := \big(s_1, \ldots, s_{m+1-\ell-k}, s_{m+2-k}, \ldots, s^{\text{test}}\big),$$
$$\boldsymbol{s}_{k,\ell}^{[1]} := \big(s_{k+\ell+1}, \ldots, s^{\text{test}}, s_1, \ldots, s_k\big).$$

Define the corresponding switch coefficients by

$$\Psi_{k,\ell}(\boldsymbol{s}) := d_{\text{TV}}\big(\boldsymbol{s}_{k,\ell}^{[0]}, \boldsymbol{s}_{k,\ell}^{[1]}\big), \qquad \overline{\Psi}_\ell(\boldsymbol{s}) := \frac{1}{m+1}\sum_{k=1}^{m+1} \Psi_{k,\ell}(\boldsymbol{s}).$$

The two subvectors differ only in which entry sits last, s^{test} in $\boldsymbol{s}_{k,\ell}^{[0]}$ and s_k in $\boldsymbol{s}_{k,\ell}^{[1]}$. They agree in distribution whenever $\boldsymbol{s}$ is exchangeable. Therefore, $\overline{\Psi}_\ell(\boldsymbol{s})$ quantifies how much the test score fails to be interchangeable with a calibration score once a gap of ℓ periods is allowed. Two properties make it the right quantity here. It is a coefficient of the scores, which are scalar. Thus, by the data-processing inequality, it is bounded by a switch coefficient of the raw process at a gap enlarged by the memory (Barber & Pananjady, 2026, Prop. 4). In practice, it can be far smaller. It is also controlled by mixing. For a stationary process,

$$\Psi_{k,\ell} \leq 2\beta_{\text{mix}}(\ell)$$

(Barber & Pananjady, 2026, Prop. 3).

Finally, because θ is fitted on the periods immediately preceding the calibration block, the earliest calibration scores may remain strongly dependent on θ. We therefore also work with the truncated vector $\boldsymbol{s}^{(\ell_*)}$ obtained from $\boldsymbol{s}$ by deleting its first ℓ_* entries. This truncation is an analysis device only. The radius is still computed from all m calibration scores.

Inspecting the proofs of Theorem 3.1 and Theorem 3.2, exchangeability is invoked only to assert that the rank of s^{test} among the $m+1$ scores is uniform and hence that

$$\mathbb{P}\big(\mathcal{E}_{\text{cal}}\big) \geq 1 - (\tau - \eta), \qquad \mathcal{E}_{\text{cal}} = \big\{s^{test} \leq s_{(k)}\big\}.$$

Everything else in the proofs is untouched by temporal dependence. The sampling event $\mathcal{E}_{\text{samp}}$ of Proposition 3.1 is a statement about the S targets and M particles of the test period, conditionally on $\boldsymbol{x}^{\text{test}}$, and the composition step is a union bound. It is enough to replace the display above.

Observe that $\mathcal{E}_{\text{cal}}$ is exactly the coverage event of split conformal prediction at level $1-(\tau-\eta)$, applied to the score sequence (45). With $k = \lceil (m+1)(1-\tau+\eta)\rceil$ and $s_{(m+1)} = +\infty$, the order statistic $s_{(k)}$ is the corresponding conformal quantile of (Barber & Pananjady, 2026, Eq. 5). Barber &

Pananjady (2026, Thm. 9) applied with $n_1 = N_{\mathrm{cal}}$ calibration periods, memory L, and miscoverage $\tau - \eta$ gives

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{cal}}\big) \;\geq\; 1 - (\tau - \eta) - \min_{\substack{\ell, \ell_* \geq 0 \\ \ell + \ell_* \leq m}} \left\{ \frac{\ell + (\tau - \eta)\ell_*}{m - \ell_* + 1} + \overline{\Psi}_\ell\big(\boldsymbol{s}^{(\ell_*)}\big) \right\}, \tag{46}$$

and bounding the switch coefficients of the scores by the mixing coefficients of the raw process. Barber & Pananjady (2026, Prop. 10) turns (46) into

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{cal}}\big) \;\geq\; 1 - (\tau - \eta) - \Delta_{\mathrm{mix}} \tag{47}$$

with Δ_{mix} as in (8). Composing (47) with $\mathbb{P}(\mathcal{E}_{\mathrm{samp}}) \geq 1 - \eta$ by the union bound, exactly as in the original proofs, proves Corollary 3.1:

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{cal}} \cap \mathcal{E}_{\mathrm{samp}}\big) \;\geq\; 1 - (\tau - \eta) - \Delta_{\mathrm{mix}} - \eta \;=\; 1 - \tau - \Delta_{\mathrm{mix}}.$$

The same substitution in the proof of Theorem 3.2 gives the W_2^2 statement. Since Theorem 4.1 only invokes the coverage event of Theorem 3.1 or Theorem 3.2, the decision guarantee holds at level $1 - \tau - \Delta_{\mathrm{mix}}$ as well. Taking $\beta_{\mathrm{mix}} \equiv 0$, $L = 0$ and $\ell = \ell_* = 0$ recovers $\Delta_{\mathrm{mix}} = 0$ and the original theorems.

I Reformulations of the Wasserstein DRO Problems as Stochastic Optimization Problems

Throughout this appendix and Appendix J the feature $\boldsymbol{x}_i$ is fixed and we abbreviate $P_{\theta,i} := P_\theta(\cdot \mid \boldsymbol{x}_i)$. We write the pieces of the task loss of Assumption 2.1 as

$$\ell_k(\boldsymbol{z}, \boldsymbol{c}) := \boldsymbol{a}_k(\boldsymbol{x}_i, \boldsymbol{z})^\top \boldsymbol{c} + b_k(\boldsymbol{x}_i, \boldsymbol{z}), \qquad f(\boldsymbol{x}_i, \boldsymbol{z}, \boldsymbol{c}) = \max_{k \in [N_K]} \ell_k(\boldsymbol{z}, \boldsymbol{c}), \tag{48}$$

and we denote the largest dual norm of the pieces by

$$L(\boldsymbol{x}_i, \boldsymbol{z}) := \max_{k \in [N_K]} \big\| \boldsymbol{a}_k(\boldsymbol{x}_i, \boldsymbol{z}) \big\|_*.$$

I.1 Wasserstein-1 DRO

Proposition I.1 (W_1 reformulation)**.** *Under Assumption 2.1 and $P_{\theta,i} \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c})$. For every $\boldsymbol{z} \in \mathbb{R}^{n_z}$ and every $\rho_\theta \geq 0$,*

$$\sup_{Q\,:\,W_1(Q, P_{\theta,i}) \leq \rho_\theta} \mathbb{E}_{\tilde{\boldsymbol{c}} \sim Q}\big[f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}})\big] \;=\; \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\big[f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}})\big] + \rho_\theta\, L(\boldsymbol{x}_i, \boldsymbol{z}). \tag{49}$$

Consequently the lower level of (1) with $p = 1$ is the stochastic optimization problem

$$\boldsymbol{z}^*_\theta(\boldsymbol{x}_i) \in \operatorname*{argmin}_{\boldsymbol{z}\,:\,\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{z}) \leq \boldsymbol{0}} \Phi_1(\boldsymbol{z}, \theta) := \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\big[f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}})\big] + \rho_\theta\, L(\boldsymbol{x}_i, \boldsymbol{z}), \tag{50}$$

whose objective is convex in $\boldsymbol{z}$.

Proof. The loss $f(\boldsymbol{x}_i, \boldsymbol{z}, \cdot)$ is continuous and $P_{\theta,i}$-integrable, because $P_{\theta,i}$ has a finite first moment, so Wasserstein strong duality (Blanchet & Murthy, 2019, Thm. 1), see also (Kuhn et al., 2019), applies with the lower semicontinuous cost $c(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|$ and gives

$$\sup_{Q\,:\,W_1(Q, P_{\theta,i}) \leq \rho_\theta} \mathbb{E}_Q[f] = \inf_{\lambda \geq 0} \left\{ \lambda \rho_\theta + \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\Big[\sup_{\boldsymbol{u}} \big(f(\boldsymbol{x}_i, \boldsymbol{z}, \boldsymbol{u}) - \lambda \|\boldsymbol{u} - \tilde{\boldsymbol{c}}\|\big)\Big] \right\}.$$

Exchanging the two suprema and substituting $\boldsymbol{u} = \tilde{\boldsymbol{c}} + \boldsymbol{w}$,

$$\sup_{\boldsymbol{u}} \Big(\max_k \ell_k(\boldsymbol{z}, \boldsymbol{u}) - \lambda \|\boldsymbol{u} - \tilde{\boldsymbol{c}}\| \Big) = \max_k \Big(\ell_k(\boldsymbol{z}, \tilde{\boldsymbol{c}}) + \sup_{\boldsymbol{w}} \big(\boldsymbol{a}_k^\top \boldsymbol{w} - \lambda \|\boldsymbol{w}\| \big) \Big) = \begin{cases} f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}}), & \lambda \geq L(\boldsymbol{x}_i, \boldsymbol{z}), \\ +\infty, & \text{otherwise}, \end{cases}$$

because writing $\boldsymbol{w} = r\boldsymbol{u}$ with $r \geq 0$, $\|\boldsymbol{u}\| = 1$ gives $\sup_{\boldsymbol{w}} (\boldsymbol{a}_k^\top \boldsymbol{w} - \lambda \|\boldsymbol{w}\|) = \sup_{r \geq 0} r\big(\|\boldsymbol{a}_k\|_* - \lambda\big)$, which is 0 if $\lambda \geq \|\boldsymbol{a}_k\|_*$ and $+\infty$ otherwise. The dual therefore reduces to $\inf_{\lambda \geq L(\boldsymbol{x}_i, \boldsymbol{z})} \{\lambda \rho_\theta + \mathbb{E}_{P_{\theta,i}}[f]\}$, which is attained at $\lambda = L(\boldsymbol{x}_i, \boldsymbol{z})$ and equals the right-hand side of (49).

