

# Robust Out-of-Distribution Stochastic Optimization Based on Wasserstein-Metric Meta-Distribution Support Learning

Biao Han

hanb16@tsinghua.org.cn

## Abstract

We study stochastic optimization with a completely unobserved target distribution and samples from related source distributions. Treating the sources and the target as independent draws from a common meta-distribution, we propose Wasserstein-metric meta-distribution support learning (MDSL). The model jointly learns the center and radius of a Wasserstein ball, with an exclusion parameter controlling the fraction of sources left outside. We identify a geometric failure of kernel-embedding-based descriptions, whose enclosing balls can admit arbitrarily dispersed distributions and produce infinite worst-case costs. The proposed MDSL is convex and admits a strong-duality formulation whose inner problem is a weighted Wasserstein barycenter of the sources. For a finitely supported learned center, standard Wasserstein duality yields a convex robust decision problem with separable inner maximizations. We connect these two stages through heuristics informed by MDSL: support distributions, their weights and optimal transport plans generate candidate scenarios for a finite convex initialization problem, while transport directions and available loss gradients guide search priorities. Cached candidates seed subsequent inner searches and are augmented by newly found adverse scenarios. On a compact observation space, we establish source-estimation perturbation bounds and, under quadratic growth, stability bounds for the center and admissible radius interval. We also derive out-of-distribution coverage and decision-performance guarantees for margin-enlarged balls, separating uncertainty from finitely many sources and finitely many observations per source. Under stated complexity and sampling conditions, coverage is asymptotically at least one minus the exclusion level as the enlargement vanishes. The framework thus links distributional support learning with robust optimization through both a learned uncertainty set and reusable transport information.

## 1 Introduction

Stochastic programming addresses decisions that must be made before the relevant randomness is resolved. Let  $\Xi$  be the space in which the uncertain data  $\xi$  take their values, and let  $\mathcal{P}(\Xi)$  denote the collection of all probability measures supported on it. A stochastic program then selects a decision  $\mathbf{x} \in \mathcal{X}$  that minimizes the expected cost incurred under some law  $\mathbb{P} \in \mathcal{P}(\Xi)$ :

$$\min_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathbb{P}} \{f(\mathbf{x}, \xi)\}, \quad (1.1)$$

where  $f$  is the cost function and the expectation is taken over  $\xi \sim \mathbb{P}$  (Shapiro et al., 2021). With  $\mathbb{P}$  in hand (1.1) would be unremarkable; the difficulty is that  $\mathbb{P}$  is precisely what the decision maker does not have. The customary substitute is an empirical measure assembled from whatever samples are available, an approach known as the sample average approximation, and its weakness is well documented: the resulting in-sample optimum is systematically optimistic. Distributionally robust optimization (DRO) responds by refusing to commit to any single substitute law and hedging instead against a whole family of them (Kuhn et al., 2025),

$$\min_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{U}} \mathbb{E}_{\xi \sim \mathbb{P}} \{f(\mathbf{x}, \xi)\}, \quad (1.2)$$

where the ambiguity set  $\mathcal{U} \subseteq \mathcal{P}(\Xi)$  comprises those laws judged close to the empirical one, with closeness quantified by, e.g.,  $\phi$ -divergences (Ben-Tal et al., 2013) or the Wasserstein metric (Mohajerin Esfahani and Kuhn, 2018). Yet each of these paradigms, whatever its choice of  $\mathcal{U}$ , presumes the same thing: that the target law  $\mathbb{P} \in \mathcal{P}(\Xi)$  has been sampled at least once.

Not every decision enjoys that luxury, and the situations this paper addresses lie strictly beyond it. The target distribution  $\mathbb{P}$  is *completely unobserved*: at the moment the decision is to be made, neither  $\mathbb{P}$  itself nor a single draw from it is available. Such *cold-start* situations are common in retail, manufacturing, and public health. A shop stocking a newly launched seasonal coffee, for example, holds no sales history of that item and can look only to the records of related products. The natural recourse is to let data from other, relevant distributions speak to the unseen  $\mathbb{P}$ .

**Problem 1.** *Let  $\mathbb{P} \in \mathcal{P}(\Xi)$  be an unobserved target distribution. Available to the decision maker are empirical data from  $M$  similar source distributions  $\mathbb{P}_1, \dots, \mathbb{P}_M$ , where the  $i$ -th source resembles  $\mathbb{P}$  without being identical to it, and is observed through  $n_i$  samples drawn from it. Using these source distributions, construct an uncertainty set  $\mathcal{U} \subset \mathcal{P}(\Xi)$  that concentrates near them, and return a decision  $\mathbf{x}^* \in \mathcal{X}$  that minimizes the worst-case expected cost*

$$\sup_{\mathbb{P} \in \mathcal{U}} \mathbb{E}_{\xi \sim \mathbb{P}} \{f(\mathbf{x}, \xi)\},$$

*so that  $\mathbf{x}^*$  performs robustly when deployed under the unseen target distribution  $\mathbb{P}$ .*

**Assumption 1** (Standing regularity). *The feasible set  $\mathcal{X} \subseteq \mathbb{R}^n$  is convex and compact, and the cost  $f : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$  is continuous on  $\mathcal{X} \times \Xi$  and convex in its first argument, so that  $f(\cdot, \xi)$  is convex for every  $\xi \in \Xi$ . Compactness of  $\Xi$ , where a statement needs it, is imposed in that statement.*

## Related Work

Because the target law  $\mathbb{P}$  is never observed, the entire burden of Problem 1 falls on how the source distributions  $\mathbb{P}_1, \dots, \mathbb{P}_M$  are converted into an uncertainty set  $\mathcal{U}$ . Work on this question has developed along three main lines of thought, which differ in what  $\mathcal{U}$  is ultimately *made of*.

**Assembling  $\mathcal{U}$  from the sources by set operations.** The first family leaves the source distributions intact and builds  $\mathcal{U}$  out of them through elementary operations. A mixture model

takes convex combinations,

$$\mathcal{U}^{\text{co}} = \left\{ \sum_{i=1}^M \alpha_i \mathbb{P}_i : \alpha_i \geq 0, \sum_{i=1}^M \alpha_i = 1 \right\},$$

whose members are precisely the convex hull of the sources in  $\mathcal{P}(\Xi)$ ; this is the geometry underlying group DRO (Sagawa\* et al., 2020; Carmon and Hausler, 2022; Zhang et al., 2023), and it has been used to encode fairness requirements (Hashimoto et al., 2018; Lahoti et al., 2020) and to pursue robustness across client populations in federated learning (Mohri et al., 2019; Deng et al., 2020). A second variant instead intersects balls around the sources,

$$\mathcal{U}^\cap = \bigcap_{i=1}^M \mathbb{B}_{\epsilon_i}(\mathbb{P}_i),$$

where  $\mathbb{B}_\epsilon(\mathbb{P})$  gathers the laws within distance  $\epsilon$  of  $\mathbb{P}$  (Awasthi et al., 2022; Taskesen et al., 2021; Selvi et al., 2025; Wang et al., 2024; Rychener et al., 2024). Both variants are appealing because they assume nothing beyond the sources themselves. Both also inherit the same two defects: the set operations provide no handle on conservatism, a mixture set being typically over-pessimistic while an intersection may be empty or, in the opposite failure mode, over-confident; and the parameters they do expose resist calibration, the radii  $\epsilon_i$  calling for prior knowledge of how far each  $\mathbb{P}_i$  lies from  $\mathbb{P}$ , which is seldom at hand.

**Synthesizing a new distribution and centering  $\mathcal{U}$  on it.** The second family does not combine the sources into a set at all. It first distills them into a single *new* distribution and only then builds  $\mathcal{U}$  around that synthesized center. Pooling every sample into one aggregate dataset and applying the sample average approximation (Gupta and Kallus, 2022; Dursun et al., 2022) is the crudest instance, its weighted-average limit  $\bar{\mathbb{P}} = \sum_{i=1}^M (n_i/N) \hat{\mathbb{P}}_i$ , with  $N := \sum_{i=1}^M n_i$ , being a barycenter in the Euclidean sense. Optimal transport replaces that arithmetic mean by a geometric one. The 2-Wasserstein distance between  $\mathbb{Q}$  and  $\mathbb{P}$  is

$$\mathcal{W}_2(\mathbb{Q}, \mathbb{P}) := \left( \inf_{\pi \in \Pi(\mathbb{Q}, \mathbb{P})} \int_{\Xi \times \Xi} \|\xi - \zeta\|^2 d\pi(\xi, \zeta) \right)^{1/2}, \quad (1.3)$$

where  $\Pi(\mathbb{Q}, \mathbb{P})$  collects the couplings of  $\mathbb{Q}$  and  $\mathbb{P}$ . The *Wasserstein barycenter* of the sources is then defined as the law that minimizes the weighted sum of squared distances to them (Agueh and Carlier, 2011),

$$\bar{\mathbb{P}}^w \in \arg \min_{\mathbb{Q} \in \mathcal{P}(\Xi)} \sum_{i=1}^M \alpha_i \mathcal{W}_2^2(\mathbb{Q}, \mathbb{P}_i), \quad (1.4)$$

and the uncertainty set is the ball  $\mathbb{B}_\epsilon(\bar{\mathbb{P}}^w)$  centered on this synthesized law (Lau and Liu, 2022; Cuturi and Doucet, 2014; Srivastava et al., 2018; Schmitz et al., 2018; Montesuma and Mboula, 2021). The break with the first family is one of kind rather than degree:  $\mathcal{U}$  is no longer made of the sources but of an object generated from them. The cost is that everything now rests on the

barycenter and on  $\epsilon$  being chosen well, and neither comes with a data-driven default.

**Learning  $\mathcal{U}$  from the sources.** The third family uses the source distributions as observations from which  $\mathcal{U}$  is to be *learned*. RooDSO (Li et al., 2026) follows this approach: it regards the target  $\mathbb{P}$  and the sources  $\mathbb{P}_i$  as independent realizations of an unknown meta-distribution  $\Psi \in \mathcal{P}(\mathcal{P}(\Xi))$ , maps the observed sources into a reproducing kernel Hilbert space (RKHS)  $\mathcal{H}$  through kernel mean embeddings (Aronszajn, 1950; Muandet et al., 2017), and learns a support description using support measure clustering (Muandet and Schölkopf, 2013). The description has a single parameter controlling its conservatism and is accompanied by finite-sample out-of-distribution (OOD) guarantees. However, their geometry depends on the chosen kernel and its parameters, and proximity in the RKHS need not have a direct interpretation in terms of transport or dispersion in the observation space. The learned embedding-space ball may admit distributions with substantially different dispersion from the sources. Section 2.3 illustrates this issue on an unbounded observation space: an RBF-based enclosing ball contains a neighborhood of the RKHS origin, admits arbitrarily dispersed distributions, and produces an infinite worst-case quadratic cost. These considerations motivate learning directly in Wasserstein geometry, where, under the stated regularity conditions, MDSL admits a barycenter-based learning dual and the robust decision problem can use standard Wasserstein duality. The resulting support distributions and transport plans also provide heuristic information for initialization and subsequent adverse-scenario searches, connecting the learning and decision stages computationally.

## Contributions

Our main contributions are summarized as follows.

1. **Support learning directly in Wasserstein space.** We introduce MDSL for robust decisions without target observations, learning the center and radius of a Wasserstein uncertainty ball from source distributions. Its hard and soft formulations are convex, and the soft formulation controls source exclusion through a single parameter  $\nu$ . Under compactness, its dual optimizes source weights over weighted Wasserstein barycenter problems and identifies support distributions governing the learned description.
2. **A geometric limitation of kernel-embedding descriptions.** We exhibit a configuration in which an RKHS enclosing ball contains the origin and admits probability distributions of arbitrarily large dispersion, making the worst-case quadratic cost infinite. We characterize origin enclosure for three RBF kernel mean embeddings and give a simple sufficient condition. The corresponding Wasserstein construction yields an explicit moment budget and a finite robust value in the same example.
3. **Information reuse between learning and decision optimization.** For a finitely supported center, we apply standard Wasserstein duality to obtain a jointly convex decision formulation with separable inner maximizations. We then use MDSL weights, boundary and exterior support

distributions, and optimal transport plans to construct candidate scenarios for joint initialization of the decision and transport penalty. This finite convex initialization supplies a lower bound on the robust optimal value. Transport directions, available loss gradients and cached adverse scenarios further provide heuristics for subsequent inner searches.

4. **Statistical guarantees under two-layer sampling.** On a compact observation space, we bound perturbations caused by empirical source estimation and, under quadratic growth, control the learned center and the full interval of optimal radius powers. For exact empirical MDSL optima, we establish coverage and decision-performance guarantees for margin-enlarged balls, separating the effects of the number of sources and within-source sample sizes. Under the stated complexity and sampling conditions, coverage is asymptotically at least  $1 - \nu$  with vanishing enlargement; these coverage results require neither uniqueness nor quadratic growth.

**Structure.** Section 2 introduces the Wasserstein geometry, the MDSL formulations and the induced robust decision problem, and examines the geometric limitation of kernel-embedding descriptions. Section 3 develops the learning dual and algorithms for both stages, including the reuse of MDSL transport information in decision optimization. Section 4 establishes finite-sample stability and out-of-distribution generalization guarantees. Section 5 concludes the paper. Appendix A collects additional notation and proofs of the geometric criterion, the robust dual reformulation and the statistical results.

