A Shrinkage Path Heuristic for Wasserstein Distributionally Robust Optimization

Lingjun Meng, Ryan Cory-Wright, Wolfram Wiesemann
Imperial Business School, Imperial College London, London SW7 2AZ, United Kingdom
{l.meng23, r.cory-wright, ww}@imperial.ac.uk

Wasserstein distributionally robust optimization (DRO) is a versatile and widely adopted framework for decision-making under uncertainty, yet its standard deterministic reformulations generally contain non-convex inner subproblems that are challenging to solve. To address this issue, we propose a shrinkage path heuristic that reduces the solution of a DRO problem to a one-dimensional search over the line segment connecting the (typically benign) sample average approximation (SAA) and the (more demanding but practically solvable) classical robust optimization solution. We derive a priori suboptimality bounds in stylized settings and, for the general case, a posteriori bounds obtained by applying a similar heuristic to a dual formulation. Numerical experiments on a multi-item newsvendor and an appointment scheduling problem show that the shrinkage path heuristic attains 85–110% (resp. 45–70%) of the out-of-sample performance improvements of Wasserstein DRO over SAA, at a fraction of the computational cost.

1. Introduction

Decision problems under uncertainty are often solved via sample average approximation (SAA), which involves optimizing against the empirical distribution of these observations. However, SAA is prone to overfitting, as the resulting decisions are tailored to the sample rather than reflecting the underlying distribution that generated it, often resulting in poor out-of-sample performance. To counteract this overfitting, distributionally robust optimization (DRO) hedges against all distributions in a neighborhood of the empirical distribution. This protection comes at a price, however. Although DRO problems can be reformulated in certain special cases, they are generally more computationally challenging to solve than their SAA counterparts. Additionally, the size of the neighborhood is a hyperparameter that is rarely known in advance and usually requires tuning through cross-validation, which involves resolving the DRO problem for each candidate neighborhood size across multiple folds. This paper proposes a simple shrinkage path heuristic that replaces DRO with a one-dimensional search along the line segment connecting the solutions of two well-studied problems: the SAA problem and its classical robust optimization counterpart. Since the line segment itself does not depend on any hyperparameters, tuning the degree of robustness via cross-validation requires only the two endpoint solutions and inexpensive evaluations along

the path. The heuristic thereby captures much of the protection of DRO over SAA while reducing computational costs by orders of magnitude.

Specifically, we study DRO problems of the form

$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} & \sup\limits_{\mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\big] \\ \text{subject to} & g_m(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le \gamma_m \qquad \mathbb{Q}\text{-a.s. } \forall \mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}}),\ \forall m\in[M], \end{array} \tag{1}$$

where the decision $\boldsymbol{x}$ is selected from a closed convex set $\mathcal{X}\subseteq\mathbb{R}^n$, the objective minimizes the worst-case expectation of a loss function ℓ that is closed convex in $\boldsymbol{x}$ for every $\boldsymbol{\xi}\in\Xi$ and upper semicontinuous in $\boldsymbol{\xi}$ for every $\boldsymbol{x}\in\mathcal{X}$, and the constraints indexed by $m\in[M]:=\{1,\ldots,M\}$ are closed convex in $\boldsymbol{x}$ for every fixed realization $\boldsymbol{\xi}$ and hold $\mathbb{Q}$-almost surely. The uncertainty is captured by the random vector $\tilde{\boldsymbol{\xi}}\in\Xi$, which may be governed by any distribution $\mathbb{Q}$ from the type-1 Wasserstein ambiguity set

$$\mathbb{B}_\rho(\hat{\mathbb{P}}) \coloneqq \{\mathbb{Q}\in\mathcal{P}(\Xi) : W(\mathbb{Q},\hat{\mathbb{P}})\le\rho\} \quad \text{with} \quad W(\mathbb{Q},\hat{\mathbb{P}}) = \inf_{\pi\in\Pi(\mathbb{Q},\hat{\mathbb{P}})} \int_{\Xi\times\Xi} d(\boldsymbol{\xi},\boldsymbol{\xi}')\,\pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}').$$

Here, d is a lower semicontinuous ground metric, and the empirical distribution $\hat{\mathbb{P}}=\frac{1}{N}\sum_{i=1}^N \delta_{\hat{\boldsymbol{\xi}}_i}$ is constructed from N independent samples $\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_N$ drawn from the unknown true data-generating distribution $\mathbb{P}^\star$. We assume throughout that the support $\Xi\subseteq\mathbb{R}^k$ is non-empty, closed, and convex.

We refer to Problem (1) as the Wasserstein DRO (WDRO) problem. It arises in a wide range of applications, including inventory management (Hanasusanto et al. 2015, Xin and Goldberg 2022), machine learning (Bennouna et al. 2023), energy systems (Xie and Ahmed 2018, Esteban-Pérez and Morales 2023), and control theory (Shafieezadeh-Abadeh et al. 2018). We refer to Rahimian and Mehrotra (2022) and Kuhn et al. (2025) for detailed reviews.

Unfortunately, Problem (1) is often not *practically tractable*—that is, solvable in a practical amount of time at instance sizes relevant to the application (Bertsimas and Dunn 2019). By strong duality, the worst-case expectation objective admits a finite-dimensional dual reformulation. In this reformulation, one optimizes over the dual multiplier corresponding to the Wasserstein radius and solves N auxiliary maximization problems, one for each sample point, over the support Ξ. Each auxiliary problem maximizes a loss-like term minus a transport-cost penalty (Mohajerin Esfahani and Kuhn 2018, Theorem 4.2). The tractability of Problem (1), therefore, hinges on the structure of these inner maximization problems. When Ξ is unbounded (e.g., $\Xi=\mathbb{R}^k$ or $\Xi=\mathbb{R}^k_+$) and the ground metric is a norm, each supremum is finite only if the loss function is Lipschitz continuous on Ξ. In the special case where $\Xi=\mathbb{R}^k$, this finite problem reduces to a regularized sample average approximation (SAA) (Shafieezadeh-Abadeh et al. 2019, Theorem 4(ii)). However, many losses of practical interest—including quadratic, exponential, and other smooth superlinear losses—fail the Lipschitz condition and therefore can render the worst-case expectation infinite along unbounded

rays. When Ξ is compact, the inner suprema are always finite, but their tractability depends on the loss structure. If the loss function is *a pointwise maximum of finitely many concave functions* of $\boldsymbol{\xi}$ (Mohajerin Esfahani and Kuhn 2018, Assumption 4.1), each inner problem maximizes a piecewise concave function over a convex set, which is tractable. In contrast, when the loss function is *convex* in $\boldsymbol{\xi}$—as is the case for quadratic and other smooth losses commonly encountered in practice—each inner subproblem becomes a k-dimensional difference-of-convex program; in general, solving such problems is NP-hard, even for highly structured quadratic instances (Pardalos and Vavasis 1991).

Several strategies aim to circumvent the intractability of WDRO with general loss functions. When the loss is convex in $\boldsymbol{\xi}$, Corollary 4.3 of Mohajerin Esfahani and Kuhn (2018) yields finite convex conservative approximations of the inner suprema, although their sizes typically grow exponentially in the dimension of $\boldsymbol{\xi}$. Cheramin et al. (2022) propose inner and outer approximations for moment and Wasserstein ambiguity sets based on block decompositions of $\boldsymbol{\xi}$ and principal component reduction, together with gap bounds that quantify the resulting accuracy–runtime trade-off. Gao et al. (2024, Theorem 2) connect type-1 Wasserstein DRO to variation regularization and show that the gap between the worst-case expectation and the empirical loss is bounded by ρ times a variation norm of the loss function.

A second stream of work solves Problem (1) algorithmically. Shafiee et al. (2026) study gradient-based methods for Wasserstein DRO when Ξ is unconstrained and the loss ℓ is nonconvex in $\boldsymbol{x}$. Wang et al. (2026) replace the Wasserstein distance with its entropically regularized counterpart; the resulting Sinkhorn dual is amenable to optimization via stochastic mirror descent. Two further works propose dual decomposition algorithms for special problem classes: Li et al. (2019) propose an ADMM framework that exploits a dual decomposition for the Wasserstein distributionally robust logistic regression. Luo and Mehrotra (2019) develop a cutting-surface decomposition algorithm for the dual reformulation of the Wasserstein DRO with bounded support for a class of statistical learning problems. Finally, Blanchet et al. (2022) study optimal-transport DRO with quadratic transport costs and linear decision rules, which together reduce each inner adversarial subproblem to a line search, and develop stochastic gradient descent schemes for the outer decision with convergence guarantees. The GitHub repository accompanying this paper classifies the literature in greater depth according to the scope of each approach, the nature of the resulting reformulation or algorithm, and whether the approach is exact[1].

All of the above approaches aim to solve Problem (1)—or an approximation thereof—for a *single fixed* Wasserstein radius ρ. In contrast, we exploit the structure of the solution path *as* ρ *varies*. We adopt this perspective because, in most practical applications, ρ is not selected a priori (*e.g.*, from a concentration inequality), but rather it is tuned via cross-validation on the data. To this end, we propose a shrinkage path heuristic (Algorithm 1) that reduces Problem (1) to a one-dimensional

search over the line segment that connects two benchmark decisions: the solution of the sample average approximation (SAA), which arises at $\rho = 0$ and is typically easy to compute, and the solution of the classical robust optimization (RO) problem, which arises as $\rho \to \infty$ and—although harder—is routinely solved in practice. The line segment itself is independent of ρ: instead of solving a full DRO problem at every candidate radius, one tunes the position on this fixed path by cross-validation, saving orders of magnitude in runtime. Although simple, the heuristic admits a priori guarantees in stylized settings (Theorems 1–3 and Proposition 1 in Section 3), a posteriori guarantees in general (Theorems 4 and 5 in Section 4), and competitive empirical performance (Section 5); these three advances form the core of this paper.

More specifically, we summarize our contributions as follows.

1. Under suitable stylized assumptions on the loss function and uncertainty support, we establish *a priori suboptimality bounds* that serve as a theoretical motivation for our heuristic.

2. For the general setting of Problem (1), we apply the same shrinkage idea to a dual formulation, yielding a companion procedure (Algorithm 2) that mirrors Algorithm 1 on the dual side. The gap between the primal and dual heuristic values provides computable (per-instance) *a posteriori suboptimality bounds* without requiring an exact solution of Problem (1).

3. We demonstrate the effectiveness of our heuristic on a multi-item newsvendor and an appointment scheduling problem, where our a posteriori bounds certify that the heuristic incurs only a small in-sample suboptimality gap for the DRO objective. Out of sample, for the multi-item newsvendor setting (resp. the appointment scheduling problem), the shrinkage path heuristic attains 85–110% (resp. 45–70%) of the median out-of-sample benefits of Wasserstein DRO over SAA, at a fraction of the computational cost.

The remainder of this paper is organized as follows. Section 2 introduces our heuristic. Section 3 derives a priori suboptimality bounds under several structural assumptions on the loss function. Section 4 develops a posteriori performance guarantees for the general setting by constructing a dual shrinkage path whose objective values provide computable lower bounds on the true DRO optimum. Section 5 illustrates our heuristic on a multi-item newsvendor and an appointment scheduling problem. All proofs are deferred to the electronic companion. We relegate an extended literature review and a generalization of Problem (1) to ambiguous conditional value-at-risk constraints to the GitHub repository accompanying this work.

Remark 1 (Multi-Stage Problems). Our heuristic extends to two-stage and multi-stage variants of Problem (1) in two natural ways. The first is to restrict the recourse decisions to affine (or piecewise-affine) functions of the uncertainty, which collapses the problem to a single-stage instance of Problem (1) in an expanded first-stage decision vector. The second is to substitute the optimal recourse value function for ℓ directly, whenever this value function remains convex in $(\boldsymbol{x}, \boldsymbol{\xi})$; this is the route taken in our numerical experiments.

2. The Shrinkage Path Heuristic

We begin by simplifying the feasible set of Problem (1). For any $\rho > 0$ and each point $\boldsymbol{\xi} \in \Xi$, the Wasserstein ball $\mathbb{B}_\rho(\hat{\mathbb{P}})$ contains a distribution that places positive mass at $\boldsymbol{\xi}$, so the almost-sure constraints decouple from the radius and reduce to the support-only robust requirement that $g_m(\boldsymbol{x}, \boldsymbol{\xi}) \leq \gamma_m$ for all $\boldsymbol{\xi} \in \Xi$ and every $m \in [M]$. Accordingly, the feasible set of Problem (1) reduces to

$$\mathcal{X}(\Xi) \coloneqq \left\{ \boldsymbol{x} \in \mathcal{X} : g_m(\boldsymbol{x}, \boldsymbol{\xi}) \leq \gamma_m \ \forall \boldsymbol{\xi} \in \Xi, \forall m \in [M] \right\}. \tag{2}$$

We assume throughout that $\mathcal{X}(\Xi)$ is non-empty, otherwise, Problem (1) is trivially infeasible. Convexity of $\mathcal{X}$ together with convexity of each $g_m(\cdot, \boldsymbol{\xi})$ ensures that $\mathcal{X}(\Xi)$ is itself convex. The set $\mathcal{X}(\Xi)$ is described by semi-infinite constraints. When each $g_m(\boldsymbol{x}, \cdot)$ is concave on Ξ and Ξ admits a conic representation, classical robust counterpart techniques yield a tractable finite-dimensional reformulation (Ben-Tal et al. 2009). More broadly, $\mathcal{X}(\Xi)$ remains tractable whenever it admits a polynomial-time separation oracle, that is, whenever for any candidate $\boldsymbol{x}^\star$ a violated constraint can be identified by maximizing $g_m(\boldsymbol{x}^\star, \boldsymbol{\xi})$ over $\boldsymbol{\xi} \in \Xi$ in polynomial time.

The *sample average approximation* (SAA) solution is any optimal solution $\boldsymbol{x}(0)$ to

$$\underset{\boldsymbol{x} \in \mathcal{X}(\Xi)}{\text{minimize}} \quad \mathbb{E}_{\hat{\mathbb{P}}}\left[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]. \tag{3a}$$

We assume that Problem (3a) attains a finite optimal value, ensuring that $\boldsymbol{x}(0)$ is well-defined. Since $\hat{\mathbb{P}} \in \mathbb{B}_\rho(\hat{\mathbb{P}})$, this further implies that the WDRO problem (1) is bounded below. Whenever ℓ is convex in $\boldsymbol{x}$ and $\mathcal{X}(\Xi)$ admits a polynomial-time separation oracle, Problem (3a) is solvable in polynomial time; the Wasserstein ambiguity set plays no role at this extreme.

The *robust optimization* (RO) solution is any optimal solution $\boldsymbol{x}(\infty)$ to

$$\underset{\boldsymbol{x} \in \mathcal{X}(\Xi)}{\text{minimize}} \quad \sup_{\boldsymbol{\xi} \in \Xi} \ell(\boldsymbol{x}, \boldsymbol{\xi}). \tag{3b}$$

We assume that Problem (3b) attains a finite optimal value, ensuring that $\boldsymbol{x}(\infty)$ is well-defined. This holds, for instance, when Ξ is compact and ℓ is upper semicontinuous in $\boldsymbol{\xi}$, or more generally when $\ell(\boldsymbol{x}, \cdot)$ is bounded above on Ξ for every $\boldsymbol{x} \in \mathcal{X}(\Xi)$. Problem (3b) is tractable when $\ell(\boldsymbol{x}, \cdot)$ is concave on Ξ. In the convex-in-$\boldsymbol{\xi}$ regime of interest in this paper, the problem is NP-hard in general (Pardalos and Vavasis 1991)—the same source of intractability that afflicts the inner suprema in the dual reformulation of Problem (1). Even so, as in classical robust optimization, this worst-case separation is routinely handled in practice by cutting-plane methods that iteratively add violated worst-case scenarios (Mutapcic and Boyd 2009).

Define the *shrinkage path* as

$$\widehat{\mathcal{X}}(\Xi) \coloneqq \left\{ (1-\lambda)\boldsymbol{x}(0) + \lambda \boldsymbol{x}(\infty) : \lambda \in [0,1] \right\}. \tag{4}$$

Because the SAA and RO problems in (3) need not admit unique optimizers, we fix one solution of each; these endpoints $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$ in turn fix the shrinkage path, and we refer henceforth to *the* SAA solution, *the* RO solution, and *the* shrinkage path. The shrinkage path is a tractable proxy for the WDRO solution path $\{\boldsymbol{x}^\star(\rho) : \rho \geq 0\}$, where $\boldsymbol{x}^\star(\rho)$ denotes any optimal solution of Problem (1) at radius ρ. As shown by Hao and Zhang (2025) for robust linear optimization, $\boldsymbol{x}^\star(\rho)$ traces a one-dimensional Bregman-type regularization path; in our setting, however, it is generally intractable to compute. We refer to λ as the *shrinkage parameter*: $\lambda = 0$ corresponds to the SAA solution $\boldsymbol{x}(0)$, $\lambda = 1$ recovers the RO solution $\boldsymbol{x}(\infty)$, and intermediate values balance nominal fit against robustness. Since both endpoints lie on the shrinkage path, optimizing over it weakly dominates the *SAA–RO heuristic* of selecting the better of $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$. Figure 1 illustrates the shrinkage path alongside the true WDRO optimizer path and the resulting suboptimality gap.

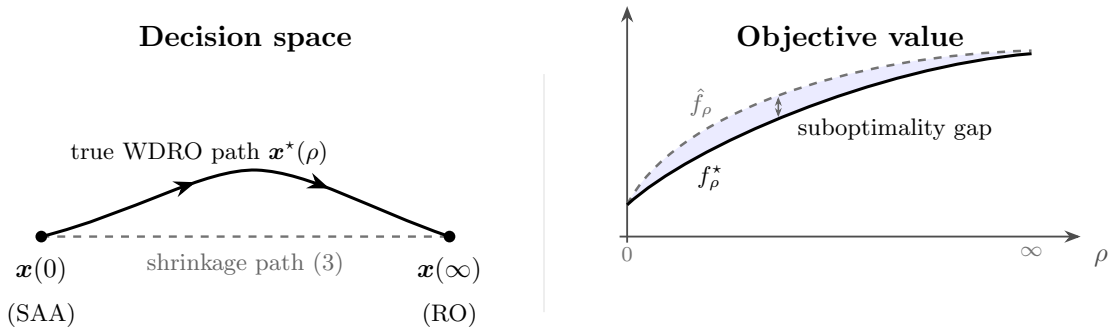

Figure 1 Schematic view of the shrinkage path heuristic. Left: geometric approximation in decision space. The dashed segment is the shrinkage path connecting the SAA and RO endpoints, and the solid curve is the WDRO optimizer path $\boldsymbol{x}^\star(\rho)$. Right: induced value gap as a function of the Wasserstein radius ρ. The dashed curve depicts the shrinkage path values $\hat{f}_\rho$, the solid curve denotes the WDRO optima $f^\star_\rho$, and the shaded region represents the induced suboptimality gaps $\hat{f}_\rho - f^\star_\rho$.

Algorithm 1 implements a one-dimensional search over the shrinkage path. Since $\mathcal{X}(\Xi)$ is convex and both endpoints $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$ reside in $\mathcal{X}(\Xi)$, every interpolant $\boldsymbol{x}$ is feasible in Problem (1).

For a *fixed* radius ρ, Algorithm 1 replaces the joint optimization in Problem (1) with two endpoint problems plus $|\Lambda|$ calls to the worst-case oracle WC-Eval$(\cdot;\rho)$. Each oracle call fixes the decision $\boldsymbol{x}$ and admits the standard Wasserstein dual reformulation (Mohajerin Esfahani and Kuhn 2018, Theorem 4.2), which decomposes the worst-case expectation into N independent k-dimensional subproblems—one per data point, each maximizing the loss minus a penalized transport cost over Ξ. These subproblems are decoupled, parallelize easily, and can be handled by standard global solvers (*e.g.*, branch-and-bound or multistart local search) for moderate k. The in-sample speedup over the joint problem is typically modest, and is not the main benefit of the heuristic. The larger savings

Algorithm 1 Shrinkage Path Heuristic for Problem (1)

Require: Empirical law $\hat{\mathbb{P}} := \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{\boldsymbol{\xi}}_i}$ with samples $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^N$; radius $\rho \geq 0$; search grid $\Lambda \subset [0,1]$; worst-case evaluation oracle WC-Eval$(\boldsymbol{x};\rho) = \sup\{\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] : \mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}})\}$.

Initialize incumbent objective $J^\star \leftarrow +\infty$ and incumbent solution $\boldsymbol{x}^\star \leftarrow \emptyset$.
for $\lambda \in \Lambda$ **do**
  Compute $\boldsymbol{x} \leftarrow (1-\lambda)\boldsymbol{x}(0) + \lambda \boldsymbol{x}(\infty)$ and $J \leftarrow$ WC-Eval$(\boldsymbol{x};\rho)$.
  if $J < J^\star$ **then**
    $J^\star \leftarrow J$ and $\boldsymbol{x}^\star \leftarrow \boldsymbol{x}$.
  end if
end for
return $\boldsymbol{x}^\star$ with objective value $J^\star$.

arise *out-of-sample*, when ρ is itself tuned by K-fold cross-validation. In practice, ρ is rarely chosen a priori; it is selected from a candidate grid $\mathcal{R} \subset \mathbb{R}_+$ to optimize cross-validated performance. For the exact WDRO problem, this requires $K|\mathcal{R}|$ solves of Problem (1), each of which is precisely the joint optimization that motivated the heuristic in the first place. The shrinkage path, by contrast, does not depend on ρ, and only one of its two endpoints depends on the data: the typically harder RO endpoint $\boldsymbol{x}(\infty)$ is computed once from $\mathcal{X}$, ℓ, and Ξ alone, while the typically easier SAA endpoint $\boldsymbol{x}(0)$ is recomputed per fold from the corresponding training sample. The entire path is then scored on the validation sample *without any worst-case evaluations*. Concretely, denote by $\hat{\mathbb{P}}^{\mathrm{tr}}$ and $\hat{\mathbb{P}}^{\mathrm{val}}$ the empirical distributions on a fold's training and validation samples, write $\boldsymbol{x}^{\mathrm{tr}}(0)$ for the SAA endpoint computed from $\hat{\mathbb{P}}^{\mathrm{tr}}$, replace the in-sample oracle in Algorithm 1 with the validation-fold sample average

$$\textsc{Val-Eval}(\boldsymbol{x}) := \mathbb{E}_{\hat{\mathbb{P}}^{\mathrm{val}}}\big[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\big] = \frac{1}{N^{\mathrm{val}}}\sum_{i=1}^{N^{\mathrm{val}}} \ell\big(\boldsymbol{x},\hat{\boldsymbol{\xi}}_i^{\mathrm{val}}\big),$$

and average Val-Eval across the K folds for each λ on the search grid. Feasibility need not be checked separately, since $\mathcal{X}(\Xi)$ does not depend on ρ and every interpolant $(1-\lambda)\boldsymbol{x}^{\mathrm{tr}}(0) + \lambda\boldsymbol{x}(\infty)$ remains in $\mathcal{X}(\Xi)$ by convexity. Let $\lambda^\star \in \Lambda$ minimize the cross-validated Val-Eval along the path; the final decision $(1-\lambda^\star)\boldsymbol{x}(0) + \lambda^\star\boldsymbol{x}(\infty)$ is then formed by recomputing the SAA endpoint $\boldsymbol{x}(0)$ on the full sample $\hat{\mathbb{P}}$, while the RO endpoint $\boldsymbol{x}(\infty)$ is reused. The heuristic thus replaces cross-validated tuning of ρ, which requires $K|\mathcal{R}|$ solves of Problem (1), with cross-validated tuning of λ, which requires one RO solve, K SAA solves, and sample-average evaluations along the path. The resulting savings are typically several orders of magnitude.

3. A Priori Performance Guarantees

To obtain a priori guarantees, this section focuses on the following structured setting.

ASSUMPTION 1 **(Lipschitz Loss)**. *Problem* (1) *satisfies the following conditions.*

(i) The ground metric is the Euclidean norm: $d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2$.

(ii) The loss is $\ell(\boldsymbol{x}, \boldsymbol{\xi}) = L(\boldsymbol{\xi}^\top \boldsymbol{x})$ *with* $L : \mathbb{R} \to \mathbb{R}_+$ *convex and Lipschitz with constant* $\mathrm{Lip}(L)$.

Pairing the Euclidean ground metric in *(i)* with the bilinear coupling $\boldsymbol{\xi}^\top \boldsymbol{x}$ in *(ii)* ensures that, for every fixed decision $\boldsymbol{x}$, the loss $\boldsymbol{\xi} \mapsto L(\boldsymbol{\xi}^\top \boldsymbol{x})$ is Lipschitz in $\boldsymbol{\xi}$ with the constant $\mathrm{Lip}(L)\|\boldsymbol{x}\|_2$, which in turn yields explicit, decision-dependent bounds on how fast the worst-case expectation can grow with ρ relative to its nominal counterpart. This implies in particular that the number of uncertain parameters k coincides with the number of decisions n in this section. The Lipschitz assumption on L in *(ii)* further keeps the worst-case expectation finite when Ξ is unbounded, and convexity of L ensures that the WDRO problem is convex in $\boldsymbol{x}$. Under these conditions, Problem (1) admits closed-form suboptimality bounds and structural insights that are difficult to obtain in the general case. In particular, when $\Xi = \mathbb{R}^k$, the WDRO problem reduces to a regularized SAA problem (Shafieezadeh-Abadeh et al. 2019, Theorem 4(ii)). In such settings, the heuristic is not needed in practice, since the WDRO problem is itself tractable. We study them because they can be analyzed in closed form, and the resulting guarantees help interpret the strong performance of the heuristic on the harder instances of Section 5.

We define the *additive suboptimality* of a candidate solution at a given Wasserstein radius ρ as the gap between its worst-case expected loss and the optimal value of Problem (1) at that radius. The worst-case additive suboptimality of a heuristic is the supremum of this gap over all $\rho \geq 0$. We say that a bound on this quantity is *tight* when equality is attained by some problem instance, and *asymptotically tight* when equality is approached in the limit along a sequence of instances parameterized by some structural quantity. When neither holds, we say the bound has *slack*.

We first bound the additive suboptimality of our shrinkage path heuristic when Ξ is bounded.