Finally, each $\ell_k(\cdot, \boldsymbol{c})$ is convex because $\boldsymbol{a}_k(\boldsymbol{x}_i, \cdot)$ is affine and $b_k(\boldsymbol{x}_i, \cdot)$ is convex, and each $\boldsymbol{z} \mapsto \|\boldsymbol{a}_k(\boldsymbol{x}_i, \boldsymbol{z})\|_*$ is a norm composed with an affine map; maxima and expectations of convex functions are convex. □

I.2 Wasserstein-2 DRO

Proposition I.2 (W_2^2 reformulation)**.** *Under Assumption 2.1 and $P_{\theta,i} \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c})$. For every $\rho_\theta > 0$ and every $\boldsymbol{z} \in \mathbb{R}^{n_z}$,*

$$\sup_{Q\,:\,W_2^2(Q, P_{\theta,i}) \leq \rho_\theta} \mathbb{E}_{\tilde{\boldsymbol{c}} \sim Q}\big[f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}})\big] \;=\; \inf_{\lambda > 0} \left\{ \lambda \rho_\theta + \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\left[\max_{k \in [N_K]} \left\{ \ell_k(\boldsymbol{z}, \tilde{\boldsymbol{c}}) + \frac{\|\boldsymbol{a}_k(\boldsymbol{x}_i, \boldsymbol{z})\|_*^2}{4\lambda} \right\} \right] \right\}. \tag{51}$$

Consequently the lower level of (1) with $p = 2$ is the stochastic optimization problem

$$\big(\boldsymbol{z}^*_\theta(\boldsymbol{x}_i), \lambda^*\big) \in \operatorname*{argmin}_{\boldsymbol{z}\,:\,\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{z}) \leq \boldsymbol{0},\ \lambda > 0} \Phi_2(\boldsymbol{z}, \lambda, \theta) := \lambda \rho_\theta + \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\big[F(\boldsymbol{z}, \lambda, \tilde{\boldsymbol{c}})\big], \tag{52}$$

with $F(\boldsymbol{z}, \lambda, \boldsymbol{c}) := \max_{k \in [N_K]} \big\{ \ell_k(\boldsymbol{z}, \boldsymbol{c}) + q_k(\boldsymbol{z})/(4\lambda) \big\}$ and $q_k(\boldsymbol{z}) := \|\boldsymbol{a}_k(\boldsymbol{x}_i, \boldsymbol{z})\|_*^2$, whose objective is jointly convex in $(\boldsymbol{z}, \lambda)$ on $\mathbb{R}^{n_z} \times (0, \infty)$.

Proof. As in the proof of Proposition I.1, now with the cost $c(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|^2$, strong duality gives

$$\sup_{Q\,:\,W_2^2(Q,P_{\theta,i})\le\rho_\theta} \mathbb{E}_Q[f] = \inf_{\lambda\ge 0}\left\{\lambda\rho_\theta + \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\Big[\sup_{\boldsymbol{u}}\big(f(\boldsymbol{x}_i,\boldsymbol{z},\boldsymbol{u}) - \lambda\|\boldsymbol{u}-\tilde{\boldsymbol{c}}\|^2\big)\Big]\right\}.$$

Exchanging the two suprema and substituting $\boldsymbol{u} = \tilde{\boldsymbol{c}} + \boldsymbol{w}$,

$$\sup_{\boldsymbol{u}}\Big(\max_k \ell_k(\boldsymbol{z},\boldsymbol{u}) - \lambda\|\boldsymbol{u}-\tilde{\boldsymbol{c}}\|^2\Big) = \max_k\Big(\ell_k(\boldsymbol{z},\tilde{\boldsymbol{c}}) + \frac{\|\boldsymbol{a}_k(\boldsymbol{x}_i,\boldsymbol{z})\|_*^2}{4\lambda}\Big), \qquad \lambda > 0,$$

because $\sup_{\boldsymbol{w}}(\boldsymbol{a}_k^\top\boldsymbol{w} - \lambda\|\boldsymbol{w}\|^2) = \sup_{r\ge 0}(r\|\boldsymbol{a}_k\|_* - \lambda r^2) = \|\boldsymbol{a}_k\|_*^2/(4\lambda)$, the Fenchel conjugate of $\lambda\|\cdot\|^2$ at $\boldsymbol{a}_k$; for $\lambda = 0$ the inner supremum is $+\infty$ unless $\boldsymbol{a}_k = \boldsymbol{0}$ for every k, so the infimum may be restricted to $\lambda > 0$. This is (51). Joint convexity holds because $(\boldsymbol{a},\lambda) \mapsto \|\boldsymbol{a}\|_*^2/(4\lambda)$ is the perspective of the convex function $\boldsymbol{a} \mapsto \|\boldsymbol{a}\|_*^2/4$, composed here with the affine map $\boldsymbol{z} \mapsto \boldsymbol{a}_k(\boldsymbol{x}_i,\boldsymbol{z})$, and adding $\ell_k(\cdot,\boldsymbol{c})$, maximizing over k, taking the expectation, and adding $\lambda\rho_\theta$ preserve it. □

J Computation of the Gradient in DRDFL

The reformulations in Appendix I turn the lower level of (1) into a smooth convex program where the dependence on θ is in two ways, namely through the predicted law $P_{\theta,i}$ that defines the expectation and through the calibrated radius ρ_θ. We keep the abbreviations of Appendix I and write $\boldsymbol{z}^* := \boldsymbol{z}^*_\theta(\boldsymbol{x}_i)$.

J.1 Computation of the Gradient of the Optimal W_1 DRO Decision with Respect to θ

Step 1: optimality conditions. By Proposition I.1 the decision solves

$$\min_{\boldsymbol{z}\,:\,\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z})\le \mathbf{0}} \Phi_1(\boldsymbol{z},\theta), \qquad \Phi_1(\boldsymbol{z},\theta) = \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\big[f(\boldsymbol{x}_i,\boldsymbol{z},\tilde{\boldsymbol{c}})\big] + \rho_\theta L(\boldsymbol{x}_i,\boldsymbol{z}), \tag{53}$$

a convex program that satisfies Slater's condition by Assumption 2.2. Its KKT conditions are therefore both necessary and sufficient: $\boldsymbol{z}^*$ is optimal if and only if there is $\boldsymbol{\nu}^* \ge \mathbf{0}$ with

$$\nabla_{\boldsymbol{z}}\Phi_1(\boldsymbol{z}^*,\theta) + \boldsymbol{J}_g(\boldsymbol{x}_i,\boldsymbol{z}^*)^\top \boldsymbol{\nu}^* = \mathbf{0}, \qquad \boldsymbol{\nu}^* \odot \boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z}^*) = \mathbf{0}, \qquad \boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z}^*) \le \mathbf{0}, \tag{54}$$

where $\boldsymbol{J}_g(\boldsymbol{x}_i,\boldsymbol{z}) := \nabla_{\boldsymbol{z}}\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z}) \in \mathbb{R}^{n_g\times n_z}$. The gradient of the objective is

$$\nabla_{\boldsymbol{z}}\Phi_1(\boldsymbol{z},\theta) = \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\big[\nabla_{\boldsymbol{z}} f(\boldsymbol{x}_i,\boldsymbol{z},\tilde{\boldsymbol{c}})\big] + \rho_\theta\, \nabla_{\boldsymbol{z}} L(\boldsymbol{x}_i,\boldsymbol{z}).$$

Differentiation and expectation commute here because for $P_{\theta,i}$-almost every $\tilde{\boldsymbol{c}}$ a single piece of (48) attains the maximum, so that $\nabla_{\boldsymbol{z}} f = \nabla_{\boldsymbol{z}}\ell_{k(\tilde{\boldsymbol{c}})}$: the two pieces agree only on the set $\{\boldsymbol{c} : (\boldsymbol{a}_2-\boldsymbol{a}_1)^\top \boldsymbol{c} + b_2 - b_1 = 0\}$, a hyperplane, which is null because $\boldsymbol{a}_1 \ne \boldsymbol{a}_2$ and $P_{\theta,i}$ has a density.

