On the Trade-Off Between Distributional Belief and Ambiguity: Conservatism, Finite-Sample Guarantees, and Asymptotic Properties

Man Yiu Tsang, Karmel S. Shehadeh*

Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA

Abstract

We propose and analyze a new data-driven trade-off (TRO) approach for modeling uncertainty that serves as a middle ground between the optimistic approach, which adopts a distributional belief, and the pessimistic distributionally robust optimization approach, which hedges against distributional ambiguity. We equip the TRO model with a TRO ambiguity set characterized by a size parameter controlling the level of optimism and a shape parameter representing distributional ambiguity. We first show that constructing the TRO ambiguity set using a general star-shaped shape parameter with the empirical distribution as its star center is necessary and sufficient to guarantee the hierarchical structure of the sequence of TRO ambiguity sets. Then, we analyze the properties of the TRO model, including quantifying conservatism, quantifying bias and generalization error, and establishing asymptotic properties. Specifically, we show that the TRO model could generate a spectrum of decisions, ranging from optimistic to conservative decisions. Additionally, we show that it could produce an unbiased estimator of the true optimal value. Furthermore, we establish the almost-sure convergence of the optimal value and the set of optimal solutions of the TRO model to their true counterparts. We exemplify our theoretical results using an inventory control problem and a portfolio optimization problem.

Keywords: Stochastic optimization, distributionally robust optimization, conservatism, bias analysis, asymptotic convergence

1. Introduction

Consider stochastic optimization problems of the following form:

$$v^\star = \inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}\big[f(\boldsymbol{x},\boldsymbol{\xi})\big]. \tag{1}$$

In (1), $\boldsymbol{x}$ is a vector of decision variables; $\mathcal{X} \subseteq \mathbb{R}^n$ is a non-empty set of deterministic constraints on $\boldsymbol{x}$; $\boldsymbol{\xi} : \Omega \to \Xi$ is a random vector defined on a measurable space $(\Omega, \mathcal{F})$ with support $\Xi \subseteq \mathbb{R}^\ell$; and $f : \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R}$ is a continuous function in $\boldsymbol{x}$ for a given $\boldsymbol{\xi} \in \Xi$ and is measurable in $\boldsymbol{\xi}$ for

*Corresponding author.

Email addresses: mat420@lehigh.edu (Man Yiu Tsang), kas720@lehigh.edu (Karmel S. Shehadeh)

given $\boldsymbol{x} \in \mathcal{X}$. Function f measures the performance of the system of interest, and $\mathbb{E}_{\mathbb{P}^\star}[\cdot]$ denotes the expectation with respect to (w.r.t.) probability measure $\mathbb{P}^\star$ defined on $(\Omega, \mathcal{F})$. We assume that (1) has a finite optimal value with a non-empty set of optimal solutions and $\mathbb{E}_{\mathbb{P}^\star}|f(\boldsymbol{x}, \boldsymbol{\xi})| < \infty$ for any $\boldsymbol{x} \in \mathcal{X}$.

In many real-life settings, the true distribution $\mathbb{P}^\star$ of $\boldsymbol{\xi}$ is fundamentally unknown. Suppose instead that we have a (potentially small) set of historical observations of $\{\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_N\}$ of $\boldsymbol{\xi}$. Moreover, suppose we *believe* this data reflects or well approximates the true distribution $\mathbb{P}^\star$. Then, we can estimate $\mathbb{E}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ for any $\boldsymbol{x} \in \mathcal{X}$ by averaging values $f(\boldsymbol{x}, \boldsymbol{\xi}_j)$, for $j \in [N] := \{1, \ldots, N\}$. This leads to the following sample average approximation (SAA)

$$\widehat{v}_N^{\mathrm{SAA}} = \inf_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\widehat{\mathbb{P}}_N}\big[f(\boldsymbol{x}, \boldsymbol{\xi})\big] = \inf_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) \tag{2}$$

of the true problem (1), where $\widehat{\mathbb{P}}_N$ is the empirical distribution of $\widehat{\boldsymbol{\xi}}$. Optimal solutions to (2) are known to be sensitive to the input data used in the model, i.e., they are a function of a particular *distributional belief* (Birge and Louveaux, 2011; Shapiro et al., 2014). Moreover, as pointed out by Kuhn et al. (2019), even if one employs sophisticated statistical techniques to estimate the uncertainty distribution using historical data, the estimated distribution $\widehat{\mathbb{P}}_N$ may differ from the actual distribution $\mathbb{P}^\star$. This is concerning because optimal solutions to (2) obtained using an estimated distribution may inherit estimation errors and bias, leading to poor performance and disappointments under unseen data. This phenomenon is known as the *optimizer's curse* (Mohajerin Esfahani and Kuhn, 2018; Smith and Winkler, 2006; Van Parys et al., 2021).

