Distributionally Robust Optimization via Targeted Integral Probability Metrics for General Data Processes

Lanran Fang*, Jianqiang Cheng†, Grani A. Hanasusanto*, and Yijie Wang‡

Distributionally robust optimization (DRO) has been successful in addressing decision-making problems under uncertainty when the underlying distribution is unknown. Existing data-driven DRO frameworks, however, often impose restrictive assumptions on the data-generating process. We propose a new DRO framework based on targeted integral probability metrics. The ambiguity set is defined directly through the loss functions induced by feasible decisions, leading to an expected hinge-constrained formulation that is equivalent to an infinitely constrained ambiguity set. This targeted construction aligns the discrepancy measure with the downstream task and yields finite-sample guarantees that bypass the curse of dimensionality: whenever a scalar pointwise concentration inequality is available, the ambiguity radius can be calibrated at the canonical $\widetilde{\mathcal{O}}(N^{-1/2})$ rate. As a result, the framework applies broadly to settings including heavier-tailed distributions, Markovian data, outlier-corrupted observations, incomplete data, and contextual optimization. We derive exact infinite-dimensional dual reformulations, establish out-of-sample and excess-risk guarantees, and develop a Monte Carlo approximation scheme that yields conservative sampled ambiguity sets together with convergence and suboptimality guarantees. For piecewise affine losses, the sampled problems admit tractable conic reformulations, and the Monte Carlo approximation converges at a provably fast rate. Numerical experiments in inventory management under heavy-tailed demand and regression with outlier corruption demonstrate the superiority of our framework over state-of-the-art approaches.

Key words: distributionally robust optimization, integral probability metrics, curse of dimensionality, heavy-tailed distributions, Markovian data, corrupted and missing data, contextual optimization

1. Introduction

Distributionally Robust Optimization (DRO) provides a principled mathematical framework for optimization under uncertainty. Rather than placing blind faith in a single nominal distribution (e.g., empirical distribution), DRO identifies a decision that minimizes the worst-case expected loss over an ambiguity set $\mathcal{P}$, which is defined as a family of probability measures that are statistically

* Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. Email: lanranf2, gah@illinois.edu.

† Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ 85721. Email: jqcheng@arizona.edu.

‡ School of Economics and Management, Tongji University, Shanghai, China. Email: yijiewang@tongji.edu.cn.

plausible given the available data. This robust approach can be formulated as the following min-max problem:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})], \tag{1}$$

where $\ell(\boldsymbol{x},\boldsymbol{\xi})$ represents the loss function parameterized by the decision vector $\boldsymbol{x}\in\mathcal{X}$ and the uncertain parameter $\boldsymbol{\xi}\in\Xi$. By optimizing over the most adverse distribution $\mathbb{Q}$ within $\mathcal{P}$, the DRO framework yields decisions that remain reliable even when deployed in unseen environments.

Early literature predominantly constructed these ambiguity sets using moment bounds or ϕ-divergence, which, despite computational advantages, exhibited irreconcilable statistical and topological flaws. Moment ambiguity sets constrain the family of probability distributions via a finite number of generalized moment conditions. A fundamental drawback of moment ambiguity sets is the lack of asymptotic consistency: gathering infinite data will not cause the DRO solution to converge to the true stochastic programming optimal solution, as low-order moments cannot uniquely identify the true distribution. To overcome the failure of moment sets to converge to the true distribution, Ben-Tal et al. (2012) introduce discrepancy-based ambiguity sets. These sets center around the empirical distribution $\hat{\mathbb{P}}_N$ and use ϕ-divergences to define a neighborhood, capturing all distributions that are statistically close to the empirical one. While ϕ-divergence sets exhibit good statistical consistency, they have severe topological limitations when applied to continuous support spaces. Specifically, if the adversarial distribution $\mathbb{Q}$ assigns positive probability to any event that has zero probability under the reference distribution, the divergence instantly becomes infinite. This prohibits the ϕ-divergence to provide robustness against unseen parameter realizations, severely crippling its out-of-sample generalization capability.

Recently, the Wasserstein distance has become the dominant metric for constructing ambiguity sets due to its nice geometric interpretation and asymptotic consistency. The p-Wasserstein distance between two probability distributions $\mathbb{Q}$ and $\mathbb{P}$ is defined as the p-th root of the minimum expected cost of 'transporting' one measure to match the other. Unlike ϕ-divergences, even if two distributions have mutually exclusive support sets (e.g., one discrete, one continuous), the Wasserstein distance between them remains finite. Therefore, the Wasserstein ambiguity sets enable to smoothly transport probability mass from historical observations to adjacent unobserved regions in the support set Ξ, enabling genuine out-of-sample robustness.

Despite its conceptual elegance, Wasserstein DRO is hindered by the curse of dimensionality. To offer finite-sample guarantees, the radius of the Wasserstein ambiguity set must be calibrated so that the unknown true distribution $\mathbb{P}^\star$ is contained within the Wasserstein ball with high probability. Using the concentration of empirical Wasserstein distance theory derived by Fournier and Guillin (2015), Mohajerin Esfahani and Kuhn (2018) show that the empirical measure converges

to the true measure in the Wasserstein metric at a slow rate. For a random variable residing in a d-dimensional space, the concentration inequality yields a radius that scales as $\mathcal{O}(N^{-1/\max\{2,d\}})$. This indicates that the decay rate of the radius is fundamentally bottlenecked by the dimension of data. For high dimensional problems, achieving a tight confidence region requires an exponentially large and practically unattainable sample size.

To bypass the curse of dimensionality inherent in Wasserstein DRO, a stream of studies has proposed a variety of advanced methodologies. Blanchet et al. (2019), Blanchet and Kang (2021), Blanchet et al. (2022) propose a novel approach where the Wasserstein ambiguity set is not required to contain the true distribution $\mathbb{P}^\star$, but rather at least one distribution that shares the true optimal solution. This targeted coverage enables the optimal Wasserstein radius to achieve an asymptotic decay rate of $\mathcal{O}(1/\sqrt{N})$. Shafieezadeh-Abadeh et al. (2019) show that 1-Wasserstein DRO applied to specific linear models (e.g., support vector machine and logistic regression) reduces to empirical risk minimization with an additive dual norm regularization penalty. Because the infinite-dimensional supremum over the Wasserstein ball analytically collapses into a simple parametric penalty, the convergence of the learning algorithm no longer depends on matching probability measures in high dimensions, allowing the radius to shrink at a non-asymptotic rate of $\mathcal{O}(1/\sqrt{N})$. Concurrently, Wu et al. (2025) extend this regularization equivalence for general risk measures and Wasserstein balls.

A crucial step toward non-asymptotic frameworks for Wasserstein DRO was achieved by Gao (2023), which offers the first finite-sample guarantees that avoid the curse of dimensionality for generic loss functions. Gao (2023) assumes that the underlying data-generating distribution satisfies a Talagrand transportation-information inequality (Villani et al. 2008, Gozlan and Léonard 2010). Through this lens, he establishes a novel variation-based concentration result, showing that the decay rate of the tail probability is controlled by the variation of the loss rather than the ambient dimension. By combining this with localized Rademacher complexity theory, the framework elegantly bridges Wasserstein DRO with variation regularization, yielding an optimal $\mathcal{O}(1/\sqrt{N})$ generalization bound for Wasserstein robust learning. Gao et al. (2024) further develop a unified variation regularization theory that establishes non-asymptotic performance guarantees for non-convex and non-smooth losses. Additionally, Azizian et al. (2023) derive exact generalization guarantees for broad model classes in both standard and regularized Wasserstein DRO. Le and Malick (2024) extend these exact guarantees to highly complex and non-smooth parametric families, covering arbitrary transport costs and modern deep learning objectives.

While state-of-the-art Wasserstein frameworks achieve optimal non-asymptotic rates, their reliance on complex analysis and certain assumptions bottlenecks extensions to a wide range of non-standard data regimes. To overcome these theoretical hurdles, this paper proposes a fundamentally different approach to constructing the ambiguity set. We situate our framework within the broader

class of Integral Probability Metrics (IPMs), following the taxonomy established by Kuhn et al. (2025). Formally, given a family of integrable functions $f \in \mathcal{F}$, an IPM measures the discrepancy between two distributions $\mathbb{P}, \hat{\mathbb{P}} \in \mathscr{P}(\Xi)$ as

$$D_{\mathcal{F}}(\mathbb{P}, \hat{\mathbb{P}}) = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{Z}} f(z) d\mathbb{P}(z) - \int_{\mathcal{Z}} f(\hat{z}) d\hat{\mathbb{P}}(\hat{z}) \right|.$$

The IPM is a highly expressive topological framework. Depending on the specific choice of the function class $\mathcal{F}$, it can recover several foundational statistical distances. For instance, when $\mathcal{F}$ consists of all indicator functions of Borel sets, then $D_{\mathcal{F}}$ recovers the total variation distance. If $\mathcal{F}$ represents the set of all 1-Lipschitz continuous functions, then IPM recovers the 1-Wasserstein distance via the Kantorovich-Rubinstein duality. Finally, if $\mathcal{F}$ is defined as the unit ball in a Reproducing Kernel Hilbert Space (RKHS), $D_{\mathcal{F}}$ yields the maximum mean discrepancy (MMD) distance, which is widely adopted in kernel methods.

Leveraging the IPM philosophy, we propose a generalized, hinge-constrained ambiguity set $\mathcal{P}_\epsilon$ designed specifically to align with the optimization objective:

$$\mathcal{P}_\epsilon := \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{c} \mathbb{E}_{\mathbb{Z}} \left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon \right]_+ \right] \leq 0 \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega \end{array} \right\}. \tag{2}$$

Here, ϵ serves as a tunable size parameter of the ambiguity set, and $\mathbb{Z}$ is an auxiliary probability distribution with full support across the entire decision space $\mathcal{X}$. The reference distribution $\hat{\mathbb{P}}_N$ is constructed from available historical data, and the second-moment bound ensures basic statistical regularity. The hinge constraint forces the inner term to be non-positive almost everywhere with respect to $\mathbb{Z}$. Since $\mathbb{Z}$ has a strictly positive density on $\mathcal{X}$, this expected hinge constraint is equivalent to enforcing a semi-infinite constraint across the entirety of the decision space.

PROPOSITION 1 **(Informal)**. *Under mild conditions, the ambiguity set* (2) *is equivalent to:*

$$\mathcal{P}_\epsilon := \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{ll} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \leq \epsilon & \forall \boldsymbol{z} \in \mathcal{X} \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega & \end{array} \right\}.$$

This reformulation reveals the fundamental nature of our proposed set: It is a *targeted* Integral Probability Metric (TIPM) ambiguity set. Our TIPM departs from classical IPM constructions in two critical ways. First, by dropping the absolute value typically required to ensure a symmetric distance metric, our TIPM fundamentally shifts from a generic "statistical distance" to a "risk-aware discrepancy." Because the inner supremum of the DRO problem seeks adversarial distributions that increase the expected loss, symmetric penalization is mathematically redundant. By employing a one-sided upper bound, we allow the ambiguity set to be more flexible: accommodating optimistic distributions while tightly constraining the exact pessimistic deviations that matter for

out-of-sample robustness. Second, unlike classical IPM ambiguity sets that define the test function class $\mathcal{F}$ using arbitrary Lipschitz functions or Reproducing Kernel Hilbert Space (RKHS) balls, our TIPM strictly defines $\mathcal{F}$ as the hypothesis class of the loss functions themselves, parameterized by the feasible decisions $\boldsymbol{z} \in \mathcal{X}$. By linking the discrepancy metric directly to the loss function of the specific optimization problem at hand, we create an ambiguity set that strictly monitors the discrepancies that actually matter to the decision-maker, ignoring irrelevant noise. As such, our proposed framework offers the following contributions:

- **A task-aware ambiguity set with exact characterization**: We introduce a new expected hinge-constrained ambiguity set based on targeted integral probability metrics (TIPMs), in which the discrepancy is defined directly through the loss functions induced by feasible decisions. We show that this ambiguity set is equivalent to an infinitely constrained loss-discrepancy ambiguity set, thereby aligning robustness with the downstream decision problem. The resulting ambiguity set is distributional in nature but task-aware in geometry. We further derive an exact infinite-dimensional dual reformulation of the resulting DRO problem.
- **Curse of dimensionality-free statistical guarantees under broad data regimes**: The TIPM construction yields finite-sample guarantees that bypass the curse of dimensionality. In particular, whenever a scalar pointwise concentration inequality is available, the ambiguity radius can be calibrated at the canonical $\widetilde{\mathcal{O}}(1/\sqrt{N})$ rate, without requiring stringent assumptions such as Talagrand-type transportation inequalities or problem-specific analyses. This allows the framework to accommodate a broad range of settings, including sub-Weibull losses, outlier-corrupted data, Markovian sequences, and incomplete data. Moreover, because the ambiguity set is defined directly in terms of the loss functions, it is agnostic to the representation of the underlying uncertainty and naturally handles settings with mixed discrete and continuous components.
- **A conservative Monte Carlo approximation with convergence and tractability guarantees**: To overcome the intractability caused by the continuous auxiliary distribution, we introduce a Monte Carlo approximation that replaces the ambiguity set with a conservative sampled counterpart. For piecewise affine losses, we establish convergence of optimal values and solutions together with a $\widetilde{\mathcal{O}}(1/\sqrt{M})$ sampling rate and suboptimality guarantees, and show that the resulting sampled problems admit tractable conic reformulations, including second-order cone approximations for two-stage linear losses.
- **Extension to unbounded support**: We extend the sampling and reformulation framework to unbounded support under a single-threshold tail bound implied by sub-Weibull tails. In this setting, we show that the same Monte Carlo principle remains valid up to explicit tail-dependent terms, while preserving tractability for piecewise affine losses.

1.1. Related Literature

IPM ambiguity sets: Among the various methods for constructing ambiguity sets, IPMs have gained increasing attention due to their rich topological properties and ability to unify classical statistical distances (Husain 2020, Birrell et al. 2022). By measuring the discrepancy using a specific class of test functions, the IPM framework flexibly recovers the Total Variation distance, the Wasserstein distance (Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Gao and Kleywegt 2023), and the Maximum Mean Discrepancy distance (Zhu et al. 2021). Consequently, IPM-based and specifically MMD-based DRO formulations have been extensively studied for their computational tractability, generalization bounds, and deep theoretical connections to regularized empirical risk minimization (Shafieezadeh-Abadeh et al. 2019, Duchi and Namkoong 2019, Husain 2020, Zeng and Lam 2022, Iyengar et al. 2023). Building upon this framework, our work departs from generically defined test spaces by restricting the $\mathcal{F}$ strictly to the hypothesis class of the loss functions induced by the feasible decisions $\boldsymbol{z} \in \mathcal{X}$. Benefiting from this targeted construction, the TIPM ambiguity sets achieve better structural compatibility for the downstream decision problem.

Generalization bounds: A central task in DRO is establishing generalization bounds that guarantee reliable out-of-sample performance. Early formulations primarily relied on moment-based ambiguity sets (Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014) and ϕ-divergence-based sets (Ben-Tal et al. 2013, Bertsimas et al. 2018). These approaches benefit from standard concentration inequalities, yielding favorable, dimension-independent $\mathcal{O}(1/\sqrt{N})$ radius decay rate. Recently, Wasserstein DRO has become increasingly prominent due to its nice topological properties (Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Gao and Kleywegt 2023). However, this topological richness comes at a statistical cost: standard concentration-of-measure results (Fournier and Guillin 2015) dictate that the convergence is at a slow rate $\mathcal{O}(N^{-1/\max\{2,d\}})$. To resolve this bottleneck, a line of research has developed novel frameworks to break the curse of dimensionality for Wasserstein DRO (Chen and Paschalidis 2018, Shafieezadeh-Abadeh et al. 2019, Blanchet et al. 2019, Blanchet and Kang 2021, Blanchet et al. 2022, Gao 2023, Gao et al. 2024, Wu et al. 2025), as detailed in our introduction. Besides, other studies have established generalization bounds for Wasserstein DRO using sample-size-independent fixed radii (Lee and Raginsky 2018, Sinha et al. 2017, Najafi et al. 2019), while parallel efforts have explored optimal radius calibration and bias-variance trade-offs within divergence-based frameworks (Duchi and Namkoong 2019, Lam 2019, Duchi et al. 2021). Furthermore, advanced mathematical tools, such as transportation-information inequalities, have also been leveraged to derive generalization bounds for general learning algorithms (Xu and Raginsky 2017, Lopez and Jog 2018, Russo and Zou 2019, Wang et al. 2019, 2021, Xie et al. 2022, He and Goldfeld 2025).

Learning with non-structural data: In practice, decision makers frequently operate in environments where the available data violate the standard independent and identically distributed (i.i.d.) assumption. For instance, time-series and sequential decision problems often involve Markovian data structures (Fan et al. 2021, Li et al. 2021, Nagaraj et al. 2020). Similarly, in selection problems such as hiring and college admissions, outcomes are only observable for accepted candidates, leading to incomplete data regimes (Najafi et al. 2019). Additionally, modern decision-making increasingly relies on auxiliary side information, necessitating optimization over conditional distributions in contextual data settings (Sadana et al. 2025). Furthermore, data collected in real-world environments are often contaminated by noise or adversarial perturbations, resulting in outlier-corrupted datasets (Nietert et al. 2023, 2024, Blanchet et al. 2024, Jiang and Xie 2024). While previous DRO models have attempted to tackle these complex data realities, existing solutions are predominantly problem-specific. They typically rely on specialized ambiguity set constructions or certain assumptions tailored to a single type of data anomaly. Our proposed framework bridges this gap. Benefited from our TIPM structure, the optimal radius convergence rate can be achieved immediately from any scalar, pointwise concentration inequality. This unified approach accommodates a remarkably broad spectrum of non-standard data regimes, offering a concise and unified DRO framework.

1.2. Paper Structure and Notation

Paper structure. The remainder of the paper is organized as follows. Section 2 formally introduces the expected hinge-constrained TIPM ambiguity set, establishes its equivalent semi-infinite loss-discrepancy representation, and derives coverage, out-of-sample, and excess-risk guarantees. It also shows how the same framework applies to several data-generating regimes through pointwise concentration inequalities, including sub-Weibull losses, Markovian samples, outlier-contaminated observations, incomplete data, and contextual information, and develops exact dual reformulations. Section 3 studies a Monte Carlo approximation of the ambiguity set, proves its conservative containment and convergence properties, and derives finite conic reformulations for piecewise affine losses. Section 4 extends these sampling and reformulation results to unbounded support under a single-threshold tail bound implied by sub-Weibull tails. Section 5 evaluates the proposed framework in multi-item newsvendor and outlier-corrupted regression experiments. The appendices collect all proofs, a two-stage linear decision-rule approximation, and detailed experimental specifications.

Notation. For a positive integer n, we write $[n] \coloneqq \{1, \ldots, n\}$. Random variables are designated by tilde signs (e.g., $\tilde{\boldsymbol{\xi}}$), while their realizations are denoted by the same symbols without tildes (e.g., $\boldsymbol{\xi}$). The true data-generating distribution is denoted by $\mathbb{P}^\star$, the data-driven reference distribution by $\hat{\mathbb{P}}_N$, a generic distribution in the ambiguity set by $\mathbb{Q}$, and the sampling distribution over decisions by $\mathbb{Z}$. We use $\mathscr{P}(\Xi)$ for the set of probability distributions on Ξ, $\mathbb{E}_{\mathbb{P}}[\cdot]$ and $\mathrm{Var}_{\mathbb{P}}[\cdot]$ for expectation

and variance under $\mathbb{P}$, and $\mathbb{I}_A$ or $\mathbb{I}[A]$ for the indicator of an event A. The feasible decision set is $\mathcal{X} \subseteq \mathbb{R}^{D_{\boldsymbol{x}}}$, the uncertainty support is $\Xi \subseteq \mathbb{R}^{D_{\boldsymbol{\xi}}}$, and $R_{\boldsymbol{x}}$ and $R_{\boldsymbol{\xi}}$ denote corresponding norm bounds when they are finite. The ambiguity radius is denoted by ϵ, while $\varepsilon(\cdot)$ denotes the nondecreasing function appearing in pointwise concentration bounds. The relaxation level is η, the second-moment budget is Ω, the data and Monte Carlo sample sizes are N and M respectively, and δ and τ denote confidence parameters. We write Δ^J for the probability simplex in $\mathbb{R}^J_+$, $\mathbf{e}$ for the all-ones vector, and $\mathbf{e}_d$ for the dth unit vector. For a scalar u, $[u]_+ \coloneqq \max\{u, 0\}$.

The Euclidean norm is denoted by $\|\cdot\|$ and the operator norm of a matrix by $\|\cdot\|_{\mathrm{op}}$. For a real-valued random variable X and $\vartheta > 0$, we denote its Orlicz norm by

$$\|X\|_{\psi_\vartheta} \coloneqq \inf\Big\{c > 0 : \mathbb{E}\big[\exp\big((|X|/c)^\vartheta\big) - 1\big] \le 1\Big\}.$$

For a set Ξ, its support function is $\sigma_\Xi(\boldsymbol{z}) \coloneqq \sup_{\boldsymbol{\xi}\in\Xi} \boldsymbol{z}^\top \boldsymbol{\xi}$. We denote by $\mathcal{C}(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$, by $\mathcal{C}_+(\mathcal{X}) \coloneqq \{f \in \mathcal{C}(\mathcal{X}) : f(\boldsymbol{x}) \ge 0 \ \forall \boldsymbol{x} \in \mathcal{X}\}$ the cone of nonnegative continuous functions, and by $\mathcal{M}_+(\mathcal{X})$ the cone of finite nonnegative Borel measures on $\mathcal{X}$. For a point $\boldsymbol{x}$ and a set $\mathcal{S}$, $\mathrm{dist}(\boldsymbol{x}, \mathcal{S})$ denotes the distance from $\boldsymbol{x}$ to $\mathcal{S}$, and $\xrightarrow{p}$ denotes convergence in probability. We use $\mathcal{O}(\cdot)$ for upper bounds up to absolute constants, and $\widetilde{\mathcal{O}}(\cdot)$ when logarithmic factors are additionally suppressed. Model-specific notation, such as Markov spectral gaps, MoM block parameters, propensity scores, kernel bandwidths, and tail constants, is introduced locally where needed.

2. Distributionally Robust Optimization with Expected Hinge Constrained Ambiguity Sets

We consider the distributionally robust optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})], \tag{3}$$

where the ambiguity set is given by

$$\mathcal{P}_\epsilon \coloneqq \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{l} \mathbb{E}_{\mathbb{Z}}\Big[\Big[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon\Big]_+\Big] \le 0 \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \le \Omega \end{array} \right\}, \tag{4}$$

and the feasible set $\mathcal{X}$ is compact and convex. Here, $\epsilon \ge 0$ is a size parameter and $\mathbb{Z}$ is any probability distribution with full support on $\mathcal{X}$. In this data-driven ambiguity set, the reference distribution $\hat{\mathbb{P}}_N$ is constructed from the available data, whose description depends on the specific applications. As we detail later, the hinge constraint ensures that the distribution $\mathbb{Q}$ is close to $\hat{\mathbb{P}}_N$ in terms of the expected loss. The second-moment constraint ensures regularity of the distributions in the ambiguity set, enabling rigorous theoretical development. This ambiguity set is designed to achieve both high-probability coverage of the underlying distribution $\mathbb{P}^\star$ and asymptotic convergence at a fast rate, thereby overcoming the curse of dimensionality of transportation-based ambiguity sets.

We consider a piecewise affine loss function of the form

$$\ell(\boldsymbol{x},\boldsymbol{\xi}) \coloneqq \max_{j\in[J]} \left(\boldsymbol{a}_j(\boldsymbol{x})^\top \boldsymbol{\xi} + b_j(\boldsymbol{x})\right), \tag{5}$$

where $\boldsymbol{a}_j(\boldsymbol{x})$ and $b_j(\boldsymbol{x})$, $j\in[J]$, are affine functions of $\boldsymbol{x}$ given by

$$\boldsymbol{a}_j(\boldsymbol{x}) \coloneqq \boldsymbol{A}_j\boldsymbol{x} + \overline{\boldsymbol{a}}_j, \qquad b_j(\boldsymbol{x}) \coloneqq \boldsymbol{b}_j^\top \boldsymbol{x} + \overline{b}_j.$$

This class is broad enough to capture many models arising in operations research and management science, and is standard in the distributionally robust optimization literature (Wiesemann et al. 2014, Mohajerin Esfahani and Kuhn 2018) because it often admits tractable conic reformulations.

It also includes two-stage linear loss functions as a special case:

$$\begin{aligned} \ell(\boldsymbol{x},\boldsymbol{\xi}) \coloneqq \boldsymbol{c}^\top\boldsymbol{x} + \inf\ & \boldsymbol{q}^\top\boldsymbol{y} \\ \text{s.t. } & \boldsymbol{y}\in\mathbb{R}^{D_{\boldsymbol{y}}} \\ & \boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x}) \le \boldsymbol{W}\boldsymbol{y}, \end{aligned} \tag{6}$$

where $\boldsymbol{T}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$ are a matrix and a vector that depend affinely on $\boldsymbol{x}$, and $\boldsymbol{W}$ is a fixed recourse matrix. In this context, $\boldsymbol{x}$ is a here-and-now decision made before the realization of uncertainty, while $\boldsymbol{y}$ is a wait-and-see recourse decision that can be chosen after the uncertainty is realized.

The linear program in (6) can be dualized, yielding the representation

$$\begin{aligned} \ell(\boldsymbol{x},\boldsymbol{\xi}) = \boldsymbol{c}^\top\boldsymbol{x} + \sup\ & (\boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x}))^\top\boldsymbol{p} \\ \text{s.t. } & \boldsymbol{p}\in\mathbb{R}^L_+ \\ & \boldsymbol{q} = \boldsymbol{W}^\top\boldsymbol{p}. \end{aligned} \tag{7}$$

Strong linear programming duality holds under the standard *relatively complete recourse* assumption that the primal problem is feasible for any $\boldsymbol{x}\in\mathcal{X}$ and $\boldsymbol{\xi}\in\Xi$. Since the feasible set of the dual problem is a polyhedron with finitely many extreme points $\{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_Q\}$, the loss function can be rewritten in the piecewise affine form (5), where $J=Q$, $\boldsymbol{a}_j(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{x})^\top\boldsymbol{p}_j$, and $b_j(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})^\top\boldsymbol{p}_j$.

We make the following assumptions:

(A) *Linear growth in* $\boldsymbol{\xi}$: There exist $c_1, c_2\in\mathbb{R}_+$ such that

$$|\ell(\boldsymbol{x},\boldsymbol{\xi})| \le c_1 + c_2\|\boldsymbol{\xi}\| \qquad \forall \boldsymbol{x}\in\mathcal{X}\ \forall\boldsymbol{\xi}\in\Xi.$$

(B) *Strict second-moment feasibility of* $\hat{\mathbb{P}}_N$: We assume that $\mathbb{E}_{\hat{\mathbb{P}}_N}[\|\tilde{\boldsymbol{\xi}}\|^2] < \Omega$.

The first assumption is trivially satisfied under the piecewise linear loss function (5). The second assumption is needed to ensure strong duality. It can be relaxed to a high-confidence requirement, so that any probabilistic guarantees in the paper hold under the event that the strict feasibility requirement holds. For clarity of exposition, we do not pursue this extension.

We first derive a fundamental property of this ambiguity set that will be useful for developing the applications and the solution schemes.