Step 2: differentiate the optimality conditions. Let $\mathcal{A} := \{j : g_j(\boldsymbol{x}_i,\boldsymbol{z}^*) = \mathbf{0}\}$ be the active set and let $\boldsymbol{g}_{\mathcal{A}}, \boldsymbol{\nu}_{\mathcal{A}}, \boldsymbol{J}_{g,\mathcal{A}} := \nabla_{\boldsymbol{z}}\boldsymbol{g}_{\mathcal{A}}$. Since the multipliers of the active constraints are strictly positive and everything in (54) is continuous, a small change of θ leaves $\mathcal{A}$ unchanged. Locally, the inequalities and the complementarity conditions can therefore be dropped, and (54) becomes the square system of equations

$$\boldsymbol{K}(\boldsymbol{y},\theta) := \begin{bmatrix} \nabla_{\boldsymbol{z}}\Phi_1(\boldsymbol{z},\theta) + \boldsymbol{J}_{g,\mathcal{A}}(\boldsymbol{x}_i,\boldsymbol{z})^\top \boldsymbol{\nu}_{\mathcal{A}} \\ \boldsymbol{g}_{\mathcal{A}}(\boldsymbol{x}_i,\boldsymbol{z}) \end{bmatrix} = \mathbf{0}, \qquad \boldsymbol{y} := (\boldsymbol{z},\boldsymbol{\nu}_{\mathcal{A}}) \in \mathbb{R}^{n_z+|\mathcal{A}|},$$

which holds identically in θ along the solution path $\boldsymbol{y}^*(\theta)$. Differentiating the identity $\boldsymbol{K}(\boldsymbol{y}^*(\theta),\theta) = \mathbf{0}$ with the chain rule gives $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{y}}\frac{\mathrm{d}\boldsymbol{y}^*}{\mathrm{d}\theta} + \frac{\partial \boldsymbol{K}}{\partial \theta} = \mathbf{0}$, and keeping the first n_z rows of $\mathrm{d}\boldsymbol{y}^*/\mathrm{d}\theta$ yields

$$\frac{\mathrm{d}\boldsymbol{z}^*_\theta}{\mathrm{d}\theta} = -\begin{bmatrix} \boldsymbol{I}_{n_z} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_1 & \boldsymbol{J}_{g,\mathcal{A}}^\top \\ \boldsymbol{J}_{g,\mathcal{A}} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{T}_1 + \boldsymbol{T}_2 \\ \mathbf{0} \end{bmatrix}. \tag{55}$$

All quantities are evaluated at $(\boldsymbol{z}^*, \boldsymbol{\nu}^*_{\mathcal{A}}, \theta)$. The matrix is the usual KKT matrix of the active problem; it is nonsingular as soon as $\boldsymbol{H}_1$ is positive definite on $\ker \boldsymbol{J}_{g,\mathcal{A}}$, since $\boldsymbol{H}_1\boldsymbol{u} + \boldsymbol{J}_{g,\mathcal{A}}^\top \boldsymbol{v} = \mathbf{0}$ together with $\boldsymbol{J}_{g,\mathcal{A}}\boldsymbol{u} = \mathbf{0}$ then forces $\boldsymbol{u} = \mathbf{0}$ and, by the linear independence of the active gradients, $\boldsymbol{v} = \mathbf{0}$.

Step 3: blocks. The block $\boldsymbol{H}_1 := \partial(\nabla_{\boldsymbol{z}}\Phi_1 + \boldsymbol{J}_{g,\mathcal{A}}^\top \boldsymbol{\nu}_{\mathcal{A}})/\partial \boldsymbol{z}$ is the Hessian of the Lagrangian of (53), and $\partial \boldsymbol{K}/\partial\theta$ collects the two channels through which θ acts:

$$\boldsymbol{H}_1 = \underbrace{\nabla^2_{\boldsymbol{z}\boldsymbol{z}}\mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\big[f(\boldsymbol{x}_i,\boldsymbol{z}^*,\tilde{\boldsymbol{c}})\big]}_{\text{Step 4}} + \rho_\theta \underbrace{\nabla^2_{\boldsymbol{z}\boldsymbol{z}} L(\boldsymbol{x}_i,\boldsymbol{z}^*)}_{\text{Step 5}} + \sum_{j\in\mathcal{A}} \nu^*_j\, \nabla^2_{\boldsymbol{z}\boldsymbol{z}} g_j(\boldsymbol{x}_i,\boldsymbol{z}^*),$$

$$\boldsymbol{T}_1 = \nabla_\theta \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\theta,i}}\big[\nabla_{\boldsymbol{z}} f(\boldsymbol{x}_i,\boldsymbol{z}^*,\tilde{\boldsymbol{c}})\big], \qquad \boldsymbol{T}_2 = \nabla_{\boldsymbol{z}} L(\boldsymbol{x}_i,\boldsymbol{z}^*)\big(\nabla_\theta \rho_\theta\big)^\top. \tag{56}$$

Only $\nabla_{\boldsymbol{z}}\Phi_1$ carries θ, so the lower block of $\partial \boldsymbol{K}/\partial\theta$ vanishes. In $\boldsymbol{T}_1$ the parameter moves the predicted law and hence the distribution of the particles at which the loss is evaluated while in $\boldsymbol{T}_2$ it moves the calibrated radius, with $\nabla_\theta\rho_\theta$ given by (13). The constraint curvature $\sum_j \nu^*_j \nabla^2_{\boldsymbol{z}\boldsymbol{z}} g_j$ vanishes for the affine simplex constraints of Section 5.

Step 4: Hessian of the expectation. This is the block that is not a routine derivative, because f is only piecewise smooth in $\boldsymbol{z}$. Write the loss as a smooth piece plus a kink,

$$f(\boldsymbol{x}_i, \boldsymbol{z}, \boldsymbol{c}) = \ell_1(\boldsymbol{z}, \boldsymbol{c}) + \big(\Delta(\boldsymbol{z}, \boldsymbol{c})\big)_+, \qquad \Delta(\boldsymbol{z}, \boldsymbol{c}) := \ell_2(\boldsymbol{z}, \boldsymbol{c}) - \ell_1(\boldsymbol{z}, \boldsymbol{c}) = \boldsymbol{\alpha}(\boldsymbol{z})^\top \boldsymbol{c} + \beta(\boldsymbol{z}), \tag{57}$$

with $\boldsymbol{\alpha} := \boldsymbol{a}_2 - \boldsymbol{a}_1$ affine in $\boldsymbol{z}$ and $\beta := b_2 - b_1$. Taking one derivative gives $\nabla_{\boldsymbol{z}} \mathbb{E}[f] = \mathbb{E}[\nabla_{\boldsymbol{z}} \ell_1] + \mathbb{E}[\mathbf{1}\{\Delta > 0\} \nabla_{\boldsymbol{z}} \Delta]$, as in Step 1. Taking a second derivative in a direction $\boldsymbol{d}$ splits the difference quotient of this expression into two effects, namely the particles that keep their active piece contribute $\mathbb{E}[\mathbf{1}\{\Delta > 0\} \nabla^2_{\boldsymbol{z}\boldsymbol{z}} \Delta\, \boldsymbol{d}]$, while the particles that switch piece contribute a term of their own, since they lie in the thin shell $\{\Delta \in (-t\, \boldsymbol{d}^\top \nabla_{\boldsymbol{z}} \Delta, 0]\}$ whose probability is $t\, p_\Delta(0)\, \mathbb{E}[\boldsymbol{d}^\top \nabla_{\boldsymbol{z}} \Delta \mid \Delta = 0] + o(t)$, with p_Δ the density of the scalar $\Delta(\boldsymbol{z}, \tilde{\boldsymbol{c}})$. Collecting the two effects,

$$\nabla^2_{\boldsymbol{z}\boldsymbol{z}} \mathbb{E}\big[f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}})\big] = \underbrace{\mathbb{E}\big[\nabla^2_{\boldsymbol{z}\boldsymbol{z}} f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}})\big]}_{\text{curvature of the pieces}} + \underbrace{p_\Delta(0)\, \mathbb{E}\big[\nabla_{\boldsymbol{z}} \Delta\, \nabla_{\boldsymbol{z}} \Delta^\top \;\big|\; \Delta = 0\big]}_{\text{curvature created by switching}}, \tag{58}$$

where $\nabla_{\boldsymbol{z}} \Delta = \boldsymbol{A}_\Delta^\top \boldsymbol{c} + \nabla_{\boldsymbol{z}} \beta(\boldsymbol{z})$ with $\boldsymbol{A}_\Delta := \nabla_{\boldsymbol{z}} \boldsymbol{\alpha}(\boldsymbol{z})$ constant, and $\nabla^2_{\boldsymbol{z}\boldsymbol{z}} f = \nabla^2_{\boldsymbol{z}\boldsymbol{z}} b_{k(\tilde{\boldsymbol{c}})}$ because $\boldsymbol{a}_k(\boldsymbol{x}_i, \cdot)$ is affine. The second term is a conditional second moment, hence positive semidefinite. So averaging over $\tilde{\boldsymbol{c}}$ smooths the kink of the maximum and the resulting rounding is genuine curvature.