**Notation.** For a positive integer  $N$ , write  $[N] := \{1, \dots, N\}$ , and let  $\mathbb{R}_+ := [0, \infty)$ . Bold symbols denote Euclidean vectors and matrices;  $\|\cdot\|$  is the Euclidean norm unless otherwise specified, and  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  and  $\|\cdot\|_{\mathcal{H}}$  denote the RKHS inner product and norm. For  $a \in \mathbb{R}$ , write  $a_+ := \max\{a, 0\}$ , also denoted by  $a^+$ . The symbols  $\mathbf{1}\{A\}$  and  $\#A$  denote the indicator of an event and the cardinality of a finite set, respectively. Let  $\delta_{\boldsymbol{\xi}}$  be the Dirac measure at  $\boldsymbol{\xi}$ . We use  $\mathcal{P}(\Xi)$  for the Borel probability measures on  $\Xi$ , and  $\mathcal{P}_p(\Xi)$  for those with finite  $p$ -th moment. Expectation and probability under  $\mathbb{P}$  are written  $\mathbb{E}_{\mathbb{P}}$  and  $\Pr_{\mathbb{P}}$ , and  $\xi_1, \dots, \xi_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$  denotes independent draws from  $\mathbb{P}$ . Hats identify quantities constructed from empirical source measures, and a superscript  $\star$  identifies optimal quantities in the mathematical formulations. In the algorithms,  $i$  indexes sources,  $j$  source samples,  $k$  center atoms and  $t$  iterations. The operator  $\text{proj}_{\mathcal{A}}$  is Euclidean projection onto a nonempty closed convex set  $\mathcal{A}$ , while  $\partial_{\mathbf{x}}f$ ,  $\nabla_{\mathbf{x}}f$  and  $\nabla_{\boldsymbol{\xi}}f$  denote the convex subdifferential in the decision variable and gradients in the indicated variables, when available. Additional conventions used in the proofs appear in Appendix A.1.

## 2 Proposed methodology

### 2.1 Wasserstein geometry on the space of distributions

Throughout,  $\Xi \subseteq \mathbb{R}^d$  is a nonempty closed set, and  $\mathcal{P}_p(\Xi)$  denotes the subset of  $\mathcal{P}(\Xi)$  consisting of the Borel probability measures with finite  $p$ -th moment,

$$\mathcal{P}_p(\Xi) := \left\{ \mathbb{P} \in \mathcal{P}(\Xi) : \int_{\Xi} \|\xi\|^p d\mathbb{P}(\xi) < \infty \right\} \subseteq \mathcal{P}(\Xi), \quad (2.1)$$

where  $p \geq 1$  is a fixed exponent. For  $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\Xi)$ , let  $\Pi(\mathbb{P}, \mathbb{Q})$  be the set of *couplings* of  $\mathbb{P}$  and  $\mathbb{Q}$ , that is, the probability measures on  $\Xi \times \Xi$  whose first and second marginals are  $\mathbb{P}$  and  $\mathbb{Q}$ . The  $p$ -Wasserstein distance, of which (1.3) is the  $p = 2$  instance already mentioned, is

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) := \left( \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\Xi \times \Xi} \|\xi - \zeta\|^p d\pi(\xi, \zeta) \right)^{1/p}. \quad (2.2)$$

Write  $\mathcal{S}_{\mathcal{W}_p} := (\mathcal{P}_p(\Xi), \mathcal{W}_p)$  for the  $p$ -Wasserstein space, and note that the definition is sound: the finite- $p$ -moment condition in (2.1) is precisely what makes the infimum in (2.2) finite, and  $\mathcal{W}_p$  is symmetric, vanishes exactly when  $\mathbb{P} = \mathbb{Q}$ , and satisfies the triangle inequality, so  $\mathcal{S}_{\mathcal{W}_p}$  is a genuine metric space. It metrizes weak convergence together with convergence of  $p$ -th moments (Villani, 2009); in particular, when  $\Xi$  is compact every probability measure on  $\Xi$  is tight, so  $\mathcal{S}_{\mathcal{W}_p}$  is a compact metric space. We shall use that compactness below when arguing that the descriptions we build are attained rather than merely approached.

Beyond these analytical properties, it is worth being explicit about what (2.2) measures, since this is where the geometry earns its place. In the Kantorovich formulation of optimal transport (Villani, 2009),  $\mathcal{W}_p^p(\mathbb{P}, \mathbb{Q})$  is the least cost of moving the mass of  $\mathbb{P}$  onto that of  $\mathbb{Q}$  when displacing a unit of mass from  $\xi$  to  $\zeta$  costs  $\|\xi - \zeta\|^p$ ; the distance is thus a transport cost, quoted in the units in which the data are themselves recorded.

### 2.2 Wasserstein-metric meta-distribution support learning

Let the source distributions be

$$\mathbb{P}_1, \dots, \mathbb{P}_M \stackrel{\text{i.i.d.}}{\sim} \Psi, \quad \Psi \in \mathcal{P}(\mathcal{P}_p(\Xi)), \quad (2.3)$$

so that the sources are themselves independent draws from a *meta-distribution*  $\Psi$  supported on the Wasserstein space, and let the unobserved target be a further independent draw  $\mathbb{P} \sim \Psi$ . This is the two-layer structure of the problem: one layer of randomness generates the distribution, the next generates the observation. Only the first layer is visible to the decision maker, and only through  $M$  draws. We write  $\mathcal{S} := \{\mathbb{P}_1, \dots, \mathbb{P}_M\}$  for the observed collection of sources.

We now give the geometry a description to learn. The description we attach to  $\mathcal{S}$  is the smallest  $\mathcal{W}_p$ -ball that contains all of them. This ball is essentially a description of the support of the meta-distribution  $\Psi$ : it is the smallest region of  $\mathcal{P}_p(\Xi)$  containing all the sources, and

the sense in which it captures  $\Psi$  itself, rather than only the observed draws from it. Writing  $\mathcal{B}_{\mathcal{W}_p}(\mathbb{Q}, R) := \{\mathbb{P} \in \mathcal{P}_p(\Xi) : \mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \leq R\}$  for the closed ball of radius  $R$  centered at  $\mathbb{Q}$ , we define the  $p$ -Wasserstein-metric meta-distribution support learning ( $\mathcal{W}_p$ -MDSL, or MDSL for short) by

$$\min_{\mathbb{Q} \in \mathcal{P}_p(\Xi), R \geq 0} R^p \quad \text{s.t.} \quad \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) \leq R^p, \quad i \in [M]. \quad (\text{MDSL-H})$$

The objective and the constraints of (MDSL-H) are written in the transport cost  $\mathcal{W}_p^p$ , whereas  $R$  remains a radius in the metric  $\mathcal{W}_p$  itself; since  $t \mapsto t^p$  is increasing on  $[0, \infty)$ , minimizing  $R^p$  is the same as minimizing  $R$ , so the two readings pick out the same ball. Because the constraints hold with equality at the optimum, (MDSL-H) is equivalent to the unconstrained minimax problem

$$\mathbb{Q}^* \in \arg \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \max_{i \in [M]} \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}), \quad R^* := \max_{i \in [M]} \mathcal{W}_p(\mathbb{P}_i, \mathbb{Q}^*). \quad (2.4)$$

This hard formulation is closely related to the Wasserstein Ball Center (WBC) problem of Wang et al. (2025), but our focus is to learn a quantile support for a latent meta-distribution.

The hard problem (MDSL-H) requires that every source be enclosed, and is therefore fragile: a single mis-specified  $\mathbb{P}_i$  forces the radius up to the diameter and inflates the description for all the others. As in the classical soft-margin construction we therefore admit a slack  $s_i \geq 0$  in each constraint and charge it linearly,

$$\min_{\substack{\mathbb{Q} \in \mathcal{P}_p(\Xi), R \geq 0 \\ \mathbf{s} \in \mathbb{R}_+^M}} R^p + \frac{1}{M\nu} \sum_{i \in [M]} s_i \quad \text{s.t.} \quad \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) \leq R^p + s_i, \quad i \in [M], \quad (\text{MDSL-S})$$

with  $\nu \in (0, 1]$  the *level of exclusion*. After eliminating the slacks, soft MDSL minimizes the empirical CVaR of the source-to-center transport costs  $\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q})$ , with  $\nu$  specifying the upper-tail fraction. As  $\nu \rightarrow 0$  the slacks are driven to zero and (MDSL-S) returns (MDSL-H); as  $\nu$  increases, sources lying far from the bulk of  $\{\mathbb{P}_i\}$  may be placed outside the ball at the price of a slack that is itself minimized. Since (MDSL-S) contains (MDSL-H) as a special case, MDSL is understood from here on as (MDSL-S) unless a distinction is drawn.

Note that posing the MDSL formulation on  $\mathcal{S}_{\mathcal{W}_p}$  is legitimate: that space is a metric space (Section 2.1), and every component the problems involve is well defined on it. What is more, although  $\mathcal{S}_{\mathcal{W}_p}$  is not a vector space, the convexity of the problem are retained.

**Proposition 1** (Convexity of the MDSL problem). *Under the substitution  $\rho := R^p$ , a bijection of  $\mathbb{R}_+$  onto itself, each of (MDSL-H) and (MDSL-S) minimises a linear function subject to convex constraints, and is therefore a convex program: every local minimizer of either is a global one.*

*Proof.* Write  $\rho := R^p$ . The constraints of (MDSL-H) then read  $\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \rho \leq 0$ , those of (MDSL-S) the same with  $-s_i$  added, and the objectives are  $\rho$  and  $\rho + \frac{1}{M\nu} \sum_{i \in [M]} s_i$ , both linear. Two facts make the constraint functions convex. First,  $\mathcal{P}_p(\Xi)$  is convex: the mixture  $\theta\mathbb{P} + (1-\theta)\mathbb{Q}$ , with  $\theta \in [0, 1]$ , of two probability measures is again a probability measure, and it retains the finite

$p$ -th moment. Second,  $\mathbb{Q} \mapsto \mathcal{W}_p^p(\mathbb{P}, \mathbb{Q})$  is convex: if  $\pi_1, \pi_2$  are optimal couplings in (2.2) for  $(\mathbb{P}, \mathbb{Q}_1)$  and  $(\mathbb{P}, \mathbb{Q}_2)$ , then  $\theta\pi_1 + (1 - \theta)\pi_2$  has marginals  $\mathbb{P}$  and  $\theta\mathbb{Q}_1 + (1 - \theta)\mathbb{Q}_2$ , so it is a coupling of those two laws and cannot beat the optimal one, whence

$$\mathcal{W}_p^p(\mathbb{P}, \theta\mathbb{Q}_1 + (1 - \theta)\mathbb{Q}_2) \leq \theta \mathcal{W}_p^p(\mathbb{P}, \mathbb{Q}_1) + (1 - \theta) \mathcal{W}_p^p(\mathbb{P}, \mathbb{Q}_2). \quad (2.5)$$

Each constraint function  $\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \rho$  is therefore convex in  $(\mathbb{Q}, \rho)$ , and the objectives being linear, both programs are convex.  $\square$

Obviously, the proposed MDSL formulation is a natural generalization of the classical support vector data description (SVDD) problem (Tax and Duin, 2004) to the Wasserstein metric space. Section 3.1 will introduce how it is solved through its dual that has a special structure. It will come as no surprise that the model inherits several of the classical properties of SVDD. The learned center  $\mathbb{Q}^*$  will be formulated by only a small fraction of non-zero dual multipliers  $\{\alpha_i^*\}$ , which in this paper are referred to as *support distributions* (SDs), with the index set

$$\text{SD} := \{i \in [M] : \alpha_i^* > 0\}. \quad (2.6)$$

SD consists of *boundary support distributions* (BSDs) and *exterior support distributions* (ESDs), the index sets of which are

$$\text{BSD} := \{i \in [M] : 0 < \alpha_i^* < \frac{1}{M\nu}\}, \quad \text{ESD} := \{i \in [M] : \alpha_i^* = \frac{1}{M\nu}\}, \quad (2.7)$$

i.e.,  $\text{SD} = \text{BSD} \cup \text{ESD}$ . We call the remaining source distributions *interior distributions* (IDs) and denote their index set by

$$\text{ID} := \{i \in [M] : \alpha_i^* = 0\}, \quad (2.8)$$

so that  $\text{ID} = [M] \setminus \text{SD}$  and BSD, ESD, ID partition  $[M]$ .

### 2.3 $\mathcal{W}_p$ -MDSL uncertainty set-induced RooDSO

Because a distribution is a point of  $\mathcal{P}_p(\Xi)$ , the ball of Section 2.2 is at once a description of the sources and a set of distributions that an unseen target may plausibly be. We therefore read off from  $\mathbb{Q}^*$  and  $R^*$  the *out-of-distribution uncertainty set* driven by  $\mathcal{P}$  under hyperparameter  $\nu$ :

$$\mathcal{U}_\nu^{\text{MDSL}}(\mathcal{P}) := \left\{ \mathbb{P} \in \mathcal{P}_p(\Xi) : \mathcal{W}_p(\mathbb{P}, \mathbb{Q}^*) \leq R^* \right\} = \mathcal{B}_{\mathcal{W}_p}(\mathbb{Q}^*, R^*), \quad (2.9)$$

where the argument  $\mathcal{P}$  is dropped when the sources are understood, writing  $\mathcal{U}_\nu^{\text{MDSL}}$  for short. Substituting this set into (1.2) gives the robustified program, which serves as the worst-case surrogate of Problem 1:

$$\min_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{U}_\nu^{\text{MDSL}}} \mathbb{E}_{\xi \sim \mathbb{P}} \{f(\mathbf{x}, \xi)\}. \quad (\text{RooDSO-MDSL})$$

Li et al. (2026) proposed a RooDSO framework based on kernel mean embedding (KME). They first map the source distributions into a reproducing kernel Hilbert space through KME, and then carry out SVDD to construct an OOD uncertainty set, which is used for the downstream distributionally robust optimization. It can be viewed as a counterpart of the method of this paper in a different operating space. This approach benefits from the Hilbert-space structure of kernel mean embeddings. However, on an unbounded observation space, an enclosing RKHS ball need not control the moments of the admissible distributions. For an RBF kernel, certain increasingly dispersed distributions have embeddings approaching the origin. If the learned ball contains the origin in its interior, such distributions can remain admissible and lead to an infinite worst-case value for an unbounded loss. As a result, the resulting RooDSO problem becomes ineffective, and no meaningful solution can be obtained. The following example illustrates this limitation.