THEOREM 1 **(Bounded Support)**. *Let* $\mathrm{diam}(\Xi) = \sup\{\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 : \boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi\} < \infty$. *Then, for any* $\rho \in [0, +\infty)$, *the additive suboptimality of the shrinkage path heuristic is at most*

$$\rho \left[1 - \frac{\rho}{\mathrm{diam}(\Xi)}\right]_+ \mathrm{Lip}(L)\|\boldsymbol{x}(0)\|_2,$$

which is maximized at $\rho = \mathrm{diam}(\Xi)/2$. *Moreover, the bound is pointwise tight: for every* $\rho \in [0, \infty)$, *the bound is attained by some problem instance.*

The bound in Theorem 1 vanishes at both endpoints of $\rho \in [0, \mathrm{diam}(\Xi)]$ and is largest near the midpoint. Thus, the heuristic is provably near-optimal for small ρ and again as ρ approaches $\mathrm{diam}(\Xi)$. The small-ρ regime is the practically relevant one: the out-of-sample performance of DRO typically improves over SAA at small radii and worsens past a critical threshold (Anderson and

Philpott 2022). The heuristic therefore suits the common goal of improving SAA with a modest amount of distributional robustness.

We now prove an analogous result for the unbounded support case.

PROPOSITION 1 **(Unbounded Support)**. *Let $\Xi = \mathbb{R}^k$, and let $\bar{r} := \frac{1}{N}\sum_{i=1}^N \|\hat{\boldsymbol{\xi}}_i\|_2$ denote the* effective diameter *of the support. The additive suboptimality of the shrinkage path heuristic is at most*

$$\mathrm{Lip}(L)\left(\rho\left(1-\frac{\rho}{\bar{r}}\right)\|\boldsymbol{x}(0)\|_2 + \frac{\rho^2}{\bar{r}}\|\boldsymbol{x}(\infty)\|_2\right) \qquad \forall \rho \in [0,\bar{r}],$$

which is maximized at $\rho = \bar{r}/2$ when $\boldsymbol{x}(\infty) = \mathbf{0}$.

Proposition 1 parallels Theorem 1, with the (finite) effective diameter $\bar{r}$ serving as a data-driven surrogate for the (infinite) support diameter $\mathrm{diam}(\Xi)$: below $\bar{r}$, the bound is a weighted sum of two terms—one proportional to the SAA endpoint $\|\boldsymbol{x}(0)\|_2$ with weight $\rho(1-\rho/\bar{r})$, and one governed by the RO endpoint $\|\boldsymbol{x}(\infty)\|_2$ with weight $\rho^2/\bar{r}$—and it reduces to a parabola with peak at $\rho = \bar{r}/2$ in the case where $\boldsymbol{x}(\infty) = \mathbf{0}$. Unlike Theorem 1, the bound is not in general tight. The bound can be extended to $\rho > \bar{r}$ and admits a closed-form worst case over all $\rho \geq 0$; we relegate these refinements to the GitHub companion.

The bounds in Theorem 1 and Proposition 1 depend only on coarse global parameters—the Lipschitz constant $\mathrm{Lip}(L)$, the support diameter $\mathrm{diam}(\Xi)$ or its effective counterpart $\bar{r}$, and the endpoint norms $\|\boldsymbol{x}(0)\|_2$ and $\|\boldsymbol{x}(\infty)\|_2$. In particular, they do not reflect the curvature of the loss function or the conditioning of the data. We now specialize to a quadratic loss setting that admits sharper, data-dependent guarantees. In addition to an improved additive bound, this setting yields a *multiplicative suboptimality* bound: a bound on the ratio of the worst-case additive suboptimality of the shrinkage path heuristic to that of the SAA–RO heuristic, which selects the better of the two endpoints $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$. This bound shows that the shrinkage path heuristic strictly improves upon the SAA–RO heuristic whenever the data are not perfectly conditioned. This justifies exploring the full shrinkage path rather than restricting attention to its two endpoints.

ASSUMPTION 2 **(Quadratic Loss)**. *Let $\mathcal{Z} := \{\boldsymbol{x}^\top \hat{\boldsymbol{\xi}}_i : \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}(0)\|_2,\ i \in [N]\}$ collect the scalar arguments along the shrinkage path. In addition to Assumption 1, the following conditions hold.*

(i) The support and feasible set are unconstrained: $\Xi = \mathbb{R}^k$ and $\mathcal{X}(\Xi) = \mathbb{R}^n$.

(ii) The loss $L : \mathbb{R} \to \mathbb{R}_+$ is differentiable, and is quadratic on $\mathcal{Z}$, that is, $L(z) = z^2 + bz + c$ on $\mathcal{Z}$ for some $b, c \in \mathbb{R}$.

(iii) The empirical second-moment matrix $\boldsymbol{H} := \frac{1}{N}\sum_{i=1}^N \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^\top$ is positive definite, with eigenvalues $0 < \lambda_{\min} \leq \cdots \leq \lambda_{\max}$ and condition number $\kappa := \lambda_{\max}/\lambda_{\min}$.

Condition *(i)* reduces the problem geometry to a form where closed-form analysis is possible: $\Xi = \mathbb{R}^k$ collapses the WDRO problem to a regularized SAA problem with a Euclidean norm penalty, and $\mathcal{X}(\Xi) = \mathbb{R}^n$ forces $\boldsymbol{x}(\infty) = \mathbf{0}$, so the shrinkage path reduces to the segment $\{\lambda \boldsymbol{x}(0) : \lambda \in [0,1]\}$. Condition *(ii)* is consistent with Assumption 1*(ii)*, provided L extends to a globally Lipschitz function on $\mathbb{R}$. This is satisfied, for example, by the Huber loss $L(z) = z^2$ for $|z| \leq \delta$ and $L(z) = 2\delta|z| - \delta^2$ otherwise: as long as $\mathcal{Z} \subseteq [-\delta, \delta]$, the loss is quadratic on $\mathcal{Z}$ and globally Lipschitz with constant 2δ. Our results readily extend to losses with a continuous, bounded second derivative on $\mathcal{Z}$, with $\lambda_{\min}$ and $\lambda_{\max}$ in the bounds replaced by the local strong convexity and smoothness constants of L on $\mathcal{Z}$; we omit the details for brevity.

The following theorem refines the additive bound of Proposition 1 under the additional structure of Assumption 2.

THEOREM 2 **(Additive Suboptimality; Quadratic Loss)**. *Let Assumption 2 hold. The worst-case additive suboptimality of the shrinkage path heuristic is bounded above by*

$$\lambda_{\max} \|\boldsymbol{x}(0)\|_2^2 \frac{(\kappa - 1)^2}{\kappa\,[\varphi^5(\kappa - 1) + 27]}, \tag{5}$$

where $\varphi = (\sqrt{5} + 1)/2$ is the golden ratio. The bound is asymptotically tight as $\kappa \to 1^+$ and $\kappa \to \infty$.

The bound (5) factors into two terms: $\lambda_{\max}\|\boldsymbol{x}(0)\|_2^2$ and a κ-dependent factor. The latter behaves like $(\kappa - 1)^2/27$ to leading order as $\kappa \to 1^+$, so the shrinkage path is provably near-optimal whenever the empirical second-moment matrix is well-conditioned; it increases monotonically with κ and approaches $1/\varphi^5 \approx 0.0902$ in the severely ill-conditioned limit $\kappa \to \infty$. The appearance of the golden ratio reflects the limiting worst-case geometry, described in the following corollary.

COROLLARY 1 **(Worst-case Geometry)**. *Let Assumption 2 hold, and let ρ_{gap}, ρ_0^{h}, and $\rho_0^\star$ denote, respectively, a radius at which the suboptimality gap is maximized, the radius at which the shrinkage path solution reaches $\mathbf{0}$, and the smallest radius at which the WDRO solution reaches $\mathbf{0}$. As $\kappa \to \infty$, the worst-case bound in Theorem 2 is approached by instances for which*

$$\rho_{\text{gap}} : \rho_0^{\text{h}} : \rho_0^\star \;\to\; 1 : \varphi : \varphi^2.$$

The corollary describes the configuration that produces the worst case of Theorem 2, in which the two forms of shrinkage are hardest to reconcile. The WDRO problem shrinks eigendirections at non-uniform rates—in the limiting worst case, the $\lambda_{\min}$ direction collapses to zero while the $\lambda_{\max}$ direction shrinks by a factor of $1/\varphi$—whereas the shrinkage path heuristic applies a single uniform factor across all coordinates. The heuristic incurs its largest gap at the configuration where this mismatch is most costly; there, the uniform factor $1/\varphi^2$ and the dominant per-coordinate factor

$1/\varphi$ sum to one: $\frac{1}{\varphi} + \frac{1}{\varphi^2} = 1$. Thus, no single direction is responsible for the bound: it arises because one common shrinkage factor cannot match the WDRO solution in directions of very different magnitude.

We now turn to the multiplicative suboptimality of our heuristic.

THEOREM 3 **(Multiplicative Suboptimality)**. *Let Assumption 2 hold. The multiplicative suboptimality of the shrinkage path heuristic relative to the SAA–RO heuristic is bounded above by*

$$\frac{(\kappa-1)^2(\kappa+1)}{(\kappa^{3/2}+1)^2}, \tag{6}$$

and the bound is strictly less than one whenever $\kappa \geq 1$.

The bound (6) equals zero at $\kappa = 1$ and increases monotonically to one as $\kappa \to \infty$, so the shrinkage path offers the largest relative improvement over the SAA–RO heuristic for well- to moderately conditioned data. The bound is generally not tight: it relaxes the worst-case suboptimality of the SAA–RO heuristic to a lower bound, and a smaller denominator enlarges the ratio.

REMARK 2 (CONNECTION TO JAMES–STEIN SHRINKAGE). Under Assumption 2, the solution of the shrinkage path heuristic admits the closed form

$$\hat{\boldsymbol{x}}(\rho) = \left[1 - \frac{\rho\,\|\boldsymbol{x}(0)\|_2}{2\,\boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0)}\right]_+ \boldsymbol{x}(0),$$

a non-negative scalar multiple of the SAA solution, with the multiplier shrinking from one toward zero as ρ grows. This is formally reminiscent of the positive-part James–Stein estimator $\hat{\boldsymbol{\theta}}^{JS+} = [1 - (n-2)/\|\boldsymbol{Y}\|_2^2]_+ \boldsymbol{Y}$ (James and Stein 1961, Baranchik 1970). In the isotropic case $\boldsymbol{H} = \boldsymbol{I}$, our shrinkage factor reduces to $1 - \rho/(2\|\boldsymbol{x}(0)\|_2)$; identifying $\boldsymbol{Y}$ with $\boldsymbol{x}(0)$, the heuristic and James–Stein shrinkage factors then coincide at $\rho = 2(n-2)/\|\boldsymbol{x}(0)\|_2$. The key difference is that the degree of shrinkage is governed by the ambiguity radius ρ rather than the sample size, suggesting that our shrinkage path heuristic implements an ambiguity-driven analogue of classical shrinkage regularization.

4. A Posteriori Performance Guarantees

We next develop computable a posteriori certificates for the suboptimality of our shrinkage path heuristic. Since the shrinkage path is a subset of the WDRO feasible set, the objective value $J^\star$ returned by Algorithm 1 *upper* bounds the optimal value of Problem (1). This section complements our upper bound with a matching *lower* bound. Our construction mirrors the primal shrinkage path on the dual side: a minimax interchange first recasts the optimal value of Problem (1) as a maximization over distributions $\mathbb{Q}$ in the Wasserstein ball $\mathbb{B}_\rho(\hat{\mathbb{P}})$, and this maximization is then restricted to the line segment joining the empirical distribution $\hat{\mathbb{P}}$ to a worst-case distribution at $\rho = \infty$. To secure the minimax interchange and the existence of a worst-case distribution at $\rho = \infty$, this section imposes the following regularity assumption on Problem (1).

ASSUMPTION 3 **(A Posteriori Regularity)**. *Problem* (1) *satisfies the following conditions.*

(i) The SAA objective $\boldsymbol{x} \mapsto \mathbb{E}_{\hat{\mathbb{P}}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big]$ *is coercive on* $\mathcal{X}$.

(ii) There exist a neighborhood $\mathcal{U}$ *of the RO solution* $\boldsymbol{x}(\infty)$ *and a compact set* $\mathcal{K} \subseteq \Xi$ *such that*

$$\sup_{\boldsymbol{\xi} \in \Xi} \ell(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{\boldsymbol{\xi} \in \mathcal{K}} \ell(\boldsymbol{x}, \boldsymbol{\xi}) \qquad \forall \boldsymbol{x} \in \mathcal{U} \cap \mathcal{X}.$$

Recall that a function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is *coercive* on $\mathcal{X}$ if all its sublevel sets on $\mathcal{X}$ are bounded, or equivalently, if $f(\boldsymbol{x}_k) \to +\infty$ along every sequence $\{\boldsymbol{x}_k\} \subseteq \mathcal{X}$ with $\|\boldsymbol{x}_k\| \to \infty$. Condition *(i)* secures the minimax interchange between the decisions $\boldsymbol{x}$ and the distributions $\mathbb{Q}$ in Problem (1), while Condition *(ii)* guarantees the existence of a worst-case distribution at $\rho = \infty$ that is supported on finitely many atoms. The latter condition holds, for example, whenever Ξ is compact.

Denote by $J^\star(\rho)$ the optimal value of Problem (1) at radius ρ. We henceforth denote by $\overline{J}(\rho)$ the optimum of Problem (1) restricted to the full shrinkage path $\widehat{\mathcal{X}}(\Xi)$, with the shrinkage parameter λ ranging over the full interval $[0, 1]$. In the notation of Section 3, $J^\star(\rho) = f_\rho^\star$ is the WDRO optimum and $\overline{J}(\rho) = \hat{f}_\rho$ is the value of the shrinkage path heuristic. The objective value returned by Algorithm 1 is a finite-grid approximation of $\overline{J}(\rho)$, since Algorithm 1 searches only over the grid $\Lambda \subset [0, 1]$. The full-path optimum itself is given by

$$\overline{J}(\rho) \coloneqq \min_{\boldsymbol{x} \in \widehat{\mathcal{X}}(\Xi)} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big] \geq J^\star(\rho),$$

where the inequality follows from $\widehat{\mathcal{X}}(\Xi) \subseteq \mathcal{X}(\Xi)$. To certify the resulting approximation gap $\overline{J}(\rho) - J^\star(\rho)$ without access to $J^\star(\rho)$, we construct a matching *lower* bound $\underline{J}(\rho)$ via a dual shrinkage path that mirrors the primal one. The construction is motivated by the minimax interchange

$$J^\star(\rho) = \min_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big] = \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}})} \inf_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big], \tag{7}$$

which holds for every $\rho \in \mathbb{R}_+ \cup \{+\infty\}$ under Assumption 3*(i)* (Kuhn et al. 2025, Corollary 5.16). Note that the inner infimum on the right-hand side may not be attained because, for a fixed $\mathbb{Q}$, the map $\boldsymbol{x} \mapsto \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big]$ need not be coercive. This min–sup = sup–inf identity instantiates the *primal worst equals dual best* principle (Beck and Ben-Tal 2009, Zhen et al. 2025); the interchange and the existence of an associated worst-case (least-favorable) distribution are well studied in this setting (Blanchet and Murthy 2019).

The dual shrinkage path is the line segment between the two saddle distributions at $\rho = 0$ and $\rho = \infty$. At $\rho = 0$, the Wasserstein ball $\mathbb{B}_\rho(\hat{\mathbb{P}})$ collapses to the empirical distribution $\hat{\mathbb{P}}$. At $\rho = \infty$, we fix any optimizer of the dual maximization in (7) as $\rho \to \infty$, namely

$$\mathbb{Q}(\infty) \in \arg\max_{\mathbb{Q} \in \mathcal{P}(\Xi)} \inf_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big], \tag{8}$$

and call it the *RO distribution* $\mathbb{Q}(\infty)$, paralleling the RO solution $\boldsymbol{x}(\infty)$ on the primal side. The GitHub companion establishes that under Assumption 3, such an optimizer exists and can be chosen to be supported on at most $n+1$ atoms. Fixing $\mathbb{Q}(\infty)$ in turn fixes the dual shrinkage path; although both are determined only up to the choice of optimizer, we henceforth speak of *the* RO distribution and *the* dual shrinkage path. The latter is defined as

$$\widehat{\mathcal{B}}(\hat{\mathbb{P}}) \coloneqq \big\{(1-\lambda)\hat{\mathbb{P}} + \lambda\mathbb{Q}(\infty) : \lambda \in [0,1]\big\}.$$

The primal and dual shrinkage paths are symmetric: the primal path $\widehat{\mathcal{X}}(\Xi)$ is the line segment between the SAA solution $\boldsymbol{x}(0)$ and the RO solution $\boldsymbol{x}(\infty)$, while the dual path $\widehat{\mathcal{B}}(\hat{\mathbb{P}})$ is the line segment between the empirical distribution $\hat{\mathbb{P}}$ and the RO distribution $\mathbb{Q}(\infty)$. By the minimax interchange (7), the endpoints of the two paths pair up: $(\boldsymbol{x}(0), \hat{\mathbb{P}})$ at $\rho = 0$ and $(\boldsymbol{x}(\infty), \mathbb{Q}(\infty))$ at $\rho = \infty$. Both pairs are in fact Nash equilibria of the game in which the decision maker minimizes and nature maximizes $\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$: at $\rho = 0$ the ambiguity set collapses to the singleton $\{\hat{\mathbb{P}}\}$, leaving neither player a profitable deviation, while at $\rho = \infty$ the minimax interchange (7) equates the min–sup and max–inf values, so their outer optimizers $\boldsymbol{x}(\infty)$ and $\mathbb{Q}(\infty)$ are mutual best responses.

Restricting the dual maximization in (7) from the Wasserstein ball $\mathbb{B}_\rho(\hat{\mathbb{P}})$ to its intersection with the dual shrinkage path $\widehat{\mathcal{B}}(\hat{\mathbb{P}})$ yields the matching a posteriori lower bound

$$\underline{J}(\rho) \coloneqq \max_{\mathbb{Q} \in \widehat{\mathcal{B}}(\hat{\mathbb{P}}) \cap \mathbb{B}_\rho(\hat{\mathbb{P}})} \inf_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big] \;\le\; J^\star(\rho). \tag{9}$$

The a posteriori gap $\overline{J}(\rho) - \underline{J}(\rho)$ then provides a suboptimality certificate for the shrinkage path heuristic in Algorithm 1.

The lower bound $\underline{J}(\rho)$ admits a reduction to a finite-scenario stochastic optimization problem.

THEOREM 4 **(Dual Shrinkage Closed Form)**. *Whenever $\hat{\mathbb{P}} \neq \mathbb{Q}(\infty)$, the a posteriori lower bound* (9) *admits the finite-scenario reformulation*

$$\underline{J}(\rho) = \min_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \mathbb{E}_{\widehat{\mathbb{Q}}(\rho)}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big] \qquad \textit{with} \qquad \widehat{\mathbb{Q}}(\rho) = \mu(\rho)\hat{\mathbb{P}} + \big(1-\mu(\rho)\big)\mathbb{Q}(\infty) \tag{10}$$

under the mixture $\mu(\rho) \coloneqq \big[1 - \rho / W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))\big]_+$. In the degenerate case where $\hat{\mathbb{P}} = \mathbb{Q}(\infty)$, the dual shrinkage path collapses to $\{\hat{\mathbb{P}}\}$, and (10) *holds with $\widehat{\mathbb{Q}}(\rho) = \hat{\mathbb{P}}$ for every $\rho \ge 0$.*

Theorem 4 identifies $\widehat{\mathbb{Q}}(\rho)$ as the distribution on the dual shrinkage path that lies as far from $\hat{\mathbb{P}}$ as the radius permits: it sits on the boundary of the Wasserstein ball, $W(\widehat{\mathbb{Q}}(\rho), \hat{\mathbb{P}}) = \rho$, whenever $\rho \le W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))$, and coincides with $\mathbb{Q}(\infty)$ otherwise; equivalently, $W(\widehat{\mathbb{Q}}(\rho), \hat{\mathbb{P}}) = \min\{\rho, W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))\}$. Geometrically, the dual shrinkage path $\widehat{\mathcal{B}}(\hat{\mathbb{P}})$ stays fixed while the Wasserstein ball $\mathbb{B}_\rho(\hat{\mathbb{P}})$ expands with ρ, so $\widehat{\mathbb{Q}}(\rho)$ slides along the path at the ball's boundary until it reaches $\mathbb{Q}(\infty)$. Like the primal

shrinkage path, the dual path does not vary with ρ: evaluating $\underline{J}(\rho)$ requires only a one-off RO distribution construction, a one-off Wasserstein-distance evaluation between two finitely supported distributions, and a single finite-scenario stochastic optimization per radius, as summarized in Algorithm 2. Sweeping a grid of radii therefore traces the entire lower-bound curve at negligible additional cost per radius. Together with $\overline{J}(\rho)$, this brackets $J^\star(\rho)$ from above and below, and so certifies the quality of the heuristic at every radius without solving Problem (1) exactly.

Algorithm 2 A Posteriori Lower Bound Computation

Require: Empirical law $\hat{\mathbb{P}} := \frac{1}{N}\sum_{i=1}^N \delta_{\hat{\boldsymbol{\xi}}_i}$ with samples $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^N$; radius $\rho \geq 0$.

1: Construct the RO distribution $\mathbb{Q}(\infty)$ as defined in (8).

2: Compute $W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))$.

3: Form the dual shrinkage distribution $\widehat{\mathbb{Q}}(\rho)$ according to (10).

4: Solve $\underline{J}(\rho) \leftarrow \min_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\widehat{\mathbb{Q}}(\rho)}\big[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\big]$.

5: **return** $\underline{J}(\rho)$.

The following theorem establishes further structural properties shared by $\overline{J}(\rho)$, $J^\star(\rho)$, and $\underline{J}(\rho)$.

THEOREM 5 **(A Posteriori Guarantee)**. *The a posteriori upper bound $\overline{J}(\rho)$, the WDRO optimum $J^\star(\rho)$, and the a posteriori lower bound $\underline{J}(\rho)$ satisfy:*

(i) **Monotonicity:** *All three values are nondecreasing in ρ.*

(ii) **Concavity:** *All three values are concave in ρ.*

(iii) **Endpoint tightness:** $\overline{J}(\rho) = J^\star(\rho) = \underline{J}(\rho)$ *at $\rho = 0$ and for every $\rho \geq W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))$.*

Theorem 5 shows that the a posteriori bounds and the exact WDRO value share the same shape: all three start from $J(0)$ (the optimal value of the SAA problem (3a)), rise concavely with ρ, and saturate at $J(\infty)$ (the optimal value of the RO problem (3b)) by $\rho = W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))$.

Algorithm 2 requires constructing the RO distribution $\mathbb{Q}(\infty)$, defined in (8). The following proposition provides three such constructions, with details deferred to Section EC.1.

PROPOSITION 2 **(RO Distribution Construction)**. *Under the standing assumptions, a finitely supported RO distribution $\mathbb{Q}(\infty)$ can be recovered from a finite-dimensional convex problem in each of the following settings.*

(i) Convex loss, polyhedral support. *The support Ξ is a pointed polyhedron, and $\ell(\boldsymbol{x}, \cdot)$ is convex on Ξ with nonpositive recession along every extreme ray for every $\boldsymbol{x} \in \mathcal{X}(\Xi)$.*

(ii) Finite-max refinement. *In addition to (i), $\mathcal{X}(\Xi)$ is polyhedral and $\ell = \max_{l\in[L]} \ell_l$, with each $\ell_l(\cdot, \boldsymbol{\xi})$ convex and differentiable. The resulting distribution has at most $n+1$ atoms.*

(iii) Convex–concave loss. *The support* Ξ *is compact and* $\ell = \max_{l\in[L]} \ell_l$*, where every* ℓ_l *is closed convex with full domain in* $\boldsymbol{x}$ *and concave and upper semicontinuous in* $\boldsymbol{\xi}$ *on* Ξ*.*

Construction *(i)* yields an RO distribution that is supported on the vertices of Ξ. Construction *(ii)* specializes *(i)* and obtains $\mathbb{Q}(\infty)$ from any basic feasible solution of a linear program by exploiting the finite-max differentiable structure of the loss. Construction *(iii)*, finally, recovers $\mathbb{Q}(\infty)$ from a Fenchel dual of the RO problem (3b) and applies to the finite-max convex–concave structure underlying tractable convex reformulations of Wasserstein DRO (Mohajerin Esfahani and Kuhn 2018, Shafiee et al. 2026). The constructions can differ in support size: the GitHub companion proves that an RO distribution on at most $n+1$ atoms always exists and describes the post-processing routine used to sparsify constructions *(i)* and *(iii)* to $n+1$ atoms. Construction *(ii)* meets this bound directly.

5. Numerical Experiments

In this section, we evaluate the effectiveness of our shrinkage path heuristic in both a multi-item newsvendor (Section 5.1) and an appointment scheduling setting (Section 5.2). All experiments were conducted on a high-performance computing environment, which hosts AMD EPYC 7742 processors with 64 GB of RAM and runs Python 3.11, CVXPY 1.7.3, MOSEK 11.0.29, and Gurobi 12.0.3. All reported runtimes correspond to single-threaded execution. All solver parameters are set to their default values except where explicitly stated otherwise. We evaluate our shrinkage heuristic along three dimensions: the quality of its a posteriori guarantees, its out-of-sample performance compared to unregularized and regularized SAA (Kleywegt et al. 2002), Wasserstein DRO (Kuhn et al. 2019), and Sinkhorn DRO (Wang et al. 2026), and its scalability relative to these methods. Supplementary implementation details are provided in the GitHub companion.

5.1. Multi-Item Newsvendor

Consider the multi-item newsvendor problem, a classical inventory control model in which the decision maker selects order quantities $\boldsymbol{x}$ under uncertain demand $\tilde{\boldsymbol{\xi}}$ supported on the box $\Xi = [\mathbf{0}, \boldsymbol{U}]$ for some $\boldsymbol{U} \in \mathbb{R}^n_{++}$, and then incurs holding costs for excess inventory and backlogging costs for unmet demand (Zipkin 2000). Given an empirical distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_{j=1}^N \delta_{\hat{\boldsymbol{\xi}}_j}$ and a type-1 Wasserstein ball with radius ρ and ℓ_2 ground metric, we benchmark the performance of different methods for (approximately) solving the WDRO newsvendor problem (Mohajerin Esfahani and Kuhn 2018, Gao and Kleywegt 2023)

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \sup\limits_{\mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\big[\boldsymbol{h}^\top[\boldsymbol{x}-\tilde{\boldsymbol{\xi}}]_+ + \boldsymbol{b}^\top[\tilde{\boldsymbol{\xi}}-\boldsymbol{x}]_+\big] \\ \text{subject to} & \mathbf{0} \le \boldsymbol{x} \le \boldsymbol{V}. \end{array} \tag{11}$$

where $[\cdot]_+ := \max\{\cdot, 0\}$ is applied componentwise, and $\boldsymbol{h}, \boldsymbol{b}, \boldsymbol{V} \in \mathbb{R}^n_{++}$ denote the marginal holding costs, the marginal backlogging costs, and the order capacities, respectively.