Step 5: radius term. With the Euclidean ground metric of Assumption 2.1 and a unique maximizing index k^*, the term $L(\boldsymbol{x}_i, \boldsymbol{z}) = \|\boldsymbol{a}_{k^*}(\boldsymbol{x}_i, \boldsymbol{z})\|_2$ is smooth near $\boldsymbol{z}^*$. Writing $\boldsymbol{a}_{k^*} := \boldsymbol{a}_{k^*}(\boldsymbol{x}_i, \boldsymbol{z}^*)$ and $\boldsymbol{A}_{k^*} := \nabla_{\boldsymbol{z}} \boldsymbol{a}_{k^*}(\boldsymbol{x}_i, \boldsymbol{z}) \in \mathbb{R}^{n_c \times n_z}$, which is constant because $\boldsymbol{a}_{k^*}(\boldsymbol{x}_i, \cdot)$ is affine,

$$\nabla_{\boldsymbol{z}} L = \boldsymbol{A}_{k^*}^\top \frac{\boldsymbol{a}_{k^*}}{\|\boldsymbol{a}_{k^*}\|_2}, \qquad \nabla^2_{\boldsymbol{z}\boldsymbol{z}} L = \frac{1}{\|\boldsymbol{a}_{k^*}\|_2} \boldsymbol{A}_{k^*}^\top \Big(\boldsymbol{I}_{n_c} - \frac{\boldsymbol{a}_{k^*} \boldsymbol{a}_{k^*}^\top}{\|\boldsymbol{a}_{k^*}\|_2^2}\Big) \boldsymbol{A}_{k^*} \;\succeq\; \mathbf{0}. \tag{59}$$

The projector in (59) shows that the radius term is flat along the directions that only rescale $\boldsymbol{a}_{k^*}$.

Step 6: computing the gradient in practice. Eliminating the multiplier block in (55) gives the equivalent projected form

$$\frac{\mathrm{d}\boldsymbol{z}_\theta^*}{\mathrm{d}\theta} = -\Big(\boldsymbol{H}_1^{-1} - \boldsymbol{H}_1^{-1} \boldsymbol{J}_{g,\mathcal{A}}^\top \big(\boldsymbol{J}_{g,\mathcal{A}} \boldsymbol{H}_1^{-1} \boldsymbol{J}_{g,\mathcal{A}}^\top\big)^{-1} \boldsymbol{J}_{g,\mathcal{A}} \boldsymbol{H}_1^{-1}\Big)\big(\boldsymbol{T}_1 + \boldsymbol{T}_2\big),$$

valid when $\boldsymbol{H}_1 \succ \mathbf{0}$, from which $\boldsymbol{J}_{g,\mathcal{A}}\, \mathrm{d}\boldsymbol{z}_\theta^*/\mathrm{d}\theta = \mathbf{0}$: the active constraints stay satisfied to first order, a useful check on an implementation. The term $\boldsymbol{T}_1$ is a derivative of an expectation under $P_{\theta,i}$ and is computed pathwise with the latent variables frozen exactly as in (62); with the particles $\tilde{\boldsymbol{c}}_i^m(\theta)$ and their Jacobians $\boldsymbol{J}^m$,

$$\boldsymbol{T}_1 = \frac{1}{M} \sum_{m=1}^{M} \nabla_{\boldsymbol{c}} \nabla_{\boldsymbol{z}} f\big(\boldsymbol{x}_i, \boldsymbol{z}^*, \tilde{\boldsymbol{c}}_i^m(\theta)\big)\, \boldsymbol{J}^m.$$

Everything above holds with $P_{\theta,i}$ replaced by the particle measure $\widehat{P}_{\theta,i}$.

J.2 Computation of the Gradient of the Optimal W_2 DRO Decision with Respect to θ

The W_2^2 case follows the same six steps once the scalar dual variable λ of Proposition I.2 is treated as part of the decision, so we only record what changes. Recall $q_k(\boldsymbol{z}) = \|\boldsymbol{a}_k(\boldsymbol{x}_i, \boldsymbol{z})\|_2^2$ and $F(\boldsymbol{z}, \lambda, \boldsymbol{c}) = \max_k \{\ell_k(\boldsymbol{z}, \boldsymbol{c}) + q_k(\boldsymbol{z})/(4\lambda)\}$, and write $\boldsymbol{w} := (\boldsymbol{z}, \lambda)$.

Optimality conditions. The decision solves $\min\{\Phi_2(\boldsymbol{z}, \lambda, \theta) : \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{z}) \le \mathbf{0},\ \lambda > 0\}$ with $\Phi_2 = \lambda \rho_\theta + \mathbb{E}[F]$, which is jointly convex in $(\boldsymbol{z}, \lambda)$. The optimal λ^* is strictly positive, so its constraint is inactive and never enters the KKT system. Since a single index attains the maximum in F for almost every $\tilde{\boldsymbol{c}}$, the first-order conditions in $\boldsymbol{z}$ and in λ are

$$\nabla_{\boldsymbol{z}} \Phi_2 = \mathbb{E}\Big[\nabla_{\boldsymbol{z}} \ell_{k(\tilde{\boldsymbol{c}})}(\boldsymbol{z}, \tilde{\boldsymbol{c}}) + \frac{\nabla_{\boldsymbol{z}} q_{k(\tilde{\boldsymbol{c}})}(\boldsymbol{z})}{4\lambda}\Big] + \boldsymbol{J}_{g,\mathcal{A}}^\top \boldsymbol{\nu}_{\mathcal{A}} = \mathbf{0}, \qquad \partial_\lambda \Phi_2 = \rho_\theta - \frac{\mathbb{E}\big[q_{k(\tilde{\boldsymbol{c}})}(\boldsymbol{z})\big]}{4\lambda^2} = 0.$$

The second equation is the scalar condition that sizes the quadratic penalty.

Differentiation. With $\boldsymbol{y} := (\boldsymbol{z}, \lambda, \boldsymbol{\nu}_{\mathcal{A}})$ and $\boldsymbol{K} := [\nabla_{\boldsymbol{z}} \Phi_2 + \boldsymbol{J}_{g,\mathcal{A}}^\top \boldsymbol{\nu}_{\mathcal{A}};\ \partial_\lambda \Phi_2;\ \boldsymbol{g}_{\mathcal{A}}]$, the same implicit differentiation gives

$$\frac{\mathrm{d}\boldsymbol{z}_\theta^*}{\mathrm{d}\theta} = -\begin{bmatrix} \boldsymbol{I}_{n_z} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_2 & \boldsymbol{h}_{\boldsymbol{z}\lambda} & \boldsymbol{J}_{g,\mathcal{A}}^\top \\ \boldsymbol{h}_{\boldsymbol{z}\lambda}^\top & h_{\lambda\lambda} & \boldsymbol{0} \\ \boldsymbol{J}_{g,\mathcal{A}} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{T}_1^{(2)} \\ \boldsymbol{T}_2^{(2)} \\ \boldsymbol{0} \end{bmatrix}, \tag{60}$$

with everything evaluated at $(\boldsymbol{z}^*, \lambda^*, \boldsymbol{\nu}_{\mathcal{A}}^*, \theta)$ and

$$\boldsymbol{H}_2 = \nabla_{\boldsymbol{z}\boldsymbol{z}}^2 \mathbb{E}\big[F\big] + \sum_{j \in \mathcal{A}} \nu_j^* \nabla_{\boldsymbol{z}\boldsymbol{z}}^2 g_j(\boldsymbol{x}_i, \boldsymbol{z}^*), \qquad \boldsymbol{h}_{\boldsymbol{z}\lambda} = \nabla_{\boldsymbol{z}\lambda}^2 \mathbb{E}\big[F\big], \qquad h_{\lambda\lambda} = \partial_{\lambda\lambda}^2 \mathbb{E}\big[F\big],$$

$$\boldsymbol{T}_1^{(2)} = \nabla_\theta \nabla_{\boldsymbol{z}} \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\big[F\big], \qquad \boldsymbol{T}_2^{(2)} = \nabla_\theta \rho_\theta - \frac{1}{4(\lambda^*)^2} \nabla_\theta \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\theta,i}}\big[q_{k(\tilde{\boldsymbol{c}})}(\boldsymbol{z}^*)\big]. \tag{61}$$

The second derivatives. The rule (58) of Step 4 applies unchanged with $\boldsymbol{w} = (\boldsymbol{z}, \lambda)$ in place of $\boldsymbol{z}$ and with the gap Δ of (57) replaced by $\Delta + \big(q_2(\boldsymbol{z}) - q_1(\boldsymbol{z})\big)/(4\lambda)$: the second derivative of $\mathbb{E}[F]$ is the expectation of the second derivative of the active piece plus the switching term. The piecewise part uses

$$\nabla_{\boldsymbol{z}\boldsymbol{z}}^2 h_k = \nabla_{\boldsymbol{z}\boldsymbol{z}}^2 b_k + \frac{\nabla_{\boldsymbol{z}\boldsymbol{z}}^2 q_k}{4\lambda}, \qquad \nabla_{\boldsymbol{z}\lambda}^2 h_k = -\frac{\nabla_{\boldsymbol{z}} q_k}{4\lambda^2}, \qquad \partial_{\lambda\lambda}^2 h_k = \frac{q_k}{2\lambda^3},$$

where $h_k(\boldsymbol{w}, \boldsymbol{c}) := \ell_k(\boldsymbol{z}, \boldsymbol{c}) + q_k(\boldsymbol{z})/(4\lambda)$ and $\nabla_{\boldsymbol{z}\boldsymbol{z}}^2 q_k = 2\boldsymbol{A}_k^\top \boldsymbol{A}_k$ for the Euclidean metric.