PROPOSITION 2. *The ambiguity set* (4) *is equivalent to the infinitely-constrained ambiguity set:*

$$\mathcal{P}_\epsilon := \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{l} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \leq \epsilon \quad \forall \boldsymbol{z} \in \mathcal{X} \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega \end{array} \right\}. \tag{8}$$

Proposition 2 clarifies the role of the expected hinge constraint in (4). The ambiguity set $\mathcal{P}_\epsilon$ can be viewed as a feasibility region defined by an infinite family of loss-discrepancy inequalities together with a moment constraint. This semi-infinite perspective highlights a key advantage of our targeted formulation: rather than controlling a problem-agnostic statistical distance, it restricts discrepancies through the lens of the specific loss function. By enforcing $\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \leq \epsilon$ for every feasible decision $\forall \boldsymbol{z} \in \mathcal{X}$, the formulation rules out adversarial distributions that could induce unexpectedly large losses for any feasible decision. Thus, the model does not require the adversarial distribution to remain close to the empirical distribution in all statistical directions, but only in those directions that are operationally relevant.

REMARK 1 (TASK-RELEVANT DISCREPANCY). Unlike classical ambiguity sets that rely on an a priori ambient-space ground metric, our targeted formulation captures task relevance natively. This structural advantage becomes particularly evident under heterogeneous uncertainty, such as when $\tilde{\boldsymbol{\xi}}$ contains both continuous and discrete components, or when the loss is much more sensitive to some components than to others. A Wasserstein formulation can also reflect such structure, but only after specifying a suitable representation and ground cost. Our formulation instead measures the induced expected-loss deviations directly.

2.1. Performance Guarantees

In the next two subsections, we show that our DRO framework can be applied to a broad spectrum of problem classes, and establish that the radius ϵ can be systematically adjusted with the number of samples N at the canonical $\widetilde{\mathcal{O}}(1/\sqrt{N})$ rate. The proofs of all subsequent coverage guarantees share the same two-step structure: (i) a *pointwise* concentration inequality for a fixed decision $\boldsymbol{z} \in \mathcal{X}$, and (ii) an extension to a *uniform* bound over all $\boldsymbol{z} \in \mathcal{X}$ via a covering number argument. We capture this common structure in the following proposition, so that each specific application needs only establish the pointwise bound in step (i).

PROPOSITION 3. *Let* $\mathcal{X} \subseteq \mathbb{R}^{D_{\boldsymbol{x}}}$ *be compact with* $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\| \leq R_{\boldsymbol{x}}$. *Assume that* $\mathbb{E}_{\mathbb{P}^\star}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega$. *Suppose for every fixed* $\boldsymbol{z} \in \mathcal{X}$, *we have the pointwise concentration inequality*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \leq \frac{1}{\sqrt{N}} \varepsilon \left(\log \left(\frac{1}{\delta} \right) \right) \tag{9}$$

with probability at least $1-\delta$*, where* $\varepsilon:\mathbb{R}_+\to\mathbb{R}_+$ *is nondecreasing and has at most polynomial growth. Let* $L \coloneqq \max_{j\in[J]}\|\boldsymbol{A}_j\|_{\mathrm{op}}\sqrt{\Omega}+\max_{j\in[J]}\|\boldsymbol{b}_j\|$ *be the Lipschitz constant from Lemma 1. Then, by setting the radius to*

$$\epsilon \coloneqq \frac{1}{\sqrt{N}}\left(\varepsilon\left(D_{\boldsymbol{x}}\log\left(1+2R_{\boldsymbol{x}}L\sqrt{N}\right)+\log\left(\frac{1}{\delta}\right)\right)+2\right), \tag{10}$$

one can ensure the distribution coverage

$$\mathbb{P}^\star\in\mathcal{P}_\epsilon$$

with probability at least $1-\delta$*.*

The proof combines the Lipschitz estimate in Lemma 1 with a covering argument; see Appendix A. The Lipschitz estimate controls nearby decisions, while monotonicity of ε lets the pointwise bound be lifted over the cover. Polynomial growth then keeps the extra covering penalty logarithmic in N and avoids a curse-of-dimensionality rate loss.

Proposition 3 provides the link between pointwise concentration and the applications below. Once (9) is established for any fixed decision $\boldsymbol{z}$, the covering argument in the proposition yields the uniform coverage condition $\mathbb{P}^\star\in\mathcal{P}_\epsilon$. Thus, in the canonical-rate cases, each application only needs to specify the reference estimator $\hat{\mathbb{P}}_N$ and the corresponding function $\varepsilon(\cdot)$ in (9); the radius is then obtained from (10). The same pointwise-to-uniform argument applies to estimators with different rates, with the rate reflected in the radius calibration.

In the following, we derive out-of-sample performance and suboptimality guarantees for the solutions to the DRO problem (3). To this end, we denote $\boldsymbol{x}^\star$ and $\hat{\boldsymbol{x}}$ as the optimal solutions of the true stochastic optimization problem and the DRO problem, respectively, i.e.,

$$\boldsymbol{x}^\star\in\arg\min_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \quad\text{and}\quad \hat{\boldsymbol{x}}\in\arg\min_{\boldsymbol{x}\in\mathcal{X}}\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]. \tag{11}$$

Note that the solution $\boldsymbol{x}^\star$ is attained since the mapping $\boldsymbol{x}\to\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$ is continuous, and the feasible set $\mathcal{X}$ is compact. Furthermore, the objective function of the DRO problem constitutes a pointwise supremum of a family of lower semicontinuous (piecewise linear) functions. Therefore, it is also lower semicontinuous, and the set of optimal solutions is nonempty. That is, $\hat{\boldsymbol{x}}$ is attained.

Corollary 1. *Let* $\hat{J}$ *be the optimal value of the DRO problem* (3)*. Then, setting the radius to* (10)*, we can ensure the out-of-sample performance guarantee*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\hat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}})]\le\hat{J}$$

with probability at least $1-\delta$*.*

Remarkably, our DRO framework enables us to establish a guarantee on the excess risk of the distributionally robust solution, as stated in the following theorem.

THEOREM 1. *Suppose the pointwise concentration inequality in Proposition 3 holds in terms of absolute error, i.e.,*

$$\left|\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]\right| \leq \frac{1}{\sqrt{N}}\varepsilon\left(\log\left(\frac{1}{\delta}\right)\right) \tag{12}$$

with probability $1-\delta$ *for any fixed* $\boldsymbol{z}\in\mathcal{X}$*. Then, setting the radius* ϵ *to* (10)*, we get*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\hat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}})] \leq \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{x}^\star,\tilde{\boldsymbol{\xi}})] + \frac{1}{\sqrt{N}}\varepsilon\left(\log\left(\frac{1}{\delta}\right)\right) + \epsilon,$$

with probability $1-2\delta$*.*

This theorem shows that the excess risk is controlled by a single pointwise estimation term together with the ambiguity radius ϵ. Therefore, whenever the radius calibration in Proposition 3 is of order $\widetilde{\mathcal{O}}(N^{-1/2})$, the same is true for the excess risk. In this sense, the distributionally robust solution achieves guaranteed robustness while paying only a near-parametric excess-risk price, without the rate deteriorating in the ambient dimension $D_{\boldsymbol{\xi}}$ of the uncertainty $\tilde{\boldsymbol{\xi}}$.

REMARK 2 (DECISION-DEPENDENT AMBIGUITY SETS). One could envisage an alternative decision-dependent ambiguity set

$$\mathcal{P}_\epsilon(\boldsymbol{x}) \coloneqq \left\{\mathbb{Q}\in\mathscr{P}(\Xi) : \begin{array}{l}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \leq \epsilon \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega\end{array}\right\}.$$

Then, for any fixed $\boldsymbol{x}\in\mathcal{X}$, (9) implies the pointwise coverage guarantee

$$\text{Prob}\left(\mathbb{P}^\star\in\mathcal{P}_\epsilon(\boldsymbol{x})\right) \geq 1-\delta$$

However, this pointwise coverage does not generally imply the uniform coverage property

$$\text{Prob}\left(\mathbb{P}^\star\in\mathcal{P}_\epsilon(\boldsymbol{x})\ \ \forall \boldsymbol{x}\in\mathcal{X}\right) \geq 1-\delta,$$

which is typically required to deduce an out-of-sample guarantee for the data-dependent optimizer $\hat{\boldsymbol{x}}$. Indeed, $\hat{\boldsymbol{x}}$ depends on the same sample used to form $\hat{\mathbb{P}}_N$, so a guarantee that holds for each fixed $\boldsymbol{x}$ need not hold at the random choice $\boldsymbol{x}=\hat{\boldsymbol{x}}$ without a uniform bound.

2.2. Applications

We now present several pertinent applications of our DRO framework. Each result below verifies the pointwise concentration condition (9) under a different data regime or reference estimator. The corresponding function $\varepsilon(\cdot)$ determines the radius in (10), which gives the coverage and out-of-sample guarantees through Proposition 3 and Corollary 1. When a two-sided version of the bound is available, Theorem 1 further yields the excess-risk guarantee.

Sub-Weibull loss

In many applications such as financial risk, insurance claims, and inventory problems, the random loss $\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})$ may exhibit substantial skewness and nontrivial tail behavior. At the same time, finite-sample DRO guarantees are often derived under light-tailed assumptions on the underlying uncertainty (Mohajerin Esfahani and Kuhn 2018, Gao 2023). It is therefore important to ask whether the same finite-sample guarantees can still be obtained under weaker tail assumptions.

Our framework extends naturally to losses with sub-Weibull tails, a broad family that unifies several tail regimes. It includes the sub-Gaussian and subexponential cases as special instances. The latter covers many familiar distributions, including the exponential, Gamma, χ^2, Poisson, and geometric distributions, as well as the square of any sub-Gaussian random variable (Vershynin 2025, Example 2.8.7)). It also extends to the heavier-than-exponential regime when $\vartheta \in (0,1)$. This allows us to treat light- and moderately heavy-tailed losses within a common framework.

Following Kuchibhotla and Chakrabortty (2022), a centered random variable X is called sub-Weibull of order $\vartheta \in (0,2]$ if $\|X\|_{\psi_\vartheta} < \infty$. Under this parameterization, $\vartheta = 2$ corresponds to the sub-Gaussian case, $\vartheta = 1$ to the subexponential case, and $0 < \vartheta < 1$ to heavier-than-exponential tails. For $\vartheta < 1$, the quantity $\|\cdot\|_{\psi_\vartheta}$ is only a quasi-norm, which makes sharp Bernstein-type concentration substantially more delicate.

The following proposition provides a pointwise concentration bound for the empirical loss under a sub-Weibull assumption.

PROPOSITION 4. *Suppose that there exist $\vartheta \in (0,2]$ and $K > 0$ such that*

$$\|\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) - \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})]\|_{\psi_\vartheta} \leq K \qquad \forall \boldsymbol{z} \in \mathcal{X}.$$

Then there exist constants $C_\vartheta, c_\vartheta > 0$, depending only on ϑ, such that the pointwise concentration (9) *holds with $\varepsilon(y) = C_\vartheta K \sqrt{1+y}$. More precisely,*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \leq \frac{C_\vartheta K}{\sqrt{N}} \sqrt{1 + \log\left(\frac{1}{\delta}\right)} \tag{13}$$

with probability at least $1-\delta$, provided that

$$\delta \in \begin{cases} \left[2\exp\left(-c_\vartheta N^{\vartheta/(2-\vartheta)}\right), 1\right), & \vartheta \in (0,1), \\ \left[2\exp\left(-c_\vartheta N\right), 1\right), & \vartheta \in [1,2), \\ (0,1), & \vartheta = 2. \end{cases}$$

REMARK 3 (SUB-WEIBULL CONFIDENCE REGIME). Proposition 4 highlights several features of the sub-Weibull setting.

(i) The restriction on δ comes from expressing the sub-Weibull concentration result in a pure $1/\sqrt{N}$ form. Over the confidence regime considered here, the generalized Bernstein inequality of Kuchibhotla and Chakrabortty (2022, Theorem 3.1) reduces to the same pointwise concentration template used in Proposition 3. More generally, if the prescribed $\delta \in (0,1)$ lies outside this regime, an additional higher-order term appears, which decays at the faster $\mathcal{O}(1/N)$ rate.

(ii) A key advantage of Proposition 4 is that it remains applicable under substantially heavier-tailed uncertainty distributions than those covered by existing finite-sample DRO guarantees. In particular, Gao (2023) relies on a transportation-information condition on the data-generating law, which implies standard sub-Gaussian-type tail assumptions, while the classical Wasserstein guarantee of Mohajerin Esfahani and Kuhn (2018), via Fournier and Guillin (2015), requires the stronger stretched-exponential moment condition $\mathbb{E}[\exp(\|\tilde{\boldsymbol{\xi}}\|^r)] < +\infty$ for some $r > 1$ on the raw uncertainty. By contrast, Proposition 4 directly covers the broader sub-Weibull regime.

Markovian data

Rather strikingly, our ambiguity set can easily accommodate circumstances when the samples are not i.i.d. Consider the setting where $\{\hat{\boldsymbol{\xi}}_i\}_{i\in[N]}$ is a sample trajectory of a Markov chain with invariant distribution $\mathbb{P}^\star$. In this case, our mean estimator $\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]$ is no longer constructed using i.i.d. samples, which prohibit the use of classical concentration inequalities. Nevertheless, using Hoeffding's inequality for general Markov chains (Fan et al. 2021), we can obtain the desired pointwise concentration inequality.

Proposition 5. *Assume the loss function is bounded: $\ell(\boldsymbol{z},\boldsymbol{\xi}) \in [\underline{\ell},\overline{\ell}]$ for all $\boldsymbol{z} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$. Suppose that the Markov chain is time-homogeneous with an absolute spectral gap of $1-\lambda > 0$, where $\lambda \in [0,1)$ is the operator norm of the Markov transition kernel acting on the Hilbert space of square-integrable mean-zero functions under $\mathbb{P}^\star$. Then, the pointwise concentration* (9) *holds with $\varepsilon(y) = \frac{\overline{\ell}-\underline{\ell}}{2}\sqrt{\frac{2(1+\lambda)}{1-\lambda}y}$, i.e.,*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \le \frac{1}{\sqrt{N}}\frac{\overline{\ell}-\underline{\ell}}{2}\sqrt{\frac{2(1+\lambda)}{1-\lambda}\log\left(\frac{1}{\delta}\right)} \tag{14}$$

with probability at least $1-\delta$. The same result holds for time-inhomogeneous chains with λ replaced by $\max_{i\in[N]}\lambda_i$, where $1-\lambda_i$ is the spectral gap of the Markov transition kernel for step i.

Remark 4 (Markovian data with general state spaces). Our proposed DRO model applies to (time-inhomogeneous) Markovian data with general state spaces. We note that DRO with time-homogeneous Markovian data has been studied in (Li et al. 2021) in a discrete-state-space setting. Unlike ours, their statistical guarantee is asymptotic as it relies on large deviation principles; hence, it does not provide a prescription for adjusting the radius with respect to the sample size N.

Outlier corrupted data

In many practical settings, the observed data may be contaminated by outliers, i.e., samples that deviate arbitrarily from the underlying distribution $\mathbb{P}^\star$. Under such corruption, the standard empirical-mean reference that uses all data points is sensitive to these atypical observations and may no longer reliably approximate $\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]$, potentially invalidating the coverage guarantee. To address this, we replace the empirical-mean reference with a *Median-of-Means* (MoM) reference functional (Laforgue et al. 2021), which retains a $1/\sqrt{N}$ convergence rate while remaining robust to a constant fraction of outliers.

MoM estimator. Given N observations $\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_N$, fix an integer $K \geq 1$ and partition the index set $[N]$ into K disjoint blocks $\mathcal{B}_1,\ldots,\mathcal{B}_K$ of equal size $B = \lfloor N/K \rfloor$, independently of the data. For each block $k \in [K]$, define the block mean

$$f_k(\boldsymbol{z}) \coloneqq \frac{1}{B}\sum_{i\in\mathcal{B}_k} \ell(\boldsymbol{z},\hat{\boldsymbol{\xi}}_i).$$

Then the MoM reference functional is defined as

$$\widehat{\mu}_{\mathrm{MoM}}(\boldsymbol{z}) \coloneqq \mathrm{median}\big(f_1(\boldsymbol{z}),\ldots,f_K(\boldsymbol{z})\big).$$

In the outlier-corrupted setting, we use $\widehat{\mu}_{\mathrm{MoM}}(\boldsymbol{z})$ in place of $\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]$ as the reference functional in the loss constraints.

Suppose the sample points $\{\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_N\}$ contains $N-n_{\mathrm{o}}$ inliers drawn i.i.d. from $\mathbb{P}^\star$, and n_{o} arbitrary outliers. We assume that the fraction of outliers is $\omega \coloneqq n_{\mathrm{o}}/N \in [0,1/2)$. Following (Laforgue et al. 2021) with the arithmetic-mean calibration, we define the inflation constant

$$\Gamma(\omega) \coloneqq \frac{\sqrt{2(1+2\omega)}}{(1-2\omega)^{3/2}},$$

which satisfies $\Gamma(\omega)\to\infty$ as $\omega \to 1/2$.

PROPOSITION 6 **(Proposition 2 (eq. (3)) of Laforgue et al. (2021))**. *Suppose the inlier loss has bounded variance: $Var_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \leq \sigma^2$ for all $\boldsymbol{z}\in\mathcal{X}$. Set $K = \left\lceil \frac{4(1+2\omega)}{(1-2\omega)^2}\log\frac{1}{\delta}\right\rceil$ and let $\delta \in \left[e^{-(1-2\omega)^2 N/(4(1+2\omega))},\, e^{-(1-2\omega)^2 N/8}\right]$. Then, the MoM reference satisfies the pointwise concentration bound*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \widehat{\mu}_{\mathrm{MoM}}(\boldsymbol{z}) \leq 4\sqrt{e}\,\sigma\,\Gamma(\omega)\sqrt{\frac{1+\log(1/\delta)}{N}}.$$

with probability at least $1-\delta$.

We note that the covering argument in Proposition 3 carries over under this replacement, because it only requires a pointwise concentration bound and a Lipschitz reference functional. The latter property holds for $\widehat{\mu}_{\mathrm{MoM}}$ with the same constant L: the median satisfies $|\mathrm{median}(\boldsymbol{a}) - \mathrm{median}(\boldsymbol{b})| \leq$

$\max_k |a_k - b_k|$ for any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^K$, and each block mean satisfies $|f_k(\boldsymbol{z}) - f_k(\boldsymbol{z}')| \leq L\|\boldsymbol{z} - \boldsymbol{z}'\|$ by the pathwise Lipschitz bound $|\ell(\boldsymbol{z}, \boldsymbol{\xi}) - \ell(\boldsymbol{z}', \boldsymbol{\xi})| \leq L\|\boldsymbol{z} - \boldsymbol{z}'\|$ (see proof of Lemma 1).

This replacement does not require interpreting $\widehat{\mu}_{\mathrm{MoM}}$ as an expectation under a new empirical distribution. When we later derive the exact dual reformulation, the same generalized moment problem argument applies to the MoM-based ambiguity set as long as this set satisfies the usual Slater condition for the chosen radius ϵ. Under this condition, strong duality and dual attainment hold with $\widehat{\mu}_{\mathrm{MoM}}(\boldsymbol{z})$ replacing $\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})]$ in the dual objective.

Remark 5 (Outlier robustness through robust estimators). Distributionally robust optimization with outlier corrupted data has gained increasing attention. Nietert et al. (2023) and Nietert et al. (2024) employ mixed Wasserstein-Total Variation ambiguity sets to capture both local geometric and global adversarial corruptions. Meanwhile, Blanchet et al. (2024) utilize optimal transport with concave cost functions to automatically discount and rectify large-distance outliers. Additionally, Jiang and Xie (2024) propose a distributionally favorable optimization model that seeks the most favorable distribution to naturally ignore outlier-induced high losses. However, these approaches rely on problem-specific ambiguity sets or non-standard optimization paradigms tailored to specific outlier structures. In contrast, our Targeted IPM framework provides a unified framework. Since our fast convergence rate only depends on a scalar concentration inequality, we can seamlessly plug in any statistically robust estimator, such as MoM, to achieve outlier resistance while preserving tractability and bypassing the curse of dimensionality.

Incomplete Data

Our framework is applicable to settings where a subset of the random parameters is not always observable. This is relevant to many selection problems, such as company hiring, college admissions, and loan approval, where outcomes (e.g., high-performing candidates, qualified students, or applicants who could repay the loan) are only observed for selected individuals. Formally, we define the random parameters as $\tilde{\boldsymbol{\xi}} = (\tilde{\boldsymbol{\chi}}, \tilde{\boldsymbol{\omega}}) \in \mathscr{X} \times \Omega$, where $\tilde{\boldsymbol{\chi}}$ are entries that are always observable, while $\tilde{\boldsymbol{\omega}}$ are those that will be observed only if the candidate is selected.

Let $\tilde{S} \in \{0, 1\}$ denote the historical selection outcome with $\tilde{S} = 1$ if the candidate is selected to advance and $\tilde{S} = 0$ otherwise. Define $\overline{\mathbb{P}}$ to be the joint distribution of $(\tilde{S}, \tilde{\boldsymbol{\chi}}, \tilde{\boldsymbol{\omega}})$. This binary random variable is governed by a logging policy

$$\pi(\boldsymbol{\chi}) := \overline{\mathbb{P}}(\tilde{S} = 1 | \tilde{\boldsymbol{\chi}} = \boldsymbol{\chi}),$$

which represents the probability that a candidate with covariate $\boldsymbol{\chi}$ is selected. Based on the logging policy, the available dataset is given by

$$\{(\boldsymbol{\chi}_n, \boldsymbol{\omega}_n)\}_{n \in \mathcal{I}} \cup \{\boldsymbol{\chi}_n\}_{n \in [N] \setminus \mathcal{I}},$$

where $\mathcal{I}$ is the set of candidates who were selected and have the outcome observed. Under the mild conditional exchangeability and positivity assumptions (Jia et al. 2024, Assumption 1 and 2), the inverse probability weighting estimator of the expected loss is given by

$$\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \coloneqq \frac{1}{N}\sum_{n\in[N]} \frac{\mathbb{I}[S_n=1]}{\pi(\boldsymbol{\chi}_n)}\ell(\boldsymbol{z},\boldsymbol{\xi}_n), \tag{15}$$

where each full sample point is weighted by the reciprocal of its propensity score $\pi(\boldsymbol{\chi}_n)$. We use this IPW estimator as the reference functional in the loss constraints and obtain the following concentration bound.

Proposition 7. *Suppose conditional exchangeability and positivity hold, with $\pi(\boldsymbol{\chi}) \geq \underline{\pi} > 0$, and suppose the losses are uniformly sub-Gaussian, i.e., $\|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\|_{\psi_2} \leq \sigma$ for all $\boldsymbol{z}\in\mathcal{X}$. The pointwise concentration* (9) *holds with $\varepsilon(y) = C_{\mathrm{IPW}}\frac{\sigma}{\underline{\pi}}\sqrt{y}$ for a universal constant $C_{\mathrm{IPW}}>0$, i.e., we have*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \leq C_{\mathrm{IPW}}\frac{\sigma}{\underline{\pi}}\sqrt{\frac{\log(1/\delta)}{N}}$$

with probability at least $1-\delta$.

Contextual optimization

Our framework can also be applied to stochastic optimization with side information. Suppose that a realization $\boldsymbol{c}\in\mathbb{R}^{D_{\mathbf{c}}}$ of some contextual covariates is observed, which changes the conditional distribution of $\tilde{\boldsymbol{\xi}}$. To exploit the side information, the decision-maker replaces the expectation in the objective with the conditional expectation

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}],$$

where $\mathbb{P}^\star$ denotes the joint distribution of $(\tilde{\boldsymbol{c}},\tilde{\boldsymbol{\xi}})$. In a data driven setting, we observe historical data $\{(\boldsymbol{c}_n,\boldsymbol{\xi}_n)\}_{n\in[N]}$ and approximate the conditional expectation using the weighted sum

$$\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}] \coloneqq \sum_{n\in[N]} w(\boldsymbol{c},\boldsymbol{c}_n)\ell(\boldsymbol{z},\boldsymbol{\xi}_n),$$

where the weights $(w(\boldsymbol{c},\boldsymbol{c}_n))_{n\in[N]}$ are designed such that they give higher values to data points that are close to the current realization $\boldsymbol{c}$. A popular scheme is the Nadaraya-Watson kernel regression, in which the weights are defined through a kernel function $\mathcal{K}_h$ with bandwidth parameter $h>0$:

$$w(\boldsymbol{c},\boldsymbol{c}_n) \coloneqq \frac{\mathcal{K}_h(\boldsymbol{c}-\boldsymbol{c}_n)}{\sum_{m\in[N]}\mathcal{K}_h(\boldsymbol{c}-\boldsymbol{c}_m)}$$

The resulting data-driven approximation is guaranteed to provide asymptotic consistency of the optimal solutions. The following proposition establishes a pointwise concentration for this data-driven conditional estimator.

PROPOSITION 8. *Assume that the loss function takes values in the interval* $[0,1]$*, the conditional expectation* $\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}]$ *is* $L_{\boldsymbol{c}}$*-Lipschitz continuous in* $\boldsymbol{c}$*, and the marginal density function* $f(\boldsymbol{c})$ *satisfies* $f(\boldsymbol{c})\geq \underline{f}>0$ *for all* $\boldsymbol{c}$ *in its support. Then, the following pointwise concentration inequality holds:*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}]-\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}]\leq \left(\frac{2\log(1/\delta)}{N c \underline{f}}\right)^{\frac{1}{D_{\boldsymbol{c}}+2}} L_{\boldsymbol{c}}^{\frac{D_{\boldsymbol{c}}}{D_{\boldsymbol{c}}+2}} \frac{D_{\boldsymbol{c}}+2}{2(D_{\boldsymbol{c}}/2)^{\frac{D_{\boldsymbol{c}}}{D_{\boldsymbol{c}}+2}}}$$

with probability at least $1-\delta$*.*

REMARK 6 (CONTEXTUAL RATES AND DIMENSIONALITY). Note that the bound suffers from the curse of dimensionality in $D_{\boldsymbol{c}}$. This dependence in the dimension is unavoidable since learning the conditional expectation in $D_{\boldsymbol{c}}$-dimensional nonparametric regression has the *minimax* squared-error rate of $O(N^{-2/(D_{\boldsymbol{c}}+2)})$ (Györfi et al. 2002), corresponding to the absolute-error rate $O(N^{-1/(D_{\boldsymbol{c}}+2)})$ in Proposition 8. Nevertheless, our DRO framework is applicable to provide a robustification with coverage guarantee $\mathbb{P}^\star(\cdot\,|\,\tilde{\boldsymbol{c}}=\boldsymbol{c})\in\mathcal{P}_{\epsilon_N}$ where the radius can be set to

$$\epsilon_N \coloneqq \frac{D_{\boldsymbol{c}}+2}{2(D_{\boldsymbol{c}}/2)^{\frac{D_{\boldsymbol{c}}}{D_{\boldsymbol{c}}+2}}} L_{\boldsymbol{c}}^{\frac{D_{\boldsymbol{c}}}{D_{\boldsymbol{c}}+2}} \left(\frac{2\left(D_{\boldsymbol{x}}\log\left(1+2R_{\boldsymbol{x}}L\sqrt{N}\right)+\log(1/\delta)\right)}{N c \underline{f}}\right)^{\frac{1}{D_{\boldsymbol{c}}+2}} + \frac{2}{\sqrt{N}}.$$

2.3. Dual Formulations

In this subsection, we derive an exact dual reformulation of the DRO problem (3) and establish convergence to empirical risk minimization as the ambiguity parameter ϵ approaches 0. We start with the reformulation.