**Example 1** (Unbounded quadratic cost under an RBF embedding ball). *Let  $\Xi = \mathbb{R}$ , and let  $\mathcal{H}$  be the RKHS of the RBF kernel  $k(x, y) = \exp(-(x - y)^2/2)$  of bandwidth  $\sigma = 1$ . The three sources  $\mathbb{P}_1 = \delta_{-2}$ ,  $\mathbb{P}_2 = \delta_0$ ,  $\mathbb{P}_3 = \delta_2$  have KMEs  $\mu_i := \mu_{\mathbb{P}_i} = \phi(x_i)$ , with  $\phi(x) = k(x, \cdot) \in \mathcal{H}$ . Since  $k(x, x) = 1$ , each  $\mu_i$  lies on the unit sphere of  $\mathcal{H}$ , and the pairwise inner products*

$$\langle \mu_1, \mu_2 \rangle_{\mathcal{H}} = \langle \mu_2, \mu_3 \rangle_{\mathcal{H}} = \kappa := e^{-2} \approx 0.1353, \quad \langle \mu_1, \mu_3 \rangle_{\mathcal{H}} = \kappa^4 \approx 0.0003.$$

*The minimum enclosing ball of  $\{\mu_1, \mu_2, \mu_3\}$  is then determined by all three points, its center being the convex combination  $\mathbf{c}^* = \lambda(\mu_1 + \mu_3) + (1 - 2\lambda)\mu_2$  that is equidistant from each of them; requiring  $\langle \mathbf{c}^*, \mu_1 \rangle_{\mathcal{H}} = \langle \mathbf{c}^*, \mu_2 \rangle_{\mathcal{H}}$  gives*

$$\lambda = \frac{1 - \kappa}{3 - 4\kappa + \kappa^4} \approx 0.3516.$$

*Because all three KME points lie on the unit sphere, equidistance from them also forces  $\|\mathbf{c}^*\|_{\mathcal{H}}^2 = \langle \mathbf{c}^*, \mu_1 \rangle_{\mathcal{H}}$ , so that*

$$\|\mathbf{c}^*\|_{\mathcal{H}}^2 = \lambda(1 + \kappa^4) + (1 - 2\lambda)\kappa \approx 0.3919, \quad r^2 = 1 - \|\mathbf{c}^*\|_{\mathcal{H}}^2 \approx 0.6081.$$

*Hence  $\|\mathbf{c}^*\|_{\mathcal{H}} \approx 0.6260 < 0.7798 \approx r$ , and the RKHS origin  $\mathbf{0}$  lies strictly inside the SVDD ball.*

*The KME uncertainty set  $\mathcal{U}^{KME} = \{\mathbb{P} : \mu_{\mathbb{P}} \in \text{SVDD ball}\}$  then contains an open neighborhood of  $\mathbf{0}$ . Since the KME of the uniform law  $\mathbb{U}_L = \mathcal{U}([-L, L])$  satisfies  $\mu_{\mathbb{U}_L} \rightarrow \mathbf{0}$  in  $\mathcal{H}$  as  $L \rightarrow \infty$ , we have  $\mathbb{U}_L \in \mathcal{U}^{KME}$  for every  $L$  large enough: the uncertainty set contains distributions of arbitrarily large support.*

*Take  $f(x, \xi) = (x - \xi)^2$  with  $x \in \mathbb{R}$ . Because  $\mathbb{E}_{\mathbb{U}_L}[(x - \xi)^2] = x^2 + L^2/3$  grows without bound in  $L$ , the inner problem is  $+\infty$ , and*

$$\min_{x \in \mathbb{R}} \max_{\mathbb{P} \in \mathcal{U}^{KME}} \mathbb{E}_{\mathbb{P}}[(x - \xi)^2]$$

*has no finite optimal value. The KME-then-SVDD construction therefore fails to define a well-posed distributionally robust optimization problem.*

The configuration above is not an isolated case assembled for the occasion. Under the RBF kernel, all distributions in  $\mathcal{P}(\Xi)$  are mapped into a convex subset of the unit ball in  $\mathcal{H}$ . It is then easy to show that the more concentrated a distribution is, the closer its kernel mean embedding lies to the unit sphere of  $\mathcal{H}$ , while the flatter it is, the closer that embedding lies to the origin. Whenever the kernel mean embeddings of the sources are dispersed in some sense, the SVDD description they induce is likely to enclose the origin as well. In fact, the geometry of origin enclosure can be examined explicitly for three source embeddings. The following proposition gives a necessary and sufficient criterion, and the subsequent corollary provides a convenient sufficient condition.

**Proposition 2** (Origin enclosure for three embedded points). *Let  $\mu_1, \mu_2, \mu_3$  be three kernel mean embeddings of different probability distributions with RBF kernel in a RKHS  $\mathcal{H}$ , with Gram matrix  $\mathbf{G} = (g_{ij})_{3 \times 3}$ ,  $g_{ij} = \langle \mu_i, \mu_j \rangle_{\mathcal{H}}$ . Then the minimum enclosing ball of  $\{\mu_1, \mu_2, \mu_3\}$  contains the origin,  $\mathbf{0} \in \mathcal{H}$ , if and only if*

$$\begin{cases} g_{ii} + g_{jk} > g_{ij} + g_{ik}, & \forall i \in \{1, 2, 3\}, \{j, k\} = \{1, 2, 3\} \setminus \{i\} \\ \mathbf{1}^\top \mathbf{G}^{-1} \mathbf{d} \geq 2, \end{cases}$$

where  $\mathbf{d} := (g_{11}, g_{22}, g_{33})^\top$ ,  $\mathbf{1} := (1, 1, 1)^\top$ . The first three inequalities are equivalent to saying that  $\mu_1, \mu_2, \mu_3$  form an acute triangle. The proof is given in Appendix A.2.

**Corollary 3** (A simple sufficient condition). *Under the assumptions of Proposition 2, suppose that*

$$\sum_{j \neq i} g_{ij} \leq \frac{1}{2} g_{ii}, \quad i = 1, 2, 3.$$

*Then the minimum enclosing ball of  $\{\mu_1, \mu_2, \mu_3\}$  contains the origin.*

**Remark 1.** *In the unit-norm case  $g_{ii} = 1$  with equal pairwise inner products  $g_{ij} = \kappa$  ( $i \neq j$ ), the row condition of Corollary 3 reads  $2\kappa \leq \frac{1}{2}$ , that is,  $\kappa \leq \frac{1}{4}$ . The proposition above then shows that if the Gram matrix produced by the method of Li et al. (2026) has a  $3 \times 3$  principal submatrix satisfying that row condition, the minimum enclosing ball of the three kernel mean embeddings it selects already encloses the origin. That small ball is very likely to be contained in the SVDD ball enclosing all the kernel mean embedding points when the three points lie in the interior of the point cloud, so that the OOD uncertainty set built from it is very likely to enclose the origin, which in turn renders the downstream distributionally robust optimization ineffective.*

Generally, under the common setting of dispersed data and a moderate or small bandwidth, the KME-then-SVDD ball very easily contains the RKHS origin, which renders the adversarial distribution meaningless. One way to avoid it is to take the bandwidth so large that all the KME points crowd together; but then the uncertainty set contracts instead, the DRO degenerates into empirical risk minimization, and robustness is lost. The median heuristic (Gretton et al., 2012) for setting the bandwidth alleviates this to some extent but does not remove the geometric flaw.

Various choices of kernels and kernel parameters affect the geometry of the learned description, but their implications for the downstream robust value require separate analysis. Traditional data-driven robust optimization, which builds its uncertainty set by this kernel method (Shang et al., 2017; Han et al., 2021), does not usually run into this situation, for the RBF kernel mapping of the data always lies on the boundary of the unit sphere of  $\mathcal{H}$ . This limitation on RooDSO motivates the use of Wasserstein geometry, which provides an explicit transport-cost budget.

We now apply our proposed methodology to the same problem as in Example 1. Since  $\mathcal{W}_p(\delta_a, \delta_b) = |a - b|$  for every  $p$ , the ball (MDSL-H) must span  $\delta_{-2}$  and  $\delta_2$ , which are 4 apart, so  $R^* \geq 2$ ; the Dirac  $\mathbb{Q}^* = \delta_0$  attains  $R^* = 2$ . Transport to a Dirac is forced, so  $\mathcal{W}_p^p(\mathbb{P}, \delta_0) = \mathbb{E}_{\mathbb{P}}[|\xi|^p]$ , and (2.9) becomes a moment budget,  $\mathcal{U}_0^{\text{MDSL}} = \{\mathbb{P} \in \mathcal{P}_p(\mathbb{R}) : \mathbb{E}_{\mathbb{P}}[|\xi|^p] \leq 2^p\}$ . The uniform law  $\mathbb{U}_L$  that left the kernel problem unbounded is therefore excluded once  $L^p/(p+1) > 2^p$ . With  $p = 2$  and the same loss  $f(x, \xi) = (x - \xi)^2$ , the inner supremum is  $(|x| + 2)^2$ , attained at  $\delta_{\pm 2}$ , so (RooDSO-MDSL) has optimal value 4 at  $x = 0$  rather than  $+\infty$ , which is a reasonable result.

### 3 Learning and optimization strategies

#### 3.1 Learning the uncertainty set

Convex programs are ordinarily solved through their duals, and the  $\mathcal{W}_p$ -MDSL problems (MDSL-H) and (MDSL-S) are convex by Proposition 1. They differ, however, from the convex programs one usually meets, which are posed over the Euclidean space  $\mathbb{R}^n$  or over a Hilbert space  $\mathcal{H}$ . Here the variable  $\mathbb{Q}$  ranges over  $\mathcal{P}_p(\Xi)$ , a convex subset of the vector space of finite signed measures, and the geometry that governs the problem is the metric space  $\mathcal{S}_{\mathcal{W}_p} = (\mathcal{P}_p(\Xi), \mathcal{W}_p)$  of Section 2.1. Because  $\mathcal{W}_p$  is not induced by an inner product,  $\mathcal{S}_{\mathcal{W}_p}$  carries no linear structure of its own. But convex analysis still applies in its abstract form,  $\mathcal{P}_p(\Xi)$  being convex and the objectives convex on it. The theorem below states the dual of (MDSL-S).

**Theorem 4** (Dual of the MDSL problem). *Suppose  $\Xi$  is compact, and write*

$$\mathcal{A} := \left\{ \alpha \in \mathbb{R}_+^M : \alpha_i \leq \frac{1}{M\nu} \text{ for every } i \in [M], \sum_{i \in [M]} \alpha_i = 1 \right\}.$$

*Then (MDSL-S) has the same optimal value as its following dual reformulation*

$$\max_{\alpha \in \mathcal{A}} \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}), \quad (\text{MDSL-D})$$

*whose inner problem is a weighted  $p$ -Wasserstein barycenter of the sources, the weights being  $\alpha_i$  (see (1.4) for the  $p = 2$  version), and whose outer problem maximizes the cost of that barycenter over the box-truncated simplex  $\mathcal{A}$ .*

*Proof.* We dualize (MDSL-S) by the Lagrangian route, working in the variable  $\rho := R^p$  as before.

Attach a multiplier  $\alpha_i \geq 0$  to each of the  $M$  constraints and a multiplier  $\eta_i \geq 0$  to each nonnegativity constraint  $s_i \geq 0$ , and set

$$\mathcal{L}(\mathbb{Q}, \rho, \mathbf{s}; \boldsymbol{\alpha}, \boldsymbol{\eta}) := \rho + \frac{1}{M\nu} \sum_{i \in [M]} s_i + \sum_{i \in [M]} \alpha_i (\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \rho - s_i) - \sum_{i \in [M]} \eta_i s_i.$$

The terms of  $\mathcal{L}$  that involve  $s_i$  are  $s_i(\frac{1}{M\nu} - \alpha_i - \eta_i)$ , so stationarity in  $s_i$  reads  $\frac{1}{M\nu} - \alpha_i - \eta_i = 0$ , which together with  $\eta_i \geq 0$  is the box constraint

$$\alpha_i \leq \frac{1}{M\nu};$$

and stationarity in  $\rho > 0$  reads  $1 - \sum_{i \in [M]} \alpha_i = 0$  (We exclude throughout the degenerate case in which the optimal ball collapses to a point,  $\rho^* = 0$ .), which is the normalization

$$\sum_{i \in [M]} \alpha_i = 1.$$

What survives in  $\mathcal{L}$  is  $\sum_i \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q})$ , and a classical treatment would now differentiate this in  $\mathbb{Q}$  and solve for it. That is impossible here:  $\mathcal{P}_p(\Xi)$  carries no linear structure, so there is no derivative to take, and no representer argument is available either to eliminate  $\mathbb{Q}$ . The Lagrangian route asks for neither, the infimum over  $\mathbb{Q}$  being carried along rather than resolved, and the dual function is

$$g(\boldsymbol{\alpha}) := \inf_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}), \quad \boldsymbol{\alpha} \in \mathcal{A}.$$

The infimum defining  $g$  is in fact attained, so it is a minimum:  $\Xi$  is compact, hence  $\mathcal{P}_p(\Xi)$  is a compact metric space, while  $\mathbb{Q} \mapsto \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q})$  is continuous, each  $\mathcal{W}_p(\mathbb{P}_i, \cdot)$  being 1-Lipschitz by the triangle inequality, and a continuous function on a compact metric space attains its minimum. Therefore the dual problem  $\max_{\boldsymbol{\alpha} \in \mathcal{A}} g(\boldsymbol{\alpha})$  of (MDSL-S) is exactly (MDSL-D).