Where the radius enters. Comparing (56) and (61), $\nabla_\theta \rho_\theta$ enters the W_1 system through the stationarity condition in $\boldsymbol{z}$, whereas in the W_2^2 system it enters only the scalar λ-row and reaches $\boldsymbol{z}^*$ through the coupling block $\boldsymbol{h}_{\boldsymbol{z}\lambda}$: a change of the calibrated radius first resizes the quadratic penalty and only then moves the decision. The decomposition of the gradient into a predicted-distribution part and a calibrated-radius part described below (12) is unaffected, with $\boldsymbol{T}_1^{(2)}$ and $\boldsymbol{T}_2^{(2)}$ in the roles of $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$. Finally, if $N_{\mathrm{K}} = 1$, the variable λ can be eliminated beforehand and (60) reduces to (55) with $L(\boldsymbol{x}_i, \boldsymbol{z}) = \|\boldsymbol{a}_1(\boldsymbol{x}_i, \boldsymbol{z})\|_2$, ρ_θ replaced by $\sqrt{\rho_\theta}$ and $\nabla_\theta \rho_\theta$ by $\nabla_\theta \rho_\theta / (2\sqrt{\rho_\theta})$.

J.3 Computation of the Gradient of the Nonconformity Score $\frac{\mathrm{d}s\left(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\right)}{\mathrm{d}\theta}$

The quantile derivative (13) requires the derivative of the nonconformity score with respect to the distribution predictor's parameter. We derive it here for SD_ϵ^1 and the SD_ϵ^2 version is recorded in Corollary J.1.

Setup. Fix a calibration point, whose feature is $\boldsymbol{x}_i$, recall $\widehat{P}_{\star,i} = \frac{1}{S}\sum_{j=1}^S \delta_{\boldsymbol{c}_i^j}$ is the empirical measure of the target distribution. Sampling from a CNF is performed by drawing a latent variable and inverting the flow, $\tilde{\boldsymbol{c}} = f_\theta^{-1}(\boldsymbol{h}; \boldsymbol{x}_i)$ with $\boldsymbol{h} \sim p_{\boldsymbol{h}}$. We draw the M latent variables once and hold them fixed during differentiation,

$$\boldsymbol{h}_i^1, \ldots, \boldsymbol{h}_i^M \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{h}}, \qquad \tilde{\boldsymbol{c}}_i^m(\theta) := f_\theta^{-1}(\boldsymbol{h}_i^m; \boldsymbol{x}_i), \qquad \widehat{P}_{\theta,i} = \frac{1}{M}\sum_{m=1}^M \delta_{\tilde{\boldsymbol{c}}_i^m(\theta)}, \tag{62}$$

Write $\boldsymbol{J}^m := \frac{\partial f_\theta^{-1}(\boldsymbol{h}_i^m; \boldsymbol{x}_i)}{\partial \theta} \in \mathbb{R}^{n_c \times n_\theta}$, where $n_\theta$is the dimension of θ. Let $\boldsymbol{a} = \boldsymbol{1}_S/S$ and $\boldsymbol{b} = \boldsymbol{1}_M/M$ be the two weight vectors.

Two observations reduce the whole computation to differentiating a cost matrix. First, the latent variables in (62) are frozen, so θ moves the locations of the samples $\widehat{P}_{\theta,i}$ is composed of, but not their weights. This means $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent of θ. Second, the term $\mathrm{OT}_\epsilon^1(\widehat{P}_{\star,i}, \widehat{P}_{\star,i})$ in (3) involves no samples and is constant in θ. Therefore, the entire dependence sits in the two cost matrices

$$\mathsf{C}_{jm}^{SM}(\theta) := \big\|\boldsymbol{c}_i^j - \tilde{\boldsymbol{c}}_i^m(\theta)\big\|_2, \qquad \mathsf{C}_{m\ell}^{MM}(\theta) := \big\|\tilde{\boldsymbol{c}}_i^m(\theta) - \tilde{\boldsymbol{c}}_i^\ell(\theta)\big\|_2.$$

Lemma J.1 (Envelope formula)**.** *Fix $\epsilon > 0$ and weight vectors $\boldsymbol{a} \in \mathbb{R}^S$, $\boldsymbol{b} \in \mathbb{R}^M$ and for $\mathsf{C} \in \mathbb{R}^{S \times M}$ define*

$$\Phi_\epsilon(\mathsf{C}) := \min_{\pi \in \Pi(\boldsymbol{a},\boldsymbol{b})} \Big\{ \langle \pi, \mathsf{C} \rangle + \epsilon \, D_{\mathrm{KL}}\big(\pi \,\|\, \boldsymbol{a}\boldsymbol{b}^\top\big) \Big\}, \qquad \Pi(\boldsymbol{a},\boldsymbol{b}) = \{\pi \geq \mathbf{0} : \pi \mathbf{1}_M = \boldsymbol{a},\ \pi^\top \mathbf{1}_S = \boldsymbol{b}\}.$$

Then the minimizer $\pi^\star(\mathsf{C})$ is unique and Φ_ϵ is differentiable on $\mathbb{R}^{S \times M}$ with

$$\frac{\partial \Phi_\epsilon}{\partial \mathsf{C}_{jm}}(\mathsf{C}) = \pi^\star_{jm}(\mathsf{C}). \tag{63}$$

Proof. The set $\Pi(\boldsymbol{a},\boldsymbol{b})$ is nonempty, convex, and compact and, on it $\pi \mapsto \langle \pi, \mathsf{C} \rangle + \epsilon D_{\mathrm{KL}}(\pi \| \boldsymbol{a}\boldsymbol{b}^\top)$ is continuous and strictly convex. Hence $\pi^\star(\mathsf{C})$ exists and is unique. Write $\pi_0 = \pi^\star(\mathsf{C})$ and $\pi_1 = \pi^\star(\mathsf{C}')$. Using π_0 as a feasible point for C' and optimality of π_1 for C' together with feasibility of π_1 for C,

$$\langle \pi_1, \mathsf{C}' - \mathsf{C} \rangle \ \leq\ \Phi_\epsilon(\mathsf{C}') - \Phi_\epsilon(\mathsf{C}) \ \leq\ \langle \pi_0, \mathsf{C}' - \mathsf{C} \rangle,$$

so that $\big|\Phi_\epsilon(\mathsf{C}') - \Phi_\epsilon(\mathsf{C}) - \langle \pi_0, \mathsf{C}' - \mathsf{C} \rangle\big| \ \leq\ \|\pi_0 - \pi_1\|_F \, \|\mathsf{C}' - \mathsf{C}\|_F$. It remains to check that $\pi_1 \to \pi_0$ as $\mathsf{C}' \to \mathsf{C}$: any accumulation point $\bar{\pi}$ of the minimizers along a sequence $\mathsf{C}' \to \mathsf{C}$ exists by compactness and by joint continuity of the objective, is a minimizer for C. Hence it equals π_0 by uniqueness. The remainder is therefore $o(\|\mathsf{C}' - \mathsf{C}\|_F)$, which is (63). □

Applying Lemma J.1 to the two θ-dependent terms of (3) gives the gradient of SD^1_ϵ.