One popular approach to address distributional ambiguity is distributionally robust optimization (DRO). In DRO, instead of faithfully adopting a specific distribution (such as the empirical distribution), we consider an ambiguity set encompassing various potential distributions of $\boldsymbol{\xi}$ against which we aim to safeguard. Let $\mathcal{P}(\Xi)$ be the set of probability measures defined on the measurable space $(\mathbb{R}^\ell, \mathcal{B})$ induced by $\boldsymbol{\xi}$ with support Ξ, where $\mathcal{B} = \mathcal{B}(\mathbb{R}^\ell)$ is the Borel σ-field. The DRO counterpart of problem (1) is as follows:

$$\widehat{v}_N^{\mathrm{DRO}} = \inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[f(\boldsymbol{x}, \boldsymbol{\xi})\big], \tag{3}$$

where $\mathcal{P}_N \subseteq \mathcal{P}(\Xi)$ is the ambiguity set. The versatility and power of DRO stem from its ability to account for one's incomplete knowledge of the distribution $\mathbb{P}$ of $\boldsymbol{\xi}$ by specifying an ambiguity set $\mathcal{P}_N$. There are two common approaches in the literature for constructing the ambiguity set. The first utilizes partial distributional information such as the moments and support (Delage and Ye, 2010; Goh and Sim, 2010; Nie et al., 2023; Postek et al., 2018, 2019; Sun et al., 2022; Wiesemann et al., 2014; Xu et al., 2018) and marginal distribution (Chen et al., 2023; Kakouris and Rustem, 2014). This approach often leads to tractable DRO formulations. However, ambiguity sets constructed using this approach do not fully capture what is known about $\boldsymbol{\xi}$ and often lead to conservative decisions. Moreover, as pointed out by Lam (2021), DRO models based on such sets do not have large-sample asymptotic convergence properties. The second approach considers distributions that are close to a reference distribution (e.g., empirical distribution) in the sense of a chosen statistical distance. Popular choices of the statistical distance are ϕ-divergence (Bayraksan and Love, 2015; Ben-Tal et al., 2013) and Wasserstein distance (Blanchet et al., 2021; Gao and Kleywegt, 2022; Mohajerin Esfahani and Kuhn, 2018; Xie, 2021). The distance-based approach also leads to tractable reformulations under mild conditions (Bayraksan and Love, 2015; Mohajerin Esfahani and Kuhn, 2018; Zhao and Guan, 2015). In addition, some distance-based DRO (e.g., Wasserstein DRO) model enjoys out-of-sample and asymptotic consistency guarantees (Duchi et al., 2021; Mohajerin Esfahani and Kuhn, 2018; Kuhn et al., 2019).

Since its inception, the DRO approach has received significant attention in operations research, economics, finance, and other fields due to its desirable theoretical properties and ability to produce decisions that maintain robust performance under various distributions. Nevertheless, there are also criticisms concerning the conservatism of the DRO approach. These concerns arise because solutions to (3) hedge against the worst-case distribution within a possibly diverse set of distributions $\mathcal{P}_N$ and hence could be overly conservative. In particular, worst-case distributions that attain $\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[f(\boldsymbol{x}, \boldsymbol{\xi})\big]$ under many celebrated (moment- and distance-based) ambiguity sets are shown to be discrete with few support points even though the true distribution $\mathbb{P}^*$ could be continuous or have a large number of other support points (Bayraksan and Love, 2015; Chen et al., 2011; Das et al., 2021; de Klerk et al., 2020; Gao and Kleywegt, 2022; Long et al., 2023). Hence, DRO often hedges against pessimistic scenarios that are less likely to be observed in practice and thus result in conservative decisions; that is, the realized objective value obtained by DRO solutions would often be better than the optimal value of problem (3).