We next establish strong duality. Split the  $\min_{\mathbb{Q}, \rho, \mathbf{s}}$  structure of the primal problem (MDSL-S) into an outer min over  $\mathbb{Q} \in \mathcal{P}_p(\Xi)$  and an inner min over  $(\rho, \mathbf{s})$ . The inner problem in  $(\rho, \mathbf{s})$  is a linear program, its objective being linear and its constraints affine, and its dual is

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\eta} \geq 0} \inf_{\rho > 0, \mathbf{s} \geq 0} \mathcal{L} = \max_{\boldsymbol{\alpha} \in \mathcal{A}} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}).$$

So we have

$$(\text{MDSL-S}) = \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}).$$

Sion's minimax theorem (Sion, 1958) permits the exchange

$$\min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \max_{\boldsymbol{\alpha} \in \mathcal{A}} = \max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)}$$

whenever one of the two sets is compact and convex and the other is convex, the objective being quasi-concave and upper semicontinuous in  $\alpha$  and quasi-convex and lower semicontinuous in  $\mathbb{Q}$ . Both sets are compact and convex here, so that the extrema on both sides are attained:  $\mathcal{P}_p(\Xi)$  is convex by Proposition 1 and compact, being a compact metric space because  $\Xi$  is;  $\mathcal{A}$  is a compact convex polytope; and  $\sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q})$  is convex in  $\mathbb{Q}$  by Proposition 1 and continuous there by the triangle inequality, which makes each  $\mathcal{W}_p(\mathbb{P}_i, \cdot)$  1-Lipschitz and hence  $\mathcal{W}_p^p(\mathbb{P}_i, \cdot)$  continuous, while it is linear in  $\alpha$ , hence at once quasi-convex, quasi-concave and continuous. The exchange is therefore valid and values of the primal and dual problems coincide via

$$(\text{MDSL-S}) = \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \max_{\alpha \in \mathcal{A}} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) = \max_{\alpha \in \mathcal{A}} \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) = (\text{MDSL-D}).$$

In other words, strong duality holds, and the theorem is proved.  $\square$

The dual (MDSL-D) is a two-level program in  $\alpha$  and  $\mathbb{Q}$ , and each of its two levels is of a kind that admits efficient solution on its own. For a fixed  $\alpha \in \mathcal{A}$  the inner problem asks for the weighted  $p$ -Wasserstein barycenter of the sources,

$$\text{Bar}_p(\mathcal{P}; \alpha) \in \arg \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}), \quad (3.1)$$

for which a mature literature provides fast algorithms: the entropic regularization of Cuturi and Doucet (2014), the iterative Bregman projections of Benamou et al. (2015), the stabilized and sparsified scaling algorithms of Schmitzer (2019), and the implementations collected in the POT library (Flamary et al., 2021) and in its Julia counterpart OptimalTransport.jl (Zhang et al., 2020), with Peyré and Cuturi (2019) surveying the field.

Writing

$$\Omega(\alpha) := \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \text{Bar}_p(\mathcal{P}; \alpha)) \quad (3.2)$$

for the resulting value, the dual reads  $\max_{\alpha \in \mathcal{A}} \Omega(\alpha)$ . The set  $\mathcal{A}$  is a compact convex polytope and  $\Omega$  is concave, being the infimum of functions affine in  $\alpha$ ; the outer problem is therefore a concave maximization, and its gradient is available in closed form. For each fixed  $\mathbb{Q}$  the map  $\alpha \mapsto \sum_{i \in [M]} \alpha_i \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q})$  is affine, and its minimizer is  $\text{Bar}_p(\mathcal{P}; \alpha)$ ; Danskin's envelope theorem (Danskin, 1967) therefore gives

$$\frac{\partial \Omega}{\partial \alpha_i}(\alpha) = \mathcal{W}_p^p(\mathbb{P}_i, \text{Bar}_p(\mathcal{P}; \alpha)). \quad (3.3)$$

The gradient is thus the vector of distances from the sources to the current barycenter, so evaluating it costs no more than the inner solve it follows. Two consequences are worth recording. Each partial derivative is nonnegative, so  $\Omega$  is nondecreasing in every weight and the maximization drives the weights upward until the box bounds or the normalization bind, in keeping with  $\sum_i \alpha_i^* = 1$  at an optimum. And  $\Omega$  is differentiable wherever the inner minimizer is unique, though not

necessarily elsewhere; where it is not, the same theorem returns the convex hull of the gradients at the minimizers, so the method below is a subgradient method in general.

Every subgradient is bounded, since  $0 \leq \partial_i \Omega = \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) \leq D^p$  for the diameter  $D$  of the sources, whence  $\|\nabla \Omega\|_2 \leq \sqrt{M} D^p$ ; but  $\Omega$  is an infimum of affine functions and is therefore non-smooth in general. We accordingly use diminishing steps  $\gamma_t = \gamma_0 t^{-1/2}$ , for which the subgradient method converges at the usual  $\mathcal{O}(t^{-1/2})$  rate, while on trajectories along which the inner minimizer remains unique  $\Omega$  is differentiable and a constant step of the order  $(\sqrt{M} D^p)^{-1}$  suffices. The projection onto  $\mathcal{A}$ , a box cut by a hyperplane, is available in closed form at the cost of a sort, and the constant  $\gamma_0$  is fixed by backtracking along the projected direction, which is reliable here because  $\Omega$  is concave.

Since the inner solve dominates the cost, the two levels are interleaved rather than nested: one barycenter solve, or at scale a single pass of the inner solver, is followed by one outer step, and the pair  $(\alpha^{(t)}, \mathbb{Q}^{(t)})$  is driven towards a saddle point of (MDSL-D). The inner update cannot be taken as a step in the direction of a gradient with respect to  $\mathbb{Q}$  and is delegated instead to a barycenter solver of the kind cited. Algorithm 1 records the resulting scheme. It is worth noting that, beyond the components the description itself requires, the algorithm also returns the optimal transport plans  $\{\pi^{(i)}\}_{i \in \text{SD}}$ ; Section 3.2 shows how this additional output is used in the subsequent distributionally robust optimization.

---

**Algorithm 1** The  $\mathcal{W}_p$ -meta-distribution support learning algorithm

---

**Input:** the source collection  $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_M\}$ ; exponent  $p$ ; level  $\nu$ ; tolerance  $\epsilon$

**Output:** center  $\mathbb{Q}^*$ , radius  $R^*$ , the index sets SD, BSD, ESD, ID, and optimal transport plans

```

     $\{\pi^{(i)}\}_{i \in \text{SD}}$ 
1:  $\alpha^{(0)} \leftarrow (1/M, \dots, 1/M)$ 
2: fix  $\gamma_0$  by backtracking along the projected direction at  $\alpha^{(0)}$ ;  $t \leftarrow 0$ 
3: repeat
4:    $\mathbb{Q}^{(t)} \leftarrow \text{Bar}_p(\mathcal{P}; \alpha^{(t)})$  ▷ inner: weighted barycenter, (3.1)
5:    $g_i \leftarrow \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^{(t)})$  for  $i \in [M]$  ▷ gradient (3.3)
6:    $\alpha^{(t+1)} \leftarrow \text{proj}_{\mathcal{A}}(\alpha^{(t)} + \gamma_t g)$  ▷ outer: weight update
7:    $\gamma_{t+1} \leftarrow \gamma_0(t+1)^{-1/2}$ 
8:    $t \leftarrow t+1$ 
9: until  $\|\alpha^{(t)} - \alpha^{(t-1)}\|_2 \leq \epsilon$ 
10:  $\mathbb{Q}^* \leftarrow \text{Bar}_p(\mathcal{P}; \alpha^{(t)})$ ; SD, BSD, ESD, ID read off from  $\alpha^{(t)}$  by (2.6), (2.7) and (2.8)
11:  $R^* \leftarrow \mathcal{W}_p(\mathbb{P}_i, \mathbb{Q}^*)$  for any  $i \in \text{BSD}$ 
12: for  $i \in \text{SD}$  do
13:    $\pi^{(i)} \leftarrow \arg \min_{\pi \in \Pi(\mathbb{Q}^*, \mathbb{P}_i)} \int_{\Xi \times \Xi} \|\zeta - \xi\|^p d\pi$  ▷ optimal plan for the returned center
14: end for
15: return  $\mathbb{Q}^*$ ,  $R^*$ , SD, BSD, ESD, ID, and  $\{\pi^{(i)}\}_{i \in \text{SD}}$ 

```

---

**Remark 2** (On the uniqueness of the center and radius). *Algorithm 1 reads the radius off a boundary support distribution, which presupposes that one exists. Complementary slackness pins the value  $R^{*p} = \rho^*$  only when some multiplier is strictly interior,  $0 < \alpha_i^* < 1/(M\nu)$ , that is, when  $\text{BSD} \neq \emptyset$ ; in its absence the radius is set-valued. Indeed, at a primal–dual optimal pair the sources*

satisfy  $\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^\star) \leq \rho^\star$  for  $i \in \text{ID}$  and  $\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^\star) \geq \rho^\star$  for  $i \in \text{ESD}$ , while  $\rho^\star = \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^\star)$  for every  $i \in \text{BSD}$ . The radius is thus confined to

$$\left[ \max_{i \in \text{ID}} \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^\star), \min_{i \in \text{ESD}} \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^\star) \right], \quad (3.4)$$

the lower endpoint being read as 0 when  $\text{ID} = \emptyset$ . When  $\text{BSD} \neq \emptyset$  the equalities on  $\text{BSD}$  pin the radius to a single value inside that interval, whereas when  $\text{BSD} = \emptyset$  the interval remains and any of its points is admissible. The classical SVDD ball obeys the same dichotomy (Wang et al., 2011; Chang et al., 2014). When  $\text{BSD} = \emptyset$  we therefore report the midpoint of the admissible interval, which balances robustness against generalization (Chang and Lin, 2011).

A second, independent source of ambiguity is the center itself. The inner problem is a relaxation of (MDSL-S): it enforces the first-order condition in  $\alpha$  but discards the pair  $(\rho, \mathbf{s})$ , and its solution set

$$\mathcal{Q}_\alpha^\star := \arg \min_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} \sum_{i \in [M]} \alpha_i^\star \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q})$$

is in general a set rather than a point—notably for  $p = 1$ , where the one-dimensional barycenter is a weighted median and hence an interval, whereas for  $p = 2$  it is unique (Agueh and Carlier, 2011). By strong duality every primal-optimal center belongs to  $\mathcal{Q}_\alpha^\star$ , but not conversely: the elements of  $\mathcal{Q}_\alpha^\star$  need not be centers that (MDSL-S) itself selects. Writing

$$\Phi(\mathbb{Q}) := \min_{\rho \geq 0} \left[ \rho + \frac{1}{M\nu} \sum_{i \in [M]} (\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \rho)^+ \right] \quad (3.5)$$

for the value of (MDSL-S) when the center is constrained to  $\mathbb{Q}$ , the two sets are related by

$$\{\mathbb{Q} : \exists \rho, \mathbf{s} \text{ with } (\mathbb{Q}, \rho, \mathbf{s}) \text{ optimal in (MDSL-S)}\} = \arg \min_{\mathbb{Q} \in \mathcal{Q}_\alpha^\star} \Phi(\mathbb{Q}) \subseteq \mathcal{Q}_\alpha^\star.$$

We therefore take the center to be a minimizer of (3.5) over  $\mathcal{Q}_\alpha^\star$ . If a unique approximate barycenter is desired, one may use the standard entropically regularized barycenter formulation on a fixed finite support with a positive regularization parameter (Cuturi and Doucet, 2014; Benamou et al., 2015).

The formulation above treats the source distributions  $\mathbb{P}_i$  as known. In practice they are not: each source is observed only through a finite sample  $\xi_{i1}, \dots, \xi_{in_i} \sim \mathbb{P}_i$ , and what is actually at hand is the empirical measure

$$\hat{\mathbb{P}}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{\xi_{ij}}, \quad i \in [M],$$

whose collection we write  $\hat{\mathcal{P}}$ . Replacing  $\mathbb{P}_i$  by  $\hat{\mathbb{P}}_i$  in (MDSL-S) gives the sample average approximation of the MDSL problem, and the substitution leaves the structure of the preceding development intact. The empirical transport cost  $\mathcal{W}_p^p(\hat{\mathbb{P}}_i, \mathbb{Q})$  is the value of the discrete transport problem between the sample and  $\mathbb{Q}$ , hence exactly the object the inner solvers cited for (3.1) evaluate, so the inner level of Algorithm 1 is solved by the same procedures, now on finite marginals; and because  $\Xi$

is compact, each  $\hat{\mathbb{P}}_i$  belongs to  $\mathcal{P}_p(\Xi)$ , so the dual (MDSL-D), the gradient (3.3) and the algorithm as a whole apply to  $\hat{\mathcal{P}}$  verbatim. We mark the quantities obtained in this way by a hat throughout.

### 3.2 Solving the $\mathcal{W}_p$ -MDSL ball-induced RooDSO problem

With the out-of-distribution uncertainty set learned through the proposed  $\mathcal{W}_p$ -MDSL method, problem (RooDSO-MDSL) becomes, in form, the distributionally robust program (1.2) with the ambiguity set specialized to the Wasserstein ball (2.9). The solution theory developed for Wasserstein distributionally robust optimization therefore applies under the regularity conditions stated below. The following lemma gives the dual reformulation established by Blanchet and Murthy (2019), Gao and Kleywegt (2023) and Mohajerin Esfahani and Kuhn (2018), which can be used to solve the present problem.

**Lemma 5** (Duality over a Wasserstein ball (Blanchet and Murthy, 2019; Gao and Kleywegt, 2023; Mohajerin Esfahani and Kuhn, 2018)). *Let  $\Xi$  be compact and  $p \geq 1$ , and let  $\ell : \Xi \rightarrow \mathbb{R}$  be continuous. Then for every  $\mathbb{Q} \in \mathcal{P}_p(\Xi)$  and every  $R \geq 0$ ,*

$$\sup_{\mathbb{P} \in \mathcal{B}_{\mathcal{W}_p}(\mathbb{Q}, R)} \mathbb{E}_{\mathbb{P}}[\ell(\boldsymbol{\xi})] = \inf_{\lambda \geq 0} \left\{ \lambda R^p + \int_{\Xi} \sup_{\boldsymbol{\xi} \in \Xi} [\ell(\boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] d\mathbb{Q}(\boldsymbol{\zeta}) \right\}, \quad (3.6)$$

For downstream computation, we use a finitely supported center selected in the learning step of Section 3.1, represented as

$$\hat{\mathbb{Q}}^* = \sum_{k=1}^K \beta_k \delta_{\boldsymbol{\zeta}_k}, \quad \boldsymbol{\zeta}_k \in \Xi, \quad \beta_k > 0, \quad \sum_{k=1}^K \beta_k = 1. \quad (3.7)$$

The following result is conditional on this finite-support representation; it does not require uniqueness of the learned center. Substituting (3.7) into Lemma 5 reduces the integral to a finite weighted sum of maximization subproblems.