Proposition J.1 (Derivative of SD^1_ϵ)**.** *Let θ make each $\theta \mapsto \tilde{\boldsymbol{c}}^m_i(\theta)$ differentiable and the $S + M$ points $\boldsymbol{c}^{1:S}_i, \tilde{\boldsymbol{c}}^{1:M}_i(\theta)$ are pairwise distinct. Let $\pi^{SM,\star} \in \Pi(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ and $\pi^{MM,\star} \in \Pi(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i})$ be the optimal plans of $\mathrm{OT}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ and $\mathrm{OT}^1_\epsilon(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i})$. Then $s(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}) = \mathrm{SD}^1_\epsilon(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})$ is differentiable at θ, with*

$$\frac{\mathrm{d}\, s\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)}{\mathrm{d}\theta} = \sum_{m=1}^{M} \big(\boldsymbol{J}^m\big)^\top \boldsymbol{g}^m, \qquad \boldsymbol{g}^m := \sum_{j=1}^{S} \pi^{SM,\star}_{jm} \frac{\tilde{\boldsymbol{c}}^m_i - \boldsymbol{c}^j_i}{\|\tilde{\boldsymbol{c}}^m_i - \boldsymbol{c}^j_i\|_2} - \sum_{\ell \neq m} \pi^{MM,\star}_{m\ell} \frac{\tilde{\boldsymbol{c}}^m_i - \tilde{\boldsymbol{c}}^\ell_i}{\|\tilde{\boldsymbol{c}}^m_i - \tilde{\boldsymbol{c}}^\ell_i\|_2}. \tag{64}$$

If in addition $p_{\boldsymbol{h}}$ is nonatomic, then for every fixed θ the hypotheses hold with probability one over the draw of $\boldsymbol{h}^{1:M}_i$.

Proof. The Euclidean norm is differentiable away from the origin. Under the stated distinctness the chain rule gives,

$$\frac{\partial \mathsf{C}^{SM}_{jm}}{\partial \theta} = \big(\boldsymbol{J}^m\big)^\top \boldsymbol{u}^{SM}_{jm}, \qquad \frac{\partial \mathsf{C}^{MM}_{m\ell}}{\partial \theta} = \big(\boldsymbol{J}^m - \boldsymbol{J}^\ell\big)^\top \boldsymbol{u}^{MM}_{m\ell} \quad (m \neq \ell),$$

where $\boldsymbol{u}^{SM}_{jm}$ and $\boldsymbol{u}^{MM}_{m\ell}$ are the two unit vectors appearing in (64). The diagonal entries $\mathsf{C}^{MM}_{mm} \equiv 0$ are constant in θ. No convention at coincident points is needed.

Composing with (63) and the fact that $\mathsf{C}(\cdot)$ is differentiable with respect to θ,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \mathrm{OT}^1_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big) = \sum_{j=1}^{S} \sum_{m=1}^{M} \pi^{SM,\star}_{jm} \big(\boldsymbol{J}^m\big)^\top \boldsymbol{u}^{SM}_{jm}, \tag{65}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \mathrm{OT}^1_\epsilon\big(\widehat{P}_{\theta,i}, \widehat{P}_{\theta,i}\big) = \sum_{m \neq \ell} \pi^{MM,\star}_{m\ell} \big(\boldsymbol{J}^m - \boldsymbol{J}^\ell\big)^\top \boldsymbol{u}^{MM}_{m\ell}. \tag{66}$$

C^{MM} is symmetric. $\Pi(\boldsymbol{b},\boldsymbol{b})$ is closed under transposition and $D_{\mathrm{KL}}(\pi^\top \| \boldsymbol{b}\boldsymbol{b}^\top) = D_{\mathrm{KL}}(\pi \| \boldsymbol{b}\boldsymbol{b}^\top)$. So $\pi \mapsto \pi^\top$ maps minimizers to minimizers and the uniqueness forces $\pi^{MM,\star} = (\pi^{MM,\star})^\top$. Since $\boldsymbol{u}^{MM}_{\ell m} = -\boldsymbol{u}^{MM}_{m\ell}$, the two halves of (66) coincide and the sum is $2\sum_{m \neq \ell} \pi^{MM,\star}_{m\ell} (\boldsymbol{J}^m)^\top \boldsymbol{u}^{MM}_{m\ell}$. Substituting (65) and (66) in (3), whose third term is constant in θ, the factor $\frac{1}{2}$ cancels the doubling and yields (64).

If $p_{\boldsymbol{h}}$ is nonatomic then so is $P_\theta(\cdot \mid \boldsymbol{x}_i)$, because $f^{-1}_\theta(\cdot\,; \boldsymbol{x}_i)$ is injective and the preimage of a point is at most a point. Hence $\mathbb{P}\{\tilde{\boldsymbol{c}}^m_i = \boldsymbol{c}^j_i\} = 0$ for all j, m and $\mathbb{P}\{\tilde{\boldsymbol{c}}^m_i = \tilde{\boldsymbol{c}}^\ell_i\} = 0$ for $m \neq \ell$ by independence. A finite union of null events is null. □

Since both terms in $\boldsymbol{g}^m$ are convex combinations of unit vectors with total mass $\sum_j \pi^{SM,\star}_{jm} = \sum_\ell \pi^{MM,\star}_{m\ell} = \frac{1}{M}$,

$$\left\|\boldsymbol{g}^m\right\|_2 \le \frac{2}{M}, \qquad \left\|\frac{\mathrm{d}\, s(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i})}{\mathrm{d}\theta}\right\|_2 \;\le\; 2 \max_{1\le m\le M} \left\|\boldsymbol{J}^m\right\|_2.$$

Therefore, the score supplies a bounded training signal on any support, with no dependence on ϵ or on the distance between $\widehat{P}_{\star,i}$ and $\widehat{P}_{\theta,i}$.

Corollary J.1 (Derivative of SD^2_ϵ)**.** *For SD^2_ϵ defined by (3) and (2), the nonconformity score is differentiable wherever each $\theta \mapsto \tilde{\boldsymbol{c}}^m_i(\theta)$ is, with no distinctness requirement, and*

$$\frac{\mathrm{d}\,\mathrm{SD}^2_\epsilon\big(\widehat{P}_{\star,i}, \widehat{P}_{\theta,i}\big)}{\mathrm{d}\theta} = \frac{2}{M}\sum_{m=1}^{M} \big(\boldsymbol{J}^m\big)^\top \big(\bar{\boldsymbol{c}}^m_M - \bar{\boldsymbol{c}}^m_S\big), \qquad \begin{aligned} \bar{\boldsymbol{c}}^m_M &:= M \sum_{\ell=1}^{M} \pi^{MM,\star}_{m\ell}\, \tilde{\boldsymbol{c}}^\ell_i, \\ \bar{\boldsymbol{c}}^m_S &:= M \sum_{j=1}^{S} \pi^{SM,\star}_{jm}\, \boldsymbol{c}^j_i. \end{aligned} \tag{67}$$

Proof. Repeat the proof with $\partial_\theta \|\boldsymbol{c}^j_i - \tilde{\boldsymbol{c}}^m_i\|_2^2 = 2(\boldsymbol{J}^m)^\top(\tilde{\boldsymbol{c}}^m_i - \boldsymbol{c}^j_i)$, which is smooth everywhere. In $\boldsymbol{g}^m$ the two copies of $\tilde{\boldsymbol{c}}^m_i$ are multiplied by the column sums $\frac{1}{M}$ of the two plans and cancel, leaving the difference of the two barycentric projections in (67). □

Remark J.1 (Implementation). Equations (64) and (67) are the form we implement, namely run the Sinkhorn iterations to convergence, detach $\pi^{SM,\star}$ and $\pi^{MM,\star}$ from the computational graph, and backpropagate the M vectors $\boldsymbol{g}^m$ through the flow. The Jacobians $\boldsymbol{J}^m$ are never formed. Only the products $(\boldsymbol{J}^m)^\top \boldsymbol{g}^m$ are needed, at the cost of one reverse pass. Differentiating through the Sinkhorn iterations returns the same limit but is exact only at convergence and stores every iterate.

Remark J.2 (Other distribution predictors). Only two properties of the CNF were used, namely the samples are generated by a map that is differentiable in θ, and the randomness entering that map is drawn once and held fixed. Therefore, the derivation applies verbatim to a CDM with a fixed noise sequence. For the GMDN (21) the component index is discrete, we keep the component assignments fixed within a gradient step and reparameterize only the Gaussian draws. If instead the K mixture weights $\pi_k(\boldsymbol{x}_i;\theta)$ are carried into $\widehat{P}_{\theta,i}$ as sample weights, then $\boldsymbol{b}$ depends on θ and (64) acquires an additional term involving the optimal dual potentials.

K Reformulation of Mean-Risk Optimization as Distributionally Robust Optimization

Fix $p \in \{1, 2\}$ and assume that $P_\theta(\cdot \mid \boldsymbol{x}) \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c})$. Define the Wasserstein ambiguity set

$$\mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta) := \left\{ Q \in \mathcal{P}_{1,2}(\mathbb{R}^{n_c}) : W_p^p\left(Q, P_\theta(\cdot \mid \boldsymbol{x})\right) \leq \rho_\theta \right\}. \tag{68}$$

Thus, $p = 1$ corresponds to a Wasserstein-1 ambiguity set, whereas $p = 2$ corresponds to an ambiguity set defined using the squared Wasserstein-2 distance.