The newsvendor problem is a useful diagnostic because both endpoints of the shrinkage path are available in closed form, while the full WDRO problem remains practically tractable. The *critical ratio* $q_i := b_i/(h_i + b_i) \in (0,1)$—the cost-optimal service level (order fractile) of the classical newsvendor—increases in the backlogging cost b_i and decreases in the holding cost h_i, so costlier shortages call for higher order quantities. Let $\widehat{\xi}_{(k),i}$ denote the kth order statistic of the ith demand coordinate, i.e., the kth smallest of its N samples. Then, at $\rho = 0$, the SAA endpoint is the empirical q_i-quantile of demand—the cost-optimal service level of the nominal newsvendor, clipped at capacity, $x_i(0) = \min\{\widehat{\xi}_{(\lceil q_i N \rceil),i}, V_i\},\ i \in [n]$. As $\rho \to \infty$, the WDRO problem reduces to the robust problem $\min_{\boldsymbol{x} \in [\boldsymbol{0}, \boldsymbol{V}]} \sum_{i=1}^n \max\{h_i x_i, b_i(U_i - x_i)\}$, whose solution is $x_i(\infty) = \min\{q_i U_i, V_i\},\ i \in [n]$. Finally, for finite ρ, Problem (11) admits an exact conic reformulation (Mohajerin Esfahani and Kuhn 2018, Corollary 5.1): the loss can be written as the maximum of 2^n affine functions of $\boldsymbol{\xi}$, and the box support and ℓ_2 ground metric yield a second-order cone program. For completeness, we provide this SOC reformulation in the GitHub companion.

Throughout this section, all experiments draw from a common master instance of $n = 20$ items. We generate the cost and support parameters by tiling a length-five pattern four times: the holding costs are $h_i = 1$ for all i, the backlogging costs $\boldsymbol{b}$ and support upper bounds $\boldsymbol{U}$ cycle through $(2, 6, 8, 10, 10)$ and $(10, 8, 6, 4, 2)$ respectively, and the order capacities are $V_i = 0.6\, U_i$; each pattern thus repeats across the four consecutive blocks of five items. Demands are sampled independently as $\tilde{\xi}_i = U_i \tilde{Z}_i$ with $\tilde{Z}_i \sim \mathrm{Beta}(\alpha_i, \beta_i)$, where (α_i, β_i) cycles through $(0.2, 2), (0.2, 1), (0.2, 3), (0.4, 2), (0.3, 2)$, and the empirical distribution $\hat{\mathbb{P}}$ is built from $N = 15$ i.i.d. samples. Experiments that use fewer than 20 items take the data of the first n items.

We first benchmark our a posteriori guarantees with $n = 15$ items, evaluating the worst-case expected loss along a logarithmic grid of 30 values of ρ between 10^{-2} and 10^2. Figure 2 (left) plots the a posteriori envelopes of Section 4 and the exact WDRO objective and demonstrates that our shrinkage heuristic is near-optimal across all radii and exact at the two endpoints. In addition, the right panel of Figure 2 reports the average per-ρ runtime of the primal approximation, the dual approximation, and the exact WDRO solve.

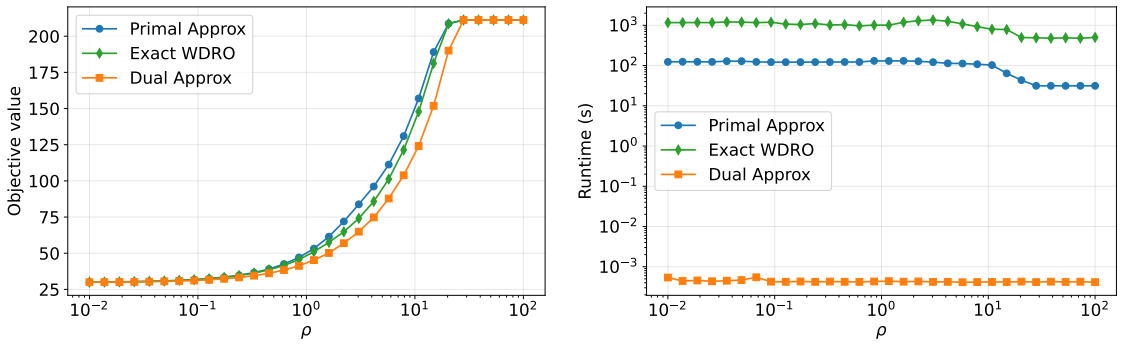

Figure 2 A posteriori upper bound (primal approximation), exact WDRO objective, and a posteriori lower bound (dual approximation) (left panel), and runtimes (right panel) for the 15-item newsvendor problem.

Next, we evaluate the out-of-sample performance of the following methods. First, we include SAA. Second, we include two regularized SAA benchmarks, "ℓ_1-RSAA-CV" and "ℓ_2-RSAA-CV", which solve $\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{j=1}^{N} \ell(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) + \nu \|\boldsymbol{x}\|_p^p$ for $p=1$ and $p=2$, respectively, where $\nu \geq 0$ is a regularization weight tuned by five-fold cross-validation over 25 values logarithmically spaced between 10^{-3} and 10^2 and augmented with $\nu = 0$. In view of the well-documented connections between Wasserstein DRO and norm regularization (Shafieezadeh-Abadeh et al. 2019, Gao et al. 2024), these benchmarks test whether shrinking the decision per se already confers the benefits of distributional robustness, or whether the direction of shrinkage toward the RO solution matters. Third, we solve the exact WDRO reformulation introduced above, with ρ tuned by five-fold cross-validation over 30 values logarithmically spaced between 0.1 and 15 and augmented with $\rho = 0$; we refer to this approach as "WDRO-CV". We additionally solve the exact WDRO problem with the cutting-plane algorithm described in the GitHub companion, denoted "WDRO-CV (CP)"; since it returns the same solutions as WDRO-CV, we report its out-of-sample performance only for instances that WDRO-CV cannot solve within our runtime and memory budget, but report its runtime for all n. Fourth, we implement the Sinkhorn DRO method of Wang et al. (2026), which we refer to as "SDRO-CV", with the nominal radius $\bar{\rho}$ and entropic strength ϵ jointly tuned by five-fold cross-validation over the $22 \times 14 = 308$ candidate $(\bar{\rho}, \epsilon)$ pairs of their default newsvendor configuration (implementation details are provided in the GitHub companion). Fifth, we evaluate our shrinkage path heuristic, with λ tuned by five-fold cross-validation over 30 uniformly spaced values in $[0,1]$; we refer to this approach as "Shrinkage-CV". Sixth, "SAA-RO-CV" uses five-fold cross-validation to choose the better of the SAA and RO decisions. Finally, we compute oracle counterparts of the shrinkage heuristic and of exact WDRO, with λ and ρ tuned directly on the test samples; these oracles, "Shrinkage-Oracle" and "WDRO-Oracle", give the best out-of-sample performance each approach can attain.

When the training and test data are drawn from the same distribution, SAA is already a strong benchmark for the data-driven newsvendor problem. The value of distributional robustness surfaces primarily under a distribution shift. Accordingly, to assess robustness to such a shift between training and deployment, the test data follow the coordinatewise contaminated mixture of Hanasusanto et al. (2015),

$$\mathbb{P}_{\text{test}} = \bigotimes_{i=1}^{n} \Big[(1-c)\mathbb{P}_{\text{train},i} + c\text{Unif}[0, U_i] \Big],$$

where $\mathbb{P}_{\text{train},i}$ is the ith training marginal, $\text{Unif}[0, U_i]$ is the uniform distribution on $[0, U_i]$, and the contamination level is $c = 0.2$. We measure out-of-sample performance through the suboptimality of each method, that is, the percentage by which the expected cost of its decision on 50,000 test samples exceeds the smallest expected cost attainable by any feasible decision under the same samples.

Figure 3 reports the out-of-sample suboptimality and average runtime across 100 random seeds, as the number of items varies over $n \in \{8, 10, 12, 15, 20\}$ with the sample size fixed at $N = 15$. The left panel reports out-of-sample suboptimality. The boxes span interquartile ranges, the whiskers reach the most extreme seeds within 1.5 interquartile ranges of the boxes, and seeds beyond the whiskers are omitted. SAA and the two regularized SAA benchmarks perform worst across all problem sizes. Both ℓ_1-RSAA-CV and ℓ_2-RSAA-CV underperform SAA; since $\nu = 0$ belongs to the tuning grid, we attribute this shortfall to cross-validation error under the small sample size $N = 15$. Wasserstein DRO and Sinkhorn DRO are marginally better than the shrinkage path heuristic for $n \leq 15$. At $n = 20$, the shrinkage path heuristic outperforms the cutting-plane variant of Wasserstein DRO but remains marginally worse than Sinkhorn DRO. The SAA-RO endpoint heuristic (SAA-RO-CV) also performs among the worst methods across all problem sizes. Overall, for the multi-item newsvendor problem, the shrinkage path heuristic captures around 85%–110% of the median out-of-sample benefit of Wasserstein DRO over SAA. The percentages are computed from the median relative suboptimalities reported in Figure 3: at each n, the gap between SAA and the shrinkage path heuristic is divided by the gap between SAA and Wasserstein DRO, and the range is taken over n.

The right panel of Figure 3 reports the runtimes for each method. Here, the (barely visible) shaded bands cover one standard deviation on either side of the mean runtime. Results for Wasserstein DRO and Sinkhorn DRO are reported only when the corresponding solve finishes within the 72-hour runtime budget. The exact reformulation of Wasserstein DRO exhausts our 64 GB memory budget at $n = 20$, where only its cutting-plane variant remains tractable. The shrinkage path heuristic runs three orders of magnitude faster than either Wasserstein DRO variant or Sinkhorn DRO. These computational savings allow the heuristic to achieve its competitive out-of-sample performance at a fraction of the computational cost and make it especially attractive in large-scale settings.

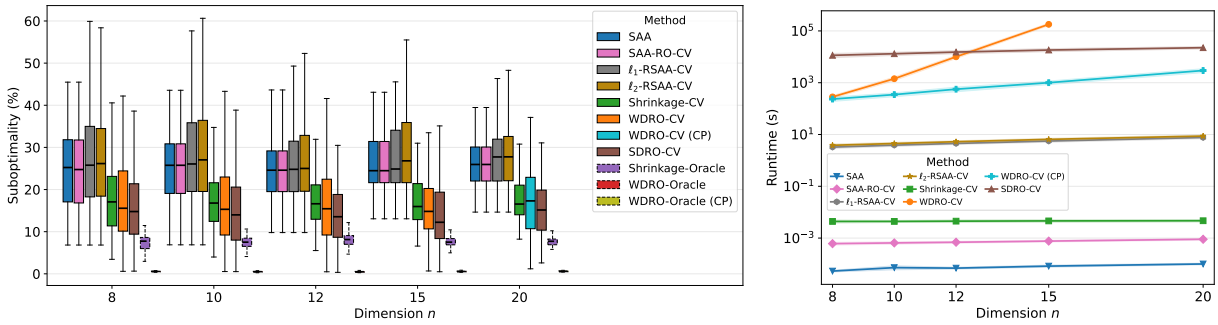

Figure 3 Out-of-sample suboptimality (left) and average runtime (right) of the different methods on the multi-item newsvendor problem with $N = 15$ samples, as the number of items varies over $n \in \{8, 10, 12, 15, 20\}$. The shrinkage path heuristic captures most of the out-of-sample benefit of Wasserstein or Sinkhorn DRO over SAA, at orders of magnitude less computational cost.

5.2. Appointment Scheduling

We consider an appointment scheduling problem in the spirit of Denton and Gupta (2003) and Jiang et al. (2019). A service provider schedules n consecutive appointment slots within a session of maximum duration D and must fix the slot lengths $\boldsymbol{x} \in \mathbb{R}^n_+$ before observing the uncertain service durations $\tilde{\boldsymbol{\xi}}$. The durations are supported on the box $\Xi = [\underline{\boldsymbol{\xi}}, \overline{\boldsymbol{\xi}}]$. Given the empirical distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_{j=1}^N \delta_{\hat{\boldsymbol{\xi}}_j}$, we study the type-1 Wasserstein DRO problem with ℓ_2 ground metric

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \sup\limits_{\mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\left[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \\ \text{subject to} & \boldsymbol{x} \geq \mathbf{0}, \quad \sum\limits_{t\in[n]} x_t \leq D, \end{array} \tag{12a}$$

where $\ell(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of the second-stage problem

$$\begin{array}{lll} \underset{\boldsymbol{w},\boldsymbol{s},o}{\text{minimize}} & \sum\limits_{t=2}^{n} \left(c_w w_t^2 + c_s s_t^2\right) + c_o o^2 & \\ \text{subject to} & w_1 = s_1 = 0, & \\ & w_{t+1} \geq w_t + \xi_t - x_t, \quad w_{t+1} \geq 0 & \forall t \in [n-1], \\ & s_{t+1} \geq x_t - w_t - \xi_t, \quad s_{t+1} \geq 0 & \forall t \in [n-1], \\ & o \geq w_n + \xi_n - x_n, \qquad o \geq 0. & \end{array} \tag{12b}$$

Here, the recourse variables w_t and s_t denote, respectively, the waiting time of patient t and the server's idle time immediately before appointment t, while o is the overtime beyond the scheduled session end $\sum_{t\in[n]} x_t$. The constants $c_w, c_s, c_o > 0$ are the penalty weights on the squared waiting times, idle times, and overtime. Since the second-stage objective is jointly convex in $(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{w}, \boldsymbol{s}, o)$ and the constraints are jointly affine, $\ell(\boldsymbol{x}, \boldsymbol{\xi})$ is jointly convex in $(\boldsymbol{x}, \boldsymbol{\xi})$.

Unlike the newsvendor loss, which is a maximum of finitely many affine functions of $\boldsymbol{\xi}$, the appointment scheduling loss $\ell(\boldsymbol{x}, \boldsymbol{\xi})$ is a maximum of *infinitely many* affine functions. The extreme-point reduction that would render it finite fails because the second-stage dual is concave-quadratic, rather than linear, in its multipliers (cf. the GitHub companion). Since no finite, exact, convex reformulation is available, we solve the exact WDRO problem using the same cutting-plane algorithm as in the previous subsection, which we denote "WDRO-CP". As the worst-case evaluation oracle, WC-EVAL$(\cdot;\rho)$, (cf. Section 2) does not admit an exact finite reformulation either, we approximate each oracle call, including the separation problems within WDRO-CP, by a multi-start alternating maximization procedure (see the GitHub companion for more details). The shrinkage path endpoints, by contrast, remain practically tractable. Writing $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n_+ : \sum_{t=1}^n x_t \leq D\}$, the SAA and RO endpoints solve the convex quadratically constrained programs

$$\boldsymbol{x}(0) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X},\ \boldsymbol{\tau}\in\mathbb{R}^N} \left\{ \tfrac{1}{N}\textstyle\sum_{j=1}^N \tau_j \ :\ \ell(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_j) \leq \tau_j \quad \forall j \in [N] \right\},$$
$$\boldsymbol{x}(\infty) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X},\ \tau\in\mathbb{R}} \left\{ \tau \ :\ \ell(\boldsymbol{x}, \boldsymbol{v}_k) \leq \tau \quad \forall k \in [2^n] \right\},$$

where $\boldsymbol{v}_k$ ranges over the vertices of Ξ. Optimizing over these vertices suffices because $\boldsymbol{\xi} \mapsto \ell(\boldsymbol{x}, \boldsymbol{\xi})$ is convex, so its maximum over the box Ξ is attained at a vertex. Since enumerating all 2^n vertices is impractical, we compute $\boldsymbol{x}(\infty)$ using a cutting-plane method. In particular, we solve a relaxation over a subset of the vertex constraints and add the most violated one at each iteration, until no vertex constraint is violated.

Throughout this section, all experiments draw from a common master instance of $n = 20$ appointments. We set the support bounds to $\underline{\boldsymbol{\xi}} = \mathbf{0}$ and $\overline{\boldsymbol{\xi}} = (2, \ldots, 2)$, the session length to $D = \sum_{i=1}^n (\underline{\xi}_i + \overline{\xi}_i)/2$, and the penalty weights to $c_w = 1.0$ and $c_s = c_o = 0.5$. Service durations are sampled as $\tilde{\xi}_t = \overline{\xi}_t \tilde{Z}_t$, where each $\tilde{Z}_t \in [0,1]$ follows a two-component Beta mixture. Its shape parameters (α_t^s, β_t^s) cycle through $\{(0.2,2),(0.2,1),(0.2,3),(0.4,2),(0.3,2)\}$, with $\alpha_t^s < \beta_t^s$ so that $\text{Beta}(\alpha_t^s, \beta_t^s)$ favors short durations. We draw $\tilde{\boldsymbol{Z}} \sim (1-c)\text{Beta}(\boldsymbol{\alpha}^s, \boldsymbol{\beta}^s) + c\text{Beta}(\boldsymbol{\beta}^s, \boldsymbol{\alpha}^s)$ with $c = 0.05$. The parameter-swapped second component favors long durations, so sessions have a 5% chance of running long. The empirical distribution $\hat{\mathbb{P}}$ is built from $N = 10$ i.i.d. samples. Experiments that use fewer than 20 appointments take the data of the first n.

We first construct the a posteriori envelopes of Section 4 for the full $n = 20$ instance and evaluate the worst-case expected loss along a geometric grid of 30 values of ρ between 10^{-3} and 30. Figure 4 reports the resulting primal and dual approximation bounds on the optimal value of (12a) and demonstrates that our shrinkage path heuristic is near-optimal across all radii and exact at the two endpoints. For comparison, it also reports the objective value of the exact WDRO problem, solved

by WDRO-CP, together with the runtime of all three methods, showing that the dual bound is two orders of magnitude cheaper than the primal.

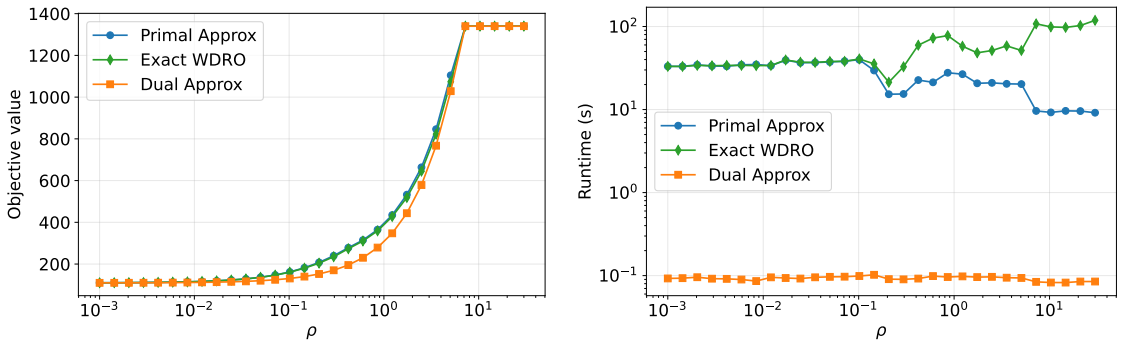

Figure 4 A posteriori upper and lower bounds (left) and average runtimes (right) with varying ρ for the appointment scheduling WDRO problem with $n = 20$ appointments and $N = 10$ samples.

Next, we compare the out-of-sample performance of our shrinkage path heuristic against the same set of methods as in the newsvendor experiment (Section 5.1), with Wasserstein DRO now solved by WDRO-CP. All hyperparameters are again tuned by five-fold cross-validation, and the number of appointments is swept over $n \in \{8, 10, 12, 15, 20\}$. The setup mirrors the newsvendor case except for the tuning grids: the shrinkage parameter λ ranges over a linear grid of 25 values in $[0, 1]$, the Wasserstein radius ρ over a logarithmic grid of 25 values in $[0.001, 5]$, and the Sinkhorn pair $(\bar{\rho}, \epsilon)$ over the 25 combinations $\bar{\rho} \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, $\epsilon \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$. Training and test samples are drawn i.i.d. from the same short/long mixture as above, with $N = 10$ held fixed.

Figure 5 (left panel) reports the out-of-sample relative suboptimality over 100 random seeds, while Figure 5 (right panel) reports the average runtime, both as functions of n. The shrinkage path heuristic outperforms SAA, both regularized SAA benchmarks, and the SAA-RO endpoint heuristic at every n, while Wasserstein DRO and Sinkhorn DRO, in turn, outperform the shrinkage path heuristic. ℓ_1-RSAA-CV and ℓ_2-RSAA-CV are nearly indistinguishable from SAA at every n, indicating that cross-validation typically selects a vanishing regularization weight; shrinking the schedule toward $\mathbf{0}$ thus confers no robustness benefit here, whereas shrinking toward the RO solution does. However, Wasserstein and Sinkhorn DRO are three to four orders of magnitude slower than the shrinkage path heuristic across all n. The appointment scheduling experiment thus complements the newsvendor one: the shrinkage path heuristic yields competitive robust decisions that capture around 45%–70% of the median out-of-sample benefits of Wasserstein DRO over SAA, at a fraction of the computational cost. These percentages follow the same convention as for Figure 3.

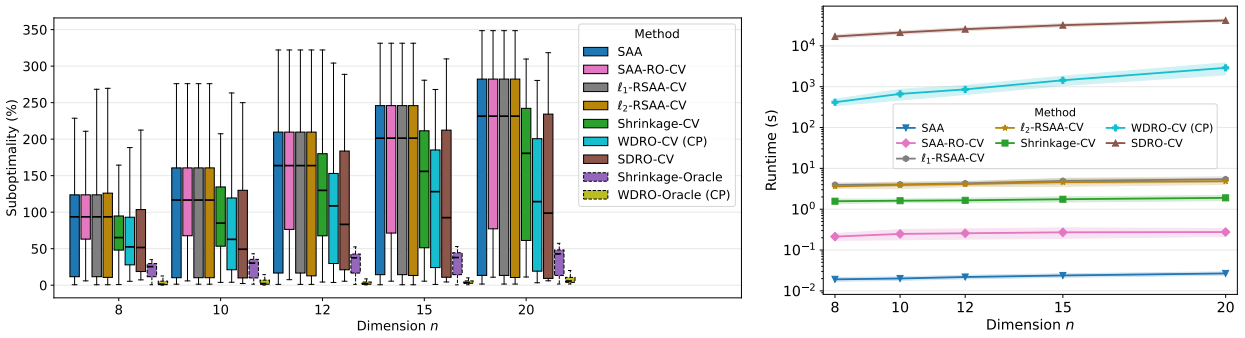

Figure 5 Out-of-sample suboptimality (left panel) and average runtime (right panel) of the different methods on the appointment scheduling problem with $N = 10$ samples, as the number of appointments varies over $n \in \{8, 10, 12, 15, 20\}$. The shrinkage path heuristic captures a significant part of the out-of-sample benefit of Wasserstein and Sinkhorn DRO over SAA, at orders of magnitude less computational cost, whereas the regularized SAA benchmarks (ℓ_1-RSAA-CV and ℓ_2-RSAA-CV) barely improve on SAA. Boxes, whiskers, and shaded bands follow the conventions of Figure 3.

Use of Large Language Models

Large language models (including Claude Fable 5 and Chat-GPT-Pro 5.6) were used to assist in writing the manuscript and implementing the numerical experiments. The authors have carefully reviewed and edited all generated content and take full responsibility for any remaining errors.

Endnotes

1. GitHub repository: `https://github.com/lingjun-meng/minimax-shrinkage`.

References

Edward Anderson and Andy Philpott. Improving sample average approximation using distributional robustness. *INFORMS Journal on Optimization*, 4(1):90–124, 2022.

Alvin J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 41(2):642–645, 1970.

Amir Beck and Aharon Ben-Tal. Duality in robust optimization: Primal worst equals dual best. *Operations Research Letters*, 37(1):1–6, 2009.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

Amine Bennouna, Ryan Lucas, and Bart Van Parys. Certified robust neural networks: Generalization and corruption resistance. In *International Conference on Machine Learning*, pages 2092–2112. PMLR, 2023.

Dimitris Bertsimas and Jack Dunn. *Machine Learning Under a Modern Optimization Lens*. Dynamic Ideas LLC, Charlestown, MA, 2019.

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

Jose H. Blanchet, Karthyek R. A. Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2): 1500–1529, 2022.

Meysam Cheramin, Jianqiang Cheng, Ruiwei Jiang, and Kai Pan. Computationally efficient approximations for distributionally robust optimization under moment and Wasserstein ambiguity. *INFORMS Journal on Computing*, 34(3):1768–1794, 2022.

Brian Denton and Diwakar Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016, 2003.

Adrián Esteban-Pérez and Juan M. Morales. Distributionally robust optimal power flow with contextual information. *European Journal of Operational Research*, 306(3):1047–1058, 2023.

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

Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

Grani A. Hanasusanto, Daniel Kuhn, Stein W. Wallace, and Steve Zymler. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, 152(1–2): 1–32, 2015.

Hao Hao and Peter Zhang. Robust paths: Geometry and computation. *arXiv preprint arXiv:2508.20039*, 2025.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, CA, 1961. University of California Press.

Ruiwei Jiang, Minseok Ryu, and Guanglin Xu. Data-driven distributionally robust appointment scheduling over Wasserstein balls. *arXiv preprint arXiv:1907.03219*, 2019.

Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, INFORMS TutORials in Operations Research, pages 130–166. INFORMS, 2019.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

Jiajin Li, Sen Huang, and Anthony Man-Cho So. A first-order algorithmic framework for distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Fengqiao Luo and Sanjay Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. *European Journal of Operational Research*, 278(1):20–35, 2019.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Almir Mutapcic and Stephen Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*, 24(3):381–406, 2009.

Panos M. Pardalos and Stephen A. Vavasis. Quadratic programming with one negative eigenvalue is NP-hard. *Journal of Global Optimization*, 1(1):15–22, 1991.

Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022. Article 4.