**Proposition 6** (The robustified program under a discrete center). *Let Assumption 1 hold, let  $\Xi$  be compact, and let  $p \geq 1$ . Let  $\hat{\mathbb{Q}}^* = \sum_{k=1}^K \beta_k \delta_{\boldsymbol{\zeta}_k}$  be a center represented as in (3.7), and let  $\hat{R}^* \geq 0$  be a prescribed radius. Then for every  $\mathbf{x} \in \mathcal{X}$ ,*

$$\sup_{\mathbb{P} \in \mathcal{B}_{\mathcal{W}_p}(\hat{\mathbb{Q}}^*, \hat{R}^*)} \mathbb{E}_{\mathbb{P}}[f(\mathbf{x}, \boldsymbol{\xi})] = \inf_{\lambda \geq 0} \left\{ \lambda \hat{R}^{*p} + \sum_{k=1}^K \beta_k \sup_{\boldsymbol{\xi} \in \Xi} [f(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \boldsymbol{\zeta}_k\|^p] \right\}, \quad (3.8)$$

and the robustified program for this ball accordingly takes the form

$$\min_{\mathbf{x} \in \mathcal{X}} \inf_{\lambda \geq 0} \left\{ \lambda \hat{R}^{*p} + \sum_{k=1}^K \beta_k \sup_{\boldsymbol{\xi} \in \Xi} [f(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \boldsymbol{\zeta}_k\|^p] \right\}, \quad (3.9)$$

in which the two infima may be taken in either order without changing the value. The objective is jointly convex in  $(\mathbf{x}, \lambda)$ , and the minimum over  $\mathbf{x}$  is attained. The proof is given in Appendix A.3.

Two features of (3.9) govern how it is solved. The objective is jointly convex in  $(\mathbf{x}, \lambda)$ : for each fixed  $\boldsymbol{\xi}$ , the function  $f(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \boldsymbol{\zeta}_k\|^p$  is jointly convex, and taking its supremum over  $\Xi$  preserves convexity. The  $K$  terms indexed by  $k$  are coupled only through the shared pair  $(\mathbf{x}, \lambda)$ : once that pair is fixed, the  $k$ -th inner supremum depends on  $\boldsymbol{\zeta}_k$  alone. The  $K$  inner problems are therefore independent and may be dispatched in parallel. This can reduce elapsed time when sufficient processors are available, although the total work still depends on  $K$ . This is the standard structure of the dual of a Wasserstein ball, the sample points of the classical empirical center being replaced here by the atoms of  $\hat{\mathbb{Q}}^*$ . Compactness and continuity ensure that each inner maximum is attained, and for the costs met in practice its supremum is available in closed form, the parallel branches then amounting to evaluating a small number of candidate maximizers rather than searching  $\Xi$ .

Although solving a Wasserstein robust program is not a new problem, the methodology proposed in this paper yields new insight into how it is solved. MDSL supplies heuristic information that can assist both initialization and the search for adversarial scenarios. Besides the learned center and radius, we retain the weights  $\hat{\boldsymbol{\alpha}}^*$ , the empirical SD, BSD and ESD index sets, and transport plans between the final center and the support distributions. Writing  $\hat{\mathbb{P}}_i = n_i^{-1} \sum_{j=1}^{n_i} \delta_{\boldsymbol{\xi}_{ij}}$ , these plans satisfy

$$\pi^{(i)} \in \Pi(\hat{\mathbb{Q}}^*, \hat{\mathbb{P}}_i), \quad \sum_j \pi_{kj}^{(i)} = \beta_k, \quad \sum_k \pi_{kj}^{(i)} = \frac{1}{n_i}, \quad i \in \widehat{\text{SD}}. \quad (3.10)$$

The plans are retained from the final transport computation, or computed for the selected final center if necessary.

For each center atom, the nonzero transport branches provide an initial candidate set

$$\mathcal{C}_k^{(0)} := \{\boldsymbol{\zeta}_k\} \cup \{\boldsymbol{\xi}_{ij} : i \in \widehat{\text{SD}}, \pi_{kj}^{(i)} > 0\}. \quad (3.11)$$

If this set is large, it may be thinned while retaining  $\boldsymbol{\zeta}_k$ , prioritizing BSD branches and larger weights  $\hat{\alpha}_i^* \pi_{kj}^{(i)} / \beta_k$ , with ESD endpoints providing additional exploratory seeds. Define  $\psi_k(\boldsymbol{\xi}; \mathbf{x}, \lambda) := f(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \boldsymbol{\zeta}_k\|^p$ . Transport displacements  $\mathbf{d}_{ikj} := \boldsymbol{\xi}_{ij} - \boldsymbol{\zeta}_k$  encode observed variation across distributions, whereas  $\nabla_{\boldsymbol{\xi}} f(\mathbf{x}, \boldsymbol{\zeta}_k)$ , where it exists, identifies local directions of increasing loss for the current decision. Their first-order scores  $\nabla_{\boldsymbol{\xi}} f(\mathbf{x}, \boldsymbol{\zeta}_k)^\top \mathbf{d}_{ikj} - \lambda \|\mathbf{d}_{ikj}\|^p$  can help prioritize transport branches that are both supported by the data and locally adverse after accounting for transport cost, providing loss-aware search seeds. We jointly initialize  $(\mathbf{x}^{(0)}, \lambda^{(0)})$  by solving

$$\begin{aligned} \min_{\substack{\mathbf{x} \in \mathcal{X}, \lambda \geq 0 \\ \mathbf{u} \in \mathbb{R}^K}} \quad & \lambda \hat{R}^{*p} + \sum_{k=1}^K \beta_k u_k \\ \text{s.t.} \quad & u_k \geq \psi_k(\boldsymbol{\xi}; \mathbf{x}, \lambda), \quad \boldsymbol{\xi} \in \mathcal{C}_k^{(0)}, \quad k \in [K]. \end{aligned} \quad (3.12)$$

This finite convex problem uses the learned radius  $\hat{R}^*$ . Restricting the inner maximizations to candidate sets makes its optimal value a lower bound on that of (3.9).

During optimization, the best cached candidate seeds each inner search, and newly found maximizers are added to the cache. Algorithm 2 retains the projected subgradient updates, assuming available convex subgradients in  $\mathbf{x}$ . The searches still range over all of  $\Xi$ ; exact dual evaluation requires global inner maximizers. The stopping rule is heuristic, and any computational benefit of this reuse remains to be assessed numerically.

---

**Algorithm 2** An illustrative framework for RooDSO using MDSL-informed heuristics

---

**Input:** MDSL center  $\hat{\mathbb{Q}}^*$ , radius  $\hat{R}^* > 0$ , weights  $\hat{\alpha}^*$ , empirical SD/BSD/ESD sets, support samples and plans  $\{\pi^{(i)}\}$ ; cost  $f$ ; set  $\mathcal{X}$ ; positive steps  $\{\gamma_t\}$ ; tolerance  $\epsilon$ ; budget  $T \geq 1$

**Output:** candidate decision  $\mathbf{x}_{\text{best}}$

```

1: Build  $\{\mathcal{C}_k^{(0)}\}_{k \in [K]}$  by (3.11), optionally thinning as described above
2: Solve (3.12) for  $(\mathbf{x}^{(0)}, \lambda^{(0)})$ ;  $G_{\text{best}} \leftarrow +\infty$ 
3: for  $t = 0, \dots, T - 1$  do
4:   for  $k = 1, \dots, K$  in parallel do
5:      $\xi_{k,\text{seed}}^{(t)} \in \arg \max_{\xi \in \mathcal{C}_k^{(t)}} \psi_k(\xi; \mathbf{x}^{(t)}, \lambda^{(t)})$ 
6:     Warm-start the maximization of  $\psi_k(\cdot; \mathbf{x}^{(t)}, \lambda^{(t)})$  over  $\Xi$  at  $\xi_{k,\text{seed}}^{(t)}$ , obtaining a global
       maximizer  $\xi_k^{(t)}$ 
7:      $\mathcal{C}_k^{(t+1)} \leftarrow \mathcal{C}_k^{(t)} \cup \{\xi_k^{(t)}\}$ 
8:     Choose  $\mathbf{v}_k^{(t)} \in \partial_{\mathbf{x}} f(\mathbf{x}^{(t)}, \xi_k^{(t)})$ 
9:   end for
10:   $G_t \leftarrow \lambda^{(t)} \hat{R}^{*p} + \sum_{k \in [K]} \beta_k \psi_k(\xi_k^{(t)}; \mathbf{x}^{(t)}, \lambda^{(t)})$ 
11:  if  $G_t < G_{\text{best}}$  then
12:     $G_{\text{best}} \leftarrow G_t$ ;  $\mathbf{x}_{\text{best}} \leftarrow \mathbf{x}^{(t)}$ 
13:  end if
14:  if  $t \geq 1$  and  $|G_t - G_{t-1}| \leq \epsilon$  then
15:    break
16:  end if
17:   $\mathbf{d}_x \leftarrow \sum_{k \in [K]} \beta_k \mathbf{v}_k^{(t)}$ ;  $d_\lambda \leftarrow \hat{R}^{*p} - \sum_{k \in [K]} \beta_k \|\xi_k^{(t)} - \zeta_k\|^p$ 
18:   $(\mathbf{x}^{(t+1)}, \lambda^{(t+1)}) \leftarrow \text{proj}_{\mathcal{X} \times \mathbb{R}_+}((\mathbf{x}^{(t)}, \lambda^{(t)}) - \gamma_t(\mathbf{d}_x, d_\lambda))$ 
19: end for
20: return  $\mathbf{x}_{\text{best}}$ 

```

---

## 4 Analysis of the proposed methodology

### 4.1 Finite-sample stability of the learned uncertainty set

We now quantify how replacing each source distribution by its empirical counterpart affects the learned center and radius. We condition first on the source distributions  $\mathbb{P}_1, \dots, \mathbb{P}_M$ ; the probability below is over the within-source samples  $\{\xi_{ij}\}_{j=1}^{n_i}$ . Throughout this subsection, suppose that  $\Xi$  is compact and write  $D := \text{diam}(\Xi)$  and  $L_p := pD^{p-1}$ , with  $L_1 := 1$ . For each source, set

$$\varepsilon_i := \mathcal{W}_p(\hat{\mathbb{P}}_i, \mathbb{P}_i), \quad i \in [M].$$

Here  $\mathbb{P}_i$  is the population distribution and hence is not decorated with a hat;  $\widehat{\mathbb{P}}_i$  is its empirical estimate.

**Lemma 7** (Perturbation of the transport costs). *For every  $i \in [M]$  and  $\mathbb{Q}, \mathbb{Q}' \in \mathcal{P}_p(\Xi)$ ,*

$$\left| \mathcal{W}_p^p(\widehat{\mathbb{P}}_i, \mathbb{Q}) - \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) \right| \leq L_p \varepsilon_i, \quad (4.1)$$

$$\left| \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}') \right| \leq L_p \mathcal{W}_p(\mathbb{Q}, \mathbb{Q}'). \quad (4.2)$$

*Proof.* The triangle inequality gives  $|\mathcal{W}_p(\widehat{\mathbb{P}}_i, \mathbb{Q}) - \mathcal{W}_p(\mathbb{P}_i, \mathbb{Q})| \leq \varepsilon_i$ , and the analogous bound for changing  $\mathbb{Q}$  to  $\mathbb{Q}'$ . All these distances are at most  $D$ . The mean value theorem applied to  $t \mapsto t^p$  on  $[0, D]$  then gives the two claims.  $\square$

Define the population and empirical profile objectives by

$$\Phi(\mathbb{Q}) := \min_{\rho \geq 0} \left\{ \rho + \frac{1}{M\nu} \sum_{i=1}^M (\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \rho)_+ \right\}, \quad (4.3)$$

$$\widehat{\Phi}(\mathbb{Q}) := \min_{\rho \geq 0} \left\{ \rho + \frac{1}{M\nu} \sum_{i=1}^M (\mathcal{W}_p^p(\widehat{\mathbb{P}}_i, \mathbb{Q}) - \rho)_+ \right\}. \quad (4.4)$$

These are the objectives in (MDSL-S) after minimizing out the slacks and writing  $\rho = R^p$ . Let  $\mathbb{Q}^* := \arg \min_{\mathbb{Q}} \Phi(\mathbb{Q})$  and  $\widehat{\mathbb{Q}}^* \in \arg \min_{\mathbb{Q}} \widehat{\Phi}(\mathbb{Q})$ .

**Lemma 8** (Uniform objective perturbation). *For every realization of the samples, with  $\eta := \frac{L_p}{M\nu} \sum_{i=1}^M \varepsilon_i$ ,*

$$\sup_{\mathbb{Q} \in \mathcal{P}_p(\Xi)} |\widehat{\Phi}(\mathbb{Q}) - \Phi(\mathbb{Q})| \leq \eta. \quad (4.5)$$

*Consequently, every empirical minimizer satisfies*

$$\Phi(\widehat{\mathbb{Q}}^*) - \min_{\mathbb{Q}} \Phi(\mathbb{Q}) \leq 2\eta. \quad (4.6)$$

*Proof.* For fixed  $(\mathbb{Q}, \rho)$ , the positive-part map is 1-Lipschitz, so Lemma 7 bounds the difference between the two profile objectives by  $\eta$ , uniformly in  $(\mathbb{Q}, \rho)$ . Taking the infimum over  $\rho$  preserves this bound and proves (4.5). Comparing the empirical minimizer with a population minimizer and applying the uniform bound twice gives (4.6).  $\square$

Uniform convergence of the objectives controls their optimal values, but by itself does not control the distance between their minimizers. We therefore state explicitly the growth condition used to obtain a center error bound.