For portfolio allocation in Section 5, we use a weighted sum of the mean and the conditional value-at-risk (CVaR) of the portfolio loss $-\boldsymbol{c}^\top \boldsymbol{z}$ as the objective function. The optimization formulation of the portfolio allocation is then

$$J_p^* := \min_{\substack{\boldsymbol{z} \\ \mathbf{1}^\top \boldsymbol{z} = 1,\ \boldsymbol{z} \geq \mathbf{0}}} \sup_{Q \in \mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta)} \left\{ \mathbb{E}_{\boldsymbol{c} \sim Q}\left[-\boldsymbol{c}^\top \boldsymbol{z}\right] + \beta \, \mathrm{CVaR}_\alpha^Q\left(-\boldsymbol{c}^\top \boldsymbol{z}\right) \right\}, \tag{69}$$

where $\alpha \in (0, 1)$ denotes the upper-tail probability in the CVaR, so that the associated confidence level is $1 - \alpha$, and $\beta > 0$ is a weight parameter controlling risk aversion. The objective in (69) balances expected performance and the risk associated with losses in the adverse tail of the candidate distributions (Rockafellar & Uryasev, 2000; Kuhn et al., 2019).

Using

$$\mathrm{CVaR}_\alpha^Q\left(-\boldsymbol{c}^\top \boldsymbol{z}\right) = \inf_{t \in \mathbb{R}} \mathbb{E}_{\boldsymbol{c} \sim Q}\left[t + \frac{1}{\alpha}\left(-\boldsymbol{c}^\top \boldsymbol{z} - t\right)_+ \right], \qquad (u)_+ := \max\{u, 0\}, \tag{70}$$

for any fixed Q and $\boldsymbol{z}$, we obtain

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{c} \sim Q}\left[-\boldsymbol{c}^\top \boldsymbol{z}\right] + \beta \, \mathrm{CVaR}_\alpha^Q\left(-\boldsymbol{c}^\top \boldsymbol{z}\right) \\ &\quad = \inf_{t \in \mathbb{R}} \mathbb{E}_{\boldsymbol{c} \sim Q}\left[-\boldsymbol{c}^\top \boldsymbol{z} + \beta t + \frac{\beta}{\alpha}\left(-\boldsymbol{c}^\top \boldsymbol{z} - t\right)_+ \right]. \end{aligned} \tag{71}$$

For fixed $\boldsymbol{z}$, define

$$\Phi_{\boldsymbol{z}}(Q, t) := \mathbb{E}_{\boldsymbol{c} \sim Q}\left[-\boldsymbol{c}^\top \boldsymbol{z} + \beta t + \frac{\beta}{\alpha}\left(-\boldsymbol{c}^\top \boldsymbol{z} - t\right)_+ \right].$$

The ambiguity set $\mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta)$ is convex, $\Phi_{\boldsymbol{z}}(Q, t)$ is affine and continuous in Q with respect to the W_p topology, and it is convex and continuous in t. Moreover, the distributions in $\mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta)$ have uniformly bounded first moments. Consequently, t can be restricted to a sufficiently large compact interval without changing either side of the minimax identity. The minimax theorem therefore gives

$$\sup_{Q \in \mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta)} \inf_{t \in \mathbb{R}} \Phi_{\boldsymbol{z}}(Q, t) = \inf_{t \in \mathbb{R}} \sup_{Q \in \mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta)} \Phi_{\boldsymbol{z}}(Q, t). \tag{72}$$

Therefore, (69) is equivalent to

$$J_p^* = \min_{\substack{\boldsymbol{z},\, t \in \mathbb{R} \\ \mathbf{1}^\top \boldsymbol{z} = 1,\ \boldsymbol{z} \geq \mathbf{0}}} \sup_{Q \in \mathcal{U}_\theta(\boldsymbol{x}; \rho_\theta)} \mathbb{E}_{\boldsymbol{c} \sim Q}\left[-\boldsymbol{c}^\top \boldsymbol{z} + \beta t + \frac{\beta}{\alpha}\left(-\boldsymbol{c}^\top \boldsymbol{z} - t\right)_+ \right]. \tag{73}$$

In portfolio allocation, define the augmented decision variable $\bar{\boldsymbol{z}} := (\boldsymbol{z}, t)$. The task loss in (1) is then

$$f(\boldsymbol{x}, \bar{\boldsymbol{z}}, \boldsymbol{c}) := -\boldsymbol{c}^\top \boldsymbol{z} + \beta t + \frac{\beta}{\alpha}\left(-\boldsymbol{c}^\top \boldsymbol{z} - t\right)_+. \tag{74}$$

This function is convex in $\bar{\boldsymbol{z}} = (\boldsymbol{z}, t)$ and piecewise affine in $\boldsymbol{c}$, since

$$f(\boldsymbol{x}, \bar{\boldsymbol{z}}, \boldsymbol{c}) = \max\left\{ -\boldsymbol{c}^\top \boldsymbol{z} + \beta t,\ -\left(1 + \frac{\beta}{\alpha}\right)\boldsymbol{c}^\top \boldsymbol{z} + \beta\left(1 - \frac{1}{\alpha}\right) t \right\}. \tag{75}$$

Finally, the portfolio constraints can be represented in the form used in (1) by defining

$$\boldsymbol{g}(\boldsymbol{x}, \bar{\boldsymbol{z}}) := \begin{bmatrix} \mathbf{1}^\top \boldsymbol{z} - 1 \\ 1 - \mathbf{1}^\top \boldsymbol{z} \\ -\boldsymbol{z} \end{bmatrix} \leq \mathbf{0}. \tag{76}$$

L Details of Experiments

We implement the lower-level DRO in (1) as a differentiable optimization layer using CVXPYLayers (Agrawal et al., 2019). Just like other DFL papers, we first pre-train the predictive model in the prediction-focused way for a number of epochs and then further train it end-to-end.

L.1 Details of the Baselines

Here we give the details of the baselines we compare against.

Distributionally robust predict-then-optimize baselines (DRPtO): We first train a distribution predictor (GMDN, CNF) to minize the prediction error, whose training loss is

$$\min_\theta G(\theta) := -\frac{1}{N_{\text{tra}}}\sum_{i=1}^{N_{\text{tra}}}\frac{1}{S}\sum_{j=1}^{S}\log p_\theta\left(\boldsymbol{c}_i^j \mid \boldsymbol{x}_i\right),$$

where p_θ is the predicted probability density. After the prediction-focused training ends, we conformally calibrate the Wasserstein ambiguity set's radius ρ_θ. On the test data, we solve the following DRO problem to obtain the decision,

$$\begin{aligned}\min_{\boldsymbol{z}} \quad & \sup_{Q:W_p^p(Q,P_\theta(\cdot|\boldsymbol{x}^{\text{test}}))\le\rho_\theta} E_{\boldsymbol{c}\sim Q}[-\boldsymbol{c}^\top\boldsymbol{z}] + \beta\,\text{CVaR}_\alpha^Q(-\boldsymbol{c}^\top\boldsymbol{z}),\\ \text{s.t.} \quad & \mathbf{1}^\top\boldsymbol{z}\le 1,\ -\mathbf{1}^\top\boldsymbol{z}\le -0.99,\ \boldsymbol{z}\ge\mathbf{0},\end{aligned} \tag{77}$$

where θ denotes the trained parameter of the predictor, $\boldsymbol{c}$ is the return vector of n_c assets, $\boldsymbol{z}$ is the allocation vector, CVaR_α represents the conditional value-at-risk with a risk level of α (Rockafellar & Uryasev, 2000) and β is a weight factor balancing the mean of the return and the risk (CVaR).

Deterministic predict-then-optimize baselines (DetPtO): We first train a multi-layer perceptron (MLP) to minimize the prediction error, whose training loss is

$$\min_\theta G(\theta) := -\frac{1}{N_{\text{tra}}}\sum_{i=1}^{N_{\text{tra}}}\frac{1}{S}\sum_{j=1}^{S}||\boldsymbol{h}_\theta(\boldsymbol{x}_i)-\boldsymbol{c}_i^j||_2^2, \tag{78}$$

where $\boldsymbol{h}_\theta(\cdot)$ is the prediction given by the MLP. After the prediction-focused training ends, we solve the following deterministic optimization problem based on the predictions given by the trained MLP,

$$\min_{\boldsymbol{z}} -\boldsymbol{h}_\theta^\top\boldsymbol{z} \tag{79}$$

$$\text{s.t.} \quad \mathbf{1}^\top\boldsymbol{z}\le 1,\ -\mathbf{1}^\top\boldsymbol{z}\le -0.99,\ \boldsymbol{z}\ge\mathbf{0} \tag{80}$$