THEOREM 2. *Fix* $\epsilon>0$*. The DRO problem* (3) *is equivalent to the infinite linear program*

$$\begin{aligned} &\min \alpha+\beta\Omega+\int_{\mathcal{X}}\left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]+\epsilon\right)\nu(\mathrm{d}\boldsymbol{z}) \\ &\text{s.t. } \boldsymbol{x}\in\mathcal{X},\ \alpha\in\mathbb{R},\ \beta\in\mathbb{R}_+,\ \nu\in\mathcal{M}_+(\mathcal{X}) \\ &\qquad \ell(\boldsymbol{x},\boldsymbol{\xi})\leq\alpha+\beta\|\boldsymbol{\xi}\|^2+\int_{\mathcal{X}}\ell(\boldsymbol{z},\boldsymbol{\xi})\nu(\mathrm{d}\boldsymbol{z}) \qquad \forall\boldsymbol{\xi}\in\Xi. \end{aligned} \tag{16}$$

The optimal solution $(\boldsymbol{x},\alpha,\beta,\nu)\in\mathcal{X}\times\mathbb{R}\times\mathbb{R}_+\times\mathcal{M}_+(\mathcal{X})$ *is attained.*

The optimization problem remains intractable because it involves an infinite number of variables and constraints. In the following section, we will develop a tractable approximation scheme with guarantees.

We close the section by showing that the optimal value and solutions of the DRO problem converge to those of the (non-robust) empirical risk minimization problem as ϵ approaches 0.

PROPOSITION 9. *As $\epsilon \to 0$, the optimal value of* (16) *converges to that of the empirical risk minimization problem*

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]. \tag{17}$$

In addition, every cluster point $\hat{\boldsymbol{x}}^\star$ of a sequence $\{\hat{\boldsymbol{x}}_\epsilon\}_{\epsilon\downarrow 0}$ of minimizers for (16) *is a minimizer for* (17).

3. A Monte Carlo Sampling Approach

Problem (16) remains difficult to solve as it involves a functional decision variable ν and an expectation over a continuous distribution $\mathbb{Z}$. By Proposition 2, our DRO problem (3) constitutes an instance of DRO problems with an infinitely constrained ambiguity set, which have been studied by Chen et al. (2019) and enable a convergent cutting-plane algorithm for settings where the support Ξ is compact. In this paper, we propose a tractable approximation via Monte Carlo sampling.

Specifically, we draw M samples from $\mathbb{Z}$ and obtain the approximate ambiguity set:

$$\mathcal{P}_\epsilon^M := \left\{ \mathbb{Q}\in\mathscr{P}(\Xi): \begin{array}{c} \frac{1}{M}\sum_{m\in[M]} \left[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+ \le 0 \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \le \Omega \end{array} \right\}. \tag{18}$$

By construction, the constraint only enforces $\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] \le \epsilon$ at finitely many sample points $\{\boldsymbol{z}_m\}_{m=1}^M$. Thus, we have a relaxed ambiguity set $\mathcal{P}_\epsilon^M \supseteq \mathcal{P}_\epsilon$, and the resulting distributionally robust optimization problem is a *conservative* approximation, and the out-of-sample performance guarantees provided in Corollary 1 remain valid. Unlike the classical sample-average approximation for stochastic programming that often introduces optimistic bias, our approximation always generates pessimistic bias and yields safe solutions. The conservativeness further endows the model with attractive theoretical properties, which we delineate in the following subsection.

Using the ambiguity set (18), we obtain the following dual reformulation:

$$\begin{array}{ll} \min & \alpha + \beta\Omega + \frac{1}{M}\sum_{m\in[M]} \nu_m(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] + \epsilon) \\ \text{s.t.} & \boldsymbol{x}\in\mathcal{X},\ \alpha\in\mathbb{R},\ \beta\in\mathbb{R}_+,\ \boldsymbol{\nu}\in\mathbb{R}_+^M \\ & \ell(\boldsymbol{x},\boldsymbol{\xi}) \le \alpha + \beta\|\boldsymbol{\xi}\|^2 + \frac{1}{M}\sum_{m\in[M]} \nu_m \ell(\boldsymbol{z}_m,\boldsymbol{\xi}) \qquad \forall \boldsymbol{\xi}\in\Xi. \end{array} \tag{19}$$

This formulation constitutes the sample-average counterpart of Theorem 2. By replacing the expectation under $\mathbb{Z}$ with an empirical average over $\{\boldsymbol{z}_m\}_{m=1}^M$, the functional multiplier $\nu(\cdot)$ reduces to a finite-dimensional vector $\boldsymbol{\nu}\in\mathbb{R}_+^M$.

In the following, we develop the theoretical properties and tractable conic reformulations of this approximation scheme. We first consider the case where the support set Ξ is compact.

3.1. Theoretical Guarantees

Our first theoretical result establishes a high-confidence guarantee for the containment of the sample-average-based ambiguity set $\mathcal{P}_\epsilon^M$ within a relaxed exact ambiguity set

$$\mathcal{P}_\epsilon(\eta) \coloneqq \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{c} \mathbb{E}_{\mathbb{Z}} \left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon \right]_+ \right] \leq \eta \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega \end{array} \right\} \tag{20}$$

parameterized by $\eta > 0$.

THEOREM 3. *Assume* $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\| \leq R_{\boldsymbol{x}}$ *and* $\sup_{\boldsymbol{\xi} \in \Xi} \|\boldsymbol{\xi}\| \leq R_{\boldsymbol{\xi}}$*, and let*

$$R \coloneqq \max_{j \in [J]} \left\| \begin{bmatrix} \boldsymbol{A}_j^\top & \boldsymbol{b}_j \\ \boldsymbol{a}_j^\top & b_j \end{bmatrix} \right\|_{\mathrm{op}} \sqrt{R_{\boldsymbol{\xi}}^2 + 1} \sqrt{R_{\boldsymbol{x}}^2 + 1}.$$

Then, setting

$$\eta \coloneqq \mathcal{O}\left(R \sqrt{\frac{J \log J (\log M)^3}{M}} \right) + 2(c_1 + c_2\sqrt{\Omega}) \sqrt{\frac{1}{2M} \log\left(\frac{1}{\tau}\right)}, \tag{21}$$

we can ensure with probability at least $1 - \tau$*:*

$$\mathcal{P}_\epsilon^M \subseteq \mathcal{P}_\epsilon(\eta).$$

This result shows that using the sample-average-based ambiguity set is akin to relaxing the expected hinge constraint by $\eta \coloneqq \widetilde{\mathcal{O}}(1/\sqrt{M})$. The result relies on a generalization bound for the sample-average approximation errors, stated as Proposition 11 in Appendix B. In the proof of that proposition, we bound the Rademacher complexity of the function class

$$\mathcal{H} \coloneqq \left\{ \boldsymbol{z} \mapsto [\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \epsilon]_+ : \mathbb{Q} \in \mathscr{P}(\Xi),\, \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega \right\},$$

which contains functions parametrized by distributions $\mathbb{Q} \in \mathscr{P}(\Xi)$ satisfying the second-moment constraint $\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega$. Unlike in traditional settings, the parameter space is infinite-dimensional, rendering classical techniques inapplicable. We employ the contraction principle to reduce the analysis to the Rademacher complexity of the functions $\boldsymbol{z} \mapsto \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})]$. Using our loss function structure, we further reduce it to the complexity of J-fold maxima of hyperplanes (Attias and Kontorovich 2024, Corollary 5). This function-class argument is the key ingredient behind Theorem 3.

REMARK 7 (ROLE OF FUNCTION-CLASS STRUCTURE). The dimension-free $\widetilde{\mathcal{O}}(1/\sqrt{M})$ sampling rate in Proposition 11 is a key benefit of exploiting the piecewise-affine structure of the loss class, rather than a generic consequence of Lipschitz continuity. Indeed, for a generic class of bounded L-Lipschitz functions on a metric space with dimension $D_{\boldsymbol{x}}$ of the decision $\boldsymbol{x}$, the empirical Rademacher complexity can scale as $\mathcal{O}(L/M^{1/(D_{\boldsymbol{x}}+1)})$; see (Gottlieb et al. 2016, Theorem 4.3). In contrast, the piecewise-affine loss class considered here reduces to the class of J-fold maxima of hyperplanes, for which the sharper Rademacher bound in Proposition 11 is available.

We next establish the convergence of the optimal value and optimal solution of the approximate DRO problem. Let $\hat{v}$ and $\hat{v}_M$ be the optimal values of the exact problem and the approximate problem, respectively, i.e.,

$$\hat{v} \coloneqq \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \quad \text{and} \quad \hat{v}_M \coloneqq \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].$$

We define $\mathcal{S}$ as the set of minimizers of the exact problem. The minimizers $\hat{\boldsymbol{x}} \in \mathcal{S}$ and $\hat{\boldsymbol{x}}_M$ of the sampled problem are attained because the corresponding objectives are lower semicontinuous on the compact set $\mathcal{X}$.

THEOREM 4. *As $M \to \infty$, the following convergences in probability hold:*

$$\hat{v}_M \xrightarrow{p} \hat{v} \quad \text{and} \quad \operatorname{dist}(\hat{\boldsymbol{x}}_M, \mathcal{S}) \xrightarrow{p} 0.$$

Furthermore, assume there exists an optimal dual measure ν in Theorem 2 such that $\gamma\mathbb{Z} - \nu \in \mathcal{M}_+(\mathcal{X})$ for some $\gamma \in \mathbb{R}_+$. Let $\gamma^\star \coloneqq \inf\{\gamma \in \mathbb{R} : \gamma\mathbb{Z} - \nu \in \mathcal{M}_+(\mathcal{X})\}$. Then, the following suboptimality bound holds with probability at least $1-\tau$:

$$\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M,\tilde{\boldsymbol{\xi}})] \le \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] + \gamma^\star \left(\mathcal{O}\left(R\sqrt{\frac{J\log J(\log M)^3}{M}} \right) + 2(c_1 + c_2\sqrt{\Omega}) \sqrt{\frac{1}{2M}\log\left(\frac{1}{\tau}\right)} \right).$$

REMARK 8 (EXISTENCE OF THE DOMINANCE CONSTANT). The domination condition used in the suboptimality bound of Theorem 4 is a condition on the auxiliary sampling distribution, not on the underlying data-generating model. Since $\mathbb{Z}$ is chosen by the modeler, it can be enriched, if needed, to dominate the relevant optimal dual measure, as we now show. Specifically, if ν is any optimal dual measure in Theorem 2, then for the case $\nu(\mathcal{X}) = 0$, any $\gamma \ge 0$ works; otherwise, for any $\lambda \in (0,1)$, the full-support mixture $\mathbb{Z}_\lambda \coloneqq (1-\lambda)\mathbb{Z} + \lambda\nu/\nu(\mathcal{X})$ satisfies $(\nu(\mathcal{X})/\lambda)\mathbb{Z}_\lambda - \nu \in \mathcal{M}_+(\mathcal{X})$, and hence $\gamma_\lambda^\star \le \nu(\mathcal{X})/\lambda < \infty$. Thus the required dominance constant can always be made finite through the modeling choice of $\mathbb{Z}$. By Proposition 2, all full-support choices of the auxiliary distribution induce the same semi-infinite ambiguity set, so replacing $\mathbb{Z}$ with $\mathbb{Z}_\lambda$ does not alter the DRO problem itself.

REMARK 9 (NUMERICAL ILLUSTRATION OF THE MONTE CARLO APPROXIMATION). We provide a simple experiment to illustrate the sampling behavior in Theorem 4. The detailed experimental design is reported in Appendix E.1. To match the bounded-support assumptions, both the decision set and the uncertainty support are Euclidean unit balls. We use a shifted piecewise-affine loss and a symmetric empirical distribution so that the exact optimizer $x_\star$ is known and nonzero. This allows us to evaluate the sampled-dual solution $\widehat{x}_M$ directly, without computing a large-reference Monte Carlo solution. We report the relative optimality gap

$$\frac{\mathbb{E}_{\widehat{P}_N}[\ell(\widehat{x}_M,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\widehat{P}_N}[\ell(x_\star,\tilde{\boldsymbol{\xi}})]}{\mathbb{E}_{\widehat{P}_N}[\ell(x_\star,\tilde{\boldsymbol{\xi}})]} \times 100\%$$

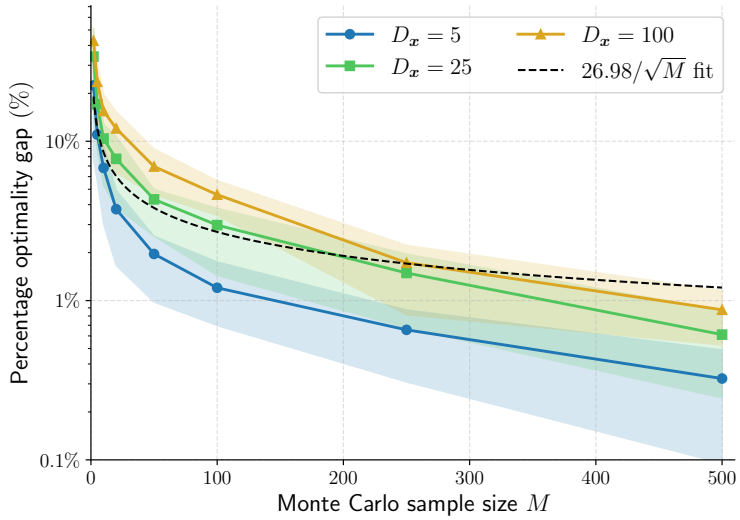


Figure 1 Percentage optimality gap of the sampled-dual solution as a function of the Monte Carlo sample size M. The dashed curve is a fitted $C/\sqrt{M}$ reference line.

across different Monte Carlo sample sizes M. Figure 1 shows that the sampled-dual solution improves rapidly as M increases. The fitted $C/\sqrt{M}$ curve provides a visual benchmark for the sampling rate in Theorem 4. All curves exhibit a similar decay pattern. At the largest tested Monte Carlo sizes, the gaps are reduced to roughly the one-percent level across all dimensions. This provides numerical evidence that the sampled-dual approximation improves consistently with M and does not suffer from the curse of dimensionality.

3.2. Conic Programming Reformulations

In this subsection, we show that the approximate problem admits a tractable conic programming reformulation. Tractable approximations for two-stage problems are provided in Appendix D.

THEOREM 5. *Suppose that $\mathcal{X}$ and Ξ are convex. Then problem* (19) *is equivalent to the following finite convex reformulation involving the support function σ_Ξ:*

$$
\begin{array}{l}
\min \alpha + \beta\Omega + \dfrac{1}{M}\displaystyle\sum_{m\in[M]} \nu_m(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] + \epsilon) \\
\text{s.t. } \boldsymbol{x}\in\mathcal{X},\ \alpha\in\mathbb{R},\ \beta\in\mathbb{R}_+,\ \boldsymbol{\nu}\in\mathbb{R}_+^M \\
\quad \boldsymbol{\lambda}_{jm}\in\mathbb{R}_+^J\ \ \forall m\in[M], \quad \boldsymbol{\theta}_j\in\mathbb{R}^{D_{\boldsymbol{\xi}}}, \zeta_j\in\mathbb{R}_+\ \ \forall j\in[J] \\
\quad \mathbf{e}^\top\boldsymbol{\lambda}_{jm} = \nu_m\ \ \forall m\in[M]\ \ \forall j\in[J] \\
\quad \zeta_j + \sigma_\Xi(\boldsymbol{\theta}_j) + b_j(\boldsymbol{x}) - \dfrac{1}{M}\displaystyle\sum_{m\in[M],k\in[J]} \lambda_{jm}^k b_k(\boldsymbol{z}_m) \le \alpha \quad \forall j\in[J] \\
\quad \left\| \begin{bmatrix} \boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M}\sum_{m\in[M],k\in[J]} \lambda_{jm}^k \boldsymbol{a}_k(\boldsymbol{z}_m) - \boldsymbol{\theta}_j \\ \zeta_j - \beta \end{bmatrix} \right\| \le \zeta_j + \beta \quad \forall j\in[J].
\end{array}
\tag{22}
$$

In particular, if $\mathcal{X}$ and Ξ are second-order cone representable, then the reformulation above is a second-order cone program.

4. Extensions to Unbounded Support

We now extend the previous results to the setting where the distributional support is $\Xi := \mathbb{R}^{D_{\boldsymbol{\xi}}}$. We assume that there exist constants $K_{\mathrm{sw}}, \vartheta > 0$ such that

$$\mathbb{P}^\star(\|\tilde{\boldsymbol{\xi}}\| \le t) \ge 1 - 2\exp\Big(-(t/K_{\mathrm{sw}})^\vartheta\Big) \qquad \forall t > 0.$$

This is a standard sub-Weibull-type tail bound (Kuchibhotla and Chakrabortty 2022, Definition 2.2). For a given threshold $t > 0$, we incorporate this tail information into the ambiguity set:

$$\mathcal{P}'_\epsilon := \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{l} \mathbb{E}_{\mathbb{Z}}\left[\Big[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon\Big]_+\right] \le 0 \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \le \Omega,\ \mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\| \le t) \ge 1 - 2\exp\Big(-(t/K_{\mathrm{sw}})^\vartheta\Big) \end{array} \right\}. \tag{23}$$

In the sequel, we will apply this bound at thresholds $t = t_M$ that vary with the Monte Carlo sample size M. Under the standing second-moment and tail assumptions on $\mathbb{P}^\star$, Proposition 3 implies that $\mathbb{P}^\star \in \mathcal{P}'_\epsilon$ with probability at least $1 - \delta$.

We next show that the proposed framework extends beyond compact support. Under the single-threshold tail bound implied by sub-Weibull tails above, we derive an exact dual reformulation, establish that the Monte Carlo approximation remains conservative, and show that the same $\widetilde{\mathcal{O}}(1/\sqrt{M})$ sampling principle continues to hold up to explicit tail-dependent terms. We also obtain a tractable conic reformulation.

The following proposition provides the dual formulation of the DRO problem under this ambiguity set.

Proposition 10. *The DRO problem* (3) *with the ambiguity set* (23) *is equivalent to the infinite linear program*

$$\begin{array}{l} \min \alpha + \beta\Omega - \kappa\left(1 - 2\exp\left(-(t/K_{\mathrm{sw}})^\vartheta\right)\right) + \displaystyle\int_{\mathcal{X}} \left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] + \epsilon\right) \nu(\mathrm{d}\boldsymbol{z}) \\ \text{s.t. } \boldsymbol{x} \in \mathcal{X},\ \alpha \in \mathbb{R},\ \beta, \kappa \in \mathbb{R}_+,\ \nu \in \mathcal{M}_+(\mathcal{X}) \\ \quad \ell(\boldsymbol{x}, \boldsymbol{\xi}) + \kappa \mathbb{I}_{\{\|\boldsymbol{\xi}\| \le t\}} \le \alpha + \beta\|\boldsymbol{\xi}\|^2 + \displaystyle\int_{\mathcal{X}} \ell(\boldsymbol{z}, \boldsymbol{\xi}) \nu(\mathrm{d}\boldsymbol{z}) \qquad \forall \boldsymbol{\xi} \in \Xi \end{array} \tag{24}$$

An optimal solution $(\boldsymbol{x}, \alpha, \beta, \kappa, \nu) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^2_+ \times \mathcal{M}_+(\mathcal{X})$ *is attained.*

Drawing M samples from $\mathbb{Z}$, we obtain an outer approximation of the ambiguity set:

$$\mathcal{P}'^M_\epsilon := \left\{ \mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{l} \dfrac{1}{M} \displaystyle\sum_{m \in [M]} \Big[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}_m, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m, \tilde{\boldsymbol{\xi}})] - \epsilon\Big]_+ \le 0 \\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \le \Omega,\ \mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\| \le t) \ge 1 - 2\exp\left(-(t/K_{\mathrm{sw}})^\vartheta\right) \end{array} \right\}. \tag{25}$$

The corresponding dual formulation is given by

$$\begin{aligned}
&\min \alpha+\beta\Omega-\kappa\left(1-2\exp\left(-\left(t/K_{\mathrm{sw}}\right)^{\vartheta}\right)\right)+\frac{1}{M}\sum_{m\in[M]}\nu_m(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})]+\epsilon)\\
&\text{s.t. } \boldsymbol{x}\in\mathcal{X},\ \alpha\in\mathbb{R},\ \beta,\kappa\in\mathbb{R}_+,\ \boldsymbol{\nu}\in\mathbb{R}_+^M \qquad (26)\\
&\qquad \ell(\boldsymbol{x},\boldsymbol{\xi})+\kappa\mathbb{I}_{\{\|\boldsymbol{\xi}\|\le t\}}\le\alpha+\beta\|\boldsymbol{\xi}\|^2+\frac{1}{M}\sum_{m\in[M]}\nu_m\ell(\boldsymbol{z}_m,\boldsymbol{\xi})\qquad\forall\boldsymbol{\xi}\in\Xi.
\end{aligned}$$

Remark 10 (Calibration of tail parameters). The results in this section provide a theoretical ambiguity-set construction with sampling error of order $\widetilde{\mathcal{O}}(1/\sqrt{M})$. In practice, however, we use the ambiguity set (18) from Section 3 regardless of whether the support is bounded or unbounded. If one nevertheless wishes to work with the ambiguity set in this section, then, as with the constant Ω, the tail parameters $(K_{\mathrm{sw}},\vartheta)$ must be chosen or calibrated in advance. The threshold t governs the tradeoff in the bound: increasing t enlarges the localized radius, while shrinking the tail correction term $\exp\left(-\left(t/K_{\mathrm{sw}}\right)^{\vartheta}\right)$.

4.1. Theoretical Guarantees

Throughout this subsection, we set $t=t_M\coloneqq K_{\mathrm{sw}}(\log M)^{1/\vartheta}$, so that $2\exp\left(-\left(t_M/K_{\mathrm{sw}}\right)^{\vartheta}\right)=2/M$. For notational clarity, we denote by $\mathcal{P}'_{\epsilon,M}$ the exact ambiguity set (23) with $t=t_M$. The corresponding relaxed moving-threshold set is

$$\mathcal{P}'_{\epsilon,M}(\eta)\coloneqq\left\{\mathbb{Q}\in\mathscr{P}(\Xi):\begin{array}{l}\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})]-\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})]-\epsilon\right]_+\right]\le\eta\\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]\le\Omega,\ \mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\|\le t_M)\ge 1-2\exp\left(-\left(t_M/K_{\mathrm{sw}}\right)^{\vartheta}\right)\end{array}\right\},\qquad(27)$$

parameterized by $\eta>0$. The first theoretical result establishes a high-confidence guarantee for the containment of the sample-average-based ambiguity set $\mathcal{P}_\epsilon^{\prime M}$ within this relaxed exact ambiguity set.

Theorem 6. *Assume* $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|\le R_{\boldsymbol{x}}$, *and let*

$$R(t)\coloneqq\left(\max_{j\in[J]}\left\|\begin{bmatrix}\boldsymbol{A}_j^\top & \boldsymbol{b}_j\\ \boldsymbol{a}_j^\top & b_j\end{bmatrix}\right\|_{\mathrm{op}}\sqrt{t^2+1}\right)\sqrt{R_{\boldsymbol{x}}^2+1},\qquad t>0.$$

Then, setting

$$\eta_M\coloneqq\mathcal{O}\left(R(t_M)\sqrt{\frac{J\log J(\log M)^3}{M}}\right)+2(c_1+c_2\sqrt{\Omega})\sqrt{\frac{1}{2M}\log\left(\frac{1}{\tau}\right)}+\frac{4c_1}{M}+2c_2\sqrt{\frac{2\Omega}{M}},$$

we can ensure with probability at least $1-\tau$*:*

$$\mathcal{P}_\epsilon^{\prime M}\subseteq\mathcal{P}'_{\epsilon,M}(\eta_M).$$

The argument mirrors the bounded-support case. The only additional ingredient is the technical generalization bound in Proposition 12 in Appendix C, which controls the contribution of the tail event after truncation at t_M. With this tail correction, the sampled ambiguity set retains the same conservative outer-approximation interpretation. Since $R(t_M) = \mathcal{O}((\log M)^{1/\vartheta})$, the relaxation level η_M vanishes at the order $\widetilde{\mathcal{O}}(M^{-1/2})$.

To compare the sampled problem with the limiting exact problem, we use Lemma 4 in Appendix C, which shows that the moving tail constraint vanishes asymptotically. Recall that $\hat{v}$ and $\mathcal{S}$ denote the optimal value and the set of minimizers of the exact problem with ambiguity set $\mathcal{P}_\epsilon$.

THEOREM 7. *Fix $\epsilon > 0$, and let $\hat{v}'_M$ and $\hat{\boldsymbol{x}}'_M$ be the optimal value and an optimal solution of the sampled problem* (26) *with $t = t_M$. Under the assumptions of Theorem 6, as $M \to \infty$,*

$$\hat{v}'_M \xrightarrow{p} \hat{v} \quad \text{and} \quad \operatorname{dist}(\hat{\boldsymbol{x}}'_M, \mathcal{S}) \xrightarrow{p} 0.$$

Furthermore, let $\boldsymbol{x}^\star_M$ be an exact optimizer of the threshold-t_M problem, and assume that there exists an associated optimal dual solution $(\boldsymbol{x}^\star_M, \alpha^\star_M, \beta^\star_M, \kappa^\star_M, \nu^\star_M)$ in Proposition 10 such that $\gamma\mathbb{Z} - \nu^\star_M \in \mathcal{M}_+(\mathcal{X})$ for some $\gamma \in \mathbb{R}_+$. Define $\gamma^\star_M := \inf\{\gamma \in \mathbb{R}_+ : \gamma\mathbb{Z} - \nu^\star_M \in \mathcal{M}_+(\mathcal{X})\}$. Then, the following suboptimality bound holds with probability at least $1 - \tau$:

$$\begin{aligned}\sup_{\mathbb{Q} \in \mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}'_M, \tilde{\boldsymbol{\xi}})] \le \min_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] \\ + \gamma^\star_M \left(\mathcal{O}\left(R(t_M) \sqrt{\frac{J \log J (\log M)^3}{M}} \right) + 2(c_1 + c_2\sqrt{\Omega}) \sqrt{\frac{1}{2M} \log\left(\frac{1}{\tau}\right)} \right. \\ \left. + \frac{4c_1}{M} + 2c_2 \sqrt{\frac{2\Omega}{M}} \right).\end{aligned}$$

4.2. Conic Programming Reformulations

We now show that the approximate problem admits a tractable conic programming reformulation.

THEOREM 8. *Suppose that $\mathcal{X}$ is convex. The problem* (26) *is equivalent to the following finite convex conic reformulation:*

$$
\begin{aligned}
\min\ & \alpha+\beta\Omega-\kappa\left(1-2\exp\left(-\left(t/K_{\mathrm{sw}}\right)^{\vartheta}\right)\right)+\frac{1}{M}\sum_{m\in[M]}\nu_m(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})]+\epsilon)\\
\text{s.t. } & \boldsymbol{x}\in\mathcal{X},\ \alpha\in\mathbb{R},\ \beta,\kappa\in\mathbb{R}_+,\ \boldsymbol{\nu}\in\mathbb{R}_+^M\\
& \boldsymbol{\lambda}_{jm},\boldsymbol{\lambda}'_{jm}\in\mathbb{R}_+^J\ \ \forall j\in[J]\,\forall m\in[M]\\
& \boldsymbol{\theta}_j\in\mathbb{R}^{D_{\boldsymbol{\xi}}},\ \zeta_j,\zeta'_j\in\mathbb{R}_+\ \ \forall j\in[J]\\
& \mathbf{e}^\top\boldsymbol{\lambda}_{jm}=\nu_m\ \ \forall m\in[M]\,\forall j\in[J]\\
& \mathbf{e}^\top\boldsymbol{\lambda}'_{jm}=\nu_m\ \ \forall m\in[M]\,\forall j\in[J]\\
& \zeta_j+b_j(\boldsymbol{x})-\frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}b_k(\boldsymbol{z}_m)\le\alpha\quad\forall j\in[J]\\
& \left\|\begin{bmatrix}\boldsymbol{a}_j(\boldsymbol{x})-\frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}\boldsymbol{a}_k(\boldsymbol{z}_m)\\ \zeta_j-\beta\end{bmatrix}\right\|\le\zeta_j+\beta\quad\forall j\in[J]\\
& \zeta'_j+t\|\boldsymbol{\theta}_j\|+b_j(\boldsymbol{x})+\kappa-\frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda'^k_{jm}b_k(\boldsymbol{z}_m)\le\alpha\quad\forall j\in[J]\\
& \left\|\begin{bmatrix}\boldsymbol{a}_j(\boldsymbol{x})-\frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda'^k_{jm}\boldsymbol{a}_k(\boldsymbol{z}_m)-\boldsymbol{\theta}_j\\ \zeta'_j-\beta\end{bmatrix}\right\|\le\zeta'_j+\beta\quad\forall j\in[J].
\end{aligned}
\tag{28}
$$

In particular, if $\mathcal{X}$ is second-order cone representable, then the reformulation above is a second-order cone program.