Recently, several authors proposed approaches to mitigate conservatism of DRO models. These include globalized DRO (Ding et al., 2023; Li and Xing, 2023; Liu et al., 2023a) and techniques to reduce the size of the ambiguity set, such as imposing additional structural properties on distributions in moment-based ambiguity set (de Klerk et al., 2020; Hanasusanto et al., 2015; Lam et al., 2021; Li et al., 2019; Van Parys et al., 2016), restricting distributions in the ambiguity set to specific parametric family (Iyengar et al., 2022; Michel et al., 2021, 2022), and developing modified Wasserstein metrics for distance-based ambiguity sets (Liu et al., 2022; Wang et al., 2021). These pioneering approaches often change the structure of the DRO problem and, as a result, require tailored reformulation and solution techniques to solve the modified problem effectively. Thus, most of these studies mainly focused on reformulating and solving their proposed models without analyzing the models' statistical properties. We refer to Rahimian and Mehrotra (2022) for a comprehensive review of the theoretical and computational developments in the DRO literature.

Inevitably, there are trade-offs among different approaches to modeling uncertainty. On the one hand, by following a blind distributional belief, one may obtain decisions based on an over-optimistic viewpoint that might lead to disappointing performance in practice. On the other hand, by focusing the optimization on the worst-case distribution, one may obtain pessimistic (conservative) solutions. In this paper, our goal is to introduce and analyze an alternative trade-off approach for modeling uncertainty that serves as a middle ground between the optimistic approach, which adopts a distributional belief, and the pessimistic approach, which protects against distributional ambiguity. We formulate the trade-off (TRO) problem as follows:

$$\widehat{v}_N(\theta) = \inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}'_{N,\theta}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})], \tag{4}$$

where $\mathcal{P}'_{N,\theta}$ is the *TRO ambiguity set* defined as

$$\mathcal{P}'_{N,\theta} = \left\{ (1-\theta)\widehat{\mathbb{P}}_N + \theta\mathbb{Q} \mid \mathbb{Q} \in \mathcal{P}_N \right\} \tag{5}$$

for some $\theta \in [0, 1]$. The TRO model (4) can be viewed as a new data-driven DRO model equipped with a TRO ambiguity set $\mathcal{P}'_{N,\theta}$. The TRO ambiguity set $\mathcal{P}'_{N,\theta}$ in (5) is characterized by two parameters: the *shape* parameter $\mathcal{P}_N$ and the *size* parameter θ. The shape parameter $\mathcal{P}_N$ represents distributional ambiguity and could be any (data-driven) ambiguity set satisfying some mild assumptions to be made precise later. The size parameter $\theta \in [0, 1]$ controls the level of optimism, i.e., it controls the trade-off between solving the problem under a distributional belief and solving it under ambiguity. When $\theta = 0$, problem (4) reduces to problem (2). In contrast, when $\theta = 1$, problem (4) reduces to problem (3). Between the two extremes, $\theta \in (0, 1)$ indicates a trade-off between optimistic and pessimistic perceptions of the objective. Thus, as we later show, by changing the value of θ in (4), one could obtain a spectrum of optimal solutions, ranging from optimistic to conservative solutions.