Stochastic predict-then-optimize baselines (StoPtO): We first train a GMDN or a CNF to minimize the prediction error, whose training loss is

$$\min_\theta G(\theta) := -\frac{1}{N_{\text{tra}}}\sum_{i=1}^{N_{\text{tra}}}\frac{1}{S}\sum_{j=1}^{S}\log p_\theta\left(\boldsymbol{c}_i^j \mid \boldsymbol{x}_i\right),$$

where p_θ is the predicted probability density. After the prediction-focused training ends, we directly solve the following stochastic optimization problem with respect to the predicted distribution to obtain the decision (no ambiguity),

$$\begin{aligned}\min_{\boldsymbol{z}} \quad & \mathbb{E}_{\boldsymbol{c}\sim P_\theta(\cdot|\boldsymbol{x}^{\text{test}})}[-\boldsymbol{c}^\top\boldsymbol{z}] + \beta\,\text{CVaR}_\alpha^{P_\theta(\cdot|\boldsymbol{x}^{\text{test}})}(-\boldsymbol{c}^\top\boldsymbol{z}),\\ \text{s.t.} \quad & \mathbf{1}^\top\boldsymbol{z}\le 1,\ -\mathbf{1}^\top\boldsymbol{z}\le -0.99,\ \boldsymbol{z}\ge\mathbf{0},\end{aligned} \tag{81}$$

Deterministic DFL (DetDFL): a MLP is trained to minimize the task regret with the lower-level as deterministic optimization,

$$\min_\theta \; F(\theta) := \frac{1}{N_{\text{tra}}} \sum_{i=1}^{N_{\text{tra}}} \frac{1}{S} \sum_{j=1}^{S} f(\boldsymbol{x}_i, \boldsymbol{z}^*_\theta(\boldsymbol{x}_i), \boldsymbol{c}^j_i) \tag{82}$$

$$\text{s.t.} \quad \boldsymbol{z}^*_\theta(\boldsymbol{x}_i) \in \underset{\boldsymbol{z}:\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z})\leq \boldsymbol{0}}{\operatorname{argmin}} \; f(\boldsymbol{x}_i, \boldsymbol{z}, \boldsymbol{h}_\theta(\boldsymbol{x}_i)), \quad \forall i \in [N_{\text{tra}}], \tag{83}$$

Stochastic DFL (SDFL): a GMDN or a CNF trained to minimize the task regret with the lower-level as stochastic optimization,

$$\min_\theta \; F(\theta) := \frac{1}{N_{\text{tra}}} \sum_{i=1}^{N_{\text{tra}}} \frac{1}{S} \sum_{j=1}^{S} f(\boldsymbol{x}_i, \boldsymbol{z}^*_\theta(\boldsymbol{x}_i), \boldsymbol{c}^j_i)$$

$$\text{s.t.} \quad \boldsymbol{z}^*_\theta(\boldsymbol{x}_i) \in \underset{\boldsymbol{z}:\boldsymbol{g}(\boldsymbol{x}_i,\boldsymbol{z})\leq \boldsymbol{0}}{\operatorname{argmin}} \; \mathbb{E}_{\tilde{\boldsymbol{c}}_i \sim P_\theta(\cdot|\boldsymbol{x}_i)} \left[f(\boldsymbol{x}_i, \boldsymbol{z}, \tilde{\boldsymbol{c}}_i) \right], \quad \forall i \in [N_{\text{tra}}].$$

L.2 Details of the Synthetic Data

We construct a dataset of the form $\left(\boldsymbol{x}_i, \{\boldsymbol{c}^j_i\}_{j=1}^S\right)$ with $i = [1, N]$ using Warped K-Sinusoidal.

First, the feature vectors are sampled independently as

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1,1]^{n_x}), \qquad i = 1, \ldots, N. \tag{84}$$

For every $\boldsymbol{x} \in [-1,1]^{n_x}$, we define K feature-dependent mixture weights

$$\pi_k(\boldsymbol{x}) = \frac{\exp\{\ell_k(\boldsymbol{x})\}}{\sum_{r=1}^{K_{\text{comp}}} \exp\{\ell_r(\boldsymbol{x})\}}, \qquad k = 1, \ldots, K_{\text{comp}}, \tag{85}$$

where

$$\ell_k(\boldsymbol{x}) = a_k + \boldsymbol{b}_k^\top \boldsymbol{x} + \gamma_k \sin(\boldsymbol{\omega}_k^\top \boldsymbol{x} + \phi_k). \tag{86}$$

The feature-dependent mean of regime k is

$$\boldsymbol{\mu}_k(\boldsymbol{x}) = s_\mu \left[\boldsymbol{q}_k + A_k \boldsymbol{x} + \boldsymbol{r}_k \odot \sin(\Omega_k \boldsymbol{x} + \boldsymbol{\varphi}_k) \right], \tag{87}$$

where $\odot$ denotes the Hadamard product and the sine is applied componentwise.

To model feature-dependent uncertainty and cross-coordinate dependence, we define

$$\boldsymbol{\Sigma}_k(\boldsymbol{x}) = \Lambda_k \Lambda_k^\top + \text{Diag}\left(\boldsymbol{\sigma}_k^2(\boldsymbol{x})\right), \tag{88}$$

where

$$\sigma_{k,\ell}(\boldsymbol{x}) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \,\text{sigmoid}\left(a^{(\sigma)}_{k,\ell} + (\boldsymbol{v}^{(\sigma)}_{k,\ell})^\top \boldsymbol{x}\right), \qquad \ell = 1, \ldots, d_c. \tag{89}$$

For each feature vector $\boldsymbol{x}_i$ and each conditional realization $j = 1, \ldots, S$, we independently sample

$$z^j_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Categorical}\left(\pi_1(\boldsymbol{x}_i), \ldots, \pi_K(\boldsymbol{x}_i)\right), \tag{90}$$

and, conditional on $z^j_i = k$,

$$\boldsymbol{u}^j_i \mid (\boldsymbol{x}_i, z^j_i = k) \overset{\text{ind}}{\sim} \mathcal{N}\left(\boldsymbol{\mu}_k(\boldsymbol{x}_i), \boldsymbol{\Sigma}_k(\boldsymbol{x}_i)\right). \tag{91}$$

To obtain a non-Gaussian conditional distribution, we apply the invertible triangular transformation

$$v^j_{i,1} = u^j_{i,1}, \tag{92}$$

$$v^j_{i,\ell} = u^j_{i,\ell} + \kappa_\ell \tanh(u^j_{i,\ell-1}), \qquad \ell = 2, \ldots, d_c. \tag{93}$$

Finally, the normalized return vector is obtained through the componentwise transformation

$$\boldsymbol{C}_i^j = r_{\max} \tanh\left(\frac{\boldsymbol{v}_i^j}{s_c}\right). \tag{94}$$

Consequently,

$$\boldsymbol{c}_i^j \in (-r_{\max}, r_{\max})^{d_c},$$

and the conditional support satisfies

$$\operatorname{diam}\left(\operatorname{supp} P(\cdot \mid \boldsymbol{x})\right) \le 2r_{\max}\sqrt{d_c} \qquad \text{for all } \boldsymbol{x}. \tag{95}$$

Since the regime variables and Gaussian innovations are independently generated for every j, the resulting samples satisfy

$$\boldsymbol{c}_i^1, \ldots, \boldsymbol{c}_i^S \mid \boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} P(\cdot \mid \boldsymbol{x}_i). \tag{96}$$

Therefore, the synthetic dataset is

$$\mathcal{D} = \left\{\left(\boldsymbol{x}_i, \{\boldsymbol{c}_i^j\}_{j=1}^S\right)\right\}_{i=1}^N. \tag{97}$$

In this paper's experiments, we set $n_{\boldsymbol{x}} = 10$, $n_{\boldsymbol{c}} = 20$ and $K = 3$.

L.3 Details of the Real-World Portfolio Allocation Data

Here we give the details of the real-world portfolio allocation dataset.

We construct the feature using the last half-year return, the last quarter return, the last month return, the last week return, the daily returns and the daily volume changes of the last 20 days and compress the feature dimension n_x to 50 using PCA. We construct samples from the return distribution on day i as $\boldsymbol{c}_i^j := \frac{\boldsymbol{p}_i^j - \boldsymbol{p}_{i-1}^j}{\boldsymbol{p}_{i-1}^j}$, where j is taken from 9:30 to 16:00 every 15 minutes. There are totally 2533 data points, which are split into $\mathcal{D}_{\text{tra}}(56\%)$, $\mathcal{D}_{\text{val}}(14\%)$, $\mathcal{D}_{\text{cal}}(15\%)$, $\mathcal{D}_{\text{test}}(15\%)$.

L.4 Further Discussion of the Experiment Results

In both the synthetic example and the real-world portfolio allocation, DRDFL CNF W_2 performs a bit better than DRDFL CNF W_1, but they are still close. For our DRDFL problem, it is difficult to give a conclusion on which of W_1 and W_2 will give a better result. Both W_1 and W_2 ambiguity sets are calibrated using noncoformity scores estimating the corresponding Wasserstein distance, which is self-consistent. The dynamics of the end-to-end training and the optimization error in the training can both influence the final task performance. Our DRDFL supports both W_1 and W_2, so that users can choose the appropriate one according to the objective function of the task.