5. Numerical Experiments

In this section, we conduct numerical experiments to validate the practical performance of our TIPM-DRO framework across two problem classes: a newsvendor problem and linear regression under adversarial outlier corruption. In each experiment, we describe the problem setting, the competing methods, and the out-of-sample (OOS) evaluation results.

5.1. Multi-Item Newsvendor Problem

We consider a multi-item newsvendor problem with order vector $\boldsymbol{x}\in\mathbb{R}_+^K$ and random demand vector $\tilde{\boldsymbol{\xi}}\in\mathbb{R}_+^K$. The loss is

$$
\ell(\boldsymbol{x},\boldsymbol{\xi})=\sum_{j=1}^{K}\left[h_j(x_j-\xi_j)_+ + b_j(\xi_j-x_j)_+\right].
$$

We optimize the $\mathbb{P}$-CVaR$_\rho$ objective of this loss and keep the formulation generic in K; in the numerical study, we set $K=3$. This design serves two purposes. First, we want to evaluate the method under both lighter-tailed and heavier-tailed demand models, since heavy tails are operationally important in inventory systems and can materially change stocking and pooling decisions (Bimpikis

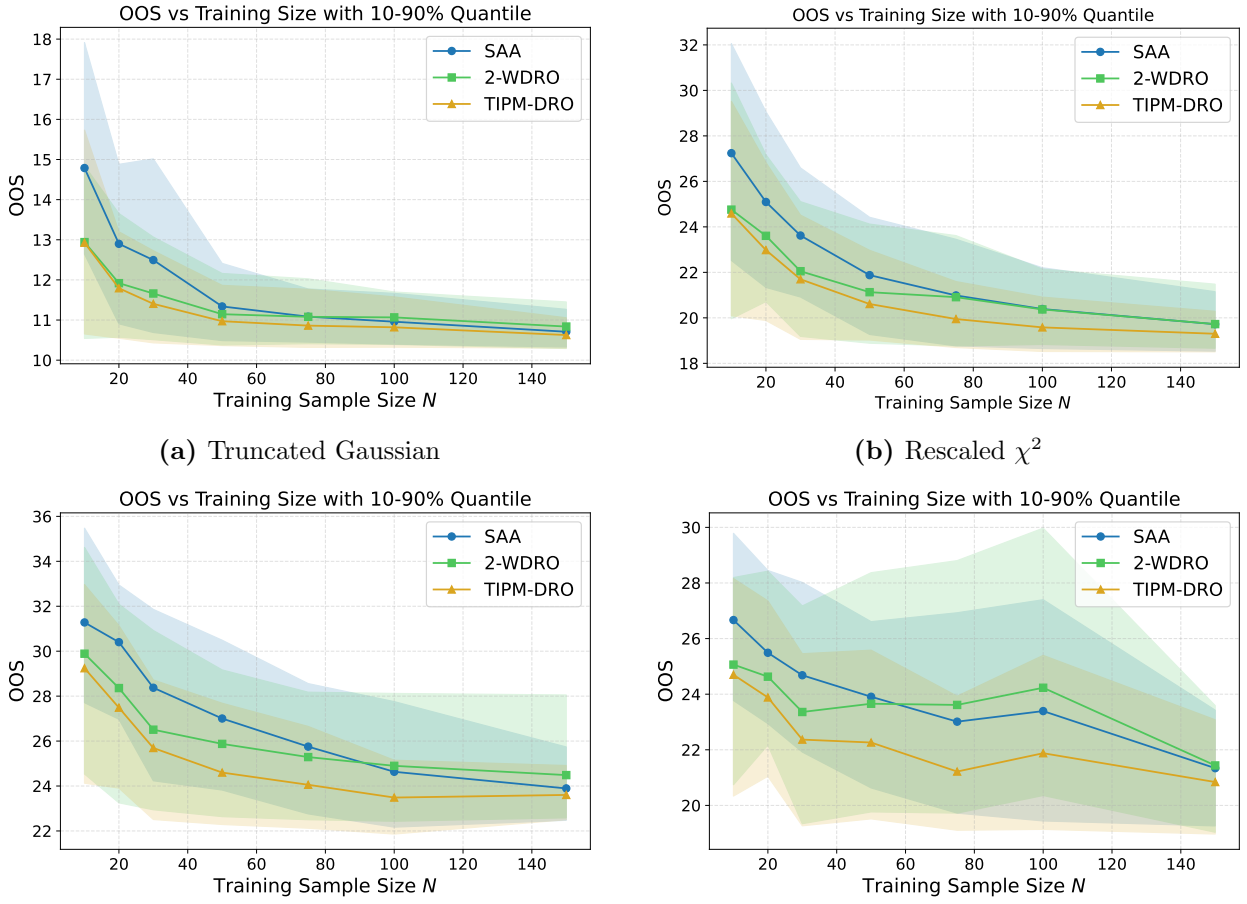
and Markakis 2016). We therefore compare four coordinatewise demand families spanning this range: truncated Gaussian, rescaled χ^2, Lognormal, and Pareto. Second, since the paper emphasizes the dimension-robust motivation behind TIPM, the experiment involves a vector-valued decision and uncertainty. In the experiment, we set $h = (0.1, 0.2, 0.3)$, $b = (1, 1, 1)$, and $\rho = 0.05$, and normalize the four demand models to share the same mean vector $\boldsymbol{\mu} = (5, 6, 6)$ and standard-deviation vector $\boldsymbol{\sigma} = (5, 6, 8)$. Full parameter settings and reformulations are deferred to Appendix E.2.

We compare SAA, TIPM-DRO, and a 2-Wasserstein DRO baseline (2-WDRO) under the same CVaR objective. We consider the 2-Wasserstein DRO, which has recently been shown to generate superior performance in out-of-sample tests (Byeon 2025). We do not separately report 1-WDRO because, for newsvendor losses of this form, 1-Wasserstein DRO shares the same optimal minimizers as SAA, see Mohajerin Esfahani and Kuhn (2018, Remark 6.7). In each trial, the in-sample data are split into training and validation subsets, the ambiguity radius of each robust method is selected from the same grid by validation, and the chosen decision is evaluated on an independent OOS sample. The TIPM-DRO model is solved through the explicit second-moment-constrained reformulation reported in Appendix E.2. Thus, the same experimental pipeline tests both stories at once: how performance changes with tail heaviness and whether the method remains effective in a non-scalar decision problem.

Figure 2 reports mean OOS cost with 10th–90th percentile bands over 50 trials, where lower values are better. TIPM-DRO achieves the best mean OOS performance in all four demand settings. Under truncated Gaussian demand, both robust methods improve on SAA and TIPM-DRO still performs slightly better than 2-WDRO, although the gap narrows as N increases. Under rescaled χ^2, Lognormal, and Pareto demand, the advantage of TIPM-DRO is more persistent, with the clearest separation appearing in the Lognormal and Pareto panels. These patterns support both design objectives. The Gaussian panel shows that TIPM-DRO is already competitive in a lighter-tailed regime, while the heavier-tailed panels show that its advantage over both SAA and 2-WDRO becomes more pronounced. At the same time, these gains are observed in a three-item vector problem rather than only in a scalar benchmark.

5.2. Outlier-Corrupted Regression

We consider linear regression with the mean absolute deviation (MAD) loss $\ell(\boldsymbol{x}, \boldsymbol{\xi}) = |\boldsymbol{x}^\top \boldsymbol{u} - v|$, where $\boldsymbol{\xi} = (\boldsymbol{u}, v) \in \mathbb{R}^{D_{\boldsymbol{\xi}}+1}$ denotes a data point with feature vector $\boldsymbol{u} \in \mathbb{R}^{D_{\boldsymbol{\xi}}}$ and response $v \in \mathbb{R}$, following the setup of Nietert et al. (2023). The true regression coefficient $\boldsymbol{x}^\star$ is drawn uniformly from $\mathbb{S}^{D_{\boldsymbol{\xi}}-1}$, clean features satisfy $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{D_{\boldsymbol{\xi}}})$, and clean responses are $v_i = \boldsymbol{x}^{\star\top} \boldsymbol{u}_i$. An adversary corrupts a fraction $\omega = 0.2$ of the training samples by replacing each corrupted pair $(\boldsymbol{u}_i, v_i)$ with $(C\boldsymbol{u}_i, -C^2 v_i + \rho)$, where $\rho = 0.1$. This creates leverage-type outliers whose severity is controlled by



(a) Truncated Gaussian

(b) Rescaled χ^2

(c) Lognormal

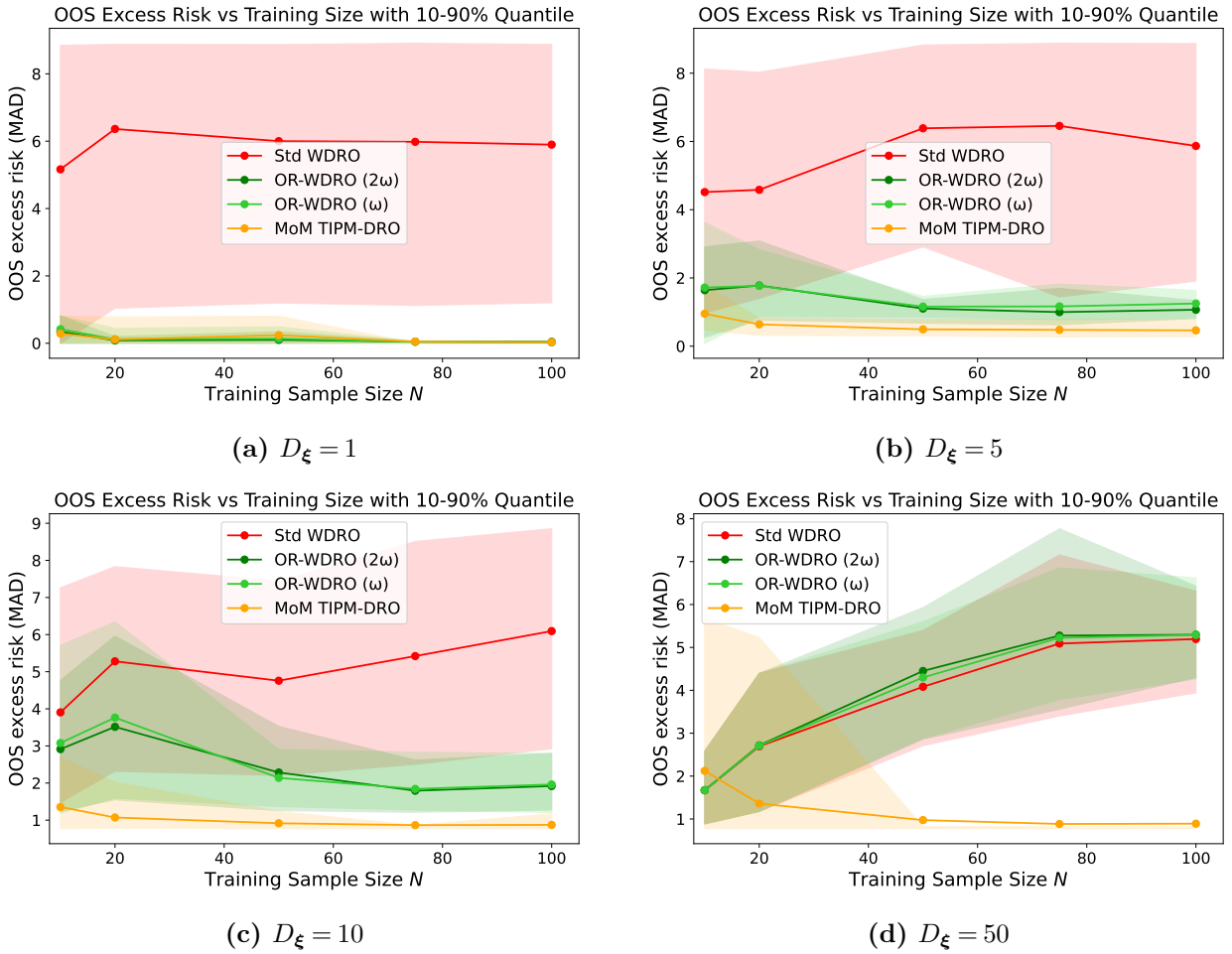
(d) Pareto

Figure 2 Mean OOS cost versus the training sample size N with 10th–90th percentile bands in the multi-item newsvendor experiment with $K = 3$.

the scale parameter C. The experiment is designed to probe two questions. First, what happens as the ambient dimension $D_{\boldsymbol{\xi}}$ grows and Wasserstein geometry becomes less informative? Second, does robust performance depend on the outliers being geometrically well separated from the clean sample? Full implementation details and tuning grids are summarized in Appendix E.3.

We benchmark standard Wasserstein DRO (Mohajerin Esfahani and Kuhn 2018) (Std WDRO), the two outlier-robust WDRO variants of Nietert et al. (2023) with TV contamination radii ω and 2ω, and our MoM TIPM-DRO estimator based on the Median-of-Means from Section 2.2. All robust models are tuned by a trimmed validation loss on a held-out split. For the three Wasserstein-based models, we cross-validate the Wasserstein radius over a common grid. For MoM TIPM-DRO, we cross-validate the number of blocks K and the TIPM radius ϵ.

Figure 3 reports the OOS excess risk under the MAD loss as a function of the training sample size N for four representative dimensions $D_{\boldsymbol{\xi}}$, based on 50 trials and a clean test set of size 1000. Standard WDRO barely improves with N, because its ambiguity set remains centered at the



(a) $D_{\boldsymbol{\xi}} = 1$ **(b)** $D_{\boldsymbol{\xi}} = 5$

(c) $D_{\boldsymbol{\xi}} = 10$ **(d)** $D_{\boldsymbol{\xi}} = 50$

Figure 3 OOS excess risk under the MAD loss versus the training sample size N, with panels ordered by increasing dimension d and 10th–90th percentile bands.

contaminated empirical distribution and therefore keeps fitting the outliers as more corrupted data are observed. OR-WDRO works reasonably well in low dimension, where transport costs can still separate inliers from outliers, but this advantage fades as $D_{\boldsymbol{\xi}}$ grows and Wasserstein distances concentrate. MoM TIPM-DRO, in contrast, stays stable across dimensions and becomes clearly best at $D_{\boldsymbol{\xi}} = 50$. The message of this figure is that transport-based outlier handling still depends on having informative geometry, whereas the MoM reference is much less sensitive to dimension. This is also consistent with our theory: the MoM contamination tolerance only requires $\omega < 1/2$, so it does not worsen with ambient dimension.

To isolate the effect of outlier severity, Figure 4 fixes $N = 50$ and $D_{\boldsymbol{\xi}} = 10$ and sweeps C from 5 to 20. For visual clarity, this figure displays only OR-WDRO(2ω), OR-WDRO(ω), and MoM TIPM-DRO. OR-WDRO shows a clear hump at moderate contamination scales: the outliers are already harmful enough to distort the empirical reference, but not yet far enough away in transport cost to be trimmed reliably. Only when C becomes very large does OR-WDRO start to recover. MoM

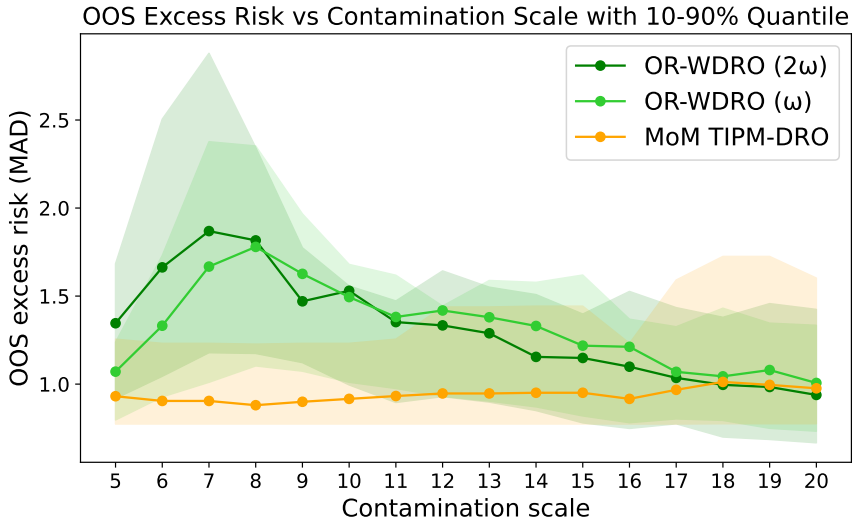


Figure 4 OOS excess risk under the MAD loss versus the contamination scale C, with 10th–90th percentile bands over the trials.

TIPM-DRO stays nearly flat throughout, showing that its performance is much less sensitive to how severe the leverage contamination is. Taken together, the two figures show that MoM TIPM-DRO does not just improve over standard WDRO. It is also more stable than outlier-aware WDRO baselines exactly when transport geometry is least reliable: at higher ambient dimension and at moderate contamination scales.

6. Concluding Remarks

We introduced a distributionally robust optimization framework based on targeted integral probability metrics. By defining the ambiguity set directly through the decision-induced loss class, the proposed expected hinge formulation yields a task-aware ambiguity set that is equivalent to a semi-infinite family of loss-discrepancy constraints. This construction enables finite-sample guarantees that bypass the curse of dimensionality whenever a scalar pointwise concentration inequality is available, thereby extending dimension-free DRO guarantees to a broad range of nonstandard data regimes. In many such settings, the framework also simplifies implementation, as it avoids the need to handcraft geometry-driven ambiguity sets. For example, one need not specify a ground cost as in Wasserstein DRO when the uncertainty contains mixed discrete and continuous features, nor combine different discrepancy measures, such as Wasserstein and total variation, to handle outlier contamination. Instead, the framework only requires a loss function that captures how uncertainty affects the decision problem and an estimator that admits an appropriate pointwise guarantee.

Our numerical results highlight the practical value of aligning the ambiguity set with the downstream loss. In both inventory management and outlier-corrupted regression, the proposed frame-

work delivers strong out-of-sample performance and robustness relative to transport-based baselines. Moreover, the experiments suggest that the framework may remain effective even beyond the sub-Weibull regime, thereby motivating further theoretical investigation. Several directions remain for future research. These include extending the framework to heavier-tailed models such as Pareto-type distributions, developing adaptive sampling schemes that may improve the Monte Carlo approximation rate, and designing scalable algorithms for richer multistage decision problems.

References

Idan Attias and Aryeh Kontorovich. Fat-shattering dimension of k-fold aggregations. *Journal of Machine Learning Research*, 25(144):1–29, 2024.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Exact generalization guarantees for (regularized) Wasserstein distributionally robust models. *Advances in Neural Information Processing Systems*, 36:14584–14596, 2023.

A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 2012. Forthcoming.

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

Aharon Ben-Tal, Omar El Housni, and Vineet Goyal. A tractable approach for designing piecewise affine policies in two-stage adjustable robust optimization. *Mathematical Programming*, 182(1):57–102, 2020.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Kostas Bimpikis and Mihalis G. Markakis. Inventory pooling under heavy-tailed demand. *Management Science*, 62(6):1800–1813, 2016. doi: 10.1287/mnsc.2015.2204.

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of machine learning research*, 23(39):1–70, 2022.

Jose Blanchet and Yang Kang. Sample out-of-sample inference based on wasserstein distance. *Operations Research*, 69(3):985–1013, 2021.

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

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in wasserstein distributionally robust estimation. *Biometrika*, 109(2):295–315, 2022.

Jose Blanchet, Jiajin Li, Markus Pelger, and Greg Zanotti. Automatic outlier rectification via optimal transport. *Advances in Neural Information Processing Systems*, 37:35313–35357, 2024.

Geunyeong Byeon. Comparative analysis of two-stage distributionally robust optimization over 1-Wasserstein and 2-Wasserstein balls. *arXiv preprint arXiv:2501.05619*, 2025.

Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018.

Zhi Chen, Melvyn Sim, and Huan Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research*, 67(5):1328–1344, 2019.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's inequality for general markov chains and its applications to statistical learning. *Journal of Machine Learning Research*, 22(139):1–35, 2021.

Xiangyi Fan and Grani A Hanasusanto. A decision rule approach for two-stage data-driven distributionally robust optimization problems with random recourse. *INFORMS Journal on Computing*, 36(2):526–542, 2024.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

Rui Gao and Anton 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.

Angelos Georghiou, Wolfram Wiesemann, and Daniel Kuhn. Generalized decision rule approximations for stochastic programming via liftings. *Mathematical Programming*, 152(1):301–338, 2015.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction. *Theoretical Computer Science*, 620:105–118, 2016.

Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A Distribution-Free Theory of Nonparametric Regression*, volume 1. Springer, 2002.

Grani A Hanasusanto and Daniel Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

Haiyun He and Ziv Goldfeld. Information-theoretic generalization bounds for deep neural networks. *IEEE Transactions on Information Theory*, 2025.

Hisham Husain. Distributional robustness with IPMs and links to regularization and GANs. *Advances in Neural Information Processing Systems*, 33:11816–11827, 2020.

Garud Iyengar, Henry Lam, and Tianyu Wang. Hedging against complexity: Distributionally robust optimization with parametric approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 9976–10011. PMLR, 2023.

Zhuangzhuang Jia, Grani A Hanasusanto, Phebe Vayanos, and Weijun Xie. Learning fair policies for multistage selection problems from observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21188–21196, 2024.

Nan Jiang and Weijun Xie. Distributionally favorable optimization: A framework for data-driven decision-making with endogenous outliers. *SIAM Journal on Optimization*, 34(1):419–458, 2024.

Arun K. Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022. doi: 10.1093/imaiai/iaac012.

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

Pierre Laforgue, Guillaume Staerman, and Stephan Clémençon. Generalization bounds in the presence of outliers: a median-of-means study. In *International conference on machine learning*, pages 5937–5947. PMLR, 2021.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Tam Le and Jérôme Malick. Universal generalization guarantees for wasserstein distributionally robust models. *arXiv preprint arXiv:2402.11981*, 2024.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

Mengmeng Li, Tobias Sutter, and Daniel Kuhn. Distributionally robust optimization with markovian data. In *International Conference on Machine Learning*, pages 6493–6503. PMLR, 2021.

Adrian Tovar Lopez and Varun Jog. Generalization error bounds using wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2018.

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

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. *Advances in Neural Information Processing Systems*, 32, 2019.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein dro. *Advances in Neural Information Processing Systems*, 36:62792–62820, 2023.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Robust distribution learning with local and global adversarial corruptions. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4007–4008. PMLR, 2024.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

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

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

A. Shapiro. On duality theory of conic linear problems. In M.A. Goberna and M.A. Lopez, editors, *Semi-Infinite Programming: Recent Advances*, pages 135–165. Kluwer Academic Publishers, 2001.

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

Roman Vershynin. High-dimensional probability, 2025.

Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Hao Wang, Mario Diaz, José Cândido S Santos Filho, and Flavio P Calmon. An information-theoretic view of generalization via wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 577–581. IEEE, 2019.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of SGLD using properties of Gaussian channels. *Advances in Neural Information Processing Systems*, 34:24222–24234, 2021.

Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho. Generalization bounds for contextual stochastic optimization using kernel regression. *arXiv preprint arXiv:2407.10764*, 2024.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations research*, 62(6):1358–1376, 2014.

Qinyu Wu, Jonathan Yu-Meng Li, and Tiantian Mao. On generalization and regularization via Wasserstein distributionally robust optimization. *Management Science*, 2025.

Weijun Xie, Jie Zhang, and Shabbir Ahmed. Distributionally robust bottleneck combinatorial problems: uncertainty quantification and robust decision making. *Mathematical Programming*, 196(1):597–640, 2022.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Guanglin Xu and Samuel Burer. A copositive approach for two-stage adjustable robust optimization with uncertain right-hand sides. *Computational Optimization and Applications*, 70(1):33–59, 2018.

Guanglin Xu and Grani A Hanasusanto. Improved decision rule approximations for multistage robust optimization via copositive programming. *Operations Research*, 73(2):842–861, 2025.

Yibo Zeng and Henry Lam. Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 35:27576–27590, 2022.

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

Appendix A: Proofs of Section 2

A.1. Ambiguity Set Characterization

Proof of Proposition 2. It suffices to show that the semi-infinite constraint

$$\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \le \epsilon \quad \forall \boldsymbol{z} \in \mathcal{X} \tag{29}$$

is equivalent to the expectation constraint

$$\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right] \le 0. \tag{30}$$

If $\mathbb{Q}$ is feasible to (29), then $[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \epsilon]_+ \le 0$ for all $\boldsymbol{z} \in \mathcal{X}$. Taking expectation with respect to $\mathbb{Z}$ yields (30).

Conversely, suppose $\mathbb{Q}$ is infeasible to (29); that is, there exists $\boldsymbol{z}' \in \mathcal{X}$ such that $[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}',\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}',\tilde{\boldsymbol{\xi}})] - \epsilon]_+ > 0$. By Assumption (A), for any $\mathbb{Q}$ feasible to (4),

$$\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \le c_1 + c_2\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|] \le c_1 + c_2\sqrt{\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]} \le c_1 + c_2\sqrt{\Omega} < +\infty \qquad \forall \boldsymbol{z} \in \mathcal{X},$$

where the second inequality follows from Jensen's inequality. Since $\hat{\mathbb{P}}_N$ is also feasible to (4) by construction and by Assumption (B), the dominated convergence theorem implies that $[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \epsilon]_+$ is continuous in $\boldsymbol{z}$. Hence there exists a neighborhood $\mathcal{Z}_\tau = \{\boldsymbol{z} \in \mathcal{X} : \|\boldsymbol{z} - \boldsymbol{z}'\| < \tau\}$ in which $[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \epsilon]_+ > 0$ for all $\boldsymbol{z} \in \mathcal{Z}_\tau$. We therefore have

$$\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right] \ge \mathbb{Z}(\boldsymbol{z} \in \mathcal{Z}_\tau)\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+ \,\middle|\, \boldsymbol{z} \in \mathcal{Z}_\tau\right] > 0,$$

because $\mathbb{Z}$ has full support on $\mathcal{X}$. Hence $\mathbb{Q}$ is infeasible to (30). Replacing the semi-infinite constraint in (4) with the equivalent expectation constraint (30) yields the claim. □

A.2. Performance Guarantees

LEMMA 1. *For any $\mathbb{Q}$ such that $\mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\xi}\|^2] \le \Omega$, the following Lipschitz condition holds:*

$$\mathbb{E}_{\mathbb{Q}}\left[|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}}) - \ell(\boldsymbol{z}',\tilde{\boldsymbol{\xi}})|\right] \le L\|\boldsymbol{z} - \boldsymbol{z}'\| \quad \forall \boldsymbol{z},\boldsymbol{z}' \in \mathcal{X},$$

where $L \coloneqq \max_{j\in[J]} \|\boldsymbol{A}_j\|_{\mathrm{op}}\sqrt{\Omega} + \max_{j\in[J]} \|\boldsymbol{b}_j\|$.