**Assumption 2** (Quadratic growth of the population objective). *There exists  $\kappa > 0$  such that, for every  $\mathbb{Q} \in \mathcal{P}_p(\Xi)$ ,*

$$\Phi(\mathbb{Q}) - \Phi^* \geq \kappa \operatorname{dist}_{\mathcal{W}_p}(\mathbb{Q}, \mathbb{Q}^*)^2, \quad \Phi^* := \min_{\mathbb{Q}'} \Phi(\mathbb{Q}'),$$

where  $\text{dist}_{\mathcal{W}_p}(\mathbb{Q}, \mathbb{Q}^*) := \inf_{\mathbb{Q}' \in \mathcal{Q}^*} \mathcal{W}_p(\mathbb{Q}, \mathbb{Q}')$ .

For a fixed center  $\mathbb{Q}$ , the optimal radius power need not be unique under the soft formulation. We therefore write its admissible set as

$$\mathcal{I}_\nu(\mathbb{Q}) := \arg \min_{\rho \geq 0} \left\{ \rho + \frac{1}{M\nu} \sum_{i=1}^M (\mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}) - \rho)_+ \right\}, \quad (4.7)$$

and denote its empirical counterpart by  $\widehat{\mathcal{I}}_\nu(\mathbb{Q})$ , formed by replacing every  $\mathbb{P}_i$  with  $\widehat{\mathbb{P}}_i$ . Both sets are closed intervals, possibly singletons. Their endpoints are order statistics of the corresponding vector of transport costs, with the ranks determined by  $M\nu$ .

**Lemma 9** (Stability of the optimal radius interval). *Suppose two vectors of costs  $\mathbf{d}, \tilde{\mathbf{d}} \in \mathbb{R}_+^M$  satisfy  $\max_i |d_i - \tilde{d}_i| \leq \tau$ . The sets of minimizers over  $\rho \geq 0$  of  $\rho + (M\nu)^{-1} \sum_i (d_i - \rho)_+$  and  $\rho + (M\nu)^{-1} \sum_i (\tilde{d}_i - \rho)_+$  then have Hausdorff distance at most  $\tau$ .*

The preceding lemmas give the following finite-sample result.

**Theorem 10** (Finite-sample error bounds for the center and radius). *Suppose  $\Xi$  is compact and Assumption 2 holds. For each  $i \in [M]$ , suppose there is a bound  $b_i(n_i, \delta_i)$  such that, conditional on  $\mathbb{P}_i$ ,*

$$\Pr \left\{ \mathcal{W}_p(\widehat{\mathbb{P}}_i, \mathbb{P}_i) > b_i(n_i, \delta_i) \right\} \leq \delta_i.$$

For  $\delta \in (0, 1)$ , set

$$b_i := b_i(n_i, \delta/M), \quad \eta_\delta := \frac{L_p}{M\nu} \sum_{i=1}^M b_i, \quad \tau_\delta := L_p \left( \max_{i \in [M]} b_i + \sqrt{\frac{2\eta_\delta}{\kappa}} \right).$$

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

$$\text{dist}_{\mathcal{W}_p}(\widehat{\mathbb{Q}}^*, \mathbb{Q}^*) \leq \sqrt{\frac{2\eta_\delta}{\kappa}}, \quad (4.8)$$

$$d_H(\widehat{\mathcal{I}}_\nu(\widehat{\mathbb{Q}}^*), \mathcal{I}_\nu(\mathbb{Q}^*)) \leq \tau_\delta, \quad (4.9)$$

for a population minimizer  $\mathbb{Q}^* \in \mathcal{Q}^*$  nearest to  $\widehat{\mathbb{Q}}^*$ . In particular, if the reported radius power is the midpoint of the corresponding optimal interval, then

$$|\widehat{R}^{*p} - R^{*p}| \leq \tau_\delta, \quad |\widehat{R}^* - R^*| \leq \tau_\delta^{1/p}.$$

The theorem separates the statistical input from the optimization argument: any valid concentration bound for the empirical Wasserstein distances can be substituted for  $b_i$ . Its rate therefore depends on the dimension, the order  $p$ , and the distributional assumptions used to control empirical Wasserstein distance; a root-sample-size rate is not asserted without such additional conditions. The radius conclusion controls the full admissible interval and so remains meaningful when the soft problem does not identify a unique radius.

## 4.2 Out-of-distribution generalization guarantees

We next study coverage of an independent target distribution  $\mathbb{P}' \sim \Psi$ . In contrast to the conditional analysis of Section 4.1, the sources  $\mathbb{P}_1, \dots, \mathbb{P}_M$  are now i.i.d. draws from  $\Psi$ . Conditional on these sources, the datasets are independent, and dataset  $i$  contains  $n_i$  i.i.d. observations from  $\mathbb{P}_i$ . All probabilities over the target below condition on the observed training data; the confidence statements quantify the randomness in those training data. Throughout this subsection,  $\Xi$  is compact,  $p \geq 1$ , and  $\nu \in (0, 1]$  is fixed. We retain  $D$  and  $L_p$  from the preceding subsection and consider  $D > 0$ ; for a singleton observation space coverage is identically one.

Writing  $\rho = R^p$ , define the score and its truncated loss by

$$s_{\mathbb{Q}, \rho}(\mathbb{P}) := \mathcal{W}_p^p(\mathbb{P}, \mathbb{Q}) - \rho, \quad \ell_\gamma(t) := \min\{1, t_+/\gamma\}, \quad \gamma > 0. \quad (4.10)$$

An optimal radius power in the soft MDL problem lies in  $[0, D^p]$ : a value above the largest source cost leaves all slacks zero and can be decreased. Let  $(\hat{\mathbb{Q}}^*, \hat{\rho}^*)$ , with  $\hat{\rho}^* = (\hat{R}^*)^p$ , be an exact empirical optimum. If an optimum is nonunique, we choose an optimal pair measurably from the training data. No uniqueness or quadratic growth assumption is needed for the results below. As in the margin construction of Li et al. (2026), we consider an enlargement in radius power:

$$\hat{\mathcal{U}}_{\nu, \gamma} := \left\{ \mathbb{P} \in \mathcal{P}_p(\Xi) : s_{\hat{\mathbb{Q}}^*, \hat{\rho}^*}(\mathbb{P}) \leq \gamma \right\}. \quad (4.11)$$

Thus its metric radius is  $(\hat{\rho}^* + \gamma)^{1/p}$ . The corresponding construction from the known sources  $\mathbb{P}_i$ , with the same measurable-selection convention, is denoted by  $\mathcal{U}_{\nu, \gamma}$ . The elementary inequalities

$$\mathbf{1}\{t > \gamma\} \leq \ell_\gamma(t) \leq \mathbf{1}\{t > 0\} \quad (4.12)$$

connect target noncoverage to training exclusion.

**Lemma 11** (Training exclusion bound). *For any source collection  $\mathbb{S}_1, \dots, \mathbb{S}_M$ , every optimal pair  $(\mathbb{Q}, \rho)$  of the soft MDL problem satisfies*

$$\frac{1}{M} \sum_{i=1}^M \mathbf{1}\{\mathcal{W}_p^p(\mathbb{S}_i, \mathbb{Q}) > \rho\} \leq \nu. \quad (4.13)$$

*This applies to both the known-source and empirical-source problems, including an optimal radius of zero.*

*Proof.* Fix an optimal center and put  $d_i := \mathcal{W}_p^p(\mathbb{S}_i, \mathbb{Q})$ . After eliminating the slacks, the radius power minimizes  $F(\rho) := \rho + (M\nu)^{-1} \sum_i (d_i - \rho)_+$  over  $\rho \geq 0$ . Its right derivative is  $F'_+(\rho) = 1 - \#\{i : d_i > \rho\}/(M\nu)$ . At a minimizer this derivative is nonnegative, since increasing  $\rho$  is feasible, even when  $\rho = 0$ . This gives (4.13).  $\square$

To control generalization uniformly over the learned center and radius, let

$$\begin{aligned}\mathcal{T} &:= \mathcal{P}_p(\Xi) \times [0, D^p], \\ d_{\mathcal{T}}((\mathbb{Q}, \rho), (\mathbb{Q}', \rho')) &:= L_p \mathcal{W}_p(\mathbb{Q}, \mathbb{Q}') + |\rho - \rho'|.\end{aligned}\tag{4.14}$$

Let  $N_{\mathcal{T}}(a)$  be the minimum number of closed  $d_{\mathcal{T}}$ -balls of radius  $a > 0$ , with centers in  $\mathcal{T}$ , needed to cover  $\mathcal{T}$ . Compactness of  $\Xi$  makes  $\mathcal{T}$  compact, so  $N_{\mathcal{T}}(a) < \infty$  for every  $a > 0$ . Define

$$\mathcal{C}_M(a, \gamma, \delta_M) := \frac{2a}{\gamma} + \sqrt{\frac{\log N_{\mathcal{T}}(a) + \log(1/\delta_M)}{2M}}.\tag{4.15}$$

Here  $a$  is an analytical covering scale. Both  $a$  and the enlargement  $\gamma$  in the statements below are fixed independently of the training data; uniformity over the fitted center and radius does not imply uniformity over these tuning parameters.

**Lemma 12** (Uniform margin generalization). *For fixed  $a, \gamma > 0$  and  $\delta_M \in (0, 1)$ , with probability at least  $1 - \delta_M$  over the i.i.d. source distributions, simultaneously for every  $(\mathbb{Q}, \rho) \in \mathcal{T}$ ,*

$$\mathbb{E}_{\mathbb{P}' \sim \Psi}[\ell_{\gamma}(s_{\mathbb{Q}, \rho}(\mathbb{P}'))] \leq \frac{1}{M} \sum_{i=1}^M \ell_{\gamma}(s_{\mathbb{Q}, \rho}(\mathbb{P}_i)) + \mathcal{C}_M(a, \gamma, \delta_M).\tag{4.16}$$

The proof, given in Appendix A.5, uses a finite cover of  $\mathcal{T}$ , Hoeffding's inequality at each cover point, and the Lipschitz property of the loss to extend the bound to every parameter pair. Consequently it also applies to a pair fitted using the within-source data. This uniformity permits a direct comparison of true and empirical source scores, without first estimating the displacement of the fitted center.

**Theorem 13** (Out-of-distribution coverage under two-layer sampling). *For each  $i$ , suppose deterministic bounds  $b_i(n_i, t) \geq 0$  satisfy, for every  $t \in (0, 1)$  and almost every source collection,*

$$\Pr \left\{ \mathcal{W}_p(\widehat{\mathbb{P}}_i, \mathbb{P}_i) > b_i(n_i, t) \mid \mathbb{P}_1, \dots, \mathbb{P}_M \right\} \leq t.\tag{4.17}$$

Fix  $a, \gamma > 0$  and  $\delta_M, \delta_n \in (0, 1)$  with  $\delta_M + \delta_n < 1$ . Set

$$\begin{aligned}b_i &:= b_i(n_i, \delta_n/M), \\ \mathcal{E}_{M, \mathbf{n}} &:= \mathcal{C}_M(a, \gamma, \delta_M) + \frac{L_p}{M\gamma} \sum_{i=1}^M b_i, \quad \mathbf{n} := (n_1, \dots, n_M).\end{aligned}\tag{4.18}$$

Then, with probability at least  $1 - \delta_M - \delta_n$  over both sampling layers, the exact empirical MDSL solution satisfies

$$\Pr_{\mathbb{P}' \sim \Psi} \{ \mathbb{P}' \notin \widehat{\mathcal{U}}_{\nu, \gamma} \} \leq \min\{1, \nu + \mathcal{E}_{M, \mathbf{n}}\}.\tag{4.19}$$

If the sources are known exactly, then with probability at least  $1 - \delta_M$  over their draws,

$$\Pr_{\mathbb{P}' \sim \Psi} \{\mathbb{P}' \notin \mathcal{U}_{\nu, \gamma}\} \leq \min\{1, \nu + \mathcal{C}_M(a, \gamma, \delta_M)\}. \quad (4.20)$$

The proof is in Appendix A.5. The two contributions in  $\mathcal{E}_{M, \mathbf{n}}$  have distinct origins. The covering term accounts for observing only  $M$  sources from  $\Psi$ ; the second term accounts for replacing those sources by empirical measures. The latter is governed by the average source estimation error and uses only Lemma 7, so Assumption 2 is not required. The dependence of  $N_{\mathcal{T}}(a)$  on  $a$ , and of  $b_i$  on  $n_i$ , determines the available rates. Compactness alone does not specify a root- $M$  rate for a vanishing covering scale.

Coverage immediately transfers to the downstream robust objective. For the enlarged ball define

$$\hat{J}_{\gamma}(\mathbf{x}) := \sup_{\mathbb{P} \in \hat{\mathcal{U}}_{\nu, \gamma}} \mathbb{E}_{\mathbb{P}}[f(\mathbf{x}, \boldsymbol{\xi})]. \quad (4.21)$$

This uses metric radius  $(\hat{\rho}^* + \gamma)^{1/p}$  in the robustified problem.