We emphasize that one can construct the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ using any shape parameter $\mathcal{P}_N$, including general moment- and distance-based ambiguity sets. This flexibility allows us to derive general results outlined in Section 1.1 on the TRO model's conservatism, finite-sample properties, and asymptotic convergence. Moreover, one can adopt the same techniques developed for reformulating and solving classical DRO problems with specific ambiguity sets $\mathcal{P}_N$ to reformulate and solve the TRO problem with $\mathcal{P}'_{N,\theta}$ constructed using $\mathcal{P}_N$ as the shape parameter. To the best of our knowledge and according to Rahimian and Mehrotra (2022)'s recent survey, our paper is the first to formally introduce the TRO approach and conduct a thorough theoretical investigation of its properties. Notably, Shehadeh and Tucker (2022) is the first to explore the use of a convex combination of SAA and DRO (with mean-support) ambiguity set to model uncertainty in the context of disaster relief. Their numerical investigations show that the trade-off approach could lead to a less conservative disaster preparation plan than the DRO approach, with better post-disaster response performance than the SAA-based plan. However, that work was problem-specific and did not analyze the theoretical properties of the model. Recently, Wang et al. (2023) analyzed the trade-off between the robustness to unseen data and the specificity of the training data in the context of machine learning. As we show in Appendix G, Wang et al. (2023)'s Bayesian distributionally robust (BDR) model is a special case of our TRO model. In particular, by replacing function f in (4) with the loss function h and choosing the shape parameter $\mathcal{P}_N$ as the distance-based ambiguity set $B_{\epsilon_N}(\widehat{\mathbb{P}}_N) = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \Delta(\mathbb{P}, \widehat{\mathbb{P}}_N) \leq \epsilon\}$ (where $\mathcal{P}(\Xi)$ is the set of probability measures on the support Ξ and Δ is a statistical distance), our TRO model reduces to the Wang et al. (2023)'s BDR model. While our theoretical investigations are valid for any shape parameter (including moment-based ambiguity sets), Wang et al. (2023)'s analyses are limited to the special case where $\mathcal{P}_N$ is a distance-based ambiguity set. In addition, different from Wang et al. (2023), we introduce and study the hierarchical properties of the TRO ambiguity $\mathcal{P}'_{N,\theta}$, analyze the conservatism of our TRO model, and provide more comprehensive analyses of the finite-sample guarantees and asymptotic properties of the TRO model; see Appendix G for detailed discussions.

1.1. Contributions

We highlight the following main contributions of our theoretical investigations:

- *Hierarchical Properties of $\mathcal{P}'_{N,\theta}$.* We establish that constructing the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ using a general star-shaped shape parameter $\mathcal{P}_N$ with a star center $\widehat{\mathbb{P}}_N$ is both necessary and sufficient for the sequence $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ to be non-decreasing (Theorem 1, part (i)). If, in addition, $\mathcal{P}_N \neq \{\widehat{\mathbb{P}}_N\}$, the sequence $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ increases with θ, i.e., $\mathcal{P}'_{N,\theta}$ contains more distributions with a larger θ (Theorem 1, part (ii)).

- *Quantifying the conservatism of the TRO model.* Prior studies have quantified the impact of variation of ambiguity set on the optimal values and solutions of DRO problems (Liu et al., 2019; Liu and Xu, 2013; Pichler and Xu, 2018; Sun and Xu, 2016). Similarly, we conduct quantitative stability analyses to quantify the difference in the optimal value $\widehat{\upsilon}_N(\theta)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta)$ and hence the conservatism incurred by perturbation in $\mathcal{P}'_{N,\theta}$. Specifically, we show the Lipschitz continuity of the optimal value function $\widehat{\upsilon}_N(\theta)$ and the Hölder continuity of $\widehat{\mathcal{X}}_N(\theta)$ (Theorem 2). In addition, we quantify the difference between the optimal value (resp. set of optimal solutions) and the convex combination of optimal values (resp. sets of optimal solutions) to the SAA and DRO problems resulting from solving each separately (Theorem 3). Together, the results of Theorems 1–3 show that by solving the TRO model with different $\theta \in [0,1]$, one can obtain a spectrum of decisions that span $[\widehat{\upsilon}_N(0), \widehat{\upsilon}_N(1)]$, representing decisions with different levels of conservatism.

- *Finite-sample properties.* The optimal value of our TRO model $\widehat{\upsilon}_N(\theta)$ is an estimator of the optimal value $\upsilon^\star$ of problem (1). We show that under some mild assumptions, there exists $\theta^{\mathrm{u}}_N \in [0,1]$ such that $\widehat{\upsilon}_N(\theta^{\mathrm{u}}_N)$ is an unbiased estimator (Theorem 4). In addition, we show that our TRO model could produce estimators with a smaller bias than the SAA estimator (Corollary 1) and derive the asymptotic convergence rate of θ^{u}_N as $N \to \infty$ (Theorems 5 and

6). These analytical results hold for TRO models with TRO ambiguity sets constructed using general shape parameters, including moment-based ambiguity sets. Moreover, we derive a bound on the generalization error of the TRO model (Theorem 7). In particular, we show that the generalization error has an exponentially decaying tail for specific choices of the shape parameter.