Proof of Lemma 1. For any $\boldsymbol{z},\boldsymbol{z}' \in \mathcal{X}$, since $\boldsymbol{a}_j(\boldsymbol{z}) - \boldsymbol{a}_j(\boldsymbol{z}') = \boldsymbol{A}_j(\boldsymbol{z} - \boldsymbol{z}')$ and $b_j(\boldsymbol{z}) - b_j(\boldsymbol{z}') = \boldsymbol{b}_j^\top(\boldsymbol{z} - \boldsymbol{z}')$, we have

$$|\ell(\boldsymbol{z},\boldsymbol{\xi}) - \ell(\boldsymbol{z}',\boldsymbol{\xi})| \le \max_{j\in[J]} \left|(\boldsymbol{A}_j(\boldsymbol{z} - \boldsymbol{z}'))^\top\boldsymbol{\xi} + \boldsymbol{b}_j^\top(\boldsymbol{z} - \boldsymbol{z}')\right|$$

$$\leq \max_{j\in[J]}\big(\|\boldsymbol{A}_j\|_{\mathrm{op}}\|\boldsymbol{\xi}\| + \|\boldsymbol{b}_j\|\big)\cdot\|\boldsymbol{z}-\boldsymbol{z}'\|$$
$$\leq \left(\max_{j\in[J]}\|\boldsymbol{A}_j\|_{\mathrm{op}}\|\boldsymbol{\xi}\| + \max_{j\in[J]}\|\boldsymbol{b}_j\|\right)\cdot\|\boldsymbol{z}-\boldsymbol{z}'\|.$$

Taking expectation under $\mathbb{Q}$ yields

$$\mathbb{E}_{\mathbb{Q}}\left[|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})-\ell(\boldsymbol{z}',\tilde{\boldsymbol{\xi}})|\right] \leq \left(\max_{j\in[J]}\|\boldsymbol{A}_j\|_{\mathrm{op}}\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|] + \max_{j\in[J]}\|\boldsymbol{b}_j\|\right)\|\boldsymbol{z}-\boldsymbol{z}'\|$$
$$\leq \left(\max_{j\in[J]}\|\boldsymbol{A}_j\|_{\mathrm{op}}\sqrt{\Omega} + \max_{j\in[J]}\|\boldsymbol{b}_j\|\right)\|\boldsymbol{z}-\boldsymbol{z}'\|.$$

This is the desired bound. □

Proof of Proposition 3. Let

$$\eta_N := \frac{1}{L\sqrt{N}},$$

and let $\mathcal{X}_{\eta_N}$ be an η_N-net of $\mathcal{X}$ under $\|\cdot\|$. Since $\mathcal{X}\subseteq\mathbb{B}(0,R_{\boldsymbol{x}})$ is compact, the standard volumetric bound gives

$$|\mathcal{X}_{\eta_N}| \leq \left(1+\frac{2R_{\boldsymbol{x}}}{\eta_N}\right)^{D_{\boldsymbol{x}}} = \left(1+2R_{\boldsymbol{x}}L\sqrt{N}\right)^{D_{\boldsymbol{x}}}.$$

For each $\boldsymbol{z}_0\in\mathcal{X}_{\eta_N}$, applying (9) with confidence level

$$\delta' := \frac{\delta}{|\mathcal{X}_{\eta_N}|}$$

and then taking a union bound over $\mathcal{X}_{\eta_N}$, we obtain with probability at least $1-\delta$,

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] \leq \frac{1}{\sqrt{N}}\varepsilon\left(\log\frac{1}{\delta'}\right) \qquad \forall \boldsymbol{z}_0\in\mathcal{X}_{\eta_N}.$$

Because ε is nondecreasing and

$$\log\frac{1}{\delta'} = \log|\mathcal{X}_{\eta_N}| + \log\frac{1}{\delta} \leq D_{\boldsymbol{x}}\log\left(1+2R_{\boldsymbol{x}}L\sqrt{N}\right) + \log\frac{1}{\delta},$$

the same event implies

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] \leq \frac{1}{\sqrt{N}}\varepsilon\left(D_{\boldsymbol{x}}\log\left(1+2R_{\boldsymbol{x}}L\sqrt{N}\right) + \log\frac{1}{\delta}\right) \qquad \forall \boldsymbol{z}_0\in\mathcal{X}_{\eta_N}. \tag{31}$$

Fix any $\boldsymbol{z}\in\mathcal{X}$ and choose $\boldsymbol{z}_0\in\mathcal{X}_{\eta_N}$ with $\|\boldsymbol{z}-\boldsymbol{z}_0\|\leq\eta_N$. By Lemma 1, applied once under $\mathbb{P}^\star$ and once under $\hat{\mathbb{P}}_N$ (the latter is valid by Assumption (B)), we have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] &\leq \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] \\ &\quad + \mathbb{E}_{\mathbb{P}^\star}[|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})-\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})|] + \mathbb{E}_{\hat{\mathbb{P}}_N}[|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})-\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})|] \\ &\leq \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_0,\tilde{\boldsymbol{\xi}})] + 2L\eta_N.\end{aligned}$$

Combining this bound with (31) and using $2L\eta_N = 2/\sqrt{N}$ yields

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \leq \epsilon \qquad \forall \boldsymbol{z}\in\mathcal{X}$$

with probability at least $1-\delta$, where ϵ is defined in (10). Proposition 2 then implies $\mathbb{P}^\star\in\mathcal{P}_\epsilon$. □

Proof of Corollary 1. On the event $\{\mathbb{P}^\star \in \mathcal{P}_\epsilon\}$, which occurs with probability at least $1-\delta$ by Proposition 3, we have

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \leq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] = \hat{J}.$$

This is precisely the claimed out-of-sample guarantee. □

Proof of Theorem 1. Let

$$\mathcal{E}_1 \coloneqq \{\mathbb{P}^\star \in \mathcal{P}_\epsilon\} \qquad \text{and} \qquad \mathcal{E}_2 \coloneqq \left\{ \left| \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] \right| \leq \frac{1}{\sqrt{N}} \varepsilon\left(\log \frac{1}{\delta}\right) \right\}.$$

By Proposition 3, $\mathrm{Prob}(\mathcal{E}_1) \geq 1-\delta$, and by (12) applied at the fixed decision $\boldsymbol{x}^\star$, $\mathrm{Prob}(\mathcal{E}_2) \geq 1-\delta$. Hence

$$\mathrm{Prob}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 2\delta.$$

On $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^\star}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] &\leq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \\
&\leq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] \\
&\leq \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] + \epsilon \\
&\leq \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] + \frac{1}{\sqrt{N}} \varepsilon\left(\log \frac{1}{\delta}\right) + \epsilon.
\end{aligned}$$

Subtracting $\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]$ from both sides yields the claim. □

A.3. Applications

Proof of Proposition 4. Fix any $\boldsymbol{z} \in \mathcal{X}$ and define the centered random variables

$$X_i \coloneqq \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \ell(\boldsymbol{z}, \hat{\boldsymbol{\xi}}_i), \qquad i \in [N].$$

Then $X_1, \ldots, X_N$ are independent, mean-zero, and satisfy $\|X_i\|_{\psi_\vartheta} \leq K$. Applying Kuchibhotla and Chakrabortty (2022, Theorem 3.1) with weights $a_i = 1/N$, we obtain

$$\mathrm{Prob}\Big(\left| \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \right| \geq 2eC(\vartheta)\|b\|\sqrt{t} + 2eL_N^*(\vartheta)\|b\|_{\beta(\vartheta)} t^{1/\vartheta} \Big) \leq 2e^{-t} \qquad \forall t \geq 0,$$

where

$$b \coloneqq \left(\frac{\|X_1\|_{\psi_\vartheta}}{N}, \ldots, \frac{\|X_N\|_{\psi_\vartheta}}{N} \right), \qquad \beta(\vartheta) \coloneqq \begin{cases} \infty, & \vartheta \leq 1, \\ \frac{\vartheta}{\vartheta - 1}, & \vartheta > 1. \end{cases}$$

Since $\|X_i\|_{\psi_\vartheta} \leq K$, we have

$$\|b\| \leq \frac{K}{\sqrt{N}}.$$

Moreover,

$$\|b\|_{\beta(\vartheta)} \leq \begin{cases} \|b\|_\infty \leq \frac{K}{N}, & \vartheta \leq 1, \\ \left(N \left(\frac{K}{N}\right)^{\beta(\vartheta)} \right)^{1/\beta(\vartheta)} = K\, N^{-1/\vartheta}, & \vartheta > 1, \end{cases}$$

and therefore

$$\|b\|_{\beta(\vartheta)} \leq \frac{K}{N^{\min\{1,1/\vartheta\}}}.$$

Finally, Theorem 3.1 gives

$$L_N^*(\vartheta) = L_N(\vartheta)\frac{C(\vartheta)\|b\|}{\|b\|_{\beta(\vartheta)}} = \begin{cases} \frac{4^{1/\vartheta}}{\sqrt{2}}C(\vartheta), & \vartheta < 1, \\ \frac{4^{1/\vartheta}}{\sqrt{2}}\,4e, & \vartheta \geq 1, \end{cases}$$

which depends only on ϑ. Hence there exist constants $c_{1,\vartheta}, c_{2,\vartheta} > 0$, depending only on ϑ, such that

$$\mathrm{Prob}\left(\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \geq c_{1,\vartheta}K\sqrt{\frac{t}{N}} + c_{2,\vartheta}K\frac{t^{1/\vartheta}}{N^{\min\{1,1/\vartheta\}}}\right) \leq 2e^{-t} \qquad \forall t \geq 0.$$

We now absorb the heavy-tail correction term into the leading root-N term.

If $\vartheta = 2$, both terms are already of the same order, so there exists $C_\vartheta > 0$ such that

$$c_{1,\vartheta}K\sqrt{\frac{t}{N}} + c_{2,\vartheta}K\frac{t^{1/\vartheta}}{N^{1/\vartheta}} \leq C_\vartheta K\sqrt{\frac{t}{N}} \qquad \forall t \geq 0.$$

If $\vartheta \in [1,2)$, then $1/\vartheta - 1/2 > 0$. Therefore, there exists $c_\vartheta > 0$, depending only on ϑ, such that whenever $t \leq c_\vartheta N$,

$$c_{2,\vartheta}K\frac{t^{1/\vartheta}}{N^{1/\vartheta}} \leq c_{1,\vartheta}K\sqrt{\frac{t}{N}}.$$

Hence, for all $t \leq c_\vartheta N$,

$$c_{1,\vartheta}K\sqrt{\frac{t}{N}} + c_{2,\vartheta}K\frac{t^{1/\vartheta}}{N^{1/\vartheta}} \leq 2c_{1,\vartheta}K\sqrt{\frac{t}{N}}.$$

If $\vartheta \in (0,1)$, then there exists $c_\vartheta > 0$, depending only on ϑ, such that whenever $t \leq c_\vartheta N^{\vartheta/(2-\vartheta)}$,

$$c_{2,\vartheta}K\frac{t^{1/\vartheta}}{N} \leq c_{1,\vartheta}K\sqrt{\frac{t}{N}}.$$

Hence, for all $t \leq c_\vartheta N^{\vartheta/(2-\vartheta)}$,

$$c_{1,\vartheta}K\sqrt{\frac{t}{N}} + c_{2,\vartheta}K\frac{t^{1/\vartheta}}{N} \leq 2c_{1,\vartheta}K\sqrt{\frac{t}{N}}.$$

Setting $t = \log(2/\delta)$, the preceding three cases imply that there exist constants $C_\vartheta, c_\vartheta > 0$, depending only on ϑ, such that

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \leq C_\vartheta K\sqrt{\frac{\log(2/\delta)}{N}}$$

with probability at least $1-\delta$, provided that

$$\delta \in \begin{cases} \left[2\exp\left(-c_\vartheta N^{\vartheta/(2-\vartheta)}\right), 1\right), & \vartheta \in (0,1), \\ \left[2\exp\left(-c_\vartheta N\right), 1\right), & \vartheta \in [1,2), \\ (0,1), & \vartheta = 2. \end{cases}$$

Since $\log(2/\delta) \leq 1 + \log(1/\delta)$ for all $\delta \in (0,1)$, enlarging C_ϑ if necessary yields (13). □

Proof of Proposition 5. By Fan et al. (2021, Theorem 1), for any $s > 0$,

$$\mathrm{Prob}\left(\sum_{i=1}^{N} \ell(\boldsymbol{z}, \hat{\boldsymbol{\xi}}_i) - N\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] < -s\right) \leq \exp\left(-\frac{1-\lambda}{1+\lambda} \cdot \frac{s^2}{N(\overline{\ell} - \underline{\ell})^2/2}\right).$$

Dividing by N and substituting $s = Nt$ gives, for any $t \geq 0$,

$$\mathrm{Prob}\Big(\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] \geq t\Big) \leq \exp\left(-\frac{2(1-\lambda)Nt^2}{(1+\lambda)(\overline{\ell} - \underline{\ell})^2}\right).$$

Equating the right-hand side to δ and solving for t yields (14). The extension to the time-inhomogeneous setting follows by applying Theorem 5 in Fan et al. (2021). □

Following Jia et al. (2024), we make the following standard assumptions regarding the logging policy:

(i) *Conditional exchangeability:* Conditional on observed covariates $\boldsymbol{\chi}$, the selection decision is independent of the outcomes $\boldsymbol{\omega}$:

$$\tilde{\boldsymbol{\omega}} \perp\!\!\!\perp \tilde{S} \mid \boldsymbol{\chi}.$$

(ii) *Positivity:* The probability of a candidate being selected is strictly positive for any given covariate values:

$$\pi(\boldsymbol{\chi}) \geq \underline{\pi} \qquad \forall \boldsymbol{\chi} \in \mathscr{X},$$

for some positive constant $\underline{\pi} > 0$.

Lemma 2. *Suppose Assumptions (i) and (ii) hold. Then the inverse probability weighting estimator is equivalent to the true expectation:*

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) \frac{\mathbb{I}[\tilde{S} = 1]}{\overline{\mathbb{P}}(\tilde{S} = 1 | \tilde{\boldsymbol{\chi}} = \boldsymbol{\chi})}\right].$$

Proof of Lemma 2. We have

$$\begin{aligned}
\mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) \frac{\mathbb{I}[\tilde{S} = 1]}{\pi(\tilde{\boldsymbol{\chi}})}\right] &= \mathbb{E}_{\overline{\mathbb{P}}}\left[\mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) \frac{\mathbb{I}[\tilde{S} = 1]}{\pi(\tilde{\boldsymbol{\chi}})} \,\Big|\, \tilde{\boldsymbol{\chi}}, \tilde{\boldsymbol{\omega}}\right]\right] \\
&= \mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) \frac{\overline{\mathbb{P}}(\tilde{S} = 1 \mid \tilde{\boldsymbol{\chi}}, \tilde{\boldsymbol{\omega}})}{\pi(\tilde{\boldsymbol{\chi}})}\right] \\
&= \mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}}) \frac{\overline{\mathbb{P}}(\tilde{S} = 1 \mid \tilde{\boldsymbol{\chi}})}{\pi(\tilde{\boldsymbol{\chi}})}\right] \\
&= \mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})\right] \\
&= \mathbb{E}_{\mathbb{P}^\star}\left[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})\right].
\end{aligned}$$

The first equality is the tower property. The third equality follows from Assumption (i), the fourth uses the definition of $\pi(\boldsymbol{\chi})$, and the last holds because $\mathbb{P}^\star$ is the marginal law of $(\tilde{\boldsymbol{\chi}}, \tilde{\boldsymbol{\omega}})$ under $\overline{\mathbb{P}}$. □

The notation $\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]$ in (15) denotes an IPW reference functional rather than an expectation under a normalized empirical probability measure. This distinction does not affect the dual reformulations: the generalized moment problem only uses these reference values on the right-hand side of the loss constraints, and the same duality argument applies under the corresponding Slater condition.

Proof of Proposition 7. By positivity,

$$\left|\frac{\mathbb{I}[\tilde{S}=1]}{\overline{\mathbb{P}}(\tilde{S}=1|\tilde{\boldsymbol{\chi}}=\boldsymbol{\chi})}\right| \le \frac{1}{\underline{\pi}}.$$

Hence, under the assumed bound $\|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\|_{\psi_2} \le \sigma$,

$$\left\|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\frac{\mathbb{I}[\tilde{S}=1]}{\overline{\mathbb{P}}(\tilde{S}=1|\tilde{\boldsymbol{\chi}}=\boldsymbol{\chi})}\right\|_{\psi_2} \le \frac{\sigma}{\underline{\pi}}.$$

Centering changes the sub-Gaussian norm by at most a factor of two, so

$$\left\|\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\frac{\mathbb{I}[\tilde{S}=1]}{\overline{\mathbb{P}}(\tilde{S}=1|\tilde{\boldsymbol{\chi}}=\boldsymbol{\chi})} - \mathbb{E}_{\overline{\mathbb{P}}}\left[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\frac{\mathbb{I}[\tilde{S}=1]}{\overline{\mathbb{P}}(\tilde{S}=1|\tilde{\boldsymbol{\chi}}=\boldsymbol{\chi})}\right]\right\|_{\psi_2} \le \frac{2\sigma}{\underline{\pi}}.$$

Let

$$Y_z \coloneqq \ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\frac{\mathbb{I}[\tilde{S}=1]}{\overline{\mathbb{P}}(\tilde{S}=1|\tilde{\boldsymbol{\chi}}=\boldsymbol{\chi})}, \qquad Y_{z,n} \coloneqq \ell(\boldsymbol{z},\boldsymbol{\xi}_n)\frac{\mathbb{I}[S_n=1]}{\pi(\boldsymbol{\chi}_n)}.$$

Then (15) equals $N^{-1}\sum_{n=1}^N Y_{z,n}$, and the preceding bound applies to the centered variable $Y_z - \mathbb{E}_{\overline{\mathbb{P}}}[Y_z]$. By Lemma 2, $\mathbb{E}_{\overline{\mathbb{P}}}[Y_z] = \mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]$. Therefore, the estimation error can be written as

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] = -\frac{1}{N}\sum_{n=1}^N \left(Y_{z,n} - \mathbb{E}_{\overline{\mathbb{P}}}[Y_z]\right).$$

The summands in this display are independent, centered, and sub-Gaussian with ψ_2 norm bounded by $2\sigma/\underline{\pi}$. The negative average satisfies the same one-sided concentration bound, so taking the deviation level proportional to $\sqrt{\log(1/\delta)/N}$ and applying the standard concentration inequality for averages of independent centered sub-Gaussian variables (Vershynin 2018, Theorem 2.6.3) yields the stated bound

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \le C_{\mathrm{IPW}}\frac{\sigma}{\underline{\pi}}\sqrt{\frac{\log(1/\delta)}{N}}$$

with the numerical constants absorbed into the universal constant C_{IPW}. □

Proof of Proposition 8. Based on (Wang et al. 2024, Proposition 1), for any $h>0$ and $t>0$ we have

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}] \le L_{\boldsymbol{c}}h + t$$

with probability at least $1-\exp(-Nc\underline{f}h^{D_{\boldsymbol{c}}}t^2/2)$. Setting $\delta := \exp(-Nc\underline{f}h^{D_{\boldsymbol{c}}}t^2/2)$ gives

$$\mathbb{E}_{\mathbb{P}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{c}}=\boldsymbol{c}] \le L_{\boldsymbol{c}}h + \sqrt{\frac{2}{Nc\underline{f}h^{D_{\boldsymbol{c}}}}\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1-\delta$. The choice

$$h := \left(\frac{2\log(1/\delta)}{Nc\underline{f}}\right)^{\frac{1}{D_{\boldsymbol{c}}+2}} \left(\frac{D_{\boldsymbol{c}}}{2L_{\boldsymbol{c}}}\right)^{\frac{2}{D_{\boldsymbol{c}}+2}}$$

minimizes the right-hand side. At this bandwidth, the two terms satisfy

$$\sqrt{\frac{2}{Nc\underline{f}h^{D_{\boldsymbol{c}}}}\log\left(\frac{1}{\delta}\right)} = \frac{2L_{\boldsymbol{c}}h}{D_{\boldsymbol{c}}},$$

and hence the right-hand side equals

$$\frac{D_{\boldsymbol{c}}+2}{2(D_{\boldsymbol{c}}/2)^{\frac{D_{\boldsymbol{c}}}{D_{\boldsymbol{c}}+2}}} L_{\boldsymbol{c}}^{\frac{D_{\boldsymbol{c}}}{D_{\boldsymbol{c}}+2}} \left(\frac{2\log(1/\delta)}{Nc\underline{f}}\right)^{\frac{1}{D_{\boldsymbol{c}}+2}},$$

which is the stated bound. □

A.4. Dual Formulations

Proof of Theorem 2. Fix $\boldsymbol{x}\in\mathcal{X}$. By Proposition 2, the inner maximization problem in (3) can be written as

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathcal{M}_+(\Xi)} \quad & \int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})\,\mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \\
\text{s.t.} \quad & \int_\Xi \ell(\boldsymbol{z},\boldsymbol{\xi})\,\mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \le \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] + \epsilon \qquad \forall \boldsymbol{z}\in\mathcal{X}, \\
& \int_\Xi \|\boldsymbol{\xi}\|^2\,\mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \le \Omega, \\
& \int_\Xi 1\,\mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) = 1.
\end{aligned}$$

This is a generalized moment problem over nonnegative measures. Its conic dual introduces a nonnegative measure $\nu\in\mathcal{M}_+(\mathcal{X})$ for the family of loss constraints, a scalar multiplier $\beta\in\mathbb{R}_+$ for the second-moment constraint, and a free scalar $\alpha\in\mathbb{R}$ for the normalization constraint, yielding

$$\begin{aligned}
\min\ & \alpha + \beta\Omega + \int_{\mathcal{X}} \left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] + \epsilon\right)\nu(\mathrm{d}\boldsymbol{z}) \\
\text{s.t.}\ & \ell(\boldsymbol{x},\boldsymbol{\xi}) \le \alpha + \beta\|\boldsymbol{\xi}\|^2 + \int_{\mathcal{X}} \ell(\boldsymbol{z},\boldsymbol{\xi})\nu(\mathrm{d}\boldsymbol{z}) \qquad \forall \boldsymbol{\xi}\in\Xi.
\end{aligned}$$

Because $\epsilon>0$ and Assumption (B) gives $\mathbb{E}_{\hat{\mathbb{P}}_N}[\|\tilde{\boldsymbol{\xi}}\|^2]<\Omega$, the measure $\hat{\mathbb{P}}_N$ is strictly feasible for the primal inner problem. Standard strong duality for generalized moment problems therefore applies, and dual attainment holds for each fixed $\boldsymbol{x}$; see Shapiro (2001, Proposition 5.2).

Now let

$$F(\boldsymbol{x}) \coloneqq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].$$

The mapping $\boldsymbol{x}\mapsto F(\boldsymbol{x})$ is lower semicontinuous as a pointwise supremum of continuous functions, and $\mathcal{X}$ is compact, so the outer minimization attains its optimum. Combining this minimizer with the attained inner dual solution yields the stated infinite linear programming reformulation and the existence of an optimal tuple $(\boldsymbol{x},\alpha,\beta,\nu)$. □

Proof of Proposition 9. Fix $\boldsymbol{x}\in\mathcal{X}$, and define

$$f_\epsilon(\boldsymbol{x}) \coloneqq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].$$

Because $\mathcal{P}_0\subseteq\mathcal{P}_\epsilon$ for every $\epsilon\geq 0$, we have $f_0(\boldsymbol{x})\leq f_\epsilon(\boldsymbol{x})$. Moreover, by Proposition 2,

$$f_0(\boldsymbol{x}) = \sup_{\mathbb{Q}\in\mathcal{P}_0} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})],$$

since $\hat{\mathbb{P}}_N\in\mathcal{P}_0$ and every $\mathbb{Q}\in\mathcal{P}_0$ satisfies $\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\leq\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$.

It remains to prove $\limsup_{\epsilon\downarrow 0} f_\epsilon(\boldsymbol{x})\leq f_0(\boldsymbol{x})$. Let $\epsilon_k\downarrow 0$ and choose $\mathbb{Q}_k\in\mathcal{P}_{\epsilon_k}$ such that

$$f_{\epsilon_k}(\boldsymbol{x}) - \frac{1}{k} \leq \mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].$$

Because every $\mathbb{Q}_k$ satisfies the common second-moment bound, Markov's inequality implies that the family $\{\mathbb{Q}_k\}_{k\in\mathbb{N}}$ is tight; see (Billingsley 2013, Section 5). After passing to a subsequence if necessary, Prohorov's theorem (Billingsley 2013, Theorem 5.1) yields $\mathbb{Q}_k\Rightarrow\mathbb{Q}^\star$ for some probability measure $\mathbb{Q}^\star$. By Assumption (A) and the uniform second-moment bound, the family $\{\ell(\boldsymbol{x},\cdot)\}$ is uniformly integrable under $\{\mathbb{Q}_k\}_{k\in\mathbb{N}}$. Since $\ell(\boldsymbol{x},\cdot)$ is continuous, the mapping theorem (Billingsley 2013, Theorem 2.7) implies weak convergence of the pushforward laws, and (Billingsley 2013, Theorem 3.5) therefore gives

$$\lim_{k\to\infty} \mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].$$

The same argument holds for every fixed $\boldsymbol{z}\in\mathcal{X}$. Since $\mathbb{Q}_k\in\mathcal{P}_{\epsilon_k}$,

$$\mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \leq \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] + \epsilon_k \qquad \forall \boldsymbol{z}\in\mathcal{X}.$$

Letting $k\to\infty$ yields

$$\mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \leq \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \qquad \forall \boldsymbol{z}\in\mathcal{X},$$

and the second-moment constraint also passes to the limit by the Portmanteau theorem (Billingsley 2013, Theorem 2.1). Hence $\mathbb{Q}^\star\in\mathcal{P}_0$, so

$$\limsup_{k\to\infty} f_{\epsilon_k}(\boldsymbol{x}) \leq \lim_{k\to\infty} \mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \leq f_0(\boldsymbol{x}).$$

Therefore $f_\epsilon(\boldsymbol{x}) \downarrow f_0(\boldsymbol{x})$ pointwise on $\mathcal{X}$.

Let $v_\epsilon := \min_{\boldsymbol{x}\in\mathcal{X}} f_\epsilon(\boldsymbol{x})$ and $v_0 := \min_{\boldsymbol{x}\in\mathcal{X}} f_0(\boldsymbol{x})$. Since $f_\epsilon \geq f_0$ pointwise, we have $v_\epsilon \geq v_0$. On the other hand, for any minimizer $\boldsymbol{x}^\star \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} f_0(\boldsymbol{x})$,

$$\limsup_{\epsilon\downarrow 0} v_\epsilon \leq \limsup_{\epsilon\downarrow 0} f_\epsilon(\boldsymbol{x}^\star) = f_0(\boldsymbol{x}^\star) = v_0.$$

Hence $v_\epsilon \to v_0$.