**Corollary 14** (Out-of-distribution performance of robust decisions). *Under Theorem 13 and Assumption 1, on the same training event of probability at least  $1 - \delta_M - \delta_n$ ,*

$$\Pr_{\mathbb{P}' \sim \Psi} \left\{ \mathbb{E}_{\mathbb{P}'}[f(\mathbf{x}, \boldsymbol{\xi})] \leq \hat{J}_{\gamma}(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X} \right\} \geq (1 - \nu - \mathcal{E}_{M, \mathbf{n}})_+. \quad (4.22)$$

*In particular, the same lower bound holds for the event that the target expected cost of a decision  $\hat{\mathbf{x}}_{\gamma} \in \arg \min_{\mathbf{x} \in \mathcal{X}} \hat{J}_{\gamma}(\mathbf{x})$  does not exceed  $\hat{J}_{\gamma}(\hat{\mathbf{x}}_{\gamma})$ .*

*Proof.* Whenever  $\mathbb{P}' \in \hat{\mathcal{U}}_{\nu, \gamma}$ , its expected cost is bounded by the supremum in (4.21), simultaneously for every  $\mathbf{x}$ . Thus coverage implies the event in (4.22). Apply Theorem 13. Boundedness and continuity of the costs follow from Assumption 1 and compactness; these also make  $\hat{J}_{\gamma}$  continuous on compact  $\mathcal{X}$ , so a minimizing decision exists.  $\square$

**Corollary 15** (Asymptotic coverage with vanishing enlargement). *Under the sampling and concentration conditions of Theorem 13, keep  $\Xi, p, \nu$  fixed. For each  $M$ , use sample sizes  $n_{i, M}$ , deterministic choices  $a_M, \gamma_M > 0$ , and confidence parameters  $\delta_M, \delta_{n, M}$ . Put  $b_{i, M} := b_i(n_{i, M}, \delta_{n, M}/M)$ . Suppose*

$$\begin{aligned} \gamma_M &\longrightarrow 0, & \delta_M + \delta_{n, M} &\longrightarrow 0, & \frac{a_M}{\gamma_M} &\longrightarrow 0, \\ \frac{\log N_{\mathcal{T}}(a_M) + \log(1/\delta_M)}{M} &\longrightarrow 0, & \frac{1}{M\gamma_M} \sum_{i=1}^M b_{i, M} &\longrightarrow 0. \end{aligned} \quad (4.23)$$

*Then, for every  $\epsilon > 0$ ,*

$$\Pr_{\text{train}} \left\{ \Psi(\hat{\mathcal{U}}_{\nu, \gamma_M}) < 1 - \nu - \epsilon \right\} \longrightarrow 0. \quad (4.24)$$

*The known-source version holds after omitting the within-source error and its confidence parameter.*

Its proof appears in Appendix A.5. This conclusion concerns the enlarged sets with vanishing

enlargement. Since their centers and radii also change with  $M$ , these sets need not be nested. An extension to the unexpanded sets would additionally require control of the boundary layer, for example

$$\Psi\left\{\mathbb{P} : 0 < s_{\hat{\mathbb{Q}}^*, \hat{\rho}^*}(\mathbb{P}) \leq \gamma_M\right\} \longrightarrow 0 \quad \text{in probability over the training data.}$$

This condition is not assumed in Corollary 15.

## 5 Concluding remarks

This paper has developed MDSL for robust stochastic optimization when no observations from the target distribution are available. The model learns a Wasserstein ball directly from related source distributions, jointly determining its center and radius while controlling source exclusion through  $\nu$ . The kernel-embedding counterexample demonstrates why the underlying geometry matters: an enclosing RKHS ball can admit arbitrarily dispersed distributions and produce an infinite robust cost, whereas the corresponding Wasserstein ball imposes a transport-cost budget and gives a finite value in that example.

The learning and decision stages are connected computationally as well as through the uncertainty set. MDSL’s dual representation relates support learning to weighted Wasserstein barycenters and provides source weights, support classifications and optimal transport plans. For a finitely supported center, the robust decision problem has a jointly convex dual objective with independent atom-specific maximizations. Transport branches from boundary and exterior support distributions supply candidate scenarios for a finite convex initialization of the decision and penalty variables. Transport directions describe observed distributional variation, while available loss gradients indicate locally adverse directions for the current decision. Their combined use, together with an evolving cache of adverse scenarios, supplies search heuristics within the original robust formulation at the learned radius. The inner searches remain over the full observation space.

The statistical results distinguish within-source estimation from generalization across distributions. Under compactness, empirical source errors yield uniform objective perturbation bounds; quadratic growth further controls the center and admissible radius interval. Separately, for exact empirical MDSL optima, margin-enlarged balls admit coverage and cost guarantees whose errors account for both sampling layers. These coverage results require neither uniqueness nor quadratic growth. Under the stated conditions, target coverage is asymptotically at least  $1 - \nu$ , in probability over training data, as the enlargement vanishes.

Further work should assess the computational benefit of the proposed initialization and transport reuse, including their additional cost and their dependence on the loss structure. Extending the statistical analysis to approximate learning solutions and developing reliable stopping certificates for decision optimization would strengthen the connection between theory and implementation.

## References

- Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- Pranjal Awasthi, Christopher Jung, and Jamie Morgenstern. Distributionally robust data join. *arXiv:2202.05797*, 2022.
- Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenare, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.
- Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.
- Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. In *Advances in Neural Information Processing Systems*, volume 35, pages 35866–35879, 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/e90b00adc3ba130eb2510d93ba3ff250-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/e90b00adc3ba130eb2510d93ba3ff250-Paper-Conference.pdf).
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. URL <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- Wei-Cheng Chang, Ching-Pei Lee, and Chih-Jen Lin. A revisit to support vector data description. Technical report, National Taiwan University, 2014. URL <https://www.csie.ntu.edu.tw/~cjlin/papers/svdd.pdf>.
- Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693. PMLR, 2014.
- John M. Danskin. *The Theory of Max-Min and Its Application to Weapons Allocation Problems*, volume 5 of *Econometrics and Operations Research*. Springer-Verlag, Berlin, 1967.
- Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.

- İpek Dursun, Alp Akçay, and Geert-Jan Van Houtum. Data pooling for multiple single-component systems under population heterogeneity. *International Journal of Production Economics*, 250: 108665, 2022.
- Rémi Flamary, Nicolas Courty, Alexandre Gramfort, et al. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.
- Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.
- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.
- Vishal Gupta and Nathan Kallus. Data pooling in stochastic optimization. *Management Science*, 68(3):1595–1615, 2022.
- Biao Han, Chao Shang, and Dexian Huang. Multiple kernel learning-aided robust optimization: Learning algorithm, computational tractability, and usage in multi-stage decision-making. *European Journal of Operational Research*, 292(3):1004–1018, 2021. doi: 10.1016/j.ejor.2020.11.005.
- Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938, 2018.
- Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.
- Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 728–740, 2020.
- Tim Tsz-Kit Lau and Han Liu. Wasserstein distributionally robust optimization with Wasserstein barycenters. *arXiv:2203.12136*, 2022.
- Xianyu Li, Huan Xu, Xiaolin Huang, and Chao Shang. Robust out-of-distribution stochastic optimization. *arXiv preprint arXiv:2604.20147*, April 2026.
- Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.
- Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625, 2019.

- Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16785–16793, 2021.
- Krikamol Muandet and Bernhard Schölkopf. One-class support measure machines for group anomaly detection. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 449–458, 2013.
- Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.
- Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
- Yves Rychener, Adrián Esteban-Pérez, Juan M Morales, and Daniel Kuhn. Wasserstein distributionally robust optimization with heterogeneous data sources. *arXiv:2407.13582*, 2024.
- Shiori Sagawa\*, Pang Wei Koh\*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=ryxGuJrFvS>.
- Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.
- Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.
- Aras Selvi, Eleonora Kreacic, Mohsen Ghassemi, Vamsi K. Potluru, Tucker Balch, and Manuela Veloso. Distributionally and adversarially robust logistic regression via intersecting Wasserstein balls. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286, pages 3641–3674, 2025. URL <https://proceedings.mlr.press/v286/selvi25a.html>.
- Chao Shang, Xiaolin Huang, and Fengqi You. Data-driven robust optimization based on kernel learning. *Computers & Chemical Engineering*, 106:464–479, 2017. doi: 10.1016/j.compchemeng.2017.07.004.
- Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.
- Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.
- Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.

- Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 10162–10172, 2021. URL <https://proceedings.mlr.press/v139/taskesen21a.html>.
- David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, January 2004. ISSN 1573-0565. doi: 10.1023/B:MACH.0000008084.60811.49.
- Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.
- Tianyu Wang, Ningyuan Chen, and Chun Wang. Contextual optimization under covariate shift: A robust approach by intersecting Wasserstein balls. *arXiv:2406.02426*, 2024.
- Xiaoming Wang, Fulai Chung, and Shitong Wang. Theoretical analysis for solution of support vector data description. *Neural Networks*, 24(4):360–369, 2011.
- Yuntao Wang, Yuxuan Li, Qingyuan Yang, and Hu Ding. Finding Wasserstein ball center: Efficient algorithm and the applications in fairness. In *Proceedings of the Forty-second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. URL <https://proceedings.mlr.press/v267/wang25be.html>.
- Lijun Zhang, Peng Zhao, Zhen-Hua Zhuang, Tianbao Yang, and Zhi-Hua Zhou. Stochastic approximation approaches to group distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 36, pages 52490–52522, 2023.
- Steven Zhang, David Widmann, and Davi Barreira. OptimalTransport.jl: optimal transport algorithms for Julia. Julia package, <https://github.com/JuliaOptimalTransport/OptimalTransport.jl>, 2020.

## Appendix A Proofs

### A.1 Additional notation

For a matrix  $\mathbf{A}$ , write  $\mathbf{A}^\top$ ,  $\det \mathbf{A}$  and  $\text{adj}(\mathbf{A})$  for its transpose, determinant and adjugate. If  $\mathbf{A}$  is invertible, then  $\mathbf{A}^{-1} = \text{adj}(\mathbf{A})/\det \mathbf{A}$ . The vector  $\mathbf{1}$  has all entries equal to one, with its dimension determined by context. For a symmetric matrix,  $\mathbf{A} \succ 0$  means positive definiteness. For embedded points  $\mu_i \in \mathcal{H}$ , their Gram matrix has entries  $g_{ij} := \langle \mu_i, \mu_j \rangle_{\mathcal{H}}$ ; an unsubscripted inner product or norm involving these points is taken in  $\mathcal{H}$ .

For a metric space  $(\mathcal{Z}, d)$ , a point  $z$  and nonempty sets  $A, B \subseteq \mathcal{Z}$ , define

$$\begin{aligned} \text{dist}_d(z, A) &:= \inf_{a \in A} d(z, a), \\ d_H(A, B) &:= \max \left\{ \sup_{a \in A} \text{dist}_d(a, B), \sup_{b \in B} \text{dist}_d(b, A) \right\}. \end{aligned}$$

The latter is the Hausdorff distance, using the ambient metric indicated by the sets. Center errors use  $d = \mathcal{W}_p$ ; radius-power intervals use the usual distance on  $\mathbb{R}$ . In particular,  $d_H([a, b], [c, e]) = \max\{|a - c|, |b - e|\}$ . An  $a$ -net is a collection of points whose closed balls of radius  $a$  cover the space. The covering number  $N_{\mathcal{T}}(a)$  is the smallest size of such a net in  $(\mathcal{T}, d_{\mathcal{T}})$ , with centers in  $\mathcal{T}$ . For a nonnegative vector  $\mathbf{d} \in \mathbb{R}_+^M$ ,  $d_{(1)} \geq \dots \geq d_{(M)}$  are its descending order statistics; we set  $d_{(M+1)} := 0$  when describing the optimal radius interval.

For an event  $A$ , let  $A^c$  denote its complement. The notation  $\Pr_{\text{train}}$  refers to randomness in both the source distributions and their observed samples. Target probabilities  $\Pr_{\mathbb{P}' \sim \Psi}$  concern a fresh, independent distribution and condition on the training data. Probabilities explicitly conditioned on the source distributions concern their within-source observations. A function  $g$  is lower semicontinuous if  $g(z) \leq \liminf_{m \rightarrow \infty} g(z_m)$  whenever  $z_m \rightarrow z$ ; upper semicontinuity is defined by the reversed inequality with  $\limsup$ .

### A.2 Proof of Proposition 2 and Corollary 3

*Proof.* Three kernel mean embeddings span an affine subspace of dimension at most two, so the geometry is that of a planar triangle. The three quantities

$$A_1 = g_{11} + g_{23} - g_{12} - g_{13}, \quad A_2 = g_{22} + g_{13} - g_{12} - g_{23}, \quad A_3 = g_{33} + g_{12} - g_{13} - g_{23}$$

are the inner products of the two edges meeting at  $\mu_1, \mu_2, \mu_3$ , respectively; hence  $A_1, A_2, A_3 > 0$  exactly when the triangle is acute.

*Non-acute triangles.* If the triangle is not acute, its minimum enclosing ball is the ball whose diameter is its longest edge. Let that edge join  $\mu_i$  and  $\mu_j$ . Its center is  $(\mu_i + \mu_j)/2$  and its squared

radius is  $\|\mu_i - \mu_j\|^2/4$ , so the power of the origin with respect to this ball is

$$\left\| \frac{\mu_i + \mu_j}{2} \right\|^2 - \frac{\|\mu_i - \mu_j\|^2}{4} = \langle \mu_i, \mu_j \rangle = g_{ij} > 0.$$

The origin therefore lies strictly outside the minimum enclosing ball, and origin enclosure forces the triangle to be acute.

*Acute triangles.* If the triangle is acute, its minimum enclosing ball is its circum-ball. Let  $c_o$  be its circumcenter,  $r_o$  its circumradius, and  $q := \|c_o\|^2 - r_o^2$  the power of the origin with respect to that ball. Then the origin lies in the minimum enclosing ball if and only if  $q \leq 0$ .

Since  $c_o$  lies in the affine hull of the three points, write  $c_o = \sum_i \alpha_i \mu_i$  with  $\sum_i \alpha_i = 1$ . The equidistance conditions  $\|c_o - \mu_i\|^2 = r_o^2$  read

$$2\langle c_o, \mu_i \rangle - g_{ii} = q, \quad i = 1, 2, 3,$$

that is,  $2\mathbf{G}\boldsymbol{\alpha} - \mathbf{d} = q\mathbf{1}$ . Multiplying by  $\mathbf{G}^{-1}$  and applying  $\mathbf{1}^\top$  gives

$$2 - \mathbf{1}^\top \mathbf{G}^{-1} \mathbf{d} = q \mathbf{1}^\top \mathbf{G}^{-1} \mathbf{1}, \quad \text{hence} \quad q = \frac{2 - \mathbf{1}^\top \mathbf{G}^{-1} \mathbf{d}}{\mathbf{1}^\top \mathbf{G}^{-1} \mathbf{1}}.$$

Since  $\mathbf{G} \succ 0$ , the denominator is positive and  $q \leq 0 \iff \mathbf{1}^\top \mathbf{G}^{-1} \mathbf{d} \geq 2$ . Together with the dichotomy above, this is the stated criterion.