- *Asymptotic properties.* We show that as the number of data points $N \to \infty$, the optimal value $\widehat{\upsilon}_N(\theta_N)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta_N)$ of the TRO problem (4) converge respectively to the true optimal value $\boldsymbol{x}^*$ and set of optimal solutions $\mathcal{X}^*$ of problem (1) almost surely (Theorem 8). In addition, we derive the asymptotic distribution of $\widehat{\upsilon}_N(\theta_N)$ (Theorem 9). These asymptotic properties hold for TRO models with TRO ambiguity sets constructed using general shape parameters, such as moment-based ambiguity sets. This differs from the existing convergence results established for data-driven DRO models, which mainly employ distance-based ambiguity sets. For the special case when the shape parameter is chosen as a distance-based ambiguity set, we can recover the asymptotics of the optimal value of classical distance-based DRO models (see, e.g., Blanchet and Shapiro, 2023). Specifically, we show that $\widehat{\upsilon}_N(\theta_N)$ converges to different distributions depending on the convergence rates of the size parameter θ_N and the radius r_N in the shape parameter (Theorem 10).

1.2. Structure of the paper

The remainder of the paper is organized as follows. In Section 2, we analyze hierarchical properties of the TRO ambiguity set $\mathcal{P}'_{N,\theta}$. In Section 3, we quantify the conservatism of our TRO model. In Section 4, we analyze the bias of the optimal value of our TRO model $\widehat{\upsilon}_N(\theta)$ as an estimator of the true optimal value $\upsilon^\star$ and the generalization error of our TRO model. In Section 5, we derive the asymptotic properties of our TRO model. Finally, in Section 6, we exemplify our theoretical results using an inventory control problem and a mean-risk portfolio optimization problem.

1.3. Notation

For a set C in a general convex space X, the convex hull of C is defined as $\text{conv}(C) = \big\{ \sum_{i=1}^k \alpha_i x_i \mid k \in \mathbb{N}, \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0, x_i \in X, \forall i \in \{1, \ldots, k\} \big\}$. We use $\|\cdot\|$ to denote a general norm defined on a vector space. For any $a \in \mathbb{R}$, we define $(a)_+ = \max\{a, 0\}$. In the Euclidean space $\mathbb{R}^n$, we use boldface letter such as $\boldsymbol{y}$ to denote a column vector in $\mathbb{R}^n$ and $\|\boldsymbol{y}\|_p$ to denote the p-norm of the vector $\boldsymbol{y}$. We define the distance between a point $\boldsymbol{y} \in \mathbb{R}^n$ and a set $\mathcal{Y} \subseteq \mathbb{R}^n$ as $d(\boldsymbol{y}, \mathcal{Y}) = \inf_{\boldsymbol{y}' \in \mathcal{Y}} \|\boldsymbol{y} - \boldsymbol{y}'\|$, and the distance between two sets $\mathcal{Y}_1 \subseteq \mathbb{R}^n$ and $\mathcal{Y}_2 \subseteq \mathbb{R}^n$ as $D(\mathcal{Y}_1, \mathcal{Y}_2) = \sup_{\boldsymbol{y}_1 \in \mathcal{Y}_1} \inf_{\boldsymbol{y}_2 \in \mathcal{Y}_2} \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|$. The sum of two sets $\mathcal{Y}_1 \subseteq \mathbb{R}^n$ and $\mathcal{Y}_2 \subseteq \mathbb{R}^n$ is interpreted in the sense of Minkowski, i.e., $\mathcal{Y}_1 + \mathcal{Y}_2 = \{\boldsymbol{y}_1 + \boldsymbol{y}_2 \mid \boldsymbol{y}_1 \in \mathcal{Y}_1, \boldsymbol{y}_2 \in \mathcal{Y}_2\}$. For any functions $g(N)$ and $h(N)$ defined on $\mathbb{N}$, we write $h(N) = O(g(N))$ (as $N \to \infty$) if there exists $M > 0$ and $N_0 \in \mathbb{N}$ such that $h(N) \leq Mg(N)$ for all $N \geq N_0$, and we write $h(N) = o(g(N))$ (as $N \to \infty$) if

$\lim_{N\to\infty} h(N)/g(N) = 0$. Two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on a measurable space $(\Omega, \mathcal{F})$ are equal, i.