Soroosh Shafiee, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. Nash equilibria, regularization, and computation in optimal transport-based distributionally robust optimization. *Operations Research*, 74 (3):1689–1709, 2026.

Soroosh Shafieezadeh-Abadeh, Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Wasserstein distributionally robust Kalman filtering. *Advances in Neural Information Processing Systems*, 31, 2018.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Operations Research*, 74(3): 1581–1603, 2026.

Geoffrey S. Watson. Serial correlation in regression analysis. I. *Biometrika*, 42(3–4):327–341, 1955.

Weijun Xie and Shabbir Ahmed. Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Transactions on Power Systems*, 33(2):1860–1867, 2018.

Linwei Xin and David Alan Goldberg. Distributionally robust inventory control when demand is a martingale. *Mathematics of Operations Research*, 47(3):2387–2414, 2022.

Jianzhe Zhen, Daniel Kuhn, and Wolfram Wiesemann. A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle. *Operations Research*, 73(2):862–878, 2025.

Paul H. Zipkin. *Foundations of Inventory Management*. McGraw-Hill, Boston, 2000.

Supplementary Material

EC.1. Proofs of Main Theoretical Results

Proof of Theorem 1. We first show that the shrinkage path heuristic is exact whenever $\rho \geq \operatorname{diam}(\Xi)$. This implies both the validity and the tightness of the stated bound for all ρ. We then establish the validity of the bound for $\rho \in [0, \operatorname{diam}(\Xi)]$. Finally, we establish its pointwise tightness on the same interval by constructing, for each radius $\rho \in [0, \operatorname{diam}(\Xi)]$, an instance which attains the bound for that ρ. Throughout the proof, we denote by $f_\rho^\star$ the optimal value of the WDRO problem (1), by $\hat{f}_\rho$ the objective value of the shrinkage path heuristic, and by $f_\rho(\boldsymbol{x})$ the objective value of (1) for a fixed decision $\boldsymbol{x}$.

First, observe that if $\operatorname{diam}(\Xi) = 0$, then Ξ is a singleton and the heuristic is exact for each radius. Therefore, we assume for the remainder of the proof that $\operatorname{diam}(\Xi) > 0$. Next, for $\rho \geq \operatorname{diam}(\Xi)$, Assumption 1(i) and the definition of the Wasserstein distance then yield

$$W(\mathbb{Q}, \hat{\mathbb{P}}) \;=\; \inf_{\pi \in \Pi(\mathbb{Q}, \hat{\mathbb{P}})} \int_{\Xi \times \Xi} \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \, \pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \;\leq\; \sup_{\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi} \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \;=\; \operatorname{diam}(\Xi).$$

Hence, when $\rho \geq \operatorname{diam}(\Xi)$ the constraint $W(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho$ holds vacuously for every $\mathbb{Q} \in \mathcal{P}(\Xi)$, and $f_\rho(\boldsymbol{x}) = \sup_{\mathbb{Q} \in \mathcal{P}(\Xi)} \mathbb{E}_{\mathbb{Q}}[L(\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x})] = \sup_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\xi}^\top \boldsymbol{x})$. The WDRO problem (1) thus reduces to the RO problem (3b), whose optimal value is attained by $\boldsymbol{x}(\infty)$. Since $\boldsymbol{x}(\infty)$ lies on the shrinkage path, the shrinkage path heuristic is exact. Likewise, our suboptimality bound yields $\rho \left[1 - \frac{\rho}{\operatorname{diam}(\Xi)}\right]_+ \operatorname{Lip}(L) \|\boldsymbol{x}(0)\|_2 = 0$ when $\rho \geq \operatorname{diam}(\Xi)$, which therefore is both valid and tight.

We now prove the validity of our bound for $\rho \in [0, \operatorname{diam}(\Xi)]$. To this end, we derive an upper bound on the heuristic's value $\hat{f}_\rho$ and a matching lower bound on the optimal value $f_\rho^\star$.

As for the upper bound on the heuristic's value $\hat{f}_\rho$, we evaluate the objective $f_\rho(\boldsymbol{x})$ of (1) at the two extreme solutions $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$ and then combine them to obtain a bound. While $\boldsymbol{x}(0)$ is optimal at $\rho = 0$ (i.e., $f_0(\boldsymbol{x}(0)) = f_0^\star$), it is generally suboptimal once $\rho > 0$ (i.e., $f_\rho(\boldsymbol{x}(0)) \neq f_\rho^\star$). To bound $f_\rho(\boldsymbol{x}(0))$, we control how fast the worst-case objective can grow as the radius increases to ρ. By Kantorovich–Rubinstein duality, $W(\mathbb{Q}, \hat{\mathbb{P}}) = \sup_{g:\, \operatorname{Lip}(g) \leq 1} \left\{ \mathbb{E}_{\mathbb{Q}}[g(\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}}[g(\tilde{\boldsymbol{\xi}})] \right\}$. Hence, every Lipschitz function g satisfies $\mathbb{E}_{\mathbb{Q}}\big[g(\tilde{\boldsymbol{\xi}})\big] - \mathbb{E}_{\hat{\mathbb{P}}}\big[g(\tilde{\boldsymbol{\xi}})\big] \leq \operatorname{Lip}(g) W(\mathbb{Q}, \hat{\mathbb{P}})$. By Assumption 1(ii), the loss function L is Lipschitz continuous with constant $\operatorname{Lip}(L)$, so for any $\boldsymbol{x}$ the map $\boldsymbol{\xi} \mapsto L(\boldsymbol{\xi}^\top \boldsymbol{x})$ is Lipschitz in $\boldsymbol{\xi}$ with constant $\operatorname{Lip}(L)\|\boldsymbol{x}\|_2$. Applying the Kantorovich–Rubinstein inequality to $g(\boldsymbol{\xi}) = L(\boldsymbol{\xi}^\top \boldsymbol{x})$ and invoking $f_0(\boldsymbol{x}) = \mathbb{E}_{\hat{\mathbb{P}}}[L(\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x})]$, we obtain $\mathbb{E}_{\mathbb{Q}}[L(\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x})] \leq f_0(\boldsymbol{x}) + \operatorname{Lip}(L)\|\boldsymbol{x}\|_2 W(\mathbb{Q}, \hat{\mathbb{P}})$. Taking the supremum over all $\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}})$ with $W(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho$ yields $f_\rho(\boldsymbol{x}) \leq f_0(\boldsymbol{x}) + \rho \operatorname{Lip}(L)\|\boldsymbol{x}\|_2$. Applying the preceding inequality at $\boldsymbol{x} = \boldsymbol{x}(0)$ and observing that $f_0(\boldsymbol{x}(0)) = f_0^\star$ by construction, we have

$$f_\rho(\boldsymbol{x}(0)) \;\leq\; f_0(\boldsymbol{x}(0)) + \rho \operatorname{Lip}(L)\|\boldsymbol{x}(0)\|_2 \;=\; f_0^\star + \rho \operatorname{Lip}(L)\|\boldsymbol{x}(0)\|_2. \tag{EC.1}$$

To bound $f_\rho(\boldsymbol{x}(\infty))$, observe that for each fixed $\boldsymbol{x}$ the map $\rho \mapsto f_\rho(\boldsymbol{x})$ is nondecreasing, because the Wasserstein ball expands as ρ increases. Hence, for all $\rho \in [0, \operatorname{diam}(\Xi)]$, we have

$$f_\rho(\boldsymbol{x}(\infty)) \;\leq\; f_{\operatorname{diam}(\Xi)}(\boldsymbol{x}(\infty)) \;=\; f^\star_{\operatorname{diam}(\Xi)}, \tag{EC.2}$$

where the final equality holds because $\boldsymbol{x}(\infty)$ optimally solves the WDRO problem (1) at $\rho = \operatorname{diam}(\Xi)$. It remains to combine the bounds (EC.1) and (EC.2). Note that $f_\rho(\boldsymbol{x})$ is convex in $\boldsymbol{x}$ since it constitutes a pointwise supremum of weighted sums of the convex functions $\boldsymbol{x} \mapsto L(\boldsymbol{\xi}^\top \boldsymbol{x})$. We fix $\lambda = \frac{\rho}{\operatorname{diam}(\Xi)} \in [0,1]$ on the shrinkage path and apply Jensen's inequality to f_ρ to obtain

$$\begin{aligned}
\hat{f}_\rho \;&\leq\; f_\rho\Big(\Big(1 - \tfrac{\rho}{\operatorname{diam}(\Xi)}\Big)\boldsymbol{x}(0) + \tfrac{\rho}{\operatorname{diam}(\Xi)}\boldsymbol{x}(\infty)\Big) \\
&\leq\; \Big(1 - \tfrac{\rho}{\operatorname{diam}(\Xi)}\Big) f_\rho(\boldsymbol{x}(0)) + \tfrac{\rho}{\operatorname{diam}(\Xi)} f_\rho(\boldsymbol{x}(\infty)) \\
&\leq\; \Big(1 - \tfrac{\rho}{\operatorname{diam}(\Xi)}\Big)\Big(f^\star_0 + \rho \operatorname{Lip}(L)\|\boldsymbol{x}(0)\|_2\Big) + \tfrac{\rho}{\operatorname{diam}(\Xi)} f^\star_{\operatorname{diam}(\Xi)},
\end{aligned} \tag{EC.3}$$

where the last inequality invokes (EC.1) and (EC.2).

As for the lower bound on the optimal value $f^\star_\rho$, we exploit the concavity of $f^\star_\rho$ in ρ: a concave function lies above the chord through any two of its points, so $f^\star_\rho$ is bounded below on $[0, \operatorname{diam}(\Xi)]$ by the chord joining its values at $\rho = 0$ and $\rho = \operatorname{diam}(\Xi)$. To establish this concavity, we invoke the dual formulation of $f_\rho(\boldsymbol{x})$ (Mohajerin Esfahani and Kuhn 2018, Theorem 4.2), namely $f_\rho(\boldsymbol{x}) = \inf_{\eta \geq 0} \Big\{\eta\rho + \frac{1}{N}\sum_{i=1}^N \sup_{\boldsymbol{\xi} \in \Xi} \big[\ell(\boldsymbol{x}, \boldsymbol{\xi}) - \eta\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\|_2\big]\Big\}$. As each $f_\rho(\boldsymbol{x})$ is an infimum of functions affine in ρ, the optimal value $f^\star_\rho = \min_{\boldsymbol{x} \in \mathcal{X}(\Xi)} f_\rho(\boldsymbol{x})$ is itself a pointwise infimum of affine functions of ρ, and therefore concave in ρ. Writing any $\rho \in [0, \operatorname{diam}(\Xi)]$ as the convex combination $\rho = \big(1 - \frac{\rho}{\operatorname{diam}(\Xi)}\big)0 + \frac{\rho}{\operatorname{diam}(\Xi)}\operatorname{diam}(\Xi)$, concavity then yields

$$f^\star_\rho \;\geq\; \left(1 - \frac{\rho}{\operatorname{diam}(\Xi)}\right) f^\star_0 + \frac{\rho}{\operatorname{diam}(\Xi)} f^\star_{\operatorname{diam}(\Xi)}. \tag{EC.4}$$

Finally, subtracting the lower bound (EC.4) on $f^\star_\rho$ from the upper bound (EC.3) on $\hat{f}_\rho$, we obtain

$$\hat{f}_\rho - f^\star_\rho \;\leq\; \left(1 - \frac{\rho}{\operatorname{diam}(\Xi)}\right) \rho \operatorname{Lip}(L)\|\boldsymbol{x}(0)\|_2.$$

The bound is parabolic in ρ and is thus maximized at $\rho = \frac{\operatorname{diam}(\Xi)}{2}$.

We close by establishing the pointwise tightness of our bound for $\rho \in [0, \operatorname{diam}(\Xi)]$. As observed above, $f^\star_0 = f_0(\boldsymbol{x}(0))$ and $f^\star_{\operatorname{diam}(\Xi)} = f_{\operatorname{diam}(\Xi)}(\boldsymbol{x}(\infty))$, with both $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$ lying on the shrinkage path. Hence, the heuristic is exact—and our bound tight—at both endpoints. For each remaining radius $\bar{\rho} \in (0, \operatorname{diam}(\Xi))$, we construct an instance for which the bound is attained at $\rho = \bar{\rho}$. This instance has loss $L(z) = |z|$ (which satisfies Assumption 1(ii) with $\operatorname{Lip}(L) = 1$), a single sample $\hat{\mathbb{P}} = \delta_{\boldsymbol{\xi}^0}$, and support $\Xi = \operatorname{conv}\{\boldsymbol{\xi}^0, \boldsymbol{\xi}^1\}$, where $\boldsymbol{\xi}^0 = (0, \operatorname{diam}(\Xi))^\top$ and $\boldsymbol{\xi}^1 = (\operatorname{diam}(\Xi), \operatorname{diam}(\Xi))^\top$. The

feasible set is $\mathcal{X}(\Xi) = \operatorname{conv}\Big\{(1,0)^\top, \big(0, \frac{\bar\rho}{\operatorname{diam}(\Xi)}\big)^\top, \big(\frac{\bar\rho}{\operatorname{diam}(\Xi)}, 0\big)^\top\Big\}$. On this instance, since $\boldsymbol{x} \geq \mathbf{0}$ and $\boldsymbol{\xi} \geq \mathbf{0}$, the loss is linear: $L(\boldsymbol{\xi}^\top \boldsymbol{x}) = \boldsymbol{\xi}^\top \boldsymbol{x}$. Because $\hat{\mathbb{P}}$ is supported on a single atom, the Wasserstein distance reduces to an expectation under $\mathbb{Q} \in \mathcal{P}(\Xi)$: $W(\mathbb{Q}, \hat{\mathbb{P}}) = \mathbb{E}_{\mathbb{Q}}\big[\|\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^0\|_2\big] = \mathbb{E}_{\mathbb{Q}}\big[\tilde{\xi}_1\big]$, where the second equality follows from the support structure of this instance. Hence, the worst-case objective $f_\rho(\boldsymbol{x})$ admits the closed form

$$\begin{aligned} f_\rho(\boldsymbol{x}) &= \sup_{\mathbb{Q}\in\mathcal{P}(\Xi)} \big\{\mathbb{E}_{\mathbb{Q}}[\tilde{\xi}_1 x_1 + \tilde{\xi}_2 x_2] \,:\, \mathbb{E}_{\mathbb{Q}}[\tilde{\xi}_1] \leq \rho\big\} \\ &= x_2 \operatorname{diam}(\Xi) + x_1 \sup_{\mathbb{Q}_1\in\mathcal{P}([0,\operatorname{diam}(\Xi)])} \big\{\mathbb{E}_{\mathbb{Q}_1}[\tilde{\xi}_1] \,:\, \mathbb{E}_{\mathbb{Q}_1}[\tilde{\xi}_1] \leq \rho\big\} \\ &= x_2 \operatorname{diam}(\Xi) + x_1 \min\{\rho, \operatorname{diam}(\Xi)\}. \end{aligned}$$

The two path endpoints are $\boldsymbol{x}(0) = (1,0)^\top$, the minimizer of the SAA objective $f_0(\boldsymbol{x}) = \operatorname{diam}(\Xi) x_2$ over $\mathcal{X}(\Xi)$, and $\boldsymbol{x}(\infty) = \big(0, \frac{\bar\rho}{\operatorname{diam}(\Xi)}\big)^\top$, the minimizer of the RO objective $f_{\operatorname{diam}(\Xi)}(\boldsymbol{x}) = \operatorname{diam}(\Xi)(x_1 + x_2)$ over $\mathcal{X}(\Xi)$. Both endpoints satisfy $f_{\bar\rho}(\boldsymbol{x}(0)) = f_{\bar\rho}(\boldsymbol{x}(\infty)) = \bar\rho$, so by linearity $f_{\bar\rho}(\boldsymbol{x}) = \bar\rho$ for every $\boldsymbol{x}$ on the shrinkage path, giving $\hat{f}_{\bar\rho} = \bar\rho$. To evaluate the optimal value $f^\star_{\bar\rho} = \min_{\boldsymbol{x}\in\mathcal{X}(\Xi)} f_{\bar\rho}(\boldsymbol{x})$, note that $f_{\bar\rho}$ is affine in $\boldsymbol{x}$ and hence attains its minimum at a vertex of $\mathcal{X}(\Xi)$. Evaluating $f_{\bar\rho}$ at the three vertices of $\mathcal{X}(\Xi)$, we find that $(1,0)^\top$ and $\big(0, \frac{\bar\rho}{\operatorname{diam}(\Xi)}\big)^\top$ both give $\bar\rho$, whereas $\big(\frac{\bar\rho}{\operatorname{diam}(\Xi)}, 0\big)^\top$ gives $\frac{\bar\rho^2}{\operatorname{diam}(\Xi)} < \bar\rho$ since $\bar\rho < \operatorname{diam}(\Xi)$. Hence $f^\star_{\bar\rho} = \frac{\bar\rho^2}{\operatorname{diam}(\Xi)}$, and therefore $\hat{f}_{\bar\rho} - f^\star_{\bar\rho} = \bar\rho - \frac{\bar\rho^2}{\operatorname{diam}(\Xi)} = \bar\rho\Big(1 - \frac{\bar\rho}{\operatorname{diam}(\Xi)}\Big)\operatorname{Lip}(L)\|\boldsymbol{x}(0)\|_2$. This is exactly the bound from the statement of the theorem, thereby establishing tightness. □

Proof of Proposition 1. Throughout the proof, we use the same notation $f^\star_\rho$, $\hat{f}_\rho$, and $f_\rho(\boldsymbol{x})$ as in the proof of Theorem 1. If $\bar r = 0$, then each sample is zero and the only radius in $[0, \bar r]$ at which the gap is zero is $\rho = 0$. We therefore assume throughout the proof that $\bar r > 0$. We first recall a closed-form identity for the WDRO objective value $f_\rho(\boldsymbol{x})$. Under Assumption 1 with $\Xi = \mathbb{R}^k$, the WDRO problem reduces to a regularized SAA problem (Shafieezadeh-Abadeh et al. 2019, Theorem 4(ii)), yielding, for every $\boldsymbol{x} \in \mathcal{X}(\Xi)$,

$$f_\rho(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N L(\hat{\boldsymbol{\xi}}_i^\top \boldsymbol{x}) + \rho \operatorname{Lip}(L)\|\boldsymbol{x}\|_2. \tag{EC.5}$$

When $\rho \to \infty$, the penalty on $\|\boldsymbol{x}\|_2$ dominates $f_\rho(\boldsymbol{x})$. The WDRO endpoint solution is therefore the minimum-norm solution $\boldsymbol{x}(\infty) = \arg\min_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \|\boldsymbol{x}\|_2$, which is exists and is unique since $\mathcal{X}(\Xi)$ is nonempty, closed and convex.

To prove the bound, we combine an upper bound on the heuristic's value $\hat{f}_\rho$, obtained from the identity (EC.5) at $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$, with a lower bound on the optimal value $f^\star_\rho$, obtained from the concavity of $\rho \mapsto f^\star_\rho$ on $[0, \bar r]$. Subtracting the latter from the former bounds the suboptimality $\hat{f}_\rho - f^\star_\rho$ up to a residual term proportional to $f_0(\boldsymbol{x}(\infty)) - f^\star_{\bar r}$, which we show to be nonpositive. Together these steps give the claimed bound.

To bound the heuristic's value $\hat{f}_\rho$ from above, we evaluate (EC.5) at $\boldsymbol{x}(0)$ and $\boldsymbol{x}(\infty)$, which yields

$$f_\rho(\boldsymbol{x}(0)) \;=\; f_0^\star + \rho \mathrm{Lip}(L)\|\boldsymbol{x}(0)\|_2, \tag{EC.6}$$

$$f_\rho(\boldsymbol{x}(\infty)) \;=\; f_0(\boldsymbol{x}(\infty)) + \rho \mathrm{Lip}(L)\|\boldsymbol{x}(\infty)\|_2. \tag{EC.7}$$

where the first identity additionally invokes the optimality of $\boldsymbol{x}(0)$ at $\rho = 0$.

Fixing $\lambda = \frac{\rho}{\bar{r}} \in [0,1]$ on the shrinkage path and using the convexity of $f_\rho(\boldsymbol{x})$ in $\boldsymbol{x}$, Jensen's inequality yields an upper bound on the heuristic's value $\hat{f}_\rho$,

$$\begin{aligned}
\hat{f}_\rho \;&\leq\; f_\rho\Big(\Big(1-\tfrac{\rho}{\bar{r}}\Big)\boldsymbol{x}(0) + \tfrac{\rho}{\bar{r}}\boldsymbol{x}(\infty)\Big) \\
&\leq\; \Big(1-\tfrac{\rho}{\bar{r}}\Big) f_\rho(\boldsymbol{x}(0)) + \tfrac{\rho}{\bar{r}} f_\rho(\boldsymbol{x}(\infty)) \\
&=\; \Big(1-\tfrac{\rho}{\bar{r}}\Big)\big(f_0^\star + \rho \mathrm{Lip}(L)\|\boldsymbol{x}(0)\|_2\big) + \tfrac{\rho}{\bar{r}}\big(f_0(\boldsymbol{x}(\infty)) + \rho \mathrm{Lip}(L)\|\boldsymbol{x}(\infty)\|_2\big),
\end{aligned} \tag{EC.8}$$

where the equality invokes (EC.6) and (EC.7).

To bound the optimal value $f_\rho^\star$ from below, we note that by the identity (EC.5), $f_\rho(\boldsymbol{x})$ is affine in ρ, and therefore $\rho \mapsto f_\rho^\star = \min_{\boldsymbol{x}\in\mathcal{X}(\Xi)} f_\rho(\boldsymbol{x})$ is concave as the pointwise infimum of affine functions of ρ. Writing any $\rho \in [0,\bar{r}]$ as the convex combination $\rho = \big(1-\frac{\rho}{\bar{r}}\big)0 + \frac{\rho}{\bar{r}}\bar{r}$, the concavity of $\rho \mapsto f_\rho^\star$ yields the following lower bound on $f_\rho^\star$:

$$f_\rho^\star \;\geq\; \Big(1-\frac{\rho}{\bar{r}}\Big) f_0^\star + \frac{\rho}{\bar{r}} f_{\bar{r}}^\star. \tag{EC.9}$$

Subtracting the lower bound (EC.9) from the upper bound (EC.8) yields

$$\begin{aligned}
\hat{f}_\rho - f_\rho^\star \;&\leq\; \Big(1-\frac{\rho}{\bar{r}}\Big)\big(f_0^\star + \rho \mathrm{Lip}(L)\|\boldsymbol{x}(0)\|_2\big) + \frac{\rho}{\bar{r}}\big(f_0(\boldsymbol{x}(\infty)) + \rho \mathrm{Lip}(L)\|\boldsymbol{x}(\infty)\|_2\big) - \Big(1-\frac{\rho}{\bar{r}}\Big) f_0^\star - \frac{\rho}{\bar{r}} f_{\bar{r}}^\star \\
&=\; \Big(1-\frac{\rho}{\bar{r}}\Big)\rho \mathrm{Lip}(L)\|\boldsymbol{x}(0)\|_2 + \frac{\rho^2}{\bar{r}}\mathrm{Lip}(L)\|\boldsymbol{x}(\infty)\|_2 + \frac{\rho}{\bar{r}}\big(f_0(\boldsymbol{x}(\infty)) - f_{\bar{r}}^\star\big).
\end{aligned}$$

It remains to show that $f_0(\boldsymbol{x}(\infty)) \leq f_{\bar{r}}^\star$ to yield the claimed bound in Proposition 1. Observe that $f_0(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N L(\hat{\boldsymbol{\xi}}_i^\top \boldsymbol{x})$. By the Lipschitz continuity of L and the Cauchy–Schwarz inequality, for all $\boldsymbol{x} \in \mathcal{X}(\Xi)$, we have

$$\begin{aligned}
f_0(\boldsymbol{x}(\infty)) \;&\leq\; f_0(\boldsymbol{x}) + \frac{1}{N}\sum_{i=1}^N \mathrm{Lip}(L)\big|\hat{\boldsymbol{\xi}}_i^\top(\boldsymbol{x} - \boldsymbol{x}(\infty))\big| \\
&\leq\; f_0(\boldsymbol{x}) + \left(\frac{1}{N}\sum_{i=1}^N \|\hat{\boldsymbol{\xi}}_i\|_2\right)\mathrm{Lip}(L)\|\boldsymbol{x} - \boldsymbol{x}(\infty)\|_2 \\
&=\; f_0(\boldsymbol{x}) + \bar{r}\mathrm{Lip}(L)\|\boldsymbol{x} - \boldsymbol{x}(\infty)\|_2.
\end{aligned} \tag{EC.10}$$

Since $\boldsymbol{x}(\infty)$ minimizes $\|\cdot\|_2$—equivalently $\frac{1}{2}\|\cdot\|_2^2$—over the convex set $\mathcal{X}(\Xi)$, the first-order optimality condition $\nabla\big(\frac{1}{2}\|\boldsymbol{x}\|_2^2\big)\big|_{\boldsymbol{x}(\infty)}^\top(\boldsymbol{x} - \boldsymbol{x}(\infty)) \geq 0$ gives $\boldsymbol{x}(\infty)^\top(\boldsymbol{x} - \boldsymbol{x}(\infty)) \geq 0$ for all $\boldsymbol{x} \in \mathcal{X}(\Xi)$. Therefore,

we have $\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}-\boldsymbol{x}(\infty)\|_2^2 + 2\boldsymbol{x}(\infty)^\top(\boldsymbol{x}-\boldsymbol{x}(\infty)) + \|\boldsymbol{x}(\infty)\|_2^2 \geq \|\boldsymbol{x}-\boldsymbol{x}(\infty)\|_2^2$. Substituting $\|\boldsymbol{x}-\boldsymbol{x}(\infty)\|_2 \leq \|\boldsymbol{x}\|_2$ into (EC.10) and invoking the identity (EC.5) at $\rho=\bar{r}$, we have

$$f_0(\boldsymbol{x}(\infty)) \;\leq\; f_0(\boldsymbol{x}) + \bar{r}\mathrm{Lip}(L)\|\boldsymbol{x}\|_2 \;=\; f_{\bar{r}}(\boldsymbol{x}).$$

The inequality $f_0(\boldsymbol{x}(\infty)) \leq f_{\bar{r}}(\boldsymbol{x})$ holds for all $\boldsymbol{x}\in\mathcal{X}(\Xi)$. Minimizing the right-hand side over $\boldsymbol{x}\in\mathcal{X}(\Xi)$ yields the target inequality $f_0(\boldsymbol{x}(\infty)) \leq f^\star_{\bar{r}}$, and therefore

$$\begin{aligned}\hat{f}_\rho - f^\star_\rho \;&\leq \left(1-\frac{\rho}{\bar{r}}\right)\rho\mathrm{Lip}(L)\|\boldsymbol{x}(0)\|_2 + \frac{\rho^2}{\bar{r}}\mathrm{Lip}(L)\|\boldsymbol{x}(\infty)\|_2 + \frac{\rho}{\bar{r}}\big(f_0(\boldsymbol{x}(\infty)) - f^\star_{\bar{r}}\big)\\ &\leq \mathrm{Lip}(L)\left(\rho\left(1-\frac{\rho}{\bar{r}}\right)\|\boldsymbol{x}(0)\|_2 + \frac{\rho^2}{\bar{r}}\|\boldsymbol{x}(\infty)\|_2\right),\end{aligned}$$

which completes the proof of the claimed bound. For $\boldsymbol{x}(\infty)=\mathbf{0}$, this bound is maximized at $\rho=\bar{r}/2$.