*The row condition.* It remains to verify Corollary 3. Assume  $\sum_{j \neq i} g_{ij} \leq \frac{1}{2} g_{ii}$ . Then, for instance,

$$A_1 = g_{11} + g_{23} - g_{12} - g_{13} \geq \frac{1}{2} g_{11} + g_{23} > 0,$$

and  $A_2, A_3 > 0$  follow symmetrically, so the triangle is acute. The row condition also makes  $\mathbf{G}$  strictly diagonally dominant with positive diagonal, hence  $\mathbf{G} \succ 0$ .

Set  $p_i := g_{ii} - 2 \sum_{j \neq i} g_{ij} \geq 0$  and  $\mathbf{p} := (p_1, p_2, p_3)^\top$ . Since  $(\mathbf{G}\mathbf{1})_i = g_{ii} + \sum_{j \neq i} g_{ij}$ ,

$$\mathbf{d} = \frac{2}{3} \mathbf{G}\mathbf{1} + \frac{1}{3} \mathbf{p},$$

whence

$$\mathbf{1}^\top \mathbf{G}^{-1} \mathbf{d} = 2 + \frac{1}{3} \mathbf{1}^\top \mathbf{G}^{-1} \mathbf{p}.$$

It therefore suffices that  $\mathbf{G}^{-1} \mathbf{1}$  have nonnegative entries.

To see that it does, relabel the entries as  $u = g_{11}$ ,  $v = g_{22}$ ,  $w = g_{33}$ ,  $x = g_{12}$ ,  $y = g_{13}$ ,  $z = g_{23}$ , all positive, so that the row condition is  $u \geq 2(x + y)$ ,  $v \geq 2(x + z)$ ,  $w \geq 2(y + z)$ . The row sums of  $\text{adj}(\mathbf{G})$  are

$$\begin{aligned} R_1 &= vw - z^2 + yz - xw + xz - yv, \\ R_2 &= yz - xw + uw - y^2 + xy - uz, \\ R_3 &= xz - yv + xy - uz + uv - x^2. \end{aligned}$$

Now  $R_1$  is increasing in  $v$  and  $w$ , because  $\partial R_1/\partial v = w - y > 0$  and  $\partial R_1/\partial w = v - x > 0$ ;  $R_2$  is increasing in  $u$  and  $w$ ; and  $R_3$  is increasing in  $u$  and  $v$ . Evaluating at the smallest admissible diagonal entries  $u = 2(x + y)$ ,  $v = 2(x + z)$ ,  $w = 2(y + z)$ ,

$$R_1 \geq 3z(x + y + z) > 0, \quad R_2 \geq 3y(x + y + z) > 0, \quad R_3 \geq 3x(x + y + z) > 0.$$

Every entry of  $\mathbf{G}^{-1}\mathbf{1} = \text{adj}(\mathbf{G})\mathbf{1}/\det \mathbf{G}$  is thus positive, so  $\mathbf{1}^\top \mathbf{G}^{-1}\mathbf{p} \geq 0$  and  $\mathbf{1}^\top \mathbf{G}^{-1}\mathbf{d} \geq 2$ . With acuteness already established, Proposition 2 gives that the minimum enclosing ball of  $\{\mu_1, \mu_2, \mu_3\}$  contains the origin.  $\square$

### A.3 Proof of Proposition 6

*Proof.* Fix  $\mathbf{x} \in \mathcal{X}$  and apply Lemma 5 to  $\ell_{\mathbf{x}}(\boldsymbol{\xi}) := f(\mathbf{x}, \boldsymbol{\xi})$ , which is continuous on compact  $\Xi$  by Assumption 1. Substituting  $\hat{\mathbb{Q}}^\star = \sum_{k=1}^K \beta_k \delta_{\boldsymbol{\zeta}_k}$  into the integral in (3.6) gives (3.8). Minimizing this identity over  $\mathbf{x}$  gives (3.9).

For each fixed  $\boldsymbol{\xi}$ , the function  $f(\mathbf{x}, \boldsymbol{\xi}) - \lambda \|\boldsymbol{\xi} - \boldsymbol{\zeta}_k\|^p$  is jointly convex in  $(\mathbf{x}, \lambda)$ . Taking suprema and nonnegative weighted sums preserves this property. The infima over the independent variables  $\mathbf{x}$  and  $\lambda$  may be taken in either order.

Finally, write  $J(\mathbf{x})$  for the left-hand side of (3.8). For any  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ ,

$$|J(\mathbf{x}) - J(\mathbf{y})| \leq \sup_{\boldsymbol{\xi} \in \Xi} |f(\mathbf{x}, \boldsymbol{\xi}) - f(\mathbf{y}, \boldsymbol{\xi})|.$$

Uniform continuity of  $f$  on compact  $\mathcal{X} \times \Xi$  makes  $J$  continuous, so it attains its minimum on  $\mathcal{X}$ . For each fixed  $\lambda$ , the dual objective is likewise continuous in  $\mathbf{x}$ , hence its minimum over  $\mathcal{X}$  is attained as well.  $\square$

### A.4 Proofs for Section 4.1

*Proof of Lemma 9.* Arrange the coordinates of  $\mathbf{d}$  in nonincreasing order as  $d_{(1)} \geq \dots \geq d_{(M)}$ , and set  $d_{(M+1)} := 0$ . Write  $a := M\nu$ . The subgradient condition for minimizing the stated convex, piecewise-linear function shows that its minimizer set is  $[d_{(k+1)}, d_{(k)}]$  if  $a = k$  is an integer, and the singleton  $\{d_{(\lfloor a \rfloor + 1)}\}$  otherwise. This also covers  $a = M$  through the convention  $d_{(M+1)} = 0$ . The same description applies to  $\tilde{\mathbf{d}}$ . If  $\max_i |d_i - \tilde{d}_i| \leq \tau$ , every corresponding order statistic differs by at most  $\tau$ :  $|d_{(j)} - \tilde{d}_{(j)}| \leq \tau$  for each  $j$ . Hence the two singleton minimizers, or the corresponding endpoints of the two intervals, differ by at most  $\tau$ . Their Hausdorff distance is therefore at most  $\tau$ .  $\square$

*Proof of Theorem 10.* For each  $i$ , apply the assumed concentration bound with  $\delta_i = \delta/M$ . A union bound gives an event of probability at least  $1 - \delta$  on which  $\varepsilon_i \leq b_i$  for every  $i$ . On this event, Lemma 8 yields  $\eta \leq \eta_\delta$  and

$$\Phi(\hat{\mathbb{Q}}^\star) - \Phi^\star \leq 2\eta_\delta.$$

Assumption 2 then gives (4.8). Since  $\Xi$  is compact,  $\mathcal{P}_p(\Xi)$  is compact under  $\mathcal{W}_p$ , and  $\mathcal{Q}^*$  is a nonempty compact set. Thus a nearest  $\mathbb{Q}^* \in \mathcal{Q}^*$  exists.

For this pair of centers, Lemma 7 and the triangle inequality imply, for every  $i$ ,

$$\left| \mathcal{W}_p^p(\widehat{\mathbb{P}}_i, \widehat{\mathbb{Q}}^*) - \mathcal{W}_p^p(\mathbb{P}_i, \mathbb{Q}^*) \right| \leq L_p \left( b_i + \mathcal{W}_p(\widehat{\mathbb{Q}}^*, \mathbb{Q}^*) \right) \leq \tau_\delta.$$

Apply Lemma 9 to these two vectors of transport costs to obtain (4.9). The midpoint map on closed intervals is 1-Lipschitz under Hausdorff distance, so the stated midpoint bound follows. Finally, for nonnegative  $u, v$  and  $p \geq 1$ ,  $|u^{1/p} - v^{1/p}| \leq |u - v|^{1/p}$ , which gives the bound on the radii themselves.  $\square$

## A.5 Proofs for Section 4.2

*Proof of Lemma 12.* Write  $\theta = (\mathbb{Q}, \rho)$  and  $h_\theta(\mathbb{P}) := \ell_\gamma(s_{\mathbb{Q}, \rho}(\mathbb{P}))$ . The triangle inequality and the mean value bound on  $[0, D]$  give

$$\left| \mathcal{W}_p^p(\mathbb{P}, \mathbb{Q}) - \mathcal{W}_p^p(\mathbb{P}, \mathbb{Q}') \right| \leq L_p \mathcal{W}_p(\mathbb{Q}, \mathbb{Q}').$$

Since  $\ell_\gamma$  is  $1/\gamma$ -Lipschitz, it follows that

$$\sup_{\mathbb{P} \in \mathcal{P}_p(\Xi)} |h_\theta(\mathbb{P}) - h_{\theta'}(\mathbb{P})| \leq \frac{d_{\mathcal{T}}(\theta, \theta')}{\gamma}.$$

Choose a deterministic  $a$ -net  $\{\theta_1, \dots, \theta_N\}$  of  $\mathcal{T}$ , where  $N = N_{\mathcal{T}}(a)$ . For each fixed  $j$ , the variables  $h_{\theta_j}(\mathbb{P}_i)$ ,  $i = 1, \dots, M$ , are i.i.d. and take values in  $[0, 1]$ . The one-sided Hoeffding inequality and a union bound therefore give, with probability at least  $1 - \delta_M$ ,

$$\mathbb{E}_\Psi h_{\theta_j}(\mathbb{P}') \leq \frac{1}{M} \sum_{i=1}^M h_{\theta_j}(\mathbb{P}_i) + \sqrt{\frac{\log N + \log(1/\delta_M)}{2M}}, \quad j = 1, \dots, N.$$

On this event, choose for any  $\theta \in \mathcal{T}$  an index  $j$  with  $d_{\mathcal{T}}(\theta, \theta_j) \leq a$ . The population mean and the empirical mean each change by at most  $a/\gamma$  upon replacing  $\theta_j$  with  $\theta$ . Hence

$$\mathbb{E}_\Psi h_\theta(\mathbb{P}') \leq \frac{1}{M} \sum_{i=1}^M h_\theta(\mathbb{P}_i) + \frac{2a}{\gamma} + \sqrt{\frac{\log N + \log(1/\delta_M)}{2M}}.$$

The event does not depend on the chosen  $\theta$ , proving the asserted uniform bound.  $\square$

*Proof of Theorem 13.* Let  $A_M$  be the source event from Lemma 12, which has probability at least  $1 - \delta_M$ . Put  $\varepsilon_i := \mathcal{W}_p(\widehat{\mathbb{P}}_i, \mathbb{P}_i)$  and  $B_{M, \mathbf{n}} := \bigcap_{i=1}^M \{\varepsilon_i \leq b_i\}$ . Applying (4.17) with  $t = \delta_n/M$  and taking a conditional union bound gives

$$\Pr\{B_{M, \mathbf{n}}^c \mid \mathbb{P}_1, \dots, \mathbb{P}_M\} \leq \delta_n \quad \text{almost surely.}$$

Taking expectations yields  $\Pr(B_{M,n}) \geq 1 - \delta_n$ . Thus  $A_M \cap B_{M,n}$  has probability at least  $1 - \delta_M - \delta_n$ , without requiring independence of these events.

Work on this intersection and write  $\widehat{s}(\mathbb{P}) := s_{\widehat{\mathbb{Q}}^*, \widehat{\rho}^*}(\mathbb{P})$ . Lemma 7 and the Lipschitz property of the loss imply

$$\left| \ell_\gamma(\widehat{s}(\mathbb{P}_i)) - \ell_\gamma(\widehat{s}(\widehat{\mathbb{P}}_i)) \right| \leq \frac{L_p \varepsilon_i}{\gamma} \leq \frac{L_p b_i}{\gamma}.$$

Both scores here use the same empirical center and radius. By Lemma 11 applied to the empirical sources and the second inequality in (4.12),

$$\frac{1}{M} \sum_{i=1}^M \ell_\gamma(\widehat{s}(\widehat{\mathbb{P}}_i)) \leq \frac{1}{M} \sum_{i=1}^M \mathbf{1}\{\widehat{s}(\widehat{\mathbb{P}}_i) > 0\} \leq \nu.$$

The uniformity on  $A_M$  allows evaluation at the data-dependent pair  $(\widehat{\mathbb{Q}}^*, \widehat{\rho}^*)$ . Consequently,

$$\begin{aligned} \Pr_{\mathbb{P}' \sim \Psi} \{ \mathbb{P}' \notin \widehat{\mathcal{U}}_{\nu, \gamma} \} &= \mathbb{E}_\Psi \mathbf{1}\{\widehat{s}(\mathbb{P}') > \gamma\} \\ &\leq \mathbb{E}_\Psi \ell_\gamma(\widehat{s}(\mathbb{P}')) \\ &\leq \frac{1}{M} \sum_{i=1}^M \ell_\gamma(\widehat{s}(\mathbb{P}_i)) + \mathcal{C}_M(a, \gamma, \delta_M) \\ &\leq \nu + \frac{L_p}{M\gamma} \sum_{i=1}^M b_i + \mathcal{C}_M(a, \gamma, \delta_M). \end{aligned}$$

Combining this with the trivial bound by one proves (4.19). When the sources are known, apply the same argument at an optimal pair for those sources on  $A_M$ . The source-replacement step and the event  $B_{M,n}$  are then unnecessary, giving (4.20).  $\square$

*Proof of Corollary 15.* Let  $\mathcal{E}_M$  denote (4.18) evaluated at the sequences in the corollary. Conditions (4.23) imply  $\mathcal{E}_M \rightarrow 0$ . For any fixed  $\epsilon > 0$ , take  $M$  sufficiently large that  $\mathcal{E}_M < \epsilon$  and  $\delta_M + \delta_{n,M} < 1$ . Theorem 13 then gives

$$\Pr_{\text{train}} \left\{ \Psi(\widehat{\mathcal{U}}_{\nu, \gamma_M}) < 1 - \nu - \epsilon \right\} \leq \delta_M + \delta_{n,M} \rightarrow 0.$$

The proof for known sources uses (4.20) in exactly the same way. No nesting of the learned sets or interchange of a limit and target probability is needed.  $\square$