Finally, let $\hat{\boldsymbol{x}}_\epsilon$ be a minimizer of (16) and let $\hat{\boldsymbol{x}}^\star$ be any cluster point of a sequence $\{\hat{\boldsymbol{x}}_{\epsilon_k}\}_{k\in\mathbb{N}}$ with $\epsilon_k \downarrow 0$. By compactness of $\mathcal{X}$, after passing to a subsequence we may assume $\hat{\boldsymbol{x}}_{\epsilon_k} \to \hat{\boldsymbol{x}}^\star$. Since $f_0(\boldsymbol{x}) = \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$ is continuous in $\boldsymbol{x}$,

$$f_0(\hat{\boldsymbol{x}}^\star) = \lim_{k\to\infty} f_0(\hat{\boldsymbol{x}}_{\epsilon_k}) \leq \lim_{k\to\infty} f_{\epsilon_k}(\hat{\boldsymbol{x}}_{\epsilon_k}) = \lim_{k\to\infty} v_{\epsilon_k} = v_0.$$

Thus $\hat{\boldsymbol{x}}^\star$ is a minimizer of the empirical risk minimization problem (17). □

Appendix B: Proofs of Section 3

B.1. Theoretical Guarantees

PROPOSITION 11. *With probability at least $1-\tau$, we have:*

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right] \\
&\leq \frac{1}{M}\sum_{m\in[M]}\left[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}_m, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+ + \mathcal{O}\left(R\sqrt{\frac{J\log J(\log M)^3}{M}}\right) \\
&\quad + 2(c_1 + c_2\sqrt{\Omega})\sqrt{\frac{1}{2M}\log\left(\frac{1}{\tau}\right)} \qquad\qquad \forall \mathbb{Q}\in\mathscr{P}(\Xi) : \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega.
\end{aligned}$$

Proof of Proposition 11. We define $h_{\mathbb{Q}}(\boldsymbol{z}) := \left[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+$ and let

$$\mathcal{H} := \{h_{\mathbb{Q}} : \mathbb{Q}\in\mathscr{P}(\Xi),\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega\}.$$

Our Assumption (A) implies that

$$\bar{h} := 2(c_1 + c_2\sqrt{\Omega}) \geq h(\boldsymbol{z}) \geq 0 \qquad \forall h\in\mathcal{H}\ \forall \boldsymbol{z}\in\mathcal{X}.$$

For a given sample set $\{\tilde{\boldsymbol{z}}_m\}_{m\in[M]}$, we define the uniform approximation error as:

$$e(\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_M) := \sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathbb{Z}}[h(\tilde{\boldsymbol{z}})] - \frac{1}{M}\sum_{m\in[M]} h(\tilde{\boldsymbol{z}}_m)\right).$$

Note that changing one data point $\tilde{\boldsymbol{z}}_m$ changes e by at most $\bar{h}/M$. McDiarmid's inequality (Mohri et al. 2018, Appendix D) therefore yields

$$e(\tilde{\boldsymbol{z}}_1,\ldots,\tilde{\boldsymbol{z}}_M) \leq \mathbb{E}[e(\tilde{\boldsymbol{z}}_1,\ldots,\tilde{\boldsymbol{z}}_M)] + \bar{h}\sqrt{\frac{1}{2M}\log\left(\frac{1}{\tau}\right)}$$

with probability at least $1-\tau$.

We next upper bound the expectation using Rademacher complexity. Introduce M i.i.d. ghost samples $\tilde{\boldsymbol{z}}'_1,\ldots,\tilde{\boldsymbol{z}}'_M$ drawn from $\mathbb{Z}$ and let $\tilde{s}_1,\ldots,\tilde{s}_M \in \{-1,1\}$ be i.i.d. Rademacher random variables. A standard symmetrization inequality (Mohri et al. 2018, Theorem 3.3) yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathbb{Z}}[h(\tilde{\boldsymbol{z}})] - \frac{1}{M}\sum_{m\in[M]} h(\tilde{\boldsymbol{z}}_m)\right)\right] &\leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{M}\sum_{m\in[M]} h(\tilde{\boldsymbol{z}}'_m) - \frac{1}{M}\sum_{m\in[M]} h(\tilde{\boldsymbol{z}}_m)\right)\right] \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{M}\sum_{m\in[M]} \tilde{s}_m(h(\tilde{\boldsymbol{z}}'_m) - h(\tilde{\boldsymbol{z}}_m))\right)\right] \\
&\leq 2\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{M}\sum_{m\in[M]} \tilde{s}_m h(\tilde{\boldsymbol{z}}_m)\right)\right] = 2\mathcal{R}_M(\mathcal{H}).
\end{aligned}$$

Since the hinge map $u \mapsto [u]_+$ is 1-Lipschitz, the contraction principle (Shalev-Shwartz and Ben-David 2014, Lemma 26.9) gives

$$\begin{aligned}
\mathcal{R}_M(\mathcal{H}) &\leq \mathbb{E}\left[\sup_{\substack{\mathbb{Q}\in\mathscr{P}(\Xi)\\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]\leq\Omega}}\left(\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\left(\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})] - \epsilon\right)\right)\right] \\
&= \mathbb{E}\left[\sup_{\substack{\mathbb{Q}\in\mathscr{P}(\Xi)\\ \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]\leq\Omega}}\mathbb{E}_{\mathbb{Q}}\left[\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})\right]\right] - \mathbb{E}\left[\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})] + \epsilon\right)\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbb{Q}\in\mathscr{P}(\Xi)}\mathbb{E}_{\mathbb{Q}}\left[\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})\right]\right] \\
&= \mathbb{E}\left[\sup_{\boldsymbol{\xi}\in\Xi}\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\boldsymbol{\xi})\right].
\end{aligned} \tag{32}$$

In the second line, we moved the term $\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})]+\epsilon\right)$ outside the supremum as it is independent of $\mathbb{Q}$. Since each $\tilde{s}_m$ is centered and independent of $\tilde{\boldsymbol{z}}_m$, this term has expectation zero.

Each affine piece in the loss function can be rewritten as

$$\boldsymbol{a}_j(\boldsymbol{x})^\top\boldsymbol{\xi} + b_j(\boldsymbol{x}) = \left(\begin{bmatrix}\boldsymbol{A}_j^\top & \boldsymbol{b}_j\\ \overline{\boldsymbol{a}}_j^\top & \bar{b}_j\end{bmatrix}\begin{bmatrix}\boldsymbol{\xi}\\ 1\end{bmatrix}\right)^\top\begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix}.$$

Let $\mathcal{B}^{D_{\boldsymbol{x}}+1}$ be the unit ball in $\mathbb{R}^{D_{\boldsymbol{x}}+1}$. For each $\boldsymbol{\xi}\in\Xi$ and $j\in[J]$, define

$$\boldsymbol{w}_j(\boldsymbol{\xi}) \coloneqq \left(\max_{i\in[J]}\left\|\begin{bmatrix}\boldsymbol{A}_i^\top & \boldsymbol{b}_i\\ \overline{\boldsymbol{a}}_i^\top & \overline{b}_i\end{bmatrix}\right\|_{\mathrm{op}}\sqrt{R_{\boldsymbol{\xi}}^2+1}\right)^{-1}\begin{bmatrix}\boldsymbol{A}_j^\top & \boldsymbol{b}_j\\ \overline{\boldsymbol{a}}_j^\top & \overline{b}_j\end{bmatrix}\begin{bmatrix}\boldsymbol{\xi}\\ 1\end{bmatrix},$$

so that $\boldsymbol{w}_j(\boldsymbol{\xi})\in\mathcal{B}^{D_{\boldsymbol{x}}+1}$ for all $j\in[J]$, and also

$$\left(\sqrt{R_{\boldsymbol{x}}^2+1}\right)^{-1}\begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix}\in\mathcal{B}^{D_{\boldsymbol{x}}+1}.$$

Hence

$$\begin{aligned}
\mathcal{R}_M(\mathcal{H}) &\le \mathbb{E}\left[\sup_{\boldsymbol{\xi}\in\Xi}\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\boldsymbol{\xi})\right]\\
&= \mathbb{E}\left[\sup_{\boldsymbol{\xi}\in\Xi}\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\max_{j\in[J]}\big(\boldsymbol{a}_j(\tilde{\boldsymbol{z}}_m)^\top\boldsymbol{\xi}+b_j(\tilde{\boldsymbol{z}}_m)\big)\right]\\
&\le \mathbb{E}\left[\sup_{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_J\in\Xi}\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\max_{j\in[J]}\big(\boldsymbol{a}_j(\tilde{\boldsymbol{z}}_m)^\top\boldsymbol{\xi}_j+b_j(\tilde{\boldsymbol{z}}_m)\big)\right] \qquad (33)\\
&\le \max_{j\in[J]}\left\|\begin{bmatrix}\boldsymbol{A}_j^\top & \boldsymbol{b}_j\\ \overline{\boldsymbol{a}}_j^\top & \overline{b}_j\end{bmatrix}\right\|_{\mathrm{op}}\sqrt{R_{\boldsymbol{\xi}}^2+1}\sqrt{R_{\boldsymbol{x}}^2+1}\\
&\qquad\times\mathbb{E}\left[\sup_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_J\in\mathcal{B}^{D_{\boldsymbol{x}}+1}}\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\max_{j\in[J]}\boldsymbol{w}_j^\top\left(\sqrt{R_{\boldsymbol{x}}^2+1}\right)^{-1}\begin{bmatrix}\tilde{\boldsymbol{z}}_m\\ 1\end{bmatrix}\right]\\
&\le \mathcal{O}\left(\max_{j\in[J]}\left\|\begin{bmatrix}\boldsymbol{A}_j^\top & \boldsymbol{b}_j\\ \overline{\boldsymbol{a}}_j^\top & \overline{b}_j\end{bmatrix}\right\|_{\mathrm{op}}\sqrt{R_{\boldsymbol{\xi}}^2+1}\sqrt{R_{\boldsymbol{x}}^2+1}\sqrt{\frac{J\log J(\log M)^3}{M}}\right),
\end{aligned}$$

where the last step uses the Rademacher complexity of J-fold maxima of hyperplanes (Attias and Kontorovich 2024, Corollary 5). Combining the McDiarmid and Rademacher bounds yields the claim. □

Proof of Theorem 3. Consider the high-probability event of Proposition 11, which occurs with probability at least $1-\tau$. Fix any $\mathbb{Q}\in\mathcal{P}_\epsilon^M$. By definition of $\mathcal{P}_\epsilon^M$,

$$\frac{1}{M}\sum_{m\in[M]}\Big[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})]-\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})]-\epsilon\Big]_+\le 0\quad\text{and}\quad\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]\le\Omega.$$

Applying Proposition 11 to this $\mathbb{Q}$ shows that

$$\mathbb{E}_{\mathbb{Z}}\left[\Big[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})]-\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})]-\epsilon\Big]_+\right]\le\eta,$$

where η is the bound in (21). Together with the shared second-moment constraint, this means $\mathbb{Q}\in\mathcal{P}_\epsilon(\eta)$. Since $\mathbb{Q}\in\mathcal{P}_\epsilon^M$ was arbitrary, we obtain $\mathcal{P}_\epsilon^M\subseteq\mathcal{P}_\epsilon(\eta)$. □

Lemma 3. *As $\eta\downarrow 0$, the optimal value of the relaxed DRO problem*

$$\min_{\boldsymbol{x}\in\mathcal{X}}\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon(\eta)}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$$

converges to that of the exact DRO problem.

Proof of Lemma 3. We follow the development of the proof of Proposition 9. Fix $\boldsymbol{x} \in \mathcal{X}$, and define

$$f_\eta(\boldsymbol{x}) \coloneqq \sup_{\mathbb{Q} \in \mathcal{P}_\epsilon(\eta)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})].$$

Since $\mathcal{P}_\epsilon = \mathcal{P}_\epsilon(0) \subseteq \mathcal{P}_\epsilon(\eta)$ for every $\eta \geq 0$, we have $f_0(\boldsymbol{x}) \leq f_\eta(\boldsymbol{x})$. Thus, it remains to prove

$$\lim_{\eta \downarrow 0} f_\eta(\boldsymbol{x}) \leq f_0(\boldsymbol{x}). \tag{34}$$

Let $\eta_k \downarrow 0$ be any decreasing sequence converging to zero. For each $k \in \mathbb{N}$, choose $\mathbb{Q}_k \in \mathcal{P}_\epsilon(\eta_k)$ such that

$$f_{\eta_k}(\boldsymbol{x}) - \frac{1}{k} \leq \mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]. \tag{35}$$

Because each $\mathbb{Q}_k$ satisfies the second-moment bound in $\mathcal{P}_\epsilon(\eta_k)$, the family $\{\mathbb{Q}_k\}_{k \in \mathbb{N}}$ is tight. Hence, after passing to a subsequence if necessary, $\mathbb{Q}_k \Rightarrow \mathbb{Q}^\star$ for some probability measure $\mathbb{Q}^\star$.

By the same argument as in Proposition 9, weak convergence together with the uniform second-moment bound implies $\lim_{k\to\infty} \mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$, and $\mathbb{Q}^\star$ satisfies the second-moment constraint in $\mathcal{P}_\epsilon$. It remains to verify the expected hinge constraint. For each fixed $\boldsymbol{z} \in \mathcal{X}$, the convergence of expectations implies

$$\lim_{k\to\infty} \left[\mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+ = \left[\mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+.$$

By Assumptions (A) and (B), the hinge term is bounded uniformly in $\boldsymbol{z} \in \mathcal{X}$ and $k \in \mathbb{N}$ by a finite constant, so dominated convergence yields

$$\lim_{k\to\infty} \mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}_k}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right] = \mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}^\star}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right].$$

Since $\mathbb{Q}_k \in \mathcal{P}_\epsilon(\eta_k)$, we have

$$\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}_k}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right] \leq \eta_k.$$

Letting $k \to \infty$, we obtain $\mathbb{Q}^\star \in \mathcal{P}_\epsilon$. Taking the limit in (35) gives

$$\lim_{k\to\infty} f_{\eta_k}(\boldsymbol{x}) \leq \lim_{k\to\infty} \mathbb{E}_{\mathbb{Q}_k}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] \leq f_0(\boldsymbol{x}).$$

Since $\eta_k \downarrow 0$ was arbitrary, (34) holds. Thus, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\lim_{\eta \downarrow 0} \sup_{\mathbb{Q} \in \mathcal{P}_\epsilon(\eta)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] = \sup_{\mathbb{Q} \in \mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]. \tag{36}$$

The optimal value convergence then follows from the monotonicity of $f_\eta(\boldsymbol{x})$ in η. □

Proof of Theorem 4. For $\eta \geq 0$, define

$$\hat{v}(\eta) \coloneqq \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon(\eta)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].$$

By Theorem 3, setting η_M to (21), we have for any $\tau > 0$,

$$\text{Prob}\left(\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \leq \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon(\eta_M)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \right) \geq 1-\tau.$$

Equivalently,

$$\text{Prob}(\hat{v}_M \leq \hat{v}(\eta_M)) \geq 1-\tau.$$

Since $\mathcal{P}_\epsilon \subseteq \mathcal{P}_\epsilon^M$, we also have $\hat{v} \leq \hat{v}_M$. Therefore,

$$\text{Prob}(\hat{v} \leq \hat{v}_M \leq \hat{v}(\eta_M)) \geq 1-\tau.$$

By Lemma 3, $\hat{v}(\eta_M) \to \hat{v}$ as $M \to \infty$. Hence, for any $\rho > 0$, there exists $M' \in \mathbb{N}$ such that

$$\hat{v}(\eta_M) - \hat{v} \leq \rho \qquad \forall M \geq M'.$$

Combining this bound with the preceding high-probability event yields

$$\text{Prob}(\hat{v}_M - \hat{v} \leq \rho) \geq 1-\tau \qquad \forall M \geq M'.$$

Since also $\hat{v}_M - \hat{v} \geq 0$, we conclude that $\hat{v}_M \xrightarrow{p} \hat{v}$.

We next establish convergence of optimal solutions. For $\delta > 0$, define

$$\mathcal{A}_\delta \coloneqq \{\boldsymbol{x} \in \mathcal{X} : \text{dist}(\boldsymbol{x}, \mathcal{S}) \geq \delta\}.$$

This set is closed because the distance function is continuous. Since $\mathcal{X}$ is compact, $\mathcal{A}_\delta$ is compact. Moreover, by definition of $\mathcal{S}$, no point in $\mathcal{A}_\delta$ is optimal. Hence,

$$\rho_\delta \coloneqq \inf_{\boldsymbol{x}\in\mathcal{A}_\delta} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] - \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] > 0.$$

Now, if $\text{dist}(\hat{\boldsymbol{x}}_M, \mathcal{S}) \geq \delta$, then $\hat{\boldsymbol{x}}_M \in \mathcal{A}_\delta$, and therefore

$$\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M,\tilde{\boldsymbol{\xi}})] - \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \geq \rho_\delta.$$

Since $\mathcal{P}_\epsilon \subseteq \mathcal{P}_\epsilon^M$, we also have

$$\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M,\tilde{\boldsymbol{\xi}})] \leq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M,\tilde{\boldsymbol{\xi}})] = \hat{v}_M.$$

Thus, $\{\text{dist}(\hat{\boldsymbol{x}}_M, \mathcal{S}) \geq \delta\} \subseteq \{\hat{v}_M - \hat{v} \geq \rho_\delta\}$, and consequently

$$\text{Prob}(\text{dist}(\hat{\boldsymbol{x}}_M, \mathcal{S}) \geq \delta) \leq \text{Prob}(\hat{v}_M - \hat{v} \geq \rho_\delta).$$

Since $\hat{v}_M \xrightarrow{p} \hat{v}$, the right-hand side converges to zero. Therefore, $\operatorname{dist}(\hat{\boldsymbol{x}}_M, \mathcal{S}) \xrightarrow{p} 0$.

Lastly, we prove the suboptimality bound. We have

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] &= \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M, \tilde{\boldsymbol{\xi}})] \\ &\quad + \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \\ &\leq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})].\end{aligned}$$

Here, the first difference is nonpositive because $\mathcal{P}_\epsilon \subseteq \mathcal{P}_\epsilon^M$, and the second difference is bounded above by replacing $\hat{\boldsymbol{x}}_M$ with $\hat{\boldsymbol{x}}$, since $\hat{\boldsymbol{x}}$ is suboptimal for the approximate problem.

Next, by Theorem 3, setting η to (21), we obtain the high-probability bound

$$\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon^M} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \leq \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon(\eta)} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})]$$

with probability at least $1-\tau$. For any fixed $\boldsymbol{x}$ and $\eta \geq 0$, the relaxed problem (20) has the dual reformulation

$$\begin{aligned}&\min \alpha + \beta\Omega + \gamma\eta + \int_{\mathcal{X}} \left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] + \epsilon\right) \nu(\mathrm{d}\boldsymbol{z}) \\ &\text{s.t. } \boldsymbol{x} \in \mathcal{X},\ \alpha \in \mathbb{R},\ \beta, \gamma \in \mathbb{R}_+,\ \nu \in \mathcal{M}_+(\mathcal{X}), \\ &\qquad \ell(\boldsymbol{x}, \boldsymbol{\xi}) \leq \alpha + \beta\|\boldsymbol{\xi}\|^2 + \int_{\mathcal{X}} \ell(\boldsymbol{z}, \boldsymbol{\xi}) \nu(\mathrm{d}\boldsymbol{z}) \qquad \forall \boldsymbol{\xi} \in \Xi, \\ &\qquad \gamma\mathbb{Z} - \nu \in \mathcal{M}_+(\mathcal{X}),\end{aligned}$$

which follows from the same generalized-moment duality argument as in Theorem 2. Let $(\hat{\boldsymbol{x}}, \alpha^\star, \beta^\star, \nu^\star)$ be an optimal solution to the exact dual problem, and let $\gamma^\star$ be as defined in the theorem statement. Then $(\hat{\boldsymbol{x}}, \alpha^\star, \beta^\star, \gamma^\star, \nu^\star)$ is feasible, though generally suboptimal, for the relaxed dual problem. We can therefore further upper bound the preceding display by

$$\begin{aligned}&\sup_{\mathbb{Q}\in\mathcal{P}_\epsilon(\eta)} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \\ &\leq \alpha^\star + \beta^\star\Omega + \int_{\mathcal{X}} \left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] + \epsilon\right) \nu^\star(\mathrm{d}\boldsymbol{z}) + \gamma^\star\eta \\ &\qquad - \left(\alpha^\star + \beta^\star\Omega + \int_{\mathcal{X}} \left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] + \epsilon\right) \nu^\star(\mathrm{d}\boldsymbol{z})\right) \\ &= \gamma^\star\eta.\end{aligned}$$

Substituting the expression for η in (21) completes the proof. □

B.2. Conic Programming Reformulations

Proof of Theorem 5. For each sampled point $\boldsymbol{z}_m$, define

$$g_m(\boldsymbol{\xi}) \coloneqq \max_{k\in[J]} \left(\boldsymbol{a}_k(\boldsymbol{z}_m)^\top \boldsymbol{\xi} + b_k(\boldsymbol{z}_m)\right).$$

The semi-infinite constraint in (19) is equivalent to

$$\boldsymbol{a}_j(\boldsymbol{x})^\top \boldsymbol{\xi} + b_j(\boldsymbol{x}) \le \alpha + \beta\|\boldsymbol{\xi}\|^2 + \frac{1}{M}\sum_{m\in[M]} \nu_m g_m(\boldsymbol{\xi}) \qquad \forall \boldsymbol{\xi}\in\Xi,\ \forall j\in[J].$$

For each pair (j,m), the term $\nu_m g_m(\boldsymbol{\xi})$ admits the simplex representation

$$\nu_m g_m(\boldsymbol{\xi}) = \max_{\substack{\boldsymbol{\lambda}_{jm}\in\mathbb{R}^J_+ \\ \mathbf{e}^\top\boldsymbol{\lambda}_{jm}=\nu_m}} \sum_{k\in[J]} \lambda^k_{jm}\big(\boldsymbol{a}_k(\boldsymbol{z}_m)^\top\boldsymbol{\xi} + b_k(\boldsymbol{z}_m)\big).$$

Hence the semi-infinite constraint is equivalent to the existence of multipliers $\boldsymbol{\lambda}_{jm}\in\mathbb{R}^J_+$ with $\mathbf{e}^\top\boldsymbol{\lambda}_{jm}=\nu_m$ such that, for every $j\in[J]$,

$$\sup_{\boldsymbol{\xi}\in\Xi}\left\{\left(\boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}\boldsymbol{a}_k(\boldsymbol{z}_m)\right)^\top\boldsymbol{\xi} - \beta\|\boldsymbol{\xi}\|^2\right\} + b_j(\boldsymbol{x}) - \frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}b_k(\boldsymbol{z}_m) \le \alpha.$$

Let

$$\boldsymbol{c}_j \coloneqq \boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}\boldsymbol{a}_k(\boldsymbol{z}_m).$$

Using Fenchel duality with the support function of Ξ and the infimal convolution identity for Fenchel conjugates (Rockafellar 1970, Theorem 16.4),

$$\sup_{\boldsymbol{\xi}\in\Xi}\big\{\boldsymbol{c}_j^\top\boldsymbol{\xi} - \beta\|\boldsymbol{\xi}\|^2\big\} = \inf_{\boldsymbol{\theta}_j\in\mathbb{R}^{D_{\boldsymbol{\xi}}}}\left\{\sigma_\Xi(\boldsymbol{\theta}_j) + \sup_{\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}}\left[(\boldsymbol{c}_j-\boldsymbol{\theta}_j)^\top\boldsymbol{\xi} - \beta\|\boldsymbol{\xi}\|^2\right]\right\}.$$

The remaining quadratic supremum equals $\|\boldsymbol{c}_j-\boldsymbol{\theta}_j\|^2/(4\beta)$ when $\beta>0$, and its epigraph admits the standard second-order-cone representation

$$\frac{\|\boldsymbol{c}_j-\boldsymbol{\theta}_j\|^2}{4\beta} \le \zeta_j \iff \left\|\begin{bmatrix}\boldsymbol{c}_j-\boldsymbol{\theta}_j \\ \zeta_j-\beta\end{bmatrix}\right\| \le \zeta_j+\beta, \qquad \zeta_j\ge 0.$$

When $\beta=0$, the semi-infinite constraint is satisfied if and only if

$$\sigma_\Xi(\boldsymbol{c}_j) + b_j(\boldsymbol{x}) - \frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}b_k(\boldsymbol{z}_m) \le \alpha \qquad \forall j\in[J].$$

Moreover, when $\beta=0$, the second-order conic constraint above reduces to

$$\left\|\begin{bmatrix}\boldsymbol{c}_j-\boldsymbol{\theta}_j \\ \zeta_j\end{bmatrix}\right\| \le \zeta_j,$$

which implies $\boldsymbol{\theta}_j=\boldsymbol{c}_j$. Substituting this identity into the scalar inequality of the conic reformulation gives

$$\zeta_j + \sigma_\Xi(\boldsymbol{c}_j) + b_j(\boldsymbol{x}) - \frac{1}{M}\sum_{m\in[M],k\in[J]}\lambda^k_{jm}b_k(\boldsymbol{z}_m) \le \alpha.$$

In this scalar inequality, smaller values of ζ_j weakly enlarge the feasible region. Since the reduced conic constraint imposes no positive lower bound on ζ_j beyond $\zeta_j\ge 0$, the least restrictive feasible

choice is $\zeta_j = 0$, which recovers exactly the support-function constraint above. Hence the conic reformulation in the theorem remains valid also in the case $\beta = 0$. Substituting $\boldsymbol{c}_j$ back into the inequality therefore produces exactly the conic reformulation stated in the theorem.

It remains to verify convexity of the displayed reformulation. The objective is affine in the decision variables, and the sign, simplex, and equality constraints are affine. Since $\boldsymbol{a}_j(\boldsymbol{x})$ and $b_j(\boldsymbol{x})$ are affine in $\boldsymbol{x}$, the scalar inequality

$$\zeta_j + \sigma_\Xi(\boldsymbol{\theta}_j) + b_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m\in[M],k\in[J]} \lambda_{jm}^k b_k(\boldsymbol{z}_m) \leq \alpha$$

has a convex left-hand side, because σ_Ξ is convex as the support function of a closed convex set. The remaining nonlinear constraint is the second-order-cone membership

$$\left(\boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m\in[M],k\in[J]} \lambda_{jm}^k \boldsymbol{a}_k(\boldsymbol{z}_m) - \boldsymbol{\theta}_j,\ \zeta_j - \beta,\ \zeta_j + \beta \right) \in \mathcal{Q},$$

where $\mathcal{Q} \coloneqq \{(\boldsymbol{u}, t) : \|\boldsymbol{u}\| \leq t\}$ denotes the standard second-order cone. Its argument is affine in the decision variables, so its inverse image is convex. Together with the convexity of the constraint $\boldsymbol{x} \in \mathcal{X}$, this shows that the finite reformulation is a convex conic program. The final claims follow because the support function of a second-order-cone representable set is second-order-cone representable. □

Appendix C: Proofs of Section 4

C.1. Theoretical Guarantees

Proof of Proposition 10. Fix $\boldsymbol{x} \in \mathcal{X}$. The inner maximization problem associated with (23) is again a generalized moment problem, now with the additional constraint

$$\mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\| \leq t) \geq 1 - 2\exp\left(-(t/K_{\mathrm{sw}})^\vartheta\right).$$

Introducing a multiplier $\kappa \in \mathbb{R}_+$ for this constraint, together with the multipliers $\nu \in \mathcal{M}_+(\mathcal{X})$, $\beta \in \mathbb{R}_+$, and $\alpha \in \mathbb{R}$ used in the proof of Theorem 2, yields the dual problem

$$\begin{array}{ll} \min & \alpha + \beta\Omega - \kappa\left(1 - 2\exp\left(-(t/K_{\mathrm{sw}})^\vartheta\right)\right) + \displaystyle\int_{\mathcal{X}} \left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}, \tilde{\boldsymbol{\xi}})] + \epsilon\right) \nu(\mathrm{d}\boldsymbol{z}) \\ \text{s.t.} & \ell(\boldsymbol{x}, \boldsymbol{\xi}) + \kappa \mathbb{I}_{\{\|\boldsymbol{\xi}\| \leq t\}} \leq \alpha + \beta\|\boldsymbol{\xi}\|^2 + \displaystyle\int_{\mathcal{X}} \ell(\boldsymbol{z}, \boldsymbol{\xi}) \nu(\mathrm{d}\boldsymbol{z}) \qquad \forall \boldsymbol{\xi} \in \Xi. \end{array}$$

The strict feasibility argument from Theorem 2 still applies, so strong duality and dual attainment hold. Minimizing over $\boldsymbol{x} \in \mathcal{X}$ gives (24) and yields the existence of an optimal tuple. □

PROPOSITION 12. *With probability at least $1-\tau$, the following bound holds for all $\mathbb{Q} \in \mathscr{P}(\Xi)$ satisfying $\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega$ and $\mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\| \leq t) \geq 1 - 2\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right)$:*

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right] \\
&\leq \frac{1}{M}\sum_{m\in[M]}\left[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+ + \mathcal{O}\left(R(t)\sqrt{\frac{J\log J(\log M)^3}{M}}\right) \\
&\qquad + 2(c_1 + c_2\sqrt{\Omega})\sqrt{\frac{1}{2M}\log\left(\frac{1}{\tau}\right)} + 4c_1\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right) + 2c_2\sqrt{2\Omega\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right)}.
\end{aligned}$$

Here $R(t)$ is defined as in Theorem 6.