□

The proof of Theorem 2 relies on the following auxiliary results, which we state and prove first. Throughout, we use the notation $f^\star_\rho$, $\hat{f}_\rho$, and $f_\rho(\boldsymbol{x})$ from the proof of Theorem 1, we write $\Delta(\rho) := \hat{f}_\rho - f^\star_\rho$ for the suboptimality gap of the shrinkage path heuristic at the radius ρ, and we denote by $\boldsymbol{x}^\star(\rho)$ and $\hat{\boldsymbol{x}}(\rho)$ the corresponding optimizers of the WDRO problem (1) and of the shrinkage path heuristic, respectively. The first lemma expresses the optimal values $f^\star_\rho$ and $\hat{f}_\rho$ through explicit quadratics; the second lemma locates a gap-maximizing radius ρ_{gap} and expresses $\Delta(\rho_{\mathrm{gap}})$ as the covariance between two functions of a random eigenvalue of $\boldsymbol{H}$; the third lemma bounds this covariance by an *endpoint envelope*, that is, an upper bound that depends on the spectrum of $\boldsymbol{H}$ only through its extreme eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$; and the fourth lemma maximizes the resulting univariate envelope in closed form.

Lemma EC.1. *Let Assumption 2 hold and suppose $\boldsymbol{x}(0)\neq\mathbf{0}$. After absorbing* $\mathrm{Lip}(L)$ *into ρ, so that $f_\rho(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \rho\|\boldsymbol{x}\|_2$, the WDRO optimal value satisfies, for every $\rho\geq 0$,*

$$f^\star_\rho = f^\star_0 + \min_{\boldsymbol{x}\in\mathbb{R}^n}\big\{(\boldsymbol{x}-\boldsymbol{x}(0))^\top\boldsymbol{H}(\boldsymbol{x}-\boldsymbol{x}(0)) + \rho\|\boldsymbol{x}\|_2\big\}. \tag{EC.11}$$

Moreover, $\boldsymbol{x}(\infty)=\mathbf{0}$, and the shrinkage path solution satisfies $\hat{\boldsymbol{x}}(\rho)=\lambda^\star(\rho)\boldsymbol{x}(0)$, where

$$\lambda^\star(\rho) = \left[1 - \frac{\rho\,\|\boldsymbol{x}(0)\|_2}{2\,\boldsymbol{x}(0)^\top\boldsymbol{H}\,\boldsymbol{x}(0)}\right]_+, \tag{EC.12}$$

and the corresponding shrinkage path value is

$$\hat{f}_\rho = f^\star_0 + \boldsymbol{x}(0)^\top\boldsymbol{H}\,\boldsymbol{x}(0) - \boldsymbol{x}(0)^\top\boldsymbol{H}\,\boldsymbol{x}(0)\left[1 - \frac{\rho\,\|\boldsymbol{x}(0)\|_2}{2\,\boldsymbol{x}(0)^\top\boldsymbol{H}\,\boldsymbol{x}(0)}\right]_+^2. \tag{EC.13}$$

Proof of Lemma EC.1. The normalization of the radius is harmless: under the present assumptions, $\mathcal{Z}$ contains a nondegenerate interval and L is a nonconstant quadratic on it, so $\mathrm{Lip}(L) > 0$ and the substitution $\rho \leftarrow \mathrm{Lip}(L)\rho$ simply relabels the radii $\rho \geq 0$. The remainder of the proof proceeds in three steps: the first step shows that on the ball $B := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}(0)\|_2\}$, the empirical loss f_0 is a quadratic centered at $\boldsymbol{x}(0)$; the second step shows that the WDRO problem and the quadratic problem in (EC.11) both attain their minima on B, where their objectives coincide up to the constant $f_0^\star$, which establishes (EC.11); and the third step determines $\boldsymbol{x}(\infty)$ and optimizes over the shrinkage path, which yields (EC.12) and (EC.13).

For the first step, the definition of $\mathcal{Z}$ ensures that $\hat{\boldsymbol{\xi}}_i^\top \boldsymbol{x} \in \mathcal{Z}$ for all $\boldsymbol{x} \in B$ and $i \in [N]$, so Assumption 2*(ii)* implies that

$$f_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x} + b\, \boldsymbol{x}^\top \left(\frac{1}{N} \sum_{i=1}^N \hat{\boldsymbol{\xi}}_i \right) + c \qquad \text{for all } \boldsymbol{x} \in B.$$

Because L is differentiable on $\mathbb{R}$ and agrees with the quadratic $z \mapsto z^2 + bz + c$ on $\mathcal{Z}$, we have $L'(z) = 2z + b$ for all $z \in \mathcal{Z}$: at interior points this is immediate, and at the endpoints $\pm R$ of $\mathcal{Z} := [-R, R]$, where $R := \|\boldsymbol{x}(0)\|_2 \max_{i \in [N]} \|\hat{\boldsymbol{\xi}}_i\|_2$, differentiability forces L' to coincide with the one-sided derivative of the quadratic taken from inside $\mathcal{Z}$. Since $\hat{\boldsymbol{\xi}}_i^\top \boldsymbol{x}(0) \in \mathcal{Z}$ for all i and $\mathcal{X}(\Xi) = \mathbb{R}^n$, the SAA first-order condition $\nabla f_0(\boldsymbol{x}(0)) = \boldsymbol{0}$ is

$$2\boldsymbol{H}\boldsymbol{x}(0) + b \left(\frac{1}{N} \sum_{i=1}^N \hat{\boldsymbol{\xi}}_i \right) = \boldsymbol{0}.$$

Substituting $b \frac{1}{N} \sum_{i=1}^N \hat{\boldsymbol{\xi}}_i = -2\boldsymbol{H}\boldsymbol{x}(0)$ into the expression for f_0 above and completing the square, we obtain

$$f_0(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}(0))^\top \boldsymbol{H} (\boldsymbol{x} - \boldsymbol{x}(0)) + c - \boldsymbol{x}(0)^\top \boldsymbol{H}\, \boldsymbol{x}(0) = f_0^\star + (\boldsymbol{x} - \boldsymbol{x}(0))^\top \boldsymbol{H} (\boldsymbol{x} - \boldsymbol{x}(0)) \qquad \text{for all } \boldsymbol{x} \in B,$$

where setting $\boldsymbol{x} = \boldsymbol{x}(0)$ in the first equality of the preceding display identifies the constant $c - \boldsymbol{x}(0)^\top \boldsymbol{H}\, \boldsymbol{x}(0)$ as $f_0(\boldsymbol{x}(0)) = f_0^\star$.

For the second step, denote by $q(\boldsymbol{x}) := (\boldsymbol{x} - \boldsymbol{x}(0))^\top \boldsymbol{H} (\boldsymbol{x} - \boldsymbol{x}(0)) + \rho \|\boldsymbol{x}\|_2$ the function minimized in (EC.11). In view of the first step and the identity (EC.5), we have $f_\rho(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \rho\|\boldsymbol{x}\|_2 = f_0^\star + q(\boldsymbol{x})$ for all $\boldsymbol{x} \in B$. Both f_ρ and q attain their minima over $\mathbb{R}^n$ on the compact set B: for $\rho = 0$, both functions are minimized by $\boldsymbol{x}(0) \in B$, while for $\rho > 0$, every $\boldsymbol{x} \notin B$ satisfies $\|\boldsymbol{x}\|_2 > \|\boldsymbol{x}(0)\|_2$ and is therefore dominated by $\boldsymbol{x}(0)$, since the optimality of $\boldsymbol{x}(0)$ in the SAA problem (3a) implies

$$f_\rho(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \rho\|\boldsymbol{x}\|_2 > f_0(\boldsymbol{x}(0)) + \rho\|\boldsymbol{x}(0)\|_2 = f_\rho(\boldsymbol{x}(0)) \quad \text{and} \quad q(\boldsymbol{x}) \geq \rho\|\boldsymbol{x}\|_2 > \rho\|\boldsymbol{x}(0)\|_2 = q(\boldsymbol{x}(0)).$$

We thus conclude that

$$f_\rho^\star = \min_{\boldsymbol{x} \in B} f_\rho(\boldsymbol{x}) = f_0^\star + \min_{\boldsymbol{x} \in B} q(\boldsymbol{x}) = f_0^\star + \min_{\boldsymbol{x} \in \mathbb{R}^n} q(\boldsymbol{x}),$$

which is (EC.11); in particular, the minimizers of the WDRO problem coincide with those of q.

For the third step, $\mathcal{X}(\Xi) = \mathbb{R}^n$ implies that the minimum-norm endpoint is $\boldsymbol{x}(\infty) = \mathbf{0}$, so the shrinkage path is $\{\lambda \boldsymbol{x}(0) : \lambda \in [0,1]\} \subseteq B$, where λ denotes the coefficient of $\boldsymbol{x}(0)$ and thus corresponds to $1-\lambda$ in the parameterization (4). The first step and the identity (EC.5) then give

$$f_\rho(\lambda \boldsymbol{x}(0)) = f_0^\star + (1-\lambda)^2 \boldsymbol{x}(0)^\top \boldsymbol{H}\, \boldsymbol{x}(0) + \rho\lambda \|\boldsymbol{x}(0)\|_2.$$

Minimizing this strongly convex univariate quadratic over $\lambda \in [0,1]$ yields the minimizer $\lambda^\star(\rho)$ in (EC.12), so that $\hat{\boldsymbol{x}}(\rho) = \lambda^\star(\rho)\boldsymbol{x}(0)$; substituting this minimizer into the path objective gives (EC.13). □

Lemma EC.2. *Let the conditions of Lemma EC.1 hold, and let ρ_{gap} be a maximizer of $\Delta(\rho)$ over $\rho \geq 0$ with $\Delta(\rho_{\mathrm{gap}}) > 0$. Then $\rho_{\mathrm{gap}} \in (0, \rho_0^{\mathrm{h}})$, where*

$$\rho_0^{\mathrm{h}} := \frac{2\, \boldsymbol{x}(0)^\top \boldsymbol{H}\, \boldsymbol{x}(0)}{\|\boldsymbol{x}(0)\|_2}.$$

Moreover,

$$\|\hat{\boldsymbol{x}}(\rho_{\mathrm{gap}})\|_2 = \|\boldsymbol{x}^\star(\rho_{\mathrm{gap}})\|_2. \tag{EC.14}$$

If $\boldsymbol{H} = \boldsymbol{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{U}^\top$ with $\boldsymbol{U}$ orthogonal, $\tilde{\boldsymbol{x}}(0) = \boldsymbol{U}^\top \boldsymbol{x}(0)$, $p_i = \tilde{x}(0)_i^2 / \|\boldsymbol{x}(0)\|_2^2$, and $\phi_\gamma(\lambda) = (\lambda/(\lambda+\gamma))^2$, then, with $\mathbb{P}(\Lambda = \lambda_i) = p_i$ and $\gamma_{\mathrm{gap}} = \rho_{\mathrm{gap}}/(2\|\boldsymbol{x}^\star(\rho_{\mathrm{gap}})\|_2)$,

$$\frac{\Delta(\rho_{\mathrm{gap}})}{\|\boldsymbol{x}(0)\|_2^2} = \mathbb{E}_p[\Lambda \phi_{\gamma_{\mathrm{gap}}}(\Lambda)] - \mathbb{E}_p[\Lambda]\, \mathbb{E}_p[\phi_{\gamma_{\mathrm{gap}}}(\Lambda)], \tag{EC.15}$$

and

$$\mathbb{E}_p[\phi_{\gamma_{\mathrm{gap}}}(\Lambda)] = \left(\frac{\mathbb{E}_p[\Lambda]}{\mathbb{E}_p[\Lambda] + \gamma_{\mathrm{gap}}}\right)^2 = \phi_{\gamma_{\mathrm{gap}}}\left(\mathbb{E}_p[\Lambda]\right). \tag{EC.16}$$

Proof of Lemma EC.2. We proceed in three steps: first, we establish that $\rho_{\mathrm{gap}} \in (0, \rho_0^{\mathrm{h}})$ and the norm identity (EC.14), second, we derive closed forms for $f^\star_{\rho_{\mathrm{gap}}}$ and $\hat{f}_{\rho_{\mathrm{gap}}}$ from the first-order condition of the quadratic problem in (EC.11), which yield the covariance identity (EC.15) in the eigenbasis of $\boldsymbol{H}$, and third, we rewrite (EC.14) in this eigenbasis to obtain (EC.16).

For the first step, since $\Delta(0) = 0 < \Delta(\rho_{\mathrm{gap}})$, the maximizer ρ_{gap} satisfies $\rho_{\mathrm{gap}} > 0$. The objective in (EC.11) is strongly convex, and the path objective is strongly convex in λ, so both $\boldsymbol{x}^\star(\rho)$ and $\hat{\boldsymbol{x}}(\rho)$ are unique. Danskin's theorem states that a minimum-value function $\rho \mapsto \min_u g(u,\rho)$, whose objective is affine in ρ and whose minimum is attained at a unique $u(\rho)$, is differentiable with derivative $(\partial g/\partial \rho)(u(\rho), \rho)$. Since ρ enters both minimization problems only through the term $\rho\|\boldsymbol{x}\|_2$, which is affine in ρ, we obtain, for every $\rho > 0$,

$$\Delta'(\rho) = \|\hat{\boldsymbol{x}}(\rho)\|_2 - \|\boldsymbol{x}^\star(\rho)\|_2.$$

Fermat's rule, which states that the derivative of a differentiable function vanishes at every interior extremum, applied at the maximizer $\rho_{\text{gap}} > 0$, gives $\Delta'(\rho_{\text{gap}}) = 0$ and hence we obtain (EC.14). This rules out every $\rho_{\text{gap}} \geq \rho_0^{\text{h}}$: at any such radius (EC.12) gives $\hat{\boldsymbol{x}}(\rho_{\text{gap}}) = \mathbf{0}$, so (EC.14) would force $\boldsymbol{x}^\star(\rho_{\text{gap}}) = \mathbf{0}$, and then (EC.11) and (EC.13) would both evaluate to $f_0^\star + \boldsymbol{x}(0)^\top \boldsymbol{H}\boldsymbol{x}(0)$, making the heuristic exact and contradicting $\Delta(\rho_{\text{gap}}) > 0$. Thus $\rho_{\text{gap}} \in (0, \rho_0^{\text{h}})$; moreover, $\hat{\boldsymbol{x}}(\rho_{\text{gap}}) = \lambda^\star(\rho_{\text{gap}})\,\boldsymbol{x}(0) \neq \mathbf{0}$, and so $\boldsymbol{x}^\star(\rho_{\text{gap}}) \neq \mathbf{0}$ by (EC.14).

For the second step, recall from Lemma EC.1 that $\boldsymbol{x}^\star(\rho_{\text{gap}})$ minimizes the quadratic problem on the right-hand side of (EC.11); the associated first-order condition at $\boldsymbol{x}^\star(\rho_{\text{gap}}) \neq \mathbf{0}$ is

$$2\boldsymbol{H}(\boldsymbol{x}^\star(\rho_{\text{gap}}) - \boldsymbol{x}(0)) + \rho_{\text{gap}} \frac{\boldsymbol{x}^\star(\rho_{\text{gap}})}{\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2} = \mathbf{0}. \tag{EC.17}$$

With $\gamma_{\text{gap}} := \rho_{\text{gap}}/(2\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2)$, this is equivalently

$$(\boldsymbol{H} + \gamma_{\text{gap}}\boldsymbol{I})\boldsymbol{x}^\star(\rho_{\text{gap}}) = \boldsymbol{H}\boldsymbol{x}(0).$$

Multiplying this equation from the left by $\boldsymbol{x}^\star(\rho_{\text{gap}})^\top$ yields $\boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}^\star(\rho_{\text{gap}}) = \boldsymbol{x}^\star(\rho_{\text{gap}})^\top \boldsymbol{H}\,\boldsymbol{x}^\star(\rho_{\text{gap}}) + \gamma_{\text{gap}}\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2^2$; substituting this identity and $\rho_{\text{gap}}\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2 = 2\gamma_{\text{gap}}\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2^2$ into the objective of (EC.11), evaluated at $\boldsymbol{x}^\star(\rho_{\text{gap}})$, yields

$$f^\star_{\rho_{\text{gap}}} = f_0^\star + \boldsymbol{x}(0)^\top \boldsymbol{H}\boldsymbol{x}(0) - \boldsymbol{x}^\star(\rho_{\text{gap}})^\top \boldsymbol{H}\boldsymbol{x}^\star(\rho_{\text{gap}}).$$

Similarly, since $\hat{\boldsymbol{x}}(\rho_{\text{gap}}) = \lambda^\star(\rho_{\text{gap}})\boldsymbol{x}(0)$ implies $\hat{\boldsymbol{x}}(\rho_{\text{gap}})^\top \boldsymbol{H}\,\hat{\boldsymbol{x}}(\rho_{\text{gap}}) = \lambda^\star(\rho_{\text{gap}})^2 \boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0)$, the path value (EC.13) can be written as

$$\hat{f}_{\rho_{\text{gap}}} = f_0^\star + \boldsymbol{x}(0)^\top \boldsymbol{H}\boldsymbol{x}(0) - \hat{\boldsymbol{x}}(\rho_{\text{gap}})^\top \boldsymbol{H}\hat{\boldsymbol{x}}(\rho_{\text{gap}}).$$

Subtracting the first identity from the second and using $\hat{\boldsymbol{x}}(\rho_{\text{gap}})^\top \boldsymbol{H}\,\hat{\boldsymbol{x}}(\rho_{\text{gap}}) = \frac{\|\hat{\boldsymbol{x}}(\rho_{\text{gap}})\|_2^2}{\|\boldsymbol{x}(0)\|_2^2}\,\boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0)$ together with the norm identity (EC.14) gives

$$\Delta(\rho_{\text{gap}}) = \boldsymbol{x}^\star(\rho_{\text{gap}})^\top \boldsymbol{H}\,\boldsymbol{x}^\star(\rho_{\text{gap}}) - \frac{\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2^2}{\|\boldsymbol{x}(0)\|_2^2}\boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0). \tag{EC.18}$$

Diagonalize $\boldsymbol{H} = \boldsymbol{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{U}^\top$ with $\boldsymbol{U}$ orthogonal, set $\tilde{\boldsymbol{x}}(0) = \boldsymbol{U}^\top \boldsymbol{x}(0)$, and define $p_i := \tilde{x}(0)_i^2/\|\boldsymbol{x}(0)\|_2^2$. Then $p_i \geq 0$ and $\sum_i p_i = 1$. Writing the resolvent equation $(\boldsymbol{H} + \gamma_{\text{gap}}\boldsymbol{I})\,\boldsymbol{x}^\star(\rho_{\text{gap}}) = \boldsymbol{H}\boldsymbol{x}(0)$ in this eigenbasis gives

$$\tilde{x}_i^\star(\rho_{\text{gap}}) = \frac{\lambda_i}{\lambda_i + \gamma_{\text{gap}}}\,\tilde{x}_i(0).$$

Substituting this expression into (EC.18) yields

$$\Delta(\rho_{\text{gap}}) = \|\boldsymbol{x}(0)\|_2^2 \left[\sum_{i=1}^n \lambda_i \left(\frac{\lambda_i}{\lambda_i + \gamma_{\text{gap}}}\right)^2 p_i - \sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_i + \gamma_{\text{gap}}}\right)^2 p_i \sum_{j=1}^n \lambda_j p_j\right]. \tag{EC.19}$$

Define the random variable Λ by $\mathbb{P}(\Lambda = \lambda_i) = p_i$, and set $\phi_\gamma(\lambda) := (\lambda/(\lambda+\gamma))^2$. Then (EC.19) becomes the covariance identity (EC.15). Because $\phi_{\gamma_{\text{gap}}}$ is increasing in λ, this covariance is nonnegative; the gap is precisely the cost of applying one common shrinkage factor instead of the coordinatewise factors $\lambda_i/(\lambda_i+\gamma_{\text{gap}})$.

For the third step, it remains to rewrite the norm identity (EC.14) in spectral form. Let $m := \mathbb{E}_p[\Lambda]$ and $\lambda_F := \lambda^\star(\rho_{\text{gap}})$. From the eigenbasis representation,

$$\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2^2 = \|\boldsymbol{x}(0)\|_2^2 \mathbb{E}_p[\phi_{\gamma_{\text{gap}}}(\Lambda)],$$

and (EC.14) gives $\mathbb{E}_p[\phi_{\gamma_{\text{gap}}}(\Lambda)] = \lambda_F^2$. On the other hand, since $\rho_{\text{gap}} \in (0, \rho_0^{\text{h}})$,

$$\lambda_F = 1 - \frac{\rho_{\text{gap}}\,\|\boldsymbol{x}(0)\|_2}{2\,\boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0)}.$$

Using $\rho_{\text{gap}} = 2\gamma_{\text{gap}}\lambda_F\|\boldsymbol{x}(0)\|_2$—which follows from the definition of γ_{gap} together with $\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2 = \lambda_F\|\boldsymbol{x}(0)\|_2$, that is, (EC.14)—and $m = \boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0)/\|\boldsymbol{x}(0)\|_2^2$, this becomes $\lambda_F = 1 - \gamma_{\text{gap}}\lambda_F/m$, or

$$\lambda_F = \frac{m}{m+\gamma_{\text{gap}}}.$$

Consequently, (EC.16) holds. □

LEMMA EC.3. *Under the conditions of Lemma EC.2, the gap at the maximizing radius ρ_{gap} admits a spectral upper bound that depends on $\boldsymbol{H}$ only through its condition number κ. Specifically,*

$$\frac{\Delta(\rho_{\text{gap}})}{\lambda_{\min}\|\boldsymbol{x}(0)\|_2^2} \le \sup_{t\ge 0} t^2 \left(\frac{1}{\sqrt{1+t}} - \frac{1}{\sqrt{\kappa+t}}\right)^2. \tag{EC.20}$$

Proof of Lemma EC.3. The proof proceeds in three steps: the first step rewrites the covariance in (EC.15) as the difference between $\mathbb{E}_p[r_\gamma(\Lambda)]$ and $r_\gamma(\mathbb{E}_p[\Lambda])$, that is, the Jensen gap of the resolvent $r_\gamma(\lambda) := 1/(\lambda+\gamma)$; the second step bounds this gap by relaxing the law of Λ to a mean-constrained distribution on the spectral interval $[\lambda_{\min}, \lambda_{\max}]$; and the third step maximizes the resulting endpoint expression over the mean and rescales it to obtain (EC.20).

For the first step, write γ for γ_{gap}, set $m := \mathbb{E}_p[\Lambda]$, and define $r_\gamma(\lambda) := 1/(\lambda+\gamma)$. Using the covariance identity (EC.15), the spectral stationarity condition (EC.16) in the form $\mathbb{E}_p[\phi_\gamma(\Lambda)] = \phi_\gamma(m)$, and the identity

$$(\lambda+\gamma)\phi_\gamma(\lambda) = \lambda - \gamma + \gamma^2 r_\gamma(\lambda),$$

we obtain the resolvent representation

$$\frac{\Delta(\rho_{\text{gap}})}{\|\boldsymbol{x}(0)\|_2^2} = \mathbb{E}_p[\Lambda\phi_\gamma(\Lambda)] - m\phi_\gamma(m) = \gamma^2\big(\mathbb{E}_p[r_\gamma(\Lambda)] - r_\gamma(m)\big). \tag{EC.21}$$

Indeed, the middle expression of (EC.21) equals

$$m-\gamma+\gamma^2\mathbb{E}_p[r_\gamma(\Lambda)]-(m+\gamma)\phi_\gamma(m),$$

and $(m+\gamma)\phi_\gamma(m)=m-\gamma+\gamma^2 r_\gamma(m)$. Thus the stationarity condition (EC.16) has reduced the covariance to a resolvent Jensen gap. From this point on, the argument uses only the support interval $[\lambda_{\min},\lambda_{\max}]$ and the mean m of Λ.

For the second step, we relax the requirements on the law of Λ and merely require it to be supported on $[\lambda_{\min},\lambda_{\max}]$ with mean $\mathbb{E}_p[\Lambda]=m$. Since r_γ is convex on $[\lambda_{\min},\lambda_{\max}]$, every such law satisfies the secant bound

$$\mathbb{E}_p[r_\gamma(\Lambda)]\le\frac{\lambda_{\max}-m}{\lambda_{\max}-\lambda_{\min}}r_\gamma(\lambda_{\min})+\frac{m-\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}r_\gamma(\lambda_{\max}). \tag{EC.22}$$

The right-hand side of (EC.22) equals $\mathbb{E}_p[r_\gamma(\Lambda)]$ when Λ follows the two-point law on $\{\lambda_{\min},\lambda_{\max}\}$ with mean m. The secant bound therefore holds with equality for this law. Substituting (EC.22) into (EC.21) gives the endpoint envelope

$$\frac{\Delta(\rho_{\mathrm{gap}})}{\|\boldsymbol{x}(0)\|_2^2}\le\frac{\gamma^2(m-\lambda_{\min})(\lambda_{\max}-m)}{(\lambda_{\min}+\gamma)(\lambda_{\max}+\gamma)(m+\gamma)}.$$

The resulting expression is nevertheless an upper bound, not an exact reduction of the stationarity-constrained problem, because stationarity is no longer imposed after the resolvent rewrite. Taking the supremum over $m\in[\lambda_{\min},\lambda_{\max}]$ and $\gamma\ge 0$ yields

$$\frac{\Delta(\rho_{\mathrm{gap}})}{\|\boldsymbol{x}(0)\|_2^2}\le\sup_{m\in[\lambda_{\min},\lambda_{\max}],\,\gamma\ge0}\frac{\gamma^2(m-\lambda_{\min})(\lambda_{\max}-m)}{(\lambda_{\min}+\gamma)(\lambda_{\max}+\gamma)(m+\gamma)}. \tag{EC.23}$$

For the third step, fix $\gamma>0$. The only m-dependent factor in (EC.23) is $(m-\lambda_{\min})(\lambda_{\max}-m)/(m+\gamma)$. Its derivative is

$$\frac{(\lambda_{\min}+\gamma)(\lambda_{\max}+\gamma)-(m+\gamma)^2}{(m+\gamma)^2},$$

so the maximizing mean is

$$m^\star(\gamma)=\sqrt{(\lambda_{\min}+\gamma)(\lambda_{\max}+\gamma)}-\gamma,$$

which lies in $[\lambda_{\min},\lambda_{\max}]$. Evaluating (EC.23) at $m^\star(\gamma)$ reduces the bound to

$$\frac{\Delta(\rho_{\mathrm{gap}})}{\|\boldsymbol{x}(0)\|_2^2}\le\sup_{\gamma\ge0}\gamma^2\left(\frac{1}{\sqrt{\lambda_{\min}+\gamma}}-\frac{1}{\sqrt{\lambda_{\max}+\gamma}}\right)^2.$$

With $\gamma=\lambda_{\min}t$ and $\kappa=\lambda_{\max}/\lambda_{\min}$, this is (EC.20). □

LEMMA EC.4. *The scalar envelope in Lemma EC.3 admits the following closed-form bound: for every $\kappa \geq 1$ and $t \geq 0$,*

$$t^2 \left(\frac{1}{\sqrt{1+t}} - \frac{1}{\sqrt{\kappa+t}} \right)^2 \leq \frac{(\kappa-1)^2}{\varphi^5(\kappa-1)+27}. \tag{EC.24}$$

Proof of Lemma EC.4. The claim is immediate when $\kappa = 1$, so take $\kappa > 1$. The proof proceeds in two steps: the first step rationalizes the difference of square roots, then rescales the desired inequality through variables u and z; the second step maximizes the resulting left-hand side over u and compares the maximum with the right-hand side.

For the first step, we invoke the identity $\left(\sqrt{\kappa+t} - \sqrt{1+t}\right)\left(\sqrt{\kappa+t} + \sqrt{1+t}\right) = \kappa - 1$ to rewrite the left-hand side of (EC.24) as $t^2(\kappa-1)^2 / \left[(1+t)(\kappa+t)\left(\sqrt{1+t} + \sqrt{\kappa+t}\right)^2\right]$. Clearing the positive denominators and dividing by $(\kappa-1)^2 > 0$, we obtain the equivalent inequality

$$\left(\varphi^5(\kappa-1)+27\right) t^2 \leq (1+t)(\kappa+t)\left(\sqrt{1+t} + \sqrt{\kappa+t}\right)^2.$$

Now set $u := t/(1+t) \in [0,1)$ and $z := (\kappa-1)/(1+t) \geq 0$, and define $R(z) := (1+z)\left(1+\sqrt{1+z}\right)^2$. Dividing the preceding inequality by $(1+t)^3$ shows that it suffices to prove, for all $u \in [0,1]$ and $z \geq 0$,

$$u^2 \left[27(1-u) + \varphi^5 z\right] \leq R(z). \tag{EC.25}$$

For the second step, fix z. The derivative of the left-hand side with respect to u is $u(54 + 2\varphi^5 z - 81u)$. Hence, with $z_0 := 27/(2\varphi^5)$,

$$\max_{u \in [0,1]} u^2 \left[27(1-u) + \varphi^5 z\right] = \begin{cases} 4\left(1 + \varphi^5 z/27\right)^3, & 0 \leq z \leq z_0, \\ \varphi^5 z, & z \geq z_0, \end{cases}$$

where the first branch is attained at the interior critical point and the second at $u = 1$. We now compare both branches with $R(z)$. For the second branch, writing $s = \sqrt{1+z} \geq 1$ gives

$$R(z) - \varphi^5 z = (s+1)(s-\varphi)^2(s+2+\sqrt{5}) \geq 0.$$

For the first branch, set $g(z) := (R(z)/4)^{1/3}$. We claim that g is concave. Indeed, writing $s = \sqrt{1+z}$ gives $g(z) = 2^{-2/3}(s(s+1))^{2/3}$ and

$$g''(z) = -\frac{2^{-2/3}(4s^2+7s+4)}{18 s^{10/3}(s+1)^{4/3}} < 0.$$

Moreover $g(0) = 1$, and the preceding branch bound at z_0 gives $g(z_0) \geq (\varphi^5 z_0/4)^{1/3} = 3/2$. Therefore, for $z \in [0, z_0]$,

$$g(z) \geq \left(1 - \frac{z}{z_0}\right) g(0) + \frac{z}{z_0} g(z_0) \geq 1 + \frac{\varphi^5 z}{27}.$$

Cubing both sides and invoking $g(z) = (R(z)/4)^{1/3}$ yields $R(z)/4 \geq (1 + \varphi^5 z/27)^3$, hence $R(z) \geq 4(1+\varphi^5 z/27)^3$ on $[0, z_0]$. The two branch bounds $R(z) \geq 4(1+\varphi^5 z/27)^3$ on $[0, z_0]$ and $R(z) \geq \varphi^5 z$

on $[z_0,\infty)$ together give $\max_{u\in[0,1]} u^2\,[27(1-u)+\varphi^5 z]\le R(z)$ for every $z\ge 0$, which proves (EC.25) and hence (EC.24). □

Proof of Theorem 2. We use the notation introduced before Lemma EC.1. The proof proceeds in three steps: the first step handles the trivial case $\boldsymbol{x}(0)=\mathbf{0}$ and otherwise combines the auxiliary lemmas to prove the additive bound; the second step builds two-dimensional instances and stationary radii at which the gap can be evaluated explicitly; and the third step evaluates these stationary radii in the regimes $\kappa\to\infty$ and $\kappa\to 1^+$ to prove asymptotic tightness.

For the first step, if $\boldsymbol{x}(0)=\mathbf{0}$, then $\mathbf{0}$ minimizes both the empirical loss and the norm regularizer in (EC.5), so it is a WDRO optimizer for every $\rho\ge 0$; the shrinkage path is also identically zero, and hence $\Delta(\rho)=0$ for every ρ. We therefore assume $\boldsymbol{x}(0)\neq\mathbf{0}$ in the remainder of the proof and, as in Lemma EC.1, absorb the constant $\mathrm{Lip}(L)$ in (EC.5) into the radius ρ, which leaves the worst-case suboptimality $\sup_{\rho\ge 0}\Delta(\rho)$ unchanged.

Define $\rho_0^{\mathrm{h}} := 2\,\boldsymbol{x}(0)^\top \boldsymbol{H}\,\boldsymbol{x}(0)/\|\boldsymbol{x}(0)\|_2$, and let ρ_{gap} maximize $\Delta(\rho)$ over $\rho\ge 0$. Such a maximizer exists because Lemma EC.1 gives continuity of both $\hat{f}_\rho$ and $f_\rho^\star$ in ρ, while $\hat{f}_\rho$ is constant for $\rho\ge\rho_0^{\mathrm{h}}$ by (EC.13) and $f_\rho^\star$ is nondecreasing in ρ, so that Δ is nonincreasing on $[\rho_0^{\mathrm{h}},\infty)$ and $\sup_{\rho\ge 0}\Delta(\rho)$ is attained on the compact interval $[0,\rho_0^{\mathrm{h}}]$. If $\Delta(\rho_{\mathrm{gap}})=0$, the theorem is immediate. Otherwise, Lemmas EC.2, EC.3, and EC.4 give

$$\Delta(\rho_{\mathrm{gap}}) \le \lambda_{\min}\|\boldsymbol{x}(0)\|_2^2 \frac{(\kappa-1)^2}{\varphi^5(\kappa-1)+27} = \lambda_{\max}\|\boldsymbol{x}(0)\|_2^2 \frac{(\kappa-1)^2}{\kappa[\varphi^5(\kappa-1)+27]},$$

which is the bound (5).

For the second step, we set up the tightness construction. Consider the matrix $\boldsymbol{H}_\kappa$ and the SAA solution $\boldsymbol{x}_\kappa(0)$ given by

$$\boldsymbol{H}_\kappa = \mathrm{diag}(1,\kappa), \qquad \boldsymbol{x}_\kappa(0) = \|\boldsymbol{x}_\kappa(0)\|_2\,(\sqrt{1-\theta},\sqrt{\theta})^\top, \qquad \theta\in(0,1), \tag{EC.26}$$

with spectral law $p_\theta=(1-\theta)\delta_1+\theta\delta_\kappa$ and mean $m=1+\theta(\kappa-1)$, where δ_a denotes the unit point mass at a. These reduced data—in fact, any $\boldsymbol{H}\succ\mathbf{0}$ and any target $\boldsymbol{x}(0)\neq\mathbf{0}$—are realized by empirical samples satisfying Assumption 2. Fix $\beta>2\sqrt{\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)}$ and set $\bar{\boldsymbol{\xi}} := -2\boldsymbol{H}\boldsymbol{x}(0)/\beta$. Then $\bar{\boldsymbol{\xi}}^\top\boldsymbol{H}^{-1}\bar{\boldsymbol{\xi}} = 4\,\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)/\beta^2<1$, so $\boldsymbol{H}-\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}^\top\succ\mathbf{0}$ admits a factorization $\boldsymbol{H}-\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}^\top=\boldsymbol{L}\boldsymbol{L}^\top$. The $N=2n$ equally weighted samples $\bar{\boldsymbol{\xi}}\pm\sqrt{n}\,\boldsymbol{L}\boldsymbol{e}_j$, $j\in[n]$, have empirical mean $\bar{\boldsymbol{\xi}}$ and empirical second-moment matrix $\bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}^\top+\boldsymbol{L}\boldsymbol{L}^\top=\boldsymbol{H}$. Choose for L a Huber-type loss that agrees with $z^2+\beta z+c$ on an open interval containing $\mathcal{Z}$ and continues linearly, with matching value and slope, outside that interval. Here c is large enough to ensure $L\ge 0$. This loss is convex, globally Lipschitz, differentiable, and quadratic on $\mathcal{Z}$, so Assumptions 1 and 2*(ii)* hold with $b=\beta$. Finally, the choice of $\bar{\boldsymbol{\xi}}$ yields the first-order condition $\nabla f_0(\boldsymbol{x}(0))=2\boldsymbol{H}\boldsymbol{x}(0)+\beta\bar{\boldsymbol{\xi}}=\mathbf{0}$, and f_0 coincides with the strongly convex

quadratic $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \beta\bar{\boldsymbol{\xi}}^\top \boldsymbol{x} + c$ on a neighborhood of $\boldsymbol{x}(0)$. Since a convex function with a strong local minimizer admits no other minimizers, $\boldsymbol{x}(0)$ is the unique SAA solution.

To evaluate the gaps produced by this construction, we use the same covariance derivation as in Lemma EC.2: the derivation only relies on the first-order condition (EC.17) and the stationarity property (EC.14), and is therefore not specific to the maximizing radius ρ_{gap}. We now construct radii with these two properties. For any $\gamma > 0$, the point $\boldsymbol{x}_\gamma := (\boldsymbol{H}_\kappa + \gamma \boldsymbol{I})^{-1} \boldsymbol{H}_\kappa \boldsymbol{x}_\kappa(0)$ is nonzero because $\boldsymbol{H}_\kappa \succ \boldsymbol{0}$ and $\boldsymbol{x}_\kappa(0) \neq \boldsymbol{0}$. It satisfies the first-order condition (EC.17) with $\rho(\gamma) := 2\gamma \|\boldsymbol{x}_\gamma\|_2$ in place of ρ_{gap}, hence $\boldsymbol{x}^\star(\rho(\gamma)) = \boldsymbol{x}_\gamma$ by strong convexity of (EC.11). The dual multiplier $\rho(\gamma)/(2\|\boldsymbol{x}^\star(\rho(\gamma))\|_2)$ equals γ itself. If, in addition, γ satisfies the spectral stationarity condition (EC.16) for the law p_θ with mean $m = \mathbb{E}_{p_\theta}[\Lambda]$—in which case we call (θ, γ) a *stationary pair* and $\rho(\gamma)$ its *stationary radius*—then $\|\boldsymbol{x}_\gamma\|_2 = \|\boldsymbol{x}_\kappa(0)\|_2 \sqrt{\mathbb{E}_{p_\theta}[\phi_\gamma(\Lambda)]} = m(m+\gamma)^{-1} \|\boldsymbol{x}_\kappa(0)\|_2$, so that $\rho(\gamma) = 2\gamma m (m+\gamma)^{-1} \|\boldsymbol{x}_\kappa(0)\|_2$. Moreover, since $\rho(\gamma)\|\boldsymbol{x}_\kappa(0)\|_2 / (2\boldsymbol{x}_\kappa(0)^\top \boldsymbol{H}_\kappa \boldsymbol{x}_\kappa(0)) = \gamma/(m+\gamma) < 1$, the positive part in (EC.12) is inactive and $\lambda^\star(\rho(\gamma)) = m/(m+\gamma)$. Thus $\|\hat{\boldsymbol{x}}(\rho(\gamma))\|_2 = \|\boldsymbol{x}_\gamma\|_2$, and the stationarity property (EC.14) indeed holds at $\rho(\gamma)$. For the two-point law p_θ, the covariance formula at a stationary radius therefore reduces to

$$\frac{\Delta(\rho(\gamma))}{\|\boldsymbol{x}_\kappa(0)\|_2^2} = \mathbb{E}_{p_\theta}[\Lambda \phi_\gamma(\Lambda)] - \mathbb{E}_{p_\theta}[\Lambda]\, \mathbb{E}_{p_\theta}[\phi_\gamma(\Lambda)] = \theta(1-\theta)(\kappa - 1)\big[\phi_\gamma(\kappa) - \phi_\gamma(1)\big]. \tag{EC.27}$$

Since $\lambda_{\max} = \kappa$, the upper bound (5) becomes $\|\boldsymbol{x}_\kappa(0)\|_2^2 (\kappa-1)^2 / [\varphi^5(\kappa-1) + 27]$ on these instances. Thus it suffices to exhibit stationary radii whose gaps match this expression asymptotically.

For the third step, first let $\kappa \to \infty$ and set $\alpha := 1/\varphi$, so $\alpha^2 + \alpha = 1$ and $1 + \alpha = 1/\alpha = \varphi$. Take $\gamma_\kappa = \alpha\kappa$ and choose θ_κ to satisfy (EC.16), which for the law p_θ reads $(1-\theta)\,\phi_{\alpha\kappa}(1) + \theta\,\phi_{\alpha\kappa}(\kappa) = \phi_{\alpha\kappa}(m)$ with $m = 1 + \theta(\kappa - 1)$. Since $\phi_{\alpha\kappa}(\kappa) = (1+\alpha)^{-2} = \alpha^2$ for every κ, while $\phi_{\alpha\kappa}(1) \to 0$ and $\phi_{\alpha\kappa}(m) \to \theta^2/(\theta+\alpha)^2$ as $\kappa \to \infty$, the stationarity equation converges to

$$\alpha^2 \theta = \frac{\theta^2}{(\theta + \alpha)^2},$$

whose roots in $(0, 1]$ are the interior point $\theta = \alpha^2$—by the defining relation $\alpha^2 + \alpha = 1$—and the boundary point $\theta = 1$. Define the limiting stationarity defect $D(\theta) := \alpha^2\theta - \theta^2/(\theta+\alpha)^2$. Its derivative at the interior root is $D'(\alpha^2) = \alpha^2(1 - 2\alpha) \neq 0$, so D changes sign at α^2. The intermediate value theorem furnishes, for all sufficiently large κ, a root θ_κ of the exact stationarity equation with $\theta_\kappa \to \alpha^2$ and $1 - \theta_\kappa \to \alpha$. Let $\bar{\rho}_\kappa := \rho(\gamma_\kappa) = 2\gamma_\kappa m_\kappa (m_\kappa + \gamma_\kappa)^{-1} \|\boldsymbol{x}_\kappa(0)\|_2$ be the stationary radius of the pair $(\theta_\kappa, \gamma_\kappa)$, where $m_\kappa = 1 + \theta_\kappa(\kappa - 1)$. Since $\phi_{\alpha\kappa}(\kappa) = \alpha^2$ and $\phi_{\alpha\kappa}(1) \to 0$, (EC.27) gives

$$\frac{\Delta(\bar{\rho}_\kappa)}{\kappa \|\boldsymbol{x}_\kappa(0)\|_2^2} \longrightarrow \alpha^2 \alpha \alpha^2 = \frac{1}{\varphi^5}.$$

The upper bound divided by $\kappa \|\boldsymbol{x}_\kappa(0)\|_2^2$ has the same limit, proving tightness as $\kappa \to \infty$.

Finally let $\kappa = 1+\epsilon$ with $\epsilon \downarrow 0$ and choose $\theta = 1/2$. The stationarity defect

$$D_\epsilon(\gamma) := \mathbb{E}_{p_{1/2}}[\phi_\gamma(\Lambda)] - \phi_\gamma\left(1+\frac{\epsilon}{2}\right) = \frac{\epsilon^2}{8}\frac{2\gamma(\gamma-2)}{(1+\gamma)^4} + O(\epsilon^3)$$

vanishes at $\gamma = 2$ to leading order in ϵ, and the factor $2\gamma(\gamma-2)/(1+\gamma)^4$ has derivative $4/81 \neq 0$ at $\gamma = 2$. For all sufficiently small ϵ, the defect therefore changes sign near $\gamma = 2$, and the intermediate value theorem furnishes a stationary dual $\bar{\gamma}_\epsilon \to 2$. Let $\bar{\rho}_\epsilon := \rho(\bar{\gamma}_\epsilon)$ be the corresponding stationary radius. Since $\phi_2'(1) = 4/27$, (EC.27) gives

$$\frac{\Delta(\bar{\rho}_\epsilon)}{\|\boldsymbol{x}_\kappa(0)\|_2^2} = \frac{1}{4}\,\epsilon\,[\phi_{\bar{\gamma}_\epsilon}(1+\epsilon) - \phi_{\bar{\gamma}_\epsilon}(1)] = \frac{\epsilon^2}{27} + O(\epsilon^3),$$

which matches $\epsilon^2/[\varphi^5\epsilon + 27] = \epsilon^2/27 + O(\epsilon^3)$. This proves tightness as $\kappa \to 1^+$ and completes the proof. □

Proof of Corollary 1. For each κ, we use the two-point instance (EC.26) with $\theta = \theta_\kappa$, where θ_κ is the parameter constructed in the $\kappa \to \infty$ tightness step of the proof of Theorem 2 and satisfies $\theta_\kappa \to \alpha^2$ for $\alpha := 1/\varphi$. The proof proceeds in three steps: the first step shows that the maximal gap of the κth instance asymptotically attains the worst-case bound (5); the second step identifies the asymptotic scale of the dual multiplier $\gamma_{\mathrm{gap},\kappa}$ at a gap-maximizing radius; and the third step converts that scale into the three radii in the statement.

For the first step, let $\rho_{\mathrm{gap},\kappa}$ be a radius that maximizes the gap $\Delta(\rho)$ for the κth instance. The stationary radius $\bar{\rho}_\kappa$ constructed in the proof of Theorem 2 is an admissible candidate radius for the same instance, so $\Delta(\rho_{\mathrm{gap},\kappa}) \geq \Delta(\bar{\rho}_\kappa)$. The lower bound from that construction and the upper bound (5) therefore squeeze the maximizing gap:

$$\frac{\Delta(\rho_{\mathrm{gap},\kappa})}{\kappa\|\boldsymbol{x}_\kappa(0)\|_2^2} \longrightarrow \alpha^5 = \frac{1}{\varphi^5}. \tag{EC.28}$$

For all sufficiently large κ, this gap is positive. Lemma EC.2 then applies at $\rho_{\mathrm{gap},\kappa}$. We write $\gamma_{\mathrm{gap},\kappa} := \rho_{\mathrm{gap},\kappa}/(2\|\boldsymbol{x}_\kappa^\star(\rho_{\mathrm{gap},\kappa})\|_2)$ for the dual multiplier γ_{gap} from that lemma, specialized to the κth instance.

For the second step, we claim that $\gamma_{\mathrm{gap},\kappa}/\kappa \to \alpha$. For the two-point instances, $\lambda_{\min} = 1$ and $\lambda_{\max} = \kappa$. Write the scalar endpoint envelope in Lemma EC.3 as $G_\kappa(t) := t^2\big((1+t)^{-1/2} - (\kappa+t)^{-1/2}\big)^2$. The proof of that lemma gives $\Delta(\rho_{\mathrm{gap},\kappa})/\|\boldsymbol{x}_\kappa(0)\|_2^2 \leq G_\kappa(\gamma_{\mathrm{gap},\kappa})$. Together with Lemma EC.4 and (EC.28), this implies the squeeze

$$\alpha^5 \longleftarrow \frac{\Delta(\rho_{\mathrm{gap},\kappa})}{\kappa\|\boldsymbol{x}_\kappa(0)\|_2^2} \leq \frac{G_\kappa(\gamma_{\mathrm{gap},\kappa})}{\kappa} \leq \frac{(\kappa-1)^2}{\kappa[\varphi^5(\kappa-1)+27]} \longrightarrow \alpha^5. \tag{EC.29}$$

Now set $\beta_\kappa := \gamma_{\mathrm{gap},\kappa}/\kappa$ and define

$$h_\kappa(\beta) := \frac{G_\kappa(\beta\kappa)}{\kappa} = \beta^2\left(\frac{1}{\sqrt{\beta+1/\kappa}} - \frac{1}{\sqrt{1+\beta}}\right)^2, \qquad w(\beta) := \beta^2\left(\frac{1}{\sqrt{\beta}} - \frac{1}{\sqrt{1+\beta}}\right)^2.$$

Since $1/\sqrt{\beta+1/\kappa} \to 1/\sqrt{\beta}$ uniformly on every compact subset of $(0,\infty)$, we have that $h_\kappa \to w$ uniformly on every such subset. Also, the conjugate form of G_κ and the inequalities $(\sqrt{1+t}+\sqrt{\kappa+t})^2 \geq \kappa+t$ and $(\kappa-1)^2 \leq \kappa^2$ yield

$$h_\kappa(\beta) \leq \frac{\beta^2\kappa}{(1+\beta\kappa)(1+\beta)^2} \leq \frac{\beta}{(1+\beta)^2} \qquad \forall \beta > 0.$$

Since the right-hand side tends to zero as $\beta \downarrow 0$ and as $\beta \uparrow \infty$, (EC.29) rules out $\beta_\kappa \to 0$ and $\beta_\kappa \to \infty$ along any subsequence. Along any subsequence with $\beta_\kappa \to \bar{\beta} \in (0,\infty)$, the local uniform convergence gives $w(\bar{\beta}) = \lim_{\kappa\to\infty} h_\kappa(\beta_\kappa) = \alpha^5$.

It remains only to identify the maximizer of w. Since $\sqrt{w(\beta)} = \sqrt{\beta} - \beta/\sqrt{1+\beta}$, the first-order condition for an interior maximizer of w is $(1+\beta)^3 = (2+\beta)^2\beta$, equivalently $\beta^2+\beta-1=0$. Its unique positive solution is $\beta=\alpha$. Also $w(\beta)\to 0$ as $\beta\downarrow 0$ and as $\beta\uparrow\infty$, so α is the unique maximizer of w, with $w(\alpha)=\alpha^5$. Hence every convergent subsequence of $\{\beta_\kappa\}$ has limit α, and the claim follows.

For the third step, let $m_\kappa := \boldsymbol{x}_\kappa(0)^\top \boldsymbol{H}_\kappa \boldsymbol{x}_\kappa(0)/\|\boldsymbol{x}_\kappa(0)\|_2^2 = 1+\theta_\kappa(\kappa-1)$. Since $\theta_\kappa \to \alpha^2$, we have $m_\kappa/\kappa \to \alpha^2$. The stationarity relation in Lemma EC.2 and the limit $\gamma_{\mathrm{gap},\kappa}/\kappa \to \alpha$ give

$$\lambda^\star(\rho_{\mathrm{gap},\kappa}) = \frac{m_\kappa}{m_\kappa+\gamma_{\mathrm{gap},\kappa}} \longrightarrow \frac{\alpha^2}{\alpha^2+\alpha} = \alpha^2.$$

The three characteristic radii are

$$\rho_{\mathrm{gap},\kappa} = 2\gamma_{\mathrm{gap},\kappa}\lambda^\star(\rho_{\mathrm{gap},\kappa})\|\boldsymbol{x}_\kappa(0)\|_2, \qquad \rho^{\mathrm{h}}_{0,\kappa} = 2m_\kappa\|\boldsymbol{x}_\kappa(0)\|_2,$$
$$\rho^\star_{0,\kappa} = 2\|\boldsymbol{H}_\kappa\boldsymbol{x}_\kappa(0)\|_2 = 2\|\boldsymbol{x}_\kappa(0)\|_2\sqrt{(1-\theta_\kappa)+\theta_\kappa\kappa^2}.$$

The first identity combines the definition of $\gamma_{\mathrm{gap},\kappa}$ with the norm identity (EC.14) and $\hat{\boldsymbol{x}}_\kappa(\rho) = \lambda^\star(\rho)\boldsymbol{x}_\kappa(0)$; the second is the radius at which the shrinkage factor (EC.12) reaches zero; and the third holds because $\boldsymbol{x}^\star_\kappa(\rho)=\mathbf{0}$ if and only if $\mathbf{0}$ satisfies the subgradient optimality condition $\rho\,\partial\|\cdot\|_2(\mathbf{0}) \ni 2\boldsymbol{H}_\kappa\boldsymbol{x}_\kappa(0)$ in the quadratic problem of (EC.11), that is, if and only if $\rho \geq 2\|\boldsymbol{H}_\kappa\boldsymbol{x}_\kappa(0)\|_2$. After dividing by $2\kappa\|\boldsymbol{x}_\kappa(0)\|_2$, these identities yield

$$\frac{\rho_{\mathrm{gap},\kappa}}{2\kappa\|\boldsymbol{x}_\kappa(0)\|_2} \to \alpha^3, \qquad \frac{\rho^{\mathrm{h}}_{0,\kappa}}{2\kappa\|\boldsymbol{x}_\kappa(0)\|_2} \to \alpha^2, \qquad \frac{\rho^\star_{0,\kappa}}{2\kappa\|\boldsymbol{x}_\kappa(0)\|_2} \to \alpha.$$

Since $1/\alpha = \varphi$, we conclude that

$$\rho_{\mathrm{gap},\kappa} : \rho^{\mathrm{h}}_{0,\kappa} : \rho^\star_{0,\kappa} \longrightarrow \alpha^3 : \alpha^2 : \alpha = 1 : \varphi : \varphi^2.$$

□

The proof of Theorem 3 relies on the following auxiliary results, which we state and prove first. Throughout, we use the notation introduced before Lemma EC.1 and the spectral notation of Lemma EC.2, and we write $\Delta^{0,\infty}(\rho) := \min\{f_\rho(\boldsymbol{x}(0)), f_\rho(\boldsymbol{x}(\infty))\} - f^\star_\rho$ for the additive suboptimality

of the SAA–RO heuristic at the radius ρ. The first lemma locates ρ_{gap}, a maximizer of the suboptimality $\Delta(\rho)$ of the shrinkage path heuristic. The second lemma provides a correlation inequality in the spirit of Cassels' classical reverse Cauchy–Schwarz inequality (Watson 1955, Appendix 1).