Proof of Proposition 12. We define the ambiguity set

$$\mathcal{P}_{\mathrm{sw}} \coloneqq \left\{\mathbb{Q} \in \mathscr{P}(\Xi) : \begin{array}{l} \mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2] \leq \Omega \\ \mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\| \leq t) \geq 1 - 2\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right) \end{array}\right\},$$

which contains both $\mathcal{P}'_\epsilon$ and $\mathcal{P}'^M_\epsilon$. Let $h_{\mathbb{Q}}(\boldsymbol{z}) \coloneqq \left[\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+$ and $\mathcal{H} \coloneqq \{h_{\mathbb{Q}} : \mathbb{Q} \in \mathcal{P}_{\mathrm{sw}}\}$. We now repeat the steps of the proof of Proposition 11. Assumption (A) implies that $\bar{h} \coloneqq 2(c_1 + c_2\sqrt{\Omega})$ satisfies $0 \leq h(\boldsymbol{z}) \leq \bar{h}$ for all $h \in \mathcal{H}$ and $\boldsymbol{z} \in \mathcal{X}$.

By McDiarmid's inequality (Mohri et al. 2018, Appendix D), we have

$$\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathbb{Z}}[h(\tilde{\boldsymbol{z}})] - \frac{1}{M}\sum_{m\in[M]} h(\tilde{\boldsymbol{z}}_m)\right) \leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{\mathbb{Z}}[h(\tilde{\boldsymbol{z}})] - \frac{1}{M}\sum_{m\in[M]} h(\tilde{\boldsymbol{z}}_m)\right)\right] + \bar{h}\sqrt{\frac{1}{2M}\log\left(\frac{1}{\tau}\right)}$$

with probability at least $1-\tau$. The expectation term can be bounded by twice the Rademacher complexity of $\mathcal{H}$. Let $F_M(\boldsymbol{\xi}) \coloneqq M^{-1}\sum_{m\in[M]} \tilde{s}_m \ell(\tilde{\boldsymbol{z}}_m, \boldsymbol{\xi})$. Then

$$\begin{aligned}
\mathcal{R}_M(\mathcal{H}) &\leq \mathbb{E}\left[\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}} \mathbb{E}_{\mathbb{Q}}[F_M(\tilde{\boldsymbol{\xi}})]\right] \\
&= \mathbb{E}\left[\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}}\left\{\mathbb{E}_{\mathbb{Q}}\left[F_M(\tilde{\boldsymbol{\xi}})\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|\leq t\}}\right] + \mathbb{E}_{\mathbb{Q}}\left[F_M(\tilde{\boldsymbol{\xi}})\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|> t\}}\right]\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}} \mathbb{E}_{\mathbb{Q}}\left[F_M(\tilde{\boldsymbol{\xi}})\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|\leq t\}}\right]\right] \\
&\qquad + \mathbb{E}\left[\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}} \mathbb{E}_{\mathbb{Q}}\left[F_M(\tilde{\boldsymbol{\xi}})\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|> t\}}\right]\right].
\end{aligned}$$

For the first expectation, since the indicator restricts the integrand to the ball $\{\|\boldsymbol{\xi}\| \leq t\}$,

$$\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}} \mathbb{E}_{\mathbb{Q}}\left[F_M(\tilde{\boldsymbol{\xi}})\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|\leq t\}}\right] \leq \sup_{\|\boldsymbol{\xi}\|\leq t} |F_M(\boldsymbol{\xi})|.$$

By symmetry of the Rademacher signs, the expectation of the right-hand side is bounded, up to an absolute constant, by the localized version of the complexity bound in (33):

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}} \mathbb{E}_{\mathbb{Q}}\left[F_M(\tilde{\boldsymbol{\xi}})\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|\le t\}}\right]\right] \\
&\le \mathbb{E}\left[\sup_{\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}:\|\boldsymbol{\xi}\|\le t}\left|\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\boldsymbol{\xi})\right|\right] \\
&\le 2\mathbb{E}\left[\sup_{\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}:\|\boldsymbol{\xi}\|\le t}\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\boldsymbol{\xi})\right] \\
&\le \mathcal{O}\left(R(t)\sqrt{\frac{J\log J(\log M)^3}{M}}\right).
\end{aligned}$$

For the second expectation, Assumption (A), Cauchy–Schwarz, and the defining constraints of $\mathcal{P}_{\mathrm{sw}}$ yield

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}_{\mathrm{sw}}} \mathbb{E}_{\mathbb{Q}}\left[\left|\frac{1}{M}\sum_{m\in[M]}\tilde{s}_m\ell(\tilde{\boldsymbol{z}}_m,\tilde{\boldsymbol{\xi}})\right|\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\|>t\}}\right] \\
&\le c_1\,\mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\|>t)+c_2\sqrt{\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]\,\mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\|>t)} \\
&\le 2c_1\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right)+c_2\sqrt{2\Omega\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right)},
\end{aligned}$$

where the last step uses the tail-mass constraint $\mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\|>t)\le 2\exp\left(-(t/K_{\mathrm{sw}})^{\vartheta}\right)$. The factor 2 in the tail correction of the statement comes from the symmetrization step. Combining these bounds with the same symmetrization argument as in Proposition 11 yields the claim. □

Proof of Theorem 6. On the high-probability event of Proposition 12 applied with $t=t_M$, fix any $\mathbb{Q}\in\mathcal{P}_\epsilon'^M$. By definition of (25), the sample-average hinge term is nonpositive, $\mathbb{E}_{\mathbb{Q}}[\|\tilde{\boldsymbol{\xi}}\|^2]\le\Omega$, and $\mathbb{Q}(\|\tilde{\boldsymbol{\xi}}\|\le t_M)\ge 1-2/M$. Proposition 12 therefore implies

$$\mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})]-\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})]-\epsilon\right]_+\right]\le\eta,$$

where $\eta=\eta_M$ is the bound stated in Theorem 6. Thus $\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}(\eta_M)$. Since $\mathbb{Q}$ was arbitrary, we obtain $\mathcal{P}_\epsilon'^M\subseteq\mathcal{P}'_{\epsilon,M}(\eta_M)$. The inclusion $\mathcal{P}'_{\epsilon,M}\subseteq\mathcal{P}_\epsilon'^M$ is immediate from (25), because the empirical average of nonnegative terms is zero whenever the expected hinge constraint is satisfied exactly. Finally, $R(t_M)=\mathcal{O}((\log M)^{1/\vartheta})$, and hence $\eta_M\to 0$. □

Lemma 4. *Let $\mathcal{P}'_{\epsilon,M}$ denote (23) with $t=t_M$, and let $\mathcal{P}'_{\epsilon,M}(\eta)$ denote (27). Fix $\epsilon>0$. For any sequence $\eta_M\ge 0$ with $\eta_M\to 0$ as $M\to\infty$, define*

$$\begin{aligned}
f_0(\boldsymbol{x}) &:= \sup_{\mathbb{Q}\in\mathcal{P}_\epsilon}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})],\\
f_M(\boldsymbol{x}) &:= \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})],\\
g_M(\boldsymbol{x}) &:= \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}(\eta_M)}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})].
\end{aligned}$$

Then

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \to 0, \qquad \sup_{\boldsymbol{x}\in\mathcal{X}} |g_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \to 0 \qquad \text{as } M \to \infty.$$

Proof of Lemma 4. For a probability measure $\mathbb{Q}$, define

$$H(\mathbb{Q}) \coloneqq \mathbb{E}_{\mathbb{Z}}\left[\left[\mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{\xi}})] - \epsilon\right]_+\right].$$

We first show that $f_M(\boldsymbol{x}) \to f_0(\boldsymbol{x})$ and $g_M(\boldsymbol{x}) \to f_0(\boldsymbol{x})$ for each fixed $\boldsymbol{x} \in \mathcal{X}$.

Since $\mathcal{P}'_{\epsilon,M} \subseteq \mathcal{P}_\epsilon$, we have $f_M(\boldsymbol{x}) \le f_0(\boldsymbol{x})$. To prove the reverse limit inequality, fix $\delta > 0$ and choose $\mathbb{Q}^\delta \in \mathcal{P}_\epsilon$ such that

$$\mathbb{E}_{\mathbb{Q}^\delta}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \ge f_0(\boldsymbol{x}) - \delta.$$

Let $\Pi_M(\boldsymbol{\xi}) = \boldsymbol{\xi}$ if $\|\boldsymbol{\xi}\| \le t_M$ and $\Pi_M(\boldsymbol{\xi}) = t_M\boldsymbol{\xi}/\|\boldsymbol{\xi}\|$ otherwise, and let $\widetilde{\mathbb{Q}}_M = \mathbb{Q}^\delta \circ \Pi_M^{-1}$ be the push-forward of $\mathbb{Q}^\delta$ under Π_M; equivalently, if $\tilde{\boldsymbol{\xi}} \sim \mathbb{Q}^\delta$, then $\Pi_M(\tilde{\boldsymbol{\xi}}) \sim \widetilde{\mathbb{Q}}_M$. Then $\widetilde{\mathbb{Q}}_M$ is supported on $\{\boldsymbol{\xi} \in \Xi : \|\boldsymbol{\xi}\| \le t_M\}$ and satisfies the second-moment constraint. Moreover, by Assumption (A), as $M \to \infty$,

$$\begin{aligned}
\Delta_M &\coloneqq \sup_{\boldsymbol{z}\in\mathcal{X}} \left|\mathbb{E}_{\widetilde{\mathbb{Q}}_M}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\mathbb{Q}^\delta}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]\right| \\
&= \sup_{\boldsymbol{z}\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{Q}^\delta}\left[\ell(\boldsymbol{z},\Pi_M(\tilde{\boldsymbol{\xi}})) - \ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})\right]\right| \\
&\le \mathbb{E}_{\mathbb{Q}^\delta}\left[(2c_1 + 2c_2\|\tilde{\boldsymbol{\xi}}\|)\mathbb{I}_{\{\|\tilde{\boldsymbol{\xi}}\| > t_M\}}\right] \to 0.
\end{aligned}$$

Since $\epsilon > 0$, define

$$\lambda_M \coloneqq \frac{\Delta_M}{\epsilon + \Delta_M}, \qquad \mathbb{Q}_M \coloneqq (1-\lambda_M)\widetilde{\mathbb{Q}}_M + \lambda_M\hat{\mathbb{P}}_N. \tag{37}$$

Because $\hat{\mathbb{P}}_N$ is the fixed empirical reference distribution, it has finite support. Hence, for all sufficiently large M, $\hat{\mathbb{P}}_N$ is supported on $\{\boldsymbol{\xi} \in \Xi : \|\boldsymbol{\xi}\| \le t_M\}$, and therefore $\mathbb{Q}_M(\{\boldsymbol{\xi} \in \Xi : \|\boldsymbol{\xi}\| \le t_M\}) = 1$. The second-moment constraint also holds, since

$$\mathbb{E}_{\mathbb{Q}_M}[\|\tilde{\boldsymbol{\xi}}\|^2] = (1-\lambda_M)\mathbb{E}_{\widetilde{\mathbb{Q}}_M}[\|\tilde{\boldsymbol{\xi}}\|^2] + \lambda_M\mathbb{E}_{\hat{\mathbb{P}}_N}[\|\tilde{\boldsymbol{\xi}}\|^2] \le \Omega,$$

where $\widetilde{\mathbb{Q}}_M$ satisfies the second-moment constraint by construction and $\hat{\mathbb{P}}_N$ satisfies it by Assumption (B). Since $\mathbb{Q}^\delta \in \mathcal{P}_\epsilon$, feasibility gives $H(\mathbb{Q}^\delta) \le 0$. Since $H(\mathbb{Q}^\delta)$ is the expectation of a nonnegative hinge term, $H(\mathbb{Q}^\delta) \ge 0$, and hence $H(\mathbb{Q}^\delta) = 0$. Therefore, for $\mathbb{Z}$-almost every realization $\boldsymbol{z} \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}_M}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \epsilon &= (1-\lambda_M)\left(\mathbb{E}_{\widetilde{\mathbb{Q}}_M}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]\right) - \epsilon \\
&\le (1-\lambda_M)\left(\mathbb{E}_{\mathbb{Q}^\delta}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] - \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] + \Delta_M\right) - \epsilon \\
&\le (1-\lambda_M)(\epsilon + \Delta_M) - \epsilon \\
&= (1-\lambda_M)\Delta_M - \lambda_M\epsilon = 0.
\end{aligned}$$

Here, the first equality uses the definition of $\mathbb{Q}_M$ in (37); the first inequality uses the definition of Δ_M; the second inequality uses $H(\mathbb{Q}^\delta) = 0$; and the last equality follows from the definition of λ_M in (37).

Thus $\mathbb{Q}_M \in \mathcal{P}'_{\epsilon,M}$ for all sufficiently large M. Because $\lambda_M \to 0$ and $\Delta_M \to 0$,

$$\liminf_{M\to\infty} f_M(\boldsymbol{x}) \geq \liminf_{M\to\infty} \mathbb{E}_{\mathbb{Q}_M}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{Q}^\delta}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \geq f_0(\boldsymbol{x}) - \delta.$$

Letting $\delta \downarrow 0$ gives $f_M(\boldsymbol{x}) \to f_0(\boldsymbol{x})$.

Next, consider g_M. The proof follows the same compactness argument as Lemma 3. The only difference is that the sets $\mathcal{P}'_{\epsilon,M}(\eta_M)$ also impose the moving tail constraint, but this constraint does not need to be inherited by the limit because the limiting set $\mathcal{P}_\epsilon$ has no tail constraint.

The lower bound $\liminf_M g_M(\boldsymbol{x}) \geq f_0(\boldsymbol{x})$ follows from $g_M(\boldsymbol{x}) \geq f_M(\boldsymbol{x})$ and the convergence of $f_M(\boldsymbol{x})$. For the upper bound, take any subsequence and choose $\mathbb{Q}_M \in \mathcal{P}'_{\epsilon,M}(\eta_M)$ such that $g_M(\boldsymbol{x}) - \frac{1}{M} \leq \mathbb{E}_{\mathbb{Q}_M}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$. The common second-moment bound gives tightness, so along a further subsequence $\mathbb{Q}_M \Rightarrow \mathbb{Q}^\star$. By uniform integrability and Assumption (A), the loss expectations converge for every fixed $\boldsymbol{z} \in \mathcal{X}$: $\mathbb{E}_{\mathbb{Q}_M}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})] \to \mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]$. The hinge terms are uniformly bounded by Assumptions (A) and (B); hence dominated convergence gives $H(\mathbb{Q}_M) \to H(\mathbb{Q}^\star)$. Since $H(\mathbb{Q}_M) \leq \eta_M$ and $\eta_M \to 0$, we have $H(\mathbb{Q}^\star) = 0$. Lower semicontinuity preserves the second-moment constraint, and hence $\mathbb{Q}^\star \in \mathcal{P}_\epsilon$. Therefore, $\limsup_{M\to\infty} g_M(\boldsymbol{x}) \leq \mathbb{E}_{\mathbb{Q}^\star}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \leq f_0(\boldsymbol{x})$. Together with the lower bound, this proves $g_M(\boldsymbol{x}) \to f_0(\boldsymbol{x})$.

The convergence is uniform over $\mathcal{X}$. Indeed, the Lipschitz estimate in Lemma 1 holds uniformly over all measures satisfying the second-moment constraint, and taking suprema over such measures preserves the same Lipschitz constant for f_0, f_M, and g_M. Since $\mathcal{X}$ is compact, a finite-net argument upgrades the pointwise convergence above to

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \to 0, \qquad \sup_{\boldsymbol{x}\in\mathcal{X}} |g_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \to 0.$$

Thus, the claim follows. □

Proof of Theorem 7. Let f_0, f_M, and g_M be as in Lemma 4, with η_M chosen as in Theorem 6. Let

$$h_M(\boldsymbol{x}) \coloneqq \sup_{\mathbb{Q}\in\mathcal{P}'^M_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$$

be the sampled objective. The value convergence follows from the same sandwich argument used in Theorem 4, with Lemma 4 replacing Lemma 3; the solution convergence then follows from the same separation argument. Indeed, on the high-probability event of Theorem 6, we have

$$f_M(\boldsymbol{x}) \leq h_M(\boldsymbol{x}) \leq g_M(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \mathcal{X}.$$

To obtain convergence in probability, fix $\rho > 0$ and $\upsilon > 0$, and apply Theorem 6 with confidence parameter υ. The corresponding sequence η_M still satisfies $\eta_M \to 0$. Hence Lemma 4 implies that, for all sufficiently large M,

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \le \rho \quad \text{and} \quad \sup_{\boldsymbol{x}\in\mathcal{X}} |g_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \le \rho.$$

On the containment event, the sandwich then gives $\sup_{\boldsymbol{x}\in\mathcal{X}} |h_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \le \rho$. Therefore,

$$\text{Prob}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} |h_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| > \rho\right) \le \upsilon$$

for all sufficiently large M. Since $\upsilon > 0$ is arbitrary,

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |h_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \xrightarrow{p} 0.$$

Consequently,

$$|\hat{v}'_M - \hat{v}| \le \sup_{\boldsymbol{x}\in\mathcal{X}} |h_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \xrightarrow{p} 0.$$

It remains to prove convergence of minimizers. For $\delta > 0$, define

$$\mathcal{A}_\delta \coloneqq \{\boldsymbol{x} \in \mathcal{X} : \text{dist}(\boldsymbol{x}, \mathcal{S}) \ge \delta\}.$$

If $\mathcal{A}_\delta$ is empty, the claim is trivial. Otherwise, the function f_0 is continuous and $\mathcal{X}$ is compact, so

$$\rho_\delta \coloneqq \inf_{\boldsymbol{x}\in\mathcal{A}_\delta} f_0(\boldsymbol{x}) - \hat{v} > 0.$$

Whenever $\sup_{\boldsymbol{x}\in\mathcal{X}} |h_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| < \rho_\delta/3$, no minimizer of h_M can belong to $\mathcal{A}_\delta$. Therefore

$$\text{Prob}\big(\text{dist}(\hat{\boldsymbol{x}}'_M, \mathcal{S}) \ge \delta\big) \le \text{Prob}\left(\sup_{\boldsymbol{x}\in\mathcal{X}} |h_M(\boldsymbol{x}) - f_0(\boldsymbol{x})| \ge \rho_\delta/3\right) \to 0.$$

This proves $\text{dist}(\hat{\boldsymbol{x}}'_M, \mathcal{S}) \xrightarrow{p} 0$.

Lastly, we prove the suboptimality bound. Let $\boldsymbol{x}^\star_M$ be the exact optimizer from the statement. Since $\hat{\boldsymbol{x}}'_M$ minimizes the sampled objective h_M and $\mathcal{P}'_{\epsilon,M} \subseteq \mathcal{P}'^M_\epsilon$, we have

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}'_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})] \\
&\le \sup_{\mathbb{Q}\in\mathcal{P}'^M_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\hat{\boldsymbol{x}}'_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})] \\
&\le \sup_{\mathbb{Q}\in\mathcal{P}'^M_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})].
\end{aligned}$$

Next, by Theorem 6, setting $\eta = \eta_M$, we obtain the high-probability bound

$$\sup_{\mathbb{Q}\in\mathcal{P}'^M_\epsilon} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})] \le \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}(\eta_M)} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})] - \sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}^\star_M, \tilde{\boldsymbol{\xi}})]$$

with probability at least $1-\tau$. By the same moment-duality argument as in Proposition 10, the relaxed unbounded-support problem admits the dual obtained by augmenting (24) with the slack multiplier $\gamma\eta_M$ and the domination constraint $\gamma\mathbb{Z}-\nu\in\mathcal{M}_+(\mathcal{X})$. Thus, the associated exact dual optimizer $(\boldsymbol{x}_M^\star,\alpha_M^\star,\beta_M^\star,\kappa_M^\star,\nu_M^\star)$ from the statement yields a feasible, though generally suboptimal, relaxed-dual solution $(\boldsymbol{x}_M^\star,\alpha_M^\star,\beta_M^\star,\kappa_M^\star,\gamma_M^\star,\nu_M^\star)$. We can therefore further upper bound the preceding display by

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}(\eta_M)}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}_M^\star,\tilde{\boldsymbol{\xi}})]-\sup_{\mathbb{Q}\in\mathcal{P}'_{\epsilon,M}}\mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{x}_M^\star,\tilde{\boldsymbol{\xi}})]\\
&\le\alpha_M^\star+\beta_M^\star\Omega-\kappa_M^\star(1-2/M)+\int_{\mathcal{X}}\left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]+\epsilon\right)\nu_M(\mathrm{d}\boldsymbol{z})+\gamma_M^\star\eta_M\\
&\qquad-\left(\alpha_M^\star+\beta_M^\star\Omega-\kappa_M^\star(1-2/M)+\int_{\mathcal{X}}\left(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z},\tilde{\boldsymbol{\xi}})]+\epsilon\right)\nu_M(\mathrm{d}\boldsymbol{z})\right)\\
&=\gamma_M^\star\eta_M.
\end{aligned}$$

Substituting the explicit expression for η_M proves the theorem. □

C.2. Conic Programming Reformulations

Proof of Theorem 8. For each sampled point $\boldsymbol{z}_m$, define

$$g_m(\boldsymbol{\xi})\coloneqq\max_{k\in[J]}\left(\boldsymbol{a}_k(\boldsymbol{z}_m)^\top\boldsymbol{\xi}+b_k(\boldsymbol{z}_m)\right).$$

The semi-infinite constraint in (26) can be equivalently rewritten as

$$\ell(\boldsymbol{x},\boldsymbol{\xi})+\kappa\le\alpha+\beta\|\boldsymbol{\xi}\|^2+\frac{1}{M}\sum_{m\in[M]}\nu_m\ell(\boldsymbol{z}_m,\boldsymbol{\xi})\qquad\forall\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}:\|\boldsymbol{\xi}\|\le t, \tag{38}$$

$$\ell(\boldsymbol{x},\boldsymbol{\xi})\le\alpha+\beta\|\boldsymbol{\xi}\|^2+\frac{1}{M}\sum_{m\in[M]}\nu_m\ell(\boldsymbol{z}_m,\boldsymbol{\xi})\qquad\forall\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}:\|\boldsymbol{\xi}\|>t. \tag{39}$$

Since $\kappa\ge0$, (38) implies

$$\ell(\boldsymbol{x},\boldsymbol{\xi})\le\alpha+\beta\|\boldsymbol{\xi}\|^2+\frac{1}{M}\sum_{m\in[M]}\nu_m\ell(\boldsymbol{z}_m,\boldsymbol{\xi})\qquad\forall\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}:\|\boldsymbol{\xi}\|\le t.$$

Hence (39) can be simplified to

$$\ell(\boldsymbol{x},\boldsymbol{\xi})\le\alpha+\beta\|\boldsymbol{\xi}\|^2+\frac{1}{M}\sum_{m\in[M]}\nu_m\ell(\boldsymbol{z}_m,\boldsymbol{\xi})\qquad\forall\boldsymbol{\xi}\in\mathbb{R}^{D_{\boldsymbol{\xi}}}.$$

We now reformulate these two semi-infinite systems separately. Assume first that $\beta>0$; the case $\beta=0$ will be treated separately after deriving the conic reformulations for the global and ball constraints.

For the global constraint, the same simplex linearization as in Theorem 5 shows that it is equivalent to the existence of multipliers $\boldsymbol{\lambda}_{jm} \in \mathbb{R}^J_+$ with $\mathbf{e}^\top \boldsymbol{\lambda}_{jm} = \nu_m$ such that, for every $j \in [J]$,

$$\sup_{\boldsymbol{\xi} \in \mathbb{R}^{D_{\boldsymbol{\xi}}}} \left\{ \left(\boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda^k_{jm} \boldsymbol{a}_k(\boldsymbol{z}_m) \right)^\top \boldsymbol{\xi} - \beta \|\boldsymbol{\xi}\|^2 \right\} + b_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda^k_{jm} b_k(\boldsymbol{z}_m) \le \alpha.$$

Under this assumption, the quadratic supremum above is equal to $\|\boldsymbol{c}_j\|^2/(4\beta)$, where

$$\boldsymbol{c}_j \coloneqq \boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda^k_{jm} \boldsymbol{a}_k(\boldsymbol{z}_m).$$

Therefore the global constraint is equivalent to the existence of $\zeta_j \in \mathbb{R}_+$ satisfying

$$\zeta_j + b_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda^k_{jm} b_k(\boldsymbol{z}_m) \le \alpha$$

and

$$\left\| \begin{bmatrix} \boldsymbol{c}_j \\ \zeta_j - \beta \end{bmatrix} \right\| \le \zeta_j + \beta.$$

For the ball constraint (38), introduce simplex multipliers $\boldsymbol{\lambda}'_{jm} \in \mathbb{R}^J_+$ with $\mathbf{e}^\top \boldsymbol{\lambda}'_{jm} = \nu_m$. Then the constraint is equivalent to

$$\sup_{\|\boldsymbol{\xi}\| \le t} \left\{ \left(\boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda'^k_{jm} \boldsymbol{a}_k(\boldsymbol{z}_m) \right)^\top \boldsymbol{\xi} - \beta \|\boldsymbol{\xi}\|^2 \right\} + b_j(\boldsymbol{x}) + \kappa - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda'^k_{jm} b_k(\boldsymbol{z}_m) \le \alpha$$

for every $j \in [J]$. Let

$$\boldsymbol{c}'_j \coloneqq \boldsymbol{a}_j(\boldsymbol{x}) - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda'^k_{jm} \boldsymbol{a}_k(\boldsymbol{z}_m).$$

Using Fenchel duality (Rockafellar 1970, Theorem 16.4), with the support function of the Euclidean ball $\{\boldsymbol{\xi} : \|\boldsymbol{\xi}\| \le t\}$, whose support function is $t\|\cdot\|$, we obtain

$$\sup_{\|\boldsymbol{\xi}\| \le t} \left\{ \boldsymbol{c}'^\top_j \boldsymbol{\xi} - \beta \|\boldsymbol{\xi}\|^2 \right\} = \inf_{\boldsymbol{\theta}_j \in \mathbb{R}^{D_{\boldsymbol{\xi}}}} \left\{ t\|\boldsymbol{\theta}_j\| + \sup_{\boldsymbol{\xi} \in \mathbb{R}^{D_{\boldsymbol{\xi}}}} \left[(\boldsymbol{c}'_j - \boldsymbol{\theta}_j)^\top \boldsymbol{\xi} - \beta \|\boldsymbol{\xi}\|^2 \right] \right\}.$$

Since $\beta > 0$, the remaining unconstrained quadratic supremum equals $\|\boldsymbol{c}'_j - \boldsymbol{\theta}_j\|^2/(4\beta)$. Hence the ball constraint is equivalent to the existence of $(\boldsymbol{\theta}_j, \zeta'_j) \in \mathbb{R}^{D_{\boldsymbol{\xi}}} \times \mathbb{R}_+$ such that

$$\zeta'_j + t\|\boldsymbol{\theta}_j\| + b_j(\boldsymbol{x}) + \kappa - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda'^k_{jm} b_k(\boldsymbol{z}_m) \le \alpha$$

and

$$\left\| \begin{bmatrix} \boldsymbol{c}'_j - \boldsymbol{\theta}_j \\ \zeta'_j - \beta \end{bmatrix} \right\| \le \zeta'_j + \beta.$$

The case $\beta = 0$ is handled exactly as in Theorem 5. For the global constraint, finiteness of the supremum over $\mathbb{R}^{D_{\boldsymbol{\xi}}}$ requires $\boldsymbol{c}_j = \mathbf{0}$, and the conic constraint reduces to

$$\left\| \begin{bmatrix} \boldsymbol{c}_j \\ \zeta_j \end{bmatrix} \right\| \leq \zeta_j,$$

so the reduced conic constraint imposes no positive lower bound on ζ_j beyond $\zeta_j \geq 0$. Since ζ_j enters the scalar inequality for the global constraint with coefficient one, the least restrictive feasible choice is $\zeta_j = 0$. For the ball constraint, the conic constraint reduces to

$$\left\| \begin{bmatrix} \boldsymbol{c}'_j - \boldsymbol{\theta}_j \\ \zeta'_j \end{bmatrix} \right\| \leq \zeta'_j,$$

which forces $\boldsymbol{\theta}_j = \boldsymbol{c}'_j$ and imposes no positive lower bound on ζ'_j. Since ζ'_j enters the scalar inequality for the ball constraint with coefficient one, the least restrictive feasible choice is $\zeta'_j = 0$, and the scalar inequality becomes precisely the support-function condition for the Euclidean ball. Collecting the global and ball constraints yields exactly the finite conic reformulation stated in the theorem.