LEMMA EC.5. *Let the conditions of Lemma EC.2 hold. Then the gap-maximizing radius satisfies* $\rho_{\text{gap}} > \rho_{\text{end}}$, *where* $\rho_{\text{end}} := \boldsymbol{x}(0)^\top \boldsymbol{H}\, \boldsymbol{x}(0)/\|\boldsymbol{x}(0)\|_2$.

Proof of Lemma EC.5. Set $m := \mathbb{E}_p[\Lambda]$. Then $m = \boldsymbol{x}(0)^\top \boldsymbol{H}\boldsymbol{x}(0)/\|\boldsymbol{x}(0)\|_2^2 = \rho_{\text{end}}/\|\boldsymbol{x}(0)\|_2$. By Lemmas EC.1 and EC.2, $\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2 = \|\hat{\boldsymbol{x}}(\rho_{\text{gap}})\|_2 = \lambda^\star(\rho_{\text{gap}})\|\boldsymbol{x}(0)\|_2$ with $\lambda^\star(\rho_{\text{gap}}) = m/(m+\gamma_{\text{gap}})$. Hence

$$\rho_{\text{gap}} = 2\gamma_{\text{gap}}\lambda^\star(\rho_{\text{gap}})\|\boldsymbol{x}(0)\|_2, \qquad \rho_{\text{end}} = m\|\boldsymbol{x}(0)\|_2, \qquad \frac{\rho_{\text{gap}}}{\rho_{\text{end}}} = \frac{2\gamma_{\text{gap}}}{m+\gamma_{\text{gap}}}.$$

It is therefore enough to show that $\gamma_{\text{gap}} > m$.

Suppose, toward a contradiction, that $\gamma_{\text{gap}} \le m$. Let $a(\lambda) := \phi_{\gamma_{\text{gap}}}(m) + \phi'_{\gamma_{\text{gap}}}(m)(\lambda - m)$ be the tangent line to $\phi_{\gamma_{\text{gap}}}$ at m, and set $h := a - \phi_{\gamma_{\text{gap}}}$. Then $h(m) = h'(m) = 0$ and

$$h''(\lambda) = -\phi''_{\gamma_{\text{gap}}}(\lambda) = \frac{2\gamma_{\text{gap}}(2\lambda - \gamma_{\text{gap}})}{(\lambda+\gamma_{\text{gap}})^4}.$$

Thus h is concave on $(0, \gamma_{\text{gap}}/2)$ and strictly convex on $(\gamma_{\text{gap}}/2, \infty)$. Since $m \ge \gamma_{\text{gap}}$, convexity and $h(m) = h'(m) = 0$ imply $h \ge 0$ on $[\gamma_{\text{gap}}/2, \infty)$, with equality only at m. Moreover,

$$h(0) = \phi_{\gamma_{\text{gap}}}(m) - \phi'_{\gamma_{\text{gap}}}(m)m - \phi_{\gamma_{\text{gap}}}(0) = \frac{m^2}{(m+\gamma_{\text{gap}})^2} - \frac{2\gamma_{\text{gap}}m^2}{(m+\gamma_{\text{gap}})^3} = \frac{m^2(m-\gamma_{\text{gap}})}{(m+\gamma_{\text{gap}})^3} \ge 0.$$

Since also $h(\gamma_{\text{gap}}/2) > 0$ (were it zero, convexity and $h \ge 0$ on $[\gamma_{\text{gap}}/2, m]$ together with $h(m) = 0$ would force $h = 0$ on this interval, contradicting strict convexity), concavity on $[0, \gamma_{\text{gap}}/2]$ gives

$$h(\lambda) \ge \frac{\gamma_{\text{gap}}/2 - \lambda}{\gamma_{\text{gap}}/2}\, h(0) + \frac{\lambda}{\gamma_{\text{gap}}/2}\, h(\gamma_{\text{gap}}/2) > 0, \qquad \lambda \in (0, \gamma_{\text{gap}}/2].$$

Thus $h(\lambda) \ge 0$ for all $\lambda > 0$, with equality only at $\lambda = m$. Taking expectations and using the stationarity condition (EC.16) gives

$$\mathbb{E}_p[h(\Lambda)] = \mathbb{E}_p[a(\Lambda)] - \mathbb{E}_p[\phi_{\gamma_{\text{gap}}}(\Lambda)] = a(m) - \phi_{\gamma_{\text{gap}}}(m) = 0.$$

Hence $\Lambda = m$ p-almost surely. The covariance formula (EC.15) would then give $\Delta(\rho_{\text{gap}}) = 0$, contradicting $\Delta(\rho_{\text{gap}}) > 0$. Therefore $\gamma_{\text{gap}} > m$, and so $\rho_{\text{gap}} > \rho_{\text{end}}$. □

LEMMA EC.6. *Let* X *and* Y *be random variables satisfying* $X \in [a, b]$ *and* $Y \in [c, d]$ *almost surely, where* $0 < a \le b$ *and* $0 < c \le d$. *Then*

$$1 - \frac{\mathbb{E}[X]\,\mathbb{E}[Y]}{\mathbb{E}[XY]} \le \frac{(b-a)(d-c)}{(\sqrt{ac} + \sqrt{bd})^2}. \tag{EC.30}$$

Proof of Lemma EC.6. If $a = b$ or $c = d$, then X or Y is almost surely constant and both sides of (EC.30) vanish; we therefore assume $a < b$ and $c < d$. Let $x := \mathbb{E}[X]$ and $y := \mathbb{E}[Y]$. Since $(b - X)(Y - c) \geq 0$ and $(X - a)(d - Y) \geq 0$, taking expectations and dividing by xy gives

$$\frac{\mathbb{E}[XY]}{xy} \leq B_1(x, y) := \frac{by + cx - bc}{xy}, \qquad \frac{\mathbb{E}[XY]}{xy} \leq B_2(x, y) := \frac{dx + ay - ad}{xy}.$$

Thus $\mathbb{E}[XY]/xy \leq \min\{B_1(x, y), B_2(x, y)\}$. Write $u := (x - a)/(b - a)$ and $v := (y - c)/(d - c)$. A direct comparison shows that $B_1(x, y) \leq B_2(x, y)$ if $v \leq u$, and $B_2(x, y) \leq B_1(x, y)$ if $u \leq v$. Moreover, B_1 is nondecreasing in y for fixed x, since $\partial_y B_1 = c(b - x)/(xy^2) \geq 0$, and B_2 is nondecreasing in x for fixed y, since $\partial_x B_2 = a(d - y)/(x^2 y) \geq 0$. Hence the largest possible value of $\min\{B_1, B_2\}$ occurs on the diagonal $u = v = t$.

On this diagonal, $x = a + (b - a)t$ and $y = c + (d - c)t$, and the two bounds coincide:

$$\frac{\mathbb{E}[XY]}{\mathbb{E}[X]\,\mathbb{E}[Y]} \leq g(t) := \frac{ac + (bd - ac)t}{\big(a + (b - a)t\big)\big(c + (d - c)t\big)}.$$

Differentiating gives

$$g'(t) = \frac{(b - a)(d - c)\big(ac(1 - t)^2 - bd\,t^2\big)}{\big(a + (b - a)t\big)^2\big(c + (d - c)t\big)^2}.$$

Thus g is maximized at $t^\star = \sqrt{ac}/(\sqrt{ac} + \sqrt{bd})$, where

$$g(t^\star) = \frac{(\sqrt{ac} + \sqrt{bd})^2}{(\sqrt{ad} + \sqrt{bc})^2}.$$

Since $\mathbb{E}[XY]/(\mathbb{E}[X]\,\mathbb{E}[Y]) \leq g(t^\star)$, we have $\mathbb{E}[X]\,\mathbb{E}[Y]/\mathbb{E}[XY] \geq 1/g(t^\star)$. Therefore

$$1 - \frac{\mathbb{E}[X]\,\mathbb{E}[Y]}{\mathbb{E}[XY]} \leq 1 - \frac{1}{g(t^\star)} = \frac{(\sqrt{ac} + \sqrt{bd})^2 - (\sqrt{ad} + \sqrt{bc})^2}{(\sqrt{ac} + \sqrt{bd})^2} = \frac{(b - a)(d - c)}{(\sqrt{ac} + \sqrt{bd})^2}.$$

□

Proof of Theorem 3. We use the notation introduced before Lemma EC.5. The proof proceeds in two steps: the first step establishes a spectral bound on the multiplicative suboptimality using Lemmas EC.2 and EC.5; and the second step maximizes this spectral bound over admissible spectra using Lemma EC.6, which yields the bound (6).

For the first step, recall that the multiplicative suboptimality is the ratio of the worst-case additive suboptimality of the shrinkage path heuristic to that of the SAA–RO heuristic. Since ρ_{gap} maximizes $\Delta(\rho)$ over $\rho \geq 0$, this ratio is

$$\frac{\sup_{\rho \geq 0} \Delta(\rho)}{\sup_{\rho \geq 0} \Delta^{0,\infty}(\rho)} = \frac{\Delta(\rho_{\mathrm{gap}})}{\sup_{\rho \geq 0} \Delta^{0,\infty}(\rho)}, \tag{EC.31}$$

where the denominator is shown below to be positive whenever $\boldsymbol{x}(0) \neq \mathbf{0}$.

If $\boldsymbol{x}(0)=\mathbf{0}$, then the WDRO solution, the shrinkage path, and the SAA–RO heuristic are all identically $\mathbf{0}$, so both additive gaps are zero for every $\rho\geq 0$; we interpret the multiplicative suboptimality as zero in this case, and the bound holds trivially. We therefore assume $\boldsymbol{x}(0)\neq\mathbf{0}$ in the remainder.

Lemma EC.1 gives $\boldsymbol{x}(\infty)=\mathbf{0}$ and the quadratic representation of f_ρ. Evaluating it at the two endpoints gives the SAA–RO value

$$\min\{f_\rho(\boldsymbol{x}(0)), f_\rho(\boldsymbol{x}(\infty))\} = f_0^\star + \min\{\rho\|\boldsymbol{x}(0)\|_2,\ \boldsymbol{x}(0)^\top \boldsymbol{H}\boldsymbol{x}(0)\}. \tag{EC.32}$$

In particular, for $\rho_{\text{end}}=\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)/\|\boldsymbol{x}(0)\|_2$, the endpoint heuristic value is $f_0^\star+\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)$, while (EC.13) gives $\hat{f}_{\rho_{\text{end}}}=f_0^\star+\frac{3}{4}\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)$. Hence the denominator in (EC.31) is positive:

$$\begin{aligned}\sup_{\rho\geq 0}\Delta^{0,\infty}(\rho) &\geq \Delta^{0,\infty}(\rho_{\text{end}}) = \min\{f_{\rho_{\text{end}}}(\boldsymbol{x}(0)), f_{\rho_{\text{end}}}(\boldsymbol{x}(\infty))\} - f^\star_{\rho_{\text{end}}}\\ &\geq \min\{f_{\rho_{\text{end}}}(\boldsymbol{x}(0)), f_{\rho_{\text{end}}}(\boldsymbol{x}(\infty))\} - \hat{f}_{\rho_{\text{end}}} = \tfrac{1}{4}\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0) > 0,\end{aligned}$$

where the last inequality uses $\boldsymbol{H}\succ\mathbf{0}$ and $\boldsymbol{x}(0)\neq\mathbf{0}$.

If $\Delta(\rho_{\text{gap}})=0$, then the bound is immediate. We therefore assume $\Delta(\rho_{\text{gap}})>0$. Lemma EC.5 gives $\rho_{\text{gap}}>\rho_{\text{end}}$, so (EC.32) gives $\min\{f_{\rho_{\text{gap}}}(\boldsymbol{x}(0)), f_{\rho_{\text{gap}}}(\boldsymbol{x}(\infty))\} = f_0^\star+\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)$. Combining this with the identity $f^\star_{\rho_{\text{gap}}} = f_0^\star+\boldsymbol{x}(0)^\top\boldsymbol{H}\boldsymbol{x}(0)-\boldsymbol{x}^\star(\rho_{\text{gap}})^\top\boldsymbol{H}\boldsymbol{x}^\star(\rho_{\text{gap}})$ from Lemma EC.2 lower bounds the denominator of (EC.31):

$$\sup_{\rho\geq 0}\Delta^{0,\infty}(\rho) \geq \Delta^{0,\infty}(\rho_{\text{gap}}) = \boldsymbol{x}^\star(\rho_{\text{gap}})^\top\boldsymbol{H}\boldsymbol{x}^\star(\rho_{\text{gap}}).$$

Combining this lower bound with (EC.18) gives

$$\frac{\sup_{\rho\geq 0}\Delta(\rho)}{\sup_{\rho\geq 0}\Delta^{0,\infty}(\rho)} \leq \frac{\Delta(\rho_{\text{gap}})}{\Delta^{0,\infty}(\rho_{\text{gap}})} = 1-\frac{\|\boldsymbol{x}^\star(\rho_{\text{gap}})\|_2^2}{\|\boldsymbol{x}(0)\|_2^2}\frac{\boldsymbol{x}(0)^\top\boldsymbol{H}\,\boldsymbol{x}(0)}{\boldsymbol{x}^\star(\rho_{\text{gap}})^\top\boldsymbol{H}\,\boldsymbol{x}^\star(\rho_{\text{gap}})}.$$

Using the spectral notation of Lemma EC.2 and $\tilde{x}_i^\star(\rho_{\text{gap}})=\lambda_i(\lambda_i+\gamma_{\text{gap}})^{-1}\tilde{x}_i(0)$, the last bound becomes

$$\frac{\sup_{\rho\geq 0}\Delta(\rho)}{\sup_{\rho\geq 0}\Delta^{0,\infty}(\rho)} \leq 1-\frac{\mathbb{E}_p[\Lambda]\,\mathbb{E}_p[\phi_{\gamma_{\text{gap}}}(\Lambda)]}{\mathbb{E}_p[\Lambda\phi_{\gamma_{\text{gap}}}(\Lambda)]}. \tag{EC.33}$$

For the second step, we maximize the spectral representation in (EC.33). Since $\phi_{\gamma_{\text{gap}}}$ is increasing, Lemma EC.6 applies with $X=\Lambda$ and $Y=\phi_{\gamma_{\text{gap}}}(\Lambda)$. Thus, with $a=\lambda_{\min}$, $b=\lambda_{\max}$, $c=\phi_{\gamma_{\text{gap}}}(\lambda_{\min})$, and $d=\phi_{\gamma_{\text{gap}}}(\lambda_{\max})$, we obtain

$$\frac{\sup_{\rho\geq 0}\Delta(\rho)}{\sup_{\rho\geq 0}\Delta^{0,\infty}(\rho)} \leq \frac{(\lambda_{\max}-\lambda_{\min})\big[\phi_{\gamma_{\text{gap}}}(\lambda_{\max})-\phi_{\gamma_{\text{gap}}}(\lambda_{\min})\big]}{\big(\lambda_{\min}^{3/2}/(\lambda_{\min}+\gamma_{\text{gap}})+\lambda_{\max}^{3/2}/(\lambda_{\max}+\gamma_{\text{gap}})\big)^2} = \frac{\eta(\kappa-1)^2\big(2\kappa+\eta(\kappa+1)\big)}{\big(\kappa+\eta+\kappa^{3/2}(1+\eta)\big)^2}, \tag{EC.34}$$

where $\eta:=\gamma_{\text{gap}}/\lambda_{\min}$. It remains to maximize the η-parameterized bound over $\eta\geq 0$. For fixed $\kappa\geq 1$, write this bound as $R_\kappa(\eta):=\frac{\eta(\kappa-1)^2\left(2\kappa+\eta(\kappa+1)\right)}{\left(\kappa+\eta+\kappa^{3/2}(1+\eta)\right)^2}$. Differentiating $R_\kappa(\eta)$ gives

$$\frac{\mathrm{d}R_\kappa(\eta)}{\mathrm{d}\eta} = \frac{(\kappa-1)^2\left(2\kappa(\kappa+\kappa^{3/2})+2\kappa(\kappa+\sqrt{\kappa})\eta\right)}{\left(\kappa+\eta+\kappa^{3/2}(1+\eta)\right)^3} \geq 0 \qquad \forall\kappa\geq 1,\ \eta\geq 0.$$

Thus $R_\kappa(\eta)$ is nondecreasing on $[0,\infty)$, and hence

$$\sup_{\eta\geq 0} R_\kappa(\eta) = \lim_{\eta\to\infty} R_\kappa(\eta) = \frac{(\kappa-1)^2(\kappa+1)}{(\kappa^{3/2}+1)^2}.$$

Combining this supremum with (EC.34) yields the bound (6) in Theorem 3. Finally, since the denominator of (6) is larger than the numerator by

$$(\kappa^{3/2}+1)^2 - (\kappa-1)^2(\kappa+1) = \kappa(\sqrt{\kappa}+1)^2 > 0,$$

the bound (6) is strictly less than one for every $\kappa \geq 1$. □

Proof of Theorem 4. Throughout, write $D := W(\hat{\mathbb{P}}, \mathbb{Q}(\infty))$ and $\mathbb{Q}_\lambda := (1-\lambda)\hat{\mathbb{P}} + \lambda\mathbb{Q}(\infty)$, so that $\widehat{\mathcal{B}}(\hat{\mathbb{P}}) = \{\mathbb{Q}_\lambda : \lambda \in [0,1]\}$. The existence and finite support of $\mathbb{Q}(\infty)$ are established in the GitHub companion. The proof proceeds in three steps: the first step handles the trivial case $\hat{\mathbb{P}} = \mathbb{Q}(\infty)$ and otherwise shows that (9) is a maximization over those $\lambda \in [0,1]$ for which $\mathbb{Q}_\lambda$ is feasible; the second step shows that the objective of this maximization is nondecreasing in λ, so that the largest feasible λ is optimal and yields $\widehat{\mathbb{Q}}(\rho)$; and the third step verifies that the infimum in (10) is attained.

For the first step, we begin with the trivial case $\hat{\mathbb{P}} = \mathbb{Q}(\infty)$. In this case, $\mathbb{Q}_\lambda = \hat{\mathbb{P}}$ for every $\lambda \in [0,1]$, so the dual shrinkage path collapses to $\{\hat{\mathbb{P}}\}$ and $\widehat{\mathbb{Q}}(\rho) = \hat{\mathbb{P}}$ for every $\rho \geq 0$. Since $W(\hat{\mathbb{P}}, \hat{\mathbb{P}}) = 0 \leq \rho$, the intersection in (9) equals $\{\hat{\mathbb{P}}\}$, and thus $\underline{J}(\rho) = \min_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$, attained at the SAA solution $\boldsymbol{x}(0)$. We therefore assume $\hat{\mathbb{P}} \neq \mathbb{Q}(\infty)$ in the remainder of the proof, so that $D > 0$ since the Wasserstein distance is a metric.

Kantorovich–Rubinstein duality for the type-1 Wasserstein distance gives, for every $\lambda \in [0,1]$,

$$W(\hat{\mathbb{P}}, \mathbb{Q}_\lambda) = \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \left\{ \int f \,\mathrm{d}\hat{\mathbb{P}} - \int f \,\mathrm{d}\mathbb{Q}_\lambda \right\} = \lambda \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \left\{ \int f \,\mathrm{d}\hat{\mathbb{P}} - \int f \,\mathrm{d}\mathbb{Q}(\infty) \right\} = \lambda D.$$

Since $\mathcal{P}(\Xi)$ is convex and $\hat{\mathbb{P}}, \mathbb{Q}(\infty) \in \mathcal{P}(\Xi)$, every $\mathbb{Q}_\lambda$ belongs to $\mathcal{P}(\Xi)$. Hence $\mathbb{Q}_\lambda \in \mathbb{B}_\rho(\hat{\mathbb{P}})$ if and only if $W(\hat{\mathbb{P}}, \mathbb{Q}_\lambda) \leq \rho$, which is equivalent to $\lambda D \leq \rho$, that is, to $\lambda \leq \rho/D$. Intersecting with $\lambda \in [0,1]$ yields $\widehat{\mathcal{B}}(\hat{\mathbb{P}}) \cap \mathbb{B}_\rho(\hat{\mathbb{P}}) = \{\mathbb{Q}_\lambda : \lambda \in [0, \overline{\lambda}]\}$ with $\overline{\lambda} := \min\{\rho/D, 1\}$. Therefore, the a posteriori lower bound (9) is equivalent to the univariate maximization problem $\underline{J}(\rho) = \max_{\lambda\in[0,\overline{\lambda}]} \phi(\lambda)$, where $\phi(\lambda) := \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}_\lambda}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$.

For the second step, fix $\boldsymbol{x}$. The map $\lambda \mapsto \mathbb{E}_{\mathbb{Q}_\lambda}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$ is affine, so ϕ is concave on $[0,1]$ as a pointwise infimum of affine functions. Also, $\mathbb{Q}_1 = \mathbb{Q}(\infty)$ maximizes $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$ over $\mathcal{P}(\Xi)$ by (8); hence $\phi(\lambda) \leq \phi(1)$ for every $\lambda \in [0,1]$. If $0 \leq \lambda_1 < \lambda_2 \leq 1$ and $t := (\lambda_2 - \lambda_1)/(1-\lambda_1)$, then $\lambda_2 = (1-t)\lambda_1 + t$ and concavity gives $\phi(\lambda_2) \geq (1-t)\phi(\lambda_1) + t\phi(1) \geq \phi(\lambda_1)$. Thus ϕ is nondecreasing, and the maximum is attained at $\overline{\lambda} = \min\{\rho/D, 1\} = 1 - \mu(\rho)$. Substituting this value into $\mathbb{Q}_{\overline{\lambda}}$ gives $\mathbb{Q}_{\overline{\lambda}} = \mu(\rho)\hat{\mathbb{P}} + (1-\mu(\rho))\mathbb{Q}(\infty) = \widehat{\mathbb{Q}}(\rho)$, and therefore $\underline{J}(\rho) = \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\widehat{\mathbb{Q}}(\rho)}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$.

For the third step, it remains to replace the infimum in this representation of $\underline{J}(\rho)$ by a minimum. Since $\mathbb{Q}(\infty)$ is an RO distribution, an auxiliary result in the GitHub companion gives $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = J(\infty)$, where $J(\infty)$ is the finite optimal value of the RO problem (3b). Thus $\mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \geq J(\infty)$ for every $\boldsymbol{x}\in\mathcal{X}(\Xi)$. If $\overline{\lambda}=1$, then $\widehat{\mathbb{Q}}(\rho)=\mathbb{Q}(\infty)$, and the RO solution $\boldsymbol{x}(\infty)$ satisfies $\mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x}(\infty),\tilde{\boldsymbol{\xi}})] \leq \sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x}(\infty),\boldsymbol{\xi}) = J(\infty)$, so $\boldsymbol{x}(\infty)$ attains the infimum. If $\overline{\lambda}<1$, then $\mu(\rho)>0$. For every $\boldsymbol{x}\in\mathcal{X}(\Xi)$,

$$\begin{aligned}\mathbb{E}_{\widehat{\mathbb{Q}}(\rho)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] &= \mu(\rho)\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] + (1-\mu(\rho))\mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \\ &\geq \mu(\rho)\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] + (1-\mu(\rho))J(\infty).\end{aligned}$$

Assumption 3*(i)* makes the right-hand side coercive on $\mathcal{X}(\Xi)$, and hence the objective is coercive as well. The distribution $\widehat{\mathbb{Q}}(\rho)$ is finitely supported, so the objective is a finite nonnegative weighted sum of lower semicontinuous functions $\ell(\cdot,\boldsymbol{\xi})$ and is therefore lower semicontinuous. The set $\mathcal{X}(\Xi)$ is closed by the closedness of $\mathcal{X}$ and of each constraint $g_m(\cdot,\boldsymbol{\xi})\leq\gamma_m$. Weierstrass' theorem therefore gives an optimizer, and the infimum is a minimum. □

Proof of Theorem 5. We prove the three claims in order. Throughout, write $D := W(\hat{\mathbb{P}},\mathbb{Q}(\infty))$.

For monotonicity, let $0\leq\rho_1\leq\rho_2$. Then $\mathbb{B}_{\rho_1}(\hat{\mathbb{P}})\subseteq\mathbb{B}_{\rho_2}(\hat{\mathbb{P}})$. For every fixed $\boldsymbol{x}$, the inner worst-case expectation $\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}})}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$ is therefore nondecreasing in ρ. Taking the minimum over the ρ-independent sets $\mathcal{X}(\Xi)$ and $\widehat{\mathcal{X}}(\Xi)$ shows that $J^\star(\rho)$ and $\overline{J}(\rho)$ are nondecreasing. For the lower bound, the feasible sets in (9) also nest: $\widehat{\mathcal{B}}(\hat{\mathbb{P}})\cap\mathbb{B}_{\rho_1}(\hat{\mathbb{P}})\subseteq\widehat{\mathcal{B}}(\hat{\mathbb{P}})\cap\mathbb{B}_{\rho_2}(\hat{\mathbb{P}})$. Maximizing the same objective over a larger set cannot decrease the value, so $\underline{J}(\rho)$ is nondecreasing.

For concavity, we first consider $J^\star$ and $\overline{J}$. The strong duality of the Wasserstein worst-case expectation (Mohajerin Esfahani and Kuhn 2018, Theorem 4.2) gives, for every fixed $\boldsymbol{x}$ and every radius $\rho>0$,

$$\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}})}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = \inf_{\eta\geq 0}\Big\{\eta\rho + \tfrac{1}{N}\sum_{i=1}^{N}\sup_{\boldsymbol{\xi}\in\Xi}\big[\ell(\boldsymbol{x},\boldsymbol{\xi}) - \eta\, d(\boldsymbol{\xi},\hat{\boldsymbol{\xi}}_i)\big]\Big\}.$$

Thus the worst-case expectation is an infimum of functions affine in ρ. Taking the further infimum over $\boldsymbol{x}\in\mathcal{X}(\Xi)$ gives $J^\star(\rho)$, and taking it over $\boldsymbol{x}\in\widehat{\mathcal{X}}(\Xi)$ gives $\overline{J}(\rho)$. Since both feasible sets are independent of ρ, both value functions are pointwise infima of affine functions of ρ, and hence concave on $(0,\infty)$. Concavity on $[0,\infty)$ follows because both functions are nondecreasing by claim *(i)*, and a nondecreasing function on $[0,\infty)$ that is concave on $(0,\infty)$ is concave throughout. For $\underline{J}$, if $D=0$, then $\underline{J}$ is constant—and hence concave—by Theorem 4. If $D>0$, the proof of Theorem 4 shows that $\underline{J}(\rho)=\phi(\min\{\rho/D,1\})$, where $\phi(\lambda) := \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)}\mathbb{E}_{(1-\lambda)\hat{\mathbb{P}}+\lambda\mathbb{Q}(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$ is concave and nondecreasing on $[0,1]$. The map $\rho\mapsto\min\{\rho/D,1\}$ is concave, and composing a nondecreasing concave function with a concave function preserves concavity. Hence $\underline{J}$ is concave.

For endpoint tightness, first take $\rho=0$. The Wasserstein ball is then the singleton $\mathbb{B}_0(\hat{\mathbb{P}})=\{\hat{\mathbb{P}}\}$, so $J^\star(0)=\min_{\boldsymbol{x}\in\mathcal{X}(\Xi)}\mathbb{E}_{\hat{\mathbb{P}}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]=J(0)$, which equals the optimal value of the SAA problem (3a).

Since $\boldsymbol{x}(0) \in \widehat{\mathcal{X}}(\Xi)$ is SAA-optimal over the larger set $\mathcal{X}(\Xi)$, we also have $\overline{J}(0) = J(0)$. The dual shrinkage path intersects $\mathbb{B}_0(\hat{\mathbb{P}})$ only at $\hat{\mathbb{P}}$, so $\underline{J}(0) = J(0)$ as well. Thus $\overline{J}(0) = J^\star(0) = \underline{J}(0) = J(0)$.

Finally, let $\rho \geq D$. Theorem 4 gives $\widehat{\mathbb{Q}}(\rho) = \mathbb{Q}(\infty)$, with the degenerate case $D = 0$ covered by $\hat{\mathbb{P}} = \mathbb{Q}(\infty)$, so $\underline{J}(\rho) = \min_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$, which equals $J(\infty)$ by the same auxiliary result from the GitHub companion invoked in the proof of Theorem 4. Note that $\mathbb{B}_\rho(\hat{\mathbb{P}}) \subseteq \mathcal{P}(\Xi)$ implies

$$J^\star(\rho) \leq \min_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \sup_{\boldsymbol{\xi} \in \Xi} \ell(\boldsymbol{x}, \boldsymbol{\xi}) = J(\infty).$$

Since also $\underline{J}(\rho) \leq J^\star(\rho)$ by (9) and $\underline{J}(\rho) = J(\infty)$, we have $J^\star(\rho) = J(\infty)$. For the upper bound, evaluating the path problem at $\boldsymbol{x}(\infty) \in \widehat{\mathcal{X}}(\Xi)$ gives $\overline{J}(\rho) \leq \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}(\infty), \tilde{\boldsymbol{\xi}})] \leq J(\infty)$. The opposite inequality $\overline{J}(\rho) \geq J^\star(\rho) = J(\infty)$ follows from $\widehat{\mathcal{X}}(\Xi) \subseteq \mathcal{X}(\Xi)$. Hence $\overline{J}(\rho) = J^\star(\rho) = \underline{J}(\rho) = J(\infty)$ for every $\rho \geq D$. □

The proof of Proposition 2 relies on the following three lemmas, one for each of the settings *(i)–(iii)*. Each lemma exhibits a finite convex problem whose solution yields a finitely supported RO distribution $\mathbb{Q}(\infty)$ defined in (8).

LEMMA EC.7 **(Vertex Construction)**. *Let the support* $\Xi = \operatorname{conv}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) + \operatorname{cone}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_R)$ *be a pointed polyhedron with vertices* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$ *and extreme rays* $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_R$*, and let the loss* $\ell(\boldsymbol{x}, \cdot)$ *be convex on* Ξ *with nonpositive recession along every extreme ray, that is,*

$$\ell^\infty(\boldsymbol{x}, \boldsymbol{r}_r) \coloneqq \lim_{\alpha \to \infty} \alpha^{-1} \big[\ell(\boldsymbol{x}, \boldsymbol{\xi} + \alpha \boldsymbol{r}_r) - \ell(\boldsymbol{x}, \boldsymbol{\xi})\big] \leq 0 \qquad \forall \boldsymbol{x} \in \mathcal{X}(\Xi),\ \boldsymbol{\xi} \in \Xi,\ r \in [R].$$

Then the RO problem (3b) *admits the finite vertex reformulation*

$$J(\infty) = \min_{\boldsymbol{x} \in \mathcal{X}(\Xi),\, t} \big\{ t : \ell(\boldsymbol{x}, \boldsymbol{v}_j) \leq t,\ j \in [K] \big\}, \tag{EC.35}$$

and a finitely supported RO distribution can be obtained from its dual.

Proof of Lemma EC.7. Fix $\boldsymbol{x} \in \mathcal{X}(\Xi)$ and write any $\boldsymbol{\xi} \in \Xi$ as $\boldsymbol{\xi} = \bar{\boldsymbol{\xi}} + \sum_r \beta_r \boldsymbol{r}_r$ with $\bar{\boldsymbol{\xi}} = \sum_j \alpha_j \boldsymbol{v}_j$, $\boldsymbol{\alpha} \in \Delta_K$, and $\beta_r \geq 0$, where Δ_K denotes the probability simplex on $[K]$. Nonpositive recession gives $\ell(\boldsymbol{x}, \boldsymbol{\xi}) \leq \ell(\boldsymbol{x}, \bar{\boldsymbol{\xi}})$, while convexity gives $\ell(\boldsymbol{x}, \bar{\boldsymbol{\xi}}) \leq \sum_j \alpha_j \ell(\boldsymbol{x}, \boldsymbol{v}_j) \leq \max_{j \in [K]} \ell(\boldsymbol{x}, \boldsymbol{v}_j)$; since each vertex lies in Ξ, the reverse inequality is immediate, so $\sup_{\boldsymbol{\xi} \in \Xi} \ell(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{j \in [K]} \ell(\boldsymbol{x}, \boldsymbol{v}_j)$. The RO problem (3b) therefore reduces to the finite vertex reformulation (EC.35).

Dualizing the epigraph constraints of (EC.35) with multipliers $\boldsymbol{\pi} \geq \mathbf{0}$ and minimizing the Lagrangian over t forces $\sum_j \pi_j = 1$, giving the dual $\max_{\boldsymbol{\pi} \in \Delta_K} \inf_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \sum_j \pi_j \ell(\boldsymbol{x}, \boldsymbol{v}_j)$. As any $t > J(\infty)$ is strictly feasible at $\boldsymbol{x}(\infty)$, strong duality holds and an optimal $\boldsymbol{\pi}^\star \in \Delta_K$ exists with $\inf_{\boldsymbol{x} \in \mathcal{X}(\Xi)} \sum_j \pi_j^\star \ell(\boldsymbol{x}, \boldsymbol{v}_j) = J(\infty)$. Placing the optimal mass on the active vertices yields the candidate RO distribution

$$\mathbb{Q}(\infty) = \sum_{j:\, \pi_j^\star > 0} \pi_j^\star \, \delta_{\boldsymbol{v}_j}.$$

This is indeed an RO distribution: $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \sum_j \pi_j^\star \ell(\boldsymbol{x},\boldsymbol{v}_j) = J(\infty)$, whereas every $\mathbb{Q}\in\mathcal{P}(\Xi)$ satisfies $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \le \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}(\infty),\tilde{\boldsymbol{\xi}})] \le \sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x}(\infty),\boldsymbol{\xi}) = J(\infty)$, so $\mathbb{Q}(\infty)$ attains the maximum in (8) and is finitely supported. □

LEMMA EC.8 **(Finite-Max Construction)**. *In addition to the assumptions of Lemma EC.7, let the feasible set be the polyhedron $\mathcal{X}(\Xi) = \{\boldsymbol{x}\in\mathbb{R}^n : C\boldsymbol{x}\le \boldsymbol{d}\}$, and let $\ell(\boldsymbol{x},\boldsymbol{\xi}) = \max_{l\in[L]} \ell_l(\boldsymbol{x},\boldsymbol{\xi})$, where each $\ell_l(\cdot,\boldsymbol{\xi})$ is convex and differentiable. Define the active vertices, the active loss pieces at each vertex, and the active constraints at the RO solution $(\boldsymbol{x}(\infty), J(\infty))$ as*

$$\mathcal{A} = \{j : \ell(\boldsymbol{x}(\infty),\boldsymbol{v}_j) = J(\infty)\}, \quad \mathcal{L}_j = \{l : \ell_l(\boldsymbol{x}(\infty),\boldsymbol{v}_j) = J(\infty)\}, \quad \mathcal{I}_\infty = \{m : C_m\boldsymbol{x}(\infty) = d_m\}.$$

Then the feasibility linear program

$$\sum_{j\in\mathcal{A}}\sum_{l\in\mathcal{L}_j} \pi_{jl}\,\nabla_{\boldsymbol{x}}\ell_l(\boldsymbol{x}(\infty),\boldsymbol{v}_j) + {C_{\mathcal{I}_\infty}}^\top \boldsymbol{\mu} = \boldsymbol{0}, \qquad \sum_{j\in\mathcal{A}}\sum_{l\in\mathcal{L}_j} \pi_{jl} = 1, \qquad \pi_{jl}\ge 0,\ \boldsymbol{\mu}\ge\boldsymbol{0} \tag{EC.36}$$

is feasible, and every basic feasible solution induces an RO distribution supported on at most $n+1$ atoms.

Proof of Lemma EC.8. The argument in Lemma EC.7 gives $\sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x},\boldsymbol{\xi}) = \max_{j\in[K]} \max_{l\in[L]} \ell_l(\boldsymbol{x},\boldsymbol{v}_j)$, so $\boldsymbol{x}(\infty)$ minimizes this finite maximum over $\mathcal{X}(\Xi)$, and the active terms of the maximum at $\boldsymbol{x}(\infty)$ are exactly the pairs $\{(j,l) : j\in\mathcal{A},\, l\in\mathcal{L}_j\}$.

The linear program (EC.36) is feasible. Indeed, since $\boldsymbol{x}(\infty)$ minimizes the finite maximum over the polyhedron $\mathcal{X}(\Xi)$, first-order optimality gives $\boldsymbol{0}\in\partial_{\boldsymbol{x}}\big(\max_{j,l}\ell_l(\cdot,\boldsymbol{v}_j)\big)(\boldsymbol{x}(\infty)) + N_{\mathcal{X}(\Xi)}(\boldsymbol{x}(\infty))$, where $N_{\mathcal{X}(\Xi)}(\boldsymbol{x}) \coloneqq \{\boldsymbol{z}\in\mathbb{R}^n : \boldsymbol{z}^\top(\boldsymbol{y}-\boldsymbol{x})\le 0\ \forall \boldsymbol{y}\in\mathcal{X}(\Xi)\}$ denotes the normal cone to $\mathcal{X}(\Xi)$ at $\boldsymbol{x}$. The subdifferential of a finite maximum of differentiable convex functions is the convex hull of the active gradients, $\operatorname{conv}\{\nabla_{\boldsymbol{x}}\ell_l(\boldsymbol{x}(\infty),\boldsymbol{v}_j) : j\in\mathcal{A},\, l\in\mathcal{L}_j\}$, and the normal cone to the polyhedron is $\{{C_{\mathcal{I}_\infty}}^\top\boldsymbol{\mu} : \boldsymbol{\mu}\ge\boldsymbol{0}\}$; combining the two identities produces nonnegative π_{jl} summing to one together with a $\boldsymbol{\mu}\ge\boldsymbol{0}$ that satisfy (EC.36).

Any feasible solution of (EC.36) induces an RO distribution. Let $(\pi_{jl}^S, \boldsymbol{\mu}^S)$ be such a solution, put $\pi_j^S \coloneqq \sum_{l\in\mathcal{L}_j}\pi_{jl}^S$, and define $\mathbb{Q}^S(\infty) \coloneqq \sum_{j\in\mathcal{A}:\,\pi_j^S>0} \pi_j^S\,\delta_{\boldsymbol{v}_j}$. For each such j the averaged gradient $\boldsymbol{g}_j^S \coloneqq (\pi_j^S)^{-1}\sum_{l\in\mathcal{L}_j}\pi_{jl}^S\nabla_{\boldsymbol{x}}\ell_l(\boldsymbol{x}(\infty),\boldsymbol{v}_j)$ is a convex combination of active gradients and hence lies in $\partial_{\boldsymbol{x}}\ell(\boldsymbol{x}(\infty),\boldsymbol{v}_j)$, while $\boldsymbol{z}^S \coloneqq {C_{\mathcal{I}_\infty}}^\top\boldsymbol{\mu}^S \in N_{\mathcal{X}(\Xi)}(\boldsymbol{x}(\infty))$, and (EC.36) reads $\sum_{j\in\mathcal{A}}\pi_j^S\boldsymbol{g}_j^S + \boldsymbol{z}^S = \boldsymbol{0}$. For any $\boldsymbol{x}\in\mathcal{X}(\Xi)$ the subgradient inequality gives $\ell(\boldsymbol{x},\boldsymbol{v}_j)\ge J(\infty) + (\boldsymbol{g}_j^S)^\top(\boldsymbol{x}-\boldsymbol{x}(\infty))$ for every $j\in\mathcal{A}$; weighting by π_j^S and summing gives

$$\mathbb{E}_{\mathbb{Q}^S(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \;\ge\; J(\infty) + \Big(\textstyle\sum_{j\in\mathcal{A}}\pi_j^S\boldsymbol{g}_j^S\Big)^\top(\boldsymbol{x}-\boldsymbol{x}(\infty)) \;=\; J(\infty) - (\boldsymbol{z}^S)^\top(\boldsymbol{x}-\boldsymbol{x}(\infty)) \;\ge\; J(\infty),$$

the last step because $\boldsymbol{z}^S \in N_{\mathcal{X}(\Xi)}(\boldsymbol{x}(\infty))$. Equality at $\boldsymbol{x} = \boldsymbol{x}(\infty)$ gives $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}^S(\infty)}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = J(\infty)$, and since every $\mathbb{Q} \in \mathcal{P}(\Xi)$ obeys $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \le \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}(\infty),\tilde{\boldsymbol{\xi}})] \le \sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x}(\infty),\boldsymbol{\xi}) = J(\infty)$, the distribution $\mathbb{Q}^S(\infty)$ is an RO distribution.

Finally, the bound on the number of atoms follows from the structure of a basic feasible solution. The system (EC.36) has $n+1$ equality constraints—n stationarity equations and one normalization—so a basic feasible solution has at most $n+1$ positive entries among the variables $\{\pi_{jl}\} \cup \{\boldsymbol{\mu}\}$, hence at most $n+1$ positive π^S_{jl} and therefore at most $n+1$ atoms $\boldsymbol{v}_j$ in $\mathbb{Q}^S(\infty)$, matching the bound established in the GitHub companion. □

Lemma EC.9 **(Convex–Concave Construction)**. *Let the support Ξ be convex and compact, and let $\ell(\boldsymbol{x},\boldsymbol{\xi}) = \max_{j\in[K]} \ell_j(\boldsymbol{x},\boldsymbol{\xi})$, where each ℓ_j is closed convex with full domain in $\boldsymbol{x}$ and concave and upper semicontinuous in $\boldsymbol{\xi}$ on Ξ. Write $\ell_j^*(\boldsymbol{s},\boldsymbol{\xi}) \coloneqq \sup_{\boldsymbol{x}\in\mathbb{R}^n}\{\boldsymbol{s}^\top\boldsymbol{x} - \ell_j(\boldsymbol{x},\boldsymbol{\xi})\}$ for the Fenchel conjugate of each piece in its first argument, $\sigma_{\mathcal{X}(\Xi)}(\boldsymbol{z}) \coloneqq \sup_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \boldsymbol{z}^\top\boldsymbol{x}$ for the support function of $\mathcal{X}(\Xi)$, and Δ_K for the probability simplex on $[K]$. Then the RO problem* (3b) *admits the finite convex reformulation*

$$J(\infty) = \max_{\boldsymbol{\pi}\in\Delta_K,\, \boldsymbol{\zeta}_j/\pi_j\in\Xi,\, \boldsymbol{s}_j\in\mathbb{R}^n} \left\{ -\sigma_{\mathcal{X}(\Xi)}\Big(-\sum_{j=1}^K \boldsymbol{s}_j\Big) - \sum_{j=1}^K \pi_j\, \ell_j^*\big(\boldsymbol{s}_j/\pi_j, \boldsymbol{\zeta}_j/\pi_j\big) \right\}, \tag{EC.37}$$

with the standard closed-perspective convention at $\pi_j = 0$, under which the jth term of the sum equals 0 if $(\boldsymbol{s}_j,\boldsymbol{\zeta}_j) = (\boldsymbol{0},\boldsymbol{0})$ and $+\infty$ otherwise, and the constraint $\boldsymbol{\zeta}_j/\pi_j \in \Xi$ reduces to $\boldsymbol{\zeta}_j = \boldsymbol{0}$, and every maximizer of (EC.37) *induces a finitely supported RO distribution.*

Proof of Lemma EC.9. For each $\boldsymbol{x}$ we have $\sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x},\boldsymbol{\xi}) = \max_{j\in[K]} \phi_j(\boldsymbol{x})$ with $\phi_j(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{\xi}\in\Xi} \ell_j(\boldsymbol{x},\boldsymbol{\xi})$, where the maxima are attained because Ξ is compact and $\ell_j(\boldsymbol{x},\cdot)$ is upper semicontinuous, and each ϕ_j is real-valued and convex in $\boldsymbol{x}$, hence continuous. Lifting the finite maximum to the simplex gives $\max_j \phi_j(\boldsymbol{x}) = \max_{\boldsymbol{\pi}\in\Delta_K} \sum_j \pi_j\phi_j(\boldsymbol{x})$, and Sion's minimax theorem (Sion 1958) over the convex set $\mathcal{X}(\Xi)$ and the convex compact set Δ_K yields $J(\infty) = \max_{\boldsymbol{\pi}\in\Delta_K} \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \sum_j \pi_j\phi_j(\boldsymbol{x})$. For fixed $\boldsymbol{\pi}$, separability gives $\sum_j \pi_j\phi_j(\boldsymbol{x}) = \max_{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K\in\Xi} \sum_j \pi_j\ell_j(\boldsymbol{x},\boldsymbol{\xi}_j)$, an objective convex in $\boldsymbol{x}$ and concave and upper semicontinuous in $(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K)$. A second application of Sion over the convex compact set Ξ^K swaps the inner infimum and maximum, and combining the two interchanges yields

$$J(\infty) = \max_{\boldsymbol{\pi}\in\Delta_K,\, \boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K\in\Xi} \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \sum_{j=1}^K \pi_j\ell_j(\boldsymbol{x},\boldsymbol{\xi}_j). \tag{EC.38}$$

At any maximizer $(\boldsymbol{\pi}^\star,\boldsymbol{\xi}_1^\star,\ldots,\boldsymbol{\xi}_K^\star)$ the inner infimum equals $J(\infty)$ and is attained at $\boldsymbol{x}(\infty)$, since $\sum_j \pi_j^\star\ell_j(\boldsymbol{x}(\infty),\boldsymbol{\xi}_j^\star) \le \sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x}(\infty),\boldsymbol{\xi}) = J(\infty)$.

Fix $(\boldsymbol{\pi},\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K) \in \Delta_K \times \Xi^K$ with $\boldsymbol{\pi} > \boldsymbol{0}$ and dualize the inner minimization of $\sum_j \pi_j\ell_j(\boldsymbol{x},\boldsymbol{\xi}_j)$ over $\mathcal{X}(\Xi)$; zero-weight pieces drop from this sum and are reinstated below via the closed-perspective

convention with $(\boldsymbol{s}_j, \boldsymbol{\zeta}_j) = (\mathbf{0}, \mathbf{0})$. Each summand is proper closed convex with full domain $\mathbb{R}^n$, so the objective function $\sum_j \pi_j \ell_j(\cdot, \boldsymbol{\xi}_j)$ is finite convex on all of $\mathbb{R}^n$. The indicator function of $\mathcal{X}(\Xi)$ equals 0 on $\mathcal{X}(\Xi)$ and $+\infty$ outside. Adding the indicator function to the objective function enforces the constraint $\boldsymbol{x} \in \mathcal{X}(\Xi)$. The inner infimum in (EC.38) therefore equals the minimization of this sum over all of $\mathbb{R}^n$. Fenchel–Rockafellar duality applies to this sum of the objective function and the indicator function defined by $\mathcal{X}(\Xi)$. Its qualification requires the relative interiors of their domains, $\mathbb{R}^n$ and $\mathcal{X}(\Xi)$, to intersect. This holds since $\mathcal{X}(\Xi)$ is nonempty, and weak duality reads

$$-\sigma_{\mathcal{X}(\Xi)}\Big(-\textstyle\sum_j \boldsymbol{s}_j\Big) - \sum_j \pi_j \ell_j^*\big(\boldsymbol{s}_j/\pi_j, \boldsymbol{\xi}_j\big) \;\le\; \inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \textstyle\sum_j \pi_j \ell_j(\boldsymbol{x}, \boldsymbol{\xi}_j) \qquad \forall\, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_K \in \mathbb{R}^n,$$

with equality and dual attainment whenever the inner infimum is finite—in particular at a maximizer of (EC.38). Because $\ell_j(\boldsymbol{x}, \cdot)$ is concave, $\ell_j^*(\boldsymbol{s}, \boldsymbol{\xi})$ is jointly convex in $(\boldsymbol{s}, \boldsymbol{\xi})$, so its perspective $(\pi_j, \boldsymbol{s}_j, \boldsymbol{\zeta}_j) \mapsto \pi_j \ell_j^*(\boldsymbol{s}_j/\pi_j, \boldsymbol{\zeta}_j/\pi_j)$ is jointly convex; substituting $\boldsymbol{\zeta}_j = \pi_j \boldsymbol{\xi}_j$ turns the constraint $\boldsymbol{\xi}_j \in \Xi$ into $\boldsymbol{\zeta}_j/\pi_j \in \Xi$ and, maximizing the dual jointly with (EC.38), produces the convex reformulation (EC.37). The maximization problem (EC.37) has a concave objective and a convex feasible set. Weak duality bounds its objective at every feasible point by the corresponding inner infimum in (EC.38), which is at most $J(\infty)$. At a maximizer of (EC.38), this inner infimum is finite and equals $J(\infty)$. Fenchel–Rockafellar duality therefore yields multipliers $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K$ for which the objective in (EC.37) attains $J(\infty)$. Thus, the optimal value of (EC.37) is $J(\infty)$ and is attained.

Let $(\boldsymbol{\pi}^\star, \boldsymbol{\zeta}_1^\star, \ldots, \boldsymbol{\zeta}_K^\star, \boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_K^\star)$ maximize (EC.37), and for $\pi_j^\star > 0$ set $\boldsymbol{\xi}_j^\star := \boldsymbol{\zeta}_j^\star/\pi_j^\star \in \Xi$. Fenchel–Rockafellar duality gives $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \sum_j \pi_j^\star \ell_j(\boldsymbol{x}, \boldsymbol{\xi}_j^\star) = J(\infty)$. Define

$$\mathbb{Q}(\infty) = \sum_{j:\, \pi_j^\star > 0} \pi_j^\star \, \delta_{\boldsymbol{\xi}_j^\star}.$$

For any $\boldsymbol{x} \in \mathcal{X}(\Xi)$, $\ell = \max_r \ell_r \ge \ell_j$ yields $\mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] = \sum_j \pi_j^\star \ell(\boldsymbol{x}, \boldsymbol{\xi}_j^\star) \ge \sum_j \pi_j^\star \ell_j(\boldsymbol{x}, \boldsymbol{\xi}_j^\star)$. Moreover, $\sum_j \pi_j^\star \ell_j(\boldsymbol{x}, \boldsymbol{\xi}_j^\star) \ge \inf_{\boldsymbol{y}\in\mathcal{X}(\Xi)} \sum_j \pi_j^\star \ell_j(\boldsymbol{y}, \boldsymbol{\xi}_j^\star) = J(\infty)$. Thus, $\mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] \ge J(\infty)$ for every $\boldsymbol{x} \in \mathcal{X}(\Xi)$. Taking the infimum over $\boldsymbol{x} \in \mathcal{X}(\Xi)$ yields $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] \ge J(\infty)$. Conversely, the infimum is at most $\mathbb{E}_{\mathbb{Q}(\infty)}[\ell(\boldsymbol{x}(\infty), \tilde{\boldsymbol{\xi}})] \le \sup_{\boldsymbol{\xi}\in\Xi} \ell(\boldsymbol{x}(\infty), \boldsymbol{\xi}) = J(\infty)$, so it equals $J(\infty)$. Since every $\mathbb{Q} \in \mathcal{P}(\Xi)$ obeys $\inf_{\boldsymbol{x}\in\mathcal{X}(\Xi)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] \le \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}(\infty), \tilde{\boldsymbol{\xi}})] \le J(\infty)$, the distribution $\mathbb{Q}(\infty)$ attains the maximum in (8) and is a finitely supported RO distribution. □

Proof of Proposition 2. Lemmas EC.7, EC.8, and EC.9 construct a finitely supported RO distribution from the solution of a finite convex problem under the assumptions of settings *(i)*, *(ii)*, and *(iii)*, respectively. □