The displayed reformulation is convex. The objective is affine, while the sign, simplex, and equality constraints are affine. Since $\boldsymbol{a}_j(\boldsymbol{x})$ and $b_j(\boldsymbol{x})$ are affine in $\boldsymbol{x}$, the scalar global constraint is affine. The scalar ball constraint has the form

$$\zeta'_j + t\|\boldsymbol{\theta}_j\| + b_j(\boldsymbol{x}) + \kappa - \frac{1}{M} \sum_{m \in [M], k \in [J]} \lambda'^{k}_{jm} b_k(\boldsymbol{z}_m) \leq \alpha,$$

which is convex because $t \geq 0$ and the Euclidean norm is convex. The two norm inequalities are standard second-order-cone constraints with affine arguments in the decision variables, and are therefore convex. Together with the convexity of the constraint $\boldsymbol{x} \in \mathcal{X}$, this shows that the finite reformulation is a convex conic program. The final tractability claim follows as in Theorem 5, since the Euclidean norm is directly second-order cone representable. □

Appendix D: Tractable approximations for two-stage linear loss

In this section, we develop a tractable approximation to the two-stage DRO problem via linear decision rules.

THEOREM 9. *The problem* (19) *under the two-stage loss function in* (6) *can be conservatively approximated by the following conic program:*

$$
\begin{array}{ll}
\min & \alpha+\beta\Omega+\dfrac{1}{M}\displaystyle\sum_{m\in[M]}\nu_m(\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(\boldsymbol{z}_m,\tilde{\boldsymbol{\xi}})]+\epsilon)\\
\text{s.t.} & \boldsymbol{x}\in\mathcal{X},\ \alpha\in\mathbb{R},\ \beta\in\mathbb{R}_+,\ \boldsymbol{\nu}\in\mathbb{R}_+^M\\
& \boldsymbol{\lambda}_m\in\mathbb{R}_+^L\ \ \forall m\in[M],\ \boldsymbol{\theta}\in\mathbb{R}^{D_{\boldsymbol{\xi}}},\zeta\in\mathbb{R}_+\\
& \boldsymbol{Y}\in\mathbb{R}^{D_{\boldsymbol{y}}\times D_{\boldsymbol{\xi}}},\ \boldsymbol{y}_0\in\mathbb{R}^{D_{\boldsymbol{y}}}\\
& \sigma_\Xi\left([\boldsymbol{T}(\boldsymbol{x})-\boldsymbol{W}\boldsymbol{Y}]_{l:}^\top\right)\le[\boldsymbol{W}\boldsymbol{y}_0-\boldsymbol{h}(\boldsymbol{x})]_l\qquad\forall l\in[L]\\
& \zeta+\sigma_\Xi(\boldsymbol{\theta})+\boldsymbol{c}^\top\boldsymbol{x}+\boldsymbol{q}^\top\boldsymbol{y}_0-\dfrac{1}{M}\displaystyle\sum_{m\in[M]}\left(\nu_m\boldsymbol{c}^\top\boldsymbol{z}_m+\boldsymbol{h}(\boldsymbol{z}_m)^\top\boldsymbol{\lambda}_m\right)\le\alpha\\
& \left\|\begin{bmatrix}\boldsymbol{Y}^\top\boldsymbol{q}-\frac{1}{M}\sum_{m\in[M]}\boldsymbol{T}(\boldsymbol{z}_m)^\top\boldsymbol{\lambda}_m-\boldsymbol{\theta}\\ \zeta-\beta\end{bmatrix}\right\|\le\zeta+\beta\\
& \boldsymbol{W}^\top\boldsymbol{\lambda}_m=\nu_m\boldsymbol{q}\qquad\forall m\in[M].
\end{array}
$$

Proof Consider the semi-infinite constraint in (19):

$$
\sup_{\boldsymbol{\xi}\in\Xi}\left[\ell(\boldsymbol{x},\boldsymbol{\xi})-\beta\|\boldsymbol{\xi}\|^2-\frac{1}{M}\sum_{m\in[M]}\nu_m\ell(\boldsymbol{z}_m,\boldsymbol{\xi})\right]\le\alpha. \tag{40}
$$

We approximate the recourse decision in the first term by the linear decision rule

$$
\boldsymbol{y}(\boldsymbol{\xi})=\boldsymbol{Y}\boldsymbol{\xi}+\boldsymbol{y}_0,
$$

where $\boldsymbol{Y}\in\mathbb{R}^{D_{\boldsymbol{y}}\times D_{\boldsymbol{\xi}}}$ and $\boldsymbol{y}_0\in\mathbb{R}^{D_{\boldsymbol{y}}}$ satisfy

$$
\boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi}+\boldsymbol{h}(\boldsymbol{x})\le\boldsymbol{W}(\boldsymbol{Y}\boldsymbol{\xi}+\boldsymbol{y}_0)\qquad\forall\boldsymbol{\xi}\in\Xi.
$$

This robust linear constraint is equivalent, row by row, to

$$
\sup_{\boldsymbol{\xi}\in\Xi}[\boldsymbol{T}(\boldsymbol{x})-\boldsymbol{W}\boldsymbol{Y}]_{l:}\boldsymbol{\xi}\le[\boldsymbol{W}\boldsymbol{y}_0-\boldsymbol{h}(\boldsymbol{x})]_l\qquad\forall l\in[L],
$$

or, equivalently,

$$
\sigma_\Xi\left([\boldsymbol{T}(\boldsymbol{x})-\boldsymbol{W}\boldsymbol{Y}]_{l:}^\top\right)\le[\boldsymbol{W}\boldsymbol{y}_0-\boldsymbol{h}(\boldsymbol{x})]_l\qquad\forall l\in[L].
$$

Then, the two-stage loss function in (6) can be upper bounded by

$$
\ell(\boldsymbol{x},\boldsymbol{\xi})\le\boldsymbol{c}^\top\boldsymbol{x}+\boldsymbol{q}^\top(\boldsymbol{Y}\boldsymbol{\xi}+\boldsymbol{y}_0)\qquad\forall\boldsymbol{\xi}\in\Xi.
$$

Replacing only $\ell(\boldsymbol{x},\boldsymbol{\xi})$ by this upper bound, while keeping each sampled term $\ell(\boldsymbol{z}_m,\boldsymbol{\xi})$ exact, yields a conservative approximation to (40):

$$
\left.\begin{array}{ll}
\max & \boldsymbol{c}^\top\boldsymbol{x}+\boldsymbol{q}^\top(\boldsymbol{Y}\boldsymbol{\xi}+\boldsymbol{y}_0)-\beta\|\boldsymbol{\xi}\|^2-\dfrac{1}{M}\displaystyle\sum_{m\in[M]}\nu_m(\boldsymbol{c}^\top\boldsymbol{z}_m+\boldsymbol{q}^\top\boldsymbol{y}_m)\\
\text{s.t.} & \boldsymbol{\xi}\in\Xi,\ \boldsymbol{y}_m\in\mathbb{R}^{D_{\boldsymbol{y}}}\quad\forall m\in[M]\\
& \boldsymbol{T}(\boldsymbol{z}_m)\boldsymbol{\xi}+\boldsymbol{h}(\boldsymbol{z}_m)\le\boldsymbol{W}\boldsymbol{y}_m\quad\forall m\in[M]
\end{array}\right\}\le\alpha.
$$

Next, dualizing the recourse variables $\{\boldsymbol{y}_m\}_{m\in[M]}$ and applying the support-function reformulation to the remaining maximization over $\boldsymbol{\xi}$ yields the equivalent conic conditions: there exist $\boldsymbol{\lambda}_m \in \mathbb{R}^L_+ \; \forall m \in [M]$, $\boldsymbol{\theta} \in \mathbb{R}^{D_{\boldsymbol{\xi}}}$, $\zeta \in \mathbb{R}_+$, such that

$$\begin{aligned}
&\zeta + \sigma_\Xi(\boldsymbol{\theta}) + \boldsymbol{c}^\top \boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{y}_0 - \frac{1}{M}\sum_{m\in[M]} \left(\nu_m \boldsymbol{c}^\top \boldsymbol{z}_m + \boldsymbol{h}(\boldsymbol{z}_m)^\top \boldsymbol{\lambda}_m\right) \leq \alpha \\
&\left\| \begin{bmatrix} \boldsymbol{Y}^\top \boldsymbol{q} - \frac{1}{M}\sum_{m\in[M]} \boldsymbol{T}(\boldsymbol{z}_m)^\top \boldsymbol{\lambda}_m - \boldsymbol{\theta} \\ \zeta - \beta \end{bmatrix} \right\| \leq \zeta + \beta \\
&\boldsymbol{W}^\top \boldsymbol{\lambda}_m = \nu_m \boldsymbol{q} \qquad \forall m \in [M].
\end{aligned}$$

This proves the result. □

Remark 11 (Two-stage approximation alternatives). Two-stage (distributionally) robust optimization is generically NP-hard, which motivates the approximation in Theorem 9. If desired, tighter approximations via piecewise-linear decision rules (Georghiou et al. 2015, Ben-Tal et al. 2020) or quadratic decision rules (Fan and Hanasusanto 2024, Xu and Hanasusanto 2025) can also be developed. One could also, in principle, formulate a convex copositive program for the two-stage problem via the techniques of Hanasusanto and Kuhn (2018), Xu and Burer (2018). The resulting conic program admits tractable conservative semidefinite programming approximations that can be solved with standard off-the-shelf solvers.

Appendix E: Details of experiments

E.1. Numerical illustration of the Monte Carlo approximation

This experiment corresponds to Remark 9. The purpose is to isolate the sampling error in the finite-sample dual approximation, so we set $\epsilon = 0$ and evaluate the sampled-dual solution against the empirical objective $E_{\widehat{P}_N}[\ell(\boldsymbol{x}, \tilde{\xi})]$. The symmetric construction below gives a known optimizer of this benchmark, avoiding the need for a large-reference Monte Carlo solution.

For each dimension $d \in \{5, 25, 100\}$, we take $D_{\boldsymbol{x}} = D_{\boldsymbol{\xi}} = d$ and set both the decision set and uncertainty support to the Euclidean unit ball,

$$\mathcal{X} = \Xi = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| \leq 1\}.$$

The second-moment cap is set to the empirical training second moment, $\Omega = N^{-1}\sum_{i=1}^N \|\hat{\boldsymbol{\xi}}_i\|^2$. The known optimizer is chosen as $\boldsymbol{x}_\star = 0.1\mathbf{e}$, which belongs to $\mathcal{X}$ for all three dimensions. The empirical reference distribution contains $N = 100$ points and is constructed symmetrically: we draw 50 independent points uniformly from the unit ball and add their negatives. This symmetry makes $\boldsymbol{x}_\star$ an optimizer of the empirical objective for the shifted loss below.

The loss is a shifted weighted projection max-affine loss,

$$\ell(\boldsymbol{x},\boldsymbol{\xi}) = \max_{r=1,\dots,K} c_r \left|\boldsymbol{q}_r^\top((\boldsymbol{x}-\boldsymbol{x}_\star)-\boldsymbol{\xi})\right|,$$

with $K=4$. For each dimension, the directions $\boldsymbol{q}_r \in \mathbb{R}^d$ are generated as independent dense Gaussian vectors and normalized to unit length. The weights c_r are linearly spaced between 0.5 and 1.5 and then rescaled to have average one. Equivalently, the loss has $J=2K=8$ affine pieces with slopes $\{\pm c_r \boldsymbol{q}_r : r=1,\dots,K\}$, so the conic reformulation in Section 3 can be solved directly.

For each dimension, the experiment uses 50 independent replications. In each replication, it draws a pool of 500 auxiliary decisions uniformly from $\mathcal{X}$ and uses nested prefixes of this pool for $M \in \{2,5,10,20,50,100,250,500\}$. For each M, the sampled dual (19) is solved using the compact-support conic reformulation with $\sigma_\Xi(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$. As in Remark 9, the reported percentage gap is

$$\frac{E_{\widehat{P}_N}[\ell(\widehat{x}_M,\tilde{\xi})] - E_{\widehat{P}_N}[\ell(x_\star,\tilde{\xi})]}{E_{\widehat{P}_N}[\ell(x_\star,\tilde{\xi})]} \times 100\%.$$

The plotted curves report the mean percentage gap over the 10 replications for each dimension, the shaded bands show the interquartile range, and the dashed reference line is a pooled log-scale fit with fixed slope $-1/2$, i.e., $C/\sqrt{M}$.

E.2. Newsvendor

The newsvendor figures in Section 5.1 are generated by the final selected-trial plotting blocks in the four three-dimensional notebooks. Earlier exploratory cells use different parameter values, but those preliminary settings are not used for the figure files included in the manuscript.

The experiment studies a three-product newsvendor with order vector $\boldsymbol{x} \in \mathbb{R}^3_+$, demand vector $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^3_+$, and loss

$$\ell(\boldsymbol{x},\boldsymbol{\xi}) = \sum_{j=1}^{3} \left[h_j(x_j-\xi_j)_+ + b_j(\xi_j-x_j)_+\right].$$

Across all four final plotting blocks, the common decision-side parameters are

$$h=(0.1,0.2,0.3), \qquad b=(1,1,1), \qquad \rho=0.05, \qquad U=(50,50,50).$$

The demand coordinates are generated independently with mean vector $\mu=(5,6,6)$ and standard-deviation vector $\sigma=(5,6,8)$. The in-sample data are split into training and validation subsets using the ratio 0.8/0.2. We use $N \in \{10,20,30,50,75,100,150\}$ with $T=50$ independent trials for each sample size, and each trial is evaluated on an independent OOS sample of size $N_{\mathrm{OOS}}=10000$.

For both 2-WDRO and TIPM-DRO, the ambiguity radius is tuned over the common grid

$$\varepsilon \in \{a \times 10^b : a \in \{1,3,5,7,9\},\, b \in \{-3,-2,-1,0,1,2\}\}.$$

For IPM-DRO, the expectation over the auxiliary distribution is approximated by Monte Carlo sampling. The final plotting blocks use $M = 800$ sampled decision points drawn uniformly from $[0,50]^3$, and the sampled set is augmented with the SAA solution from the same trial. The second-moment cap is set to the empirical training second moment, $\Omega = \frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \|\hat{\boldsymbol{\xi}}_i\|^2$.

To state the dual reformulation used in the code, let $\mathcal{B}_{\mathrm{nv}} \coloneqq \{0,1\}^3$ denote the eight affine pieces of the separable newsvendor loss, and let $\mathcal{R}_{\mathrm{nv}} \coloneqq \mathcal{B}_{\mathrm{nv}} \cup \{T\}$ include the additional CVaR epigraph piece. For each $s \in \mathcal{B}_{\mathrm{nv}}$, define

$$p_j^s = \begin{cases} h_j, & s_j = 0, \\ -b_j, & s_j = 1, \end{cases} \qquad g_j^s = \begin{cases} -h_j, & s_j = 0, \\ b_j, & s_j = 1, \end{cases} \qquad j = 1,2,3.$$

For the epigraph piece, set $\boldsymbol{p}^T = \mathbf{0}$, $\boldsymbol{g}^T = \mathbf{0}$, and $q_T = 1$, while $q_r = 0$ for $r \in \mathcal{B}_{\mathrm{nv}}$. Then

$$\ell(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{s \in \mathcal{B}_{\mathrm{nv}}} \Big\{ (\boldsymbol{p}^s)^\top \boldsymbol{x} + (\boldsymbol{g}^s)^\top \boldsymbol{\xi} \Big\}.$$

If $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M \in [0,50]^3$ are the sampled decision points and

$$\widehat{L}_m \coloneqq \frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \ell(\boldsymbol{z}_m, \hat{\boldsymbol{\xi}}_i), \qquad m \in [M],$$

then the sampled second-moment-constrained TIPM-DRO problem solved by the implementation is

$$\begin{aligned}
\min\ & \left(1 - \frac{1}{\rho}\right) t + \frac{1}{\rho} \left(\alpha + \frac{1}{M} \sum_{m \in [M]} \beta_m (\widehat{L}_m + \epsilon) + \gamma \Omega \right) \\
\text{s.t.}\ & \boldsymbol{x} \in \mathbb{R}^3_+,\ \boldsymbol{x} \le U,\ t, \alpha \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^M_+,\ \gamma \in \mathbb{R}_+, \\
& \eta_r \in \mathbb{R}_+,\ \boldsymbol{u}_r \in \mathbb{R}^3,\ \boldsymbol{u}_r \le \mathbf{0},\ \lambda_{rms} \in \mathbb{R}_+ \qquad \forall r \in \mathcal{R}_{\mathrm{nv}}, \forall m \in [M], \forall s \in \mathcal{B}_{\mathrm{nv}}, \\
& \eta_r + (\boldsymbol{p}^r)^\top \boldsymbol{x} + q_r t - \frac{1}{M} \sum_{m \in [M]} \sum_{s \in \mathcal{B}_{\mathrm{nv}}} \lambda_{rms} (\boldsymbol{p}^s)^\top \boldsymbol{z}_m \le \alpha \qquad \forall r \in \mathcal{R}_{\mathrm{nv}}, \\
& \sum_{s \in \mathcal{B}_{\mathrm{nv}}} \lambda_{rms} = \beta_m \qquad \forall r \in \mathcal{R}_{\mathrm{nv}}, \forall m \in [M], \\
& \left\| \begin{bmatrix} \boldsymbol{g}^r - \frac{1}{M} \sum_{m \in [M]} \sum_{s \in \mathcal{B}_{\mathrm{nv}}} \lambda_{rms} \boldsymbol{g}^s - \boldsymbol{u}_r \\ \eta_r - \gamma \end{bmatrix} \right\| \le \eta_r + \gamma \qquad \forall r \in \mathcal{R}_{\mathrm{nv}}.
\end{aligned} \tag{41}$$

This is exactly the explicit three-dimensional dual/SOCP reformulation implemented in the experiment and used to generate the TIPM-DRO curves in Section 5.1.

For the four demand models, the Gaussian benchmark samples each coordinate from a Normal distribution with the target mean and variance and rejects negative draws. The rescaled χ^2 case

uses $\xi_j = s_j Y_j$, where $Y_j \sim \chi^2_{\nu_j}$, $\nu_j = 2(\mu_j/\sigma_j)^2$, and $s_j = \sigma_j^2/(2\mu_j)$. The Lognormal case uses $\xi_j \sim \mathrm{LogNormal}(\mu_{\log,j}, \sigma_{\log,j})$, where $\sigma^2_{\log,j} = \log(1+\sigma_j^2/\mu_j^2)$ and $\mu_{\log,j} = \log(\mu_j) - \frac{1}{2}\sigma^2_{\log,j}$. The Pareto case uses $\xi_j = x_{m,j}(1+Y_j)$, where $Y_j \sim \mathrm{Pareto}(\alpha_j)$, $\alpha_j = 1+\sqrt{1+(\mu_j/\sigma_j)^2}$, and $x_{m,j} = \mu_j(\alpha_j-1)/\alpha_j$.

E.3. Outlier-Corrupted Regression

The figures in Section 5.2 are generated by the final plotting blocks in our implementation, which call a common worker routine for each trial. In each trial, the code draws $\boldsymbol{x}^\star$ uniformly from the unit sphere in $\mathbb{R}^d$, generates clean covariates $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and responses $v_i = \boldsymbol{x}^{\star\top}\boldsymbol{u}_i$, and evaluates the fitted estimator on an independent clean test sample of size N_{OOS}. It then corrupts $\lfloor \omega N \rfloor$ training samples via $\boldsymbol{u}_i \leftarrow C\boldsymbol{u}_i$ and $v_i \leftarrow -C^2 v_i + \rho$, randomly shuffles the sample, and splits it into training and validation subsets with ratio 0.8/0.2. Hyperparameters are selected by a trimmed validation loss that removes the largest $\lceil \omega_{\mathrm{tv}} N_{\mathrm{val}} \rceil$ absolute residuals.

For the dimension sweep underlying Figure 3, the final plotting block uses $N \in \{10, 20, 50, 75, 100\}$ and $D_{\boldsymbol{\xi}} \in \{1, 5, 10, 50\}$ with $T = 50$, $N_{\mathrm{OOS}} = 1000$, $\omega = \omega_{\mathrm{tv}} = 0.2$, $C = 10$, $\rho = 0.1$, and $\delta = 0.1$. The OR-WDRO scale parameter is set to $\omega_{\mathrm{tv}} = 0.2$, and Std WDRO and OR-WDRO tune the Wasserstein radius over $\rho \in \{0.01, 0.1, 0.2, 0.5, 1.0\}$. The worker also computes the plain LAD estimator, but the final plotting cell reports only Std WDRO, OR-WDRO(2ω), OR-WDRO(ω), and MoM TIPM-DRO.

For MoM TIPM-DRO, the code uses $M = 800$ Monte Carlo test functions. The decision set radius is $R = \sqrt{D_{\boldsymbol{\xi}}}$, and the constraint $\|\boldsymbol{\theta}\| \le R$ is enforced with the same value of R. Let $N_{\mathrm{tr}} = \lfloor 0.8N \rfloor$. The block count is chosen from $K \in \{K_{\mathrm{theory}}, \min(\lfloor N_{\mathrm{tr}}/3 \rfloor, 2K_{\mathrm{theory}}), \lfloor N_{\mathrm{tr}}/2 \rfloor\} \cap \{1, \ldots, N_{\mathrm{tr}}\}$, where $K_{\mathrm{theory}} = \left\lceil \frac{4(1+2\omega)}{(1-2\omega)^2} \log \frac{1}{\delta} \right\rceil$. The IPM radius is tuned over $\epsilon_{\mathrm{IPM}} \in \{\epsilon_{\mathrm{theory}} : s \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1, 5, 10\}\}$ with $\epsilon_{\mathrm{theory}} = \frac{4\sqrt{e}\,\sigma_{\mathrm{loss}}\Gamma_\omega}{\sqrt{N_{\mathrm{tr}}}}$ and $\Gamma_\omega = (1-2\omega_{\mathrm{tv}})^{-1}$. The second-moment cap is set to $\Omega = 2\,\Omega_{\mathrm{tr}}$, where $\Omega_{\mathrm{tr}} = N_{\mathrm{tr}}^{-1}\sum_{i=1}^{N_{\mathrm{tr}}} \|(\boldsymbol{u}_i, v_i)\|^2$. For each candidate R, the same set of sampled test functions is reused across all $(K, \epsilon_{\mathrm{IPM}})$ candidates in that trial to stabilize validation comparisons.

For scalar-response MAD regression, with decision variable $\boldsymbol{\theta} \in \mathbb{R}^{D_{\boldsymbol{\xi}}}$, the loss admits the piecewise-affine representation

$$\ell(\boldsymbol{\theta}, \boldsymbol{\xi}) = \max\left\{ \begin{bmatrix} \boldsymbol{\theta} \\ -1 \end{bmatrix}^\top \boldsymbol{\xi}, \begin{bmatrix} -\boldsymbol{\theta} \\ 1 \end{bmatrix}^\top \boldsymbol{\xi} \right\}, \qquad \boldsymbol{\xi} = (\boldsymbol{u}, v) \in \mathbb{R}^{D_{\boldsymbol{\xi}}+1}.$$

Let $\{\boldsymbol{z}_m\}_{m=1}^M \subset \mathbb{R}^{D_{\boldsymbol{\xi}}}$ be the sampled test functions and let $\boldsymbol{Z} \in \mathbb{R}^{M \times D_{\boldsymbol{\xi}}}$ collect the row vectors $\boldsymbol{z}_m^\top$. Writing $\widehat{\mu}_{\mathrm{MoM}}(\boldsymbol{z}_m)$ for the MoM reference from Section 2.2, the implementation solves the conic program

$$
\begin{aligned}
&\min \alpha + \beta\Omega + \frac{1}{M}\sum_{m\in[M]} \nu_m\big(\widehat{\mu}_{\mathrm{MoM}}(\boldsymbol{z}_m) + \epsilon\big) \\
&\text{s.t. } \boldsymbol{\theta} \in \mathbb{R}^{D_{\boldsymbol{\xi}}},\ \|\boldsymbol{\theta}\| \le R,\ \alpha \in \mathbb{R},\ \beta \in \mathbb{R}_+,\ \boldsymbol{\nu} \in \mathbb{R}_+^M, \\
&\qquad \boldsymbol{\lambda}_1^+, \boldsymbol{\lambda}_1^-, \boldsymbol{\lambda}_2^+, \boldsymbol{\lambda}_2^- \in \mathbb{R}_+^M,\ \zeta_1, \zeta_2 \in \mathbb{R}_+, \\
&\qquad \boldsymbol{\lambda}_1^+ + \boldsymbol{\lambda}_1^- = \boldsymbol{\nu}, \qquad \boldsymbol{\lambda}_2^+ + \boldsymbol{\lambda}_2^- = \boldsymbol{\nu}, \\
&\qquad \zeta_1 \le \alpha, \qquad \zeta_2 \le \alpha, \\
&\qquad \left\| \begin{bmatrix} \boldsymbol{\theta} - \frac{1}{M}\boldsymbol{Z}^\top(\boldsymbol{\lambda}_1^+ - \boldsymbol{\lambda}_1^-) \\ -1 + \frac{1}{M}\mathbf{e}^\top(\boldsymbol{\lambda}_1^+ - \boldsymbol{\lambda}_1^-) \\ \zeta_1 - \beta \end{bmatrix} \right\| \le \zeta_1 + \beta, \\
&\qquad \left\| \begin{bmatrix} -\boldsymbol{\theta} - \frac{1}{M}\boldsymbol{Z}^\top(\boldsymbol{\lambda}_2^+ - \boldsymbol{\lambda}_2^-) \\ 1 + \frac{1}{M}\mathbf{e}^\top(\boldsymbol{\lambda}_2^+ - \boldsymbol{\lambda}_2^-) \\ \zeta_2 - \beta \end{bmatrix} \right\| \le \zeta_2 + \beta.
\end{aligned}
\tag{42}
$$

This is exactly the formulation implemented by the MoM TIPM-DRO solver used in Section 5.2, written in the notation of Section 2.

For Figure 4, the final plotting block fixes $N = 50$, $D_{\boldsymbol{\xi}} = 10$, and $C \in \{5, 6, \ldots, 20\}$ while keeping $T = 50$, $N_{\mathrm{OOS}} = 1000$, $\omega = \omega_{\mathrm{tv}} = 0.2$, $\rho = 0.1$, $\delta = 0.1$, and $M = 800$. In this block, the OR-WDRO scale parameter is fixed at $\omega_{\mathrm{tv}} = 0.2$, and the Wasserstein radius grid is expanded to $\rho \in \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0\}$. The plotting cell for Figure 4 reports only OR-WDRO(2ω), OR-WDRO(ω), and MoM TIPM-DRO. Apart from the fixed dimension and the sweep over C, these are the only substantive differences relative to the dimension-sweep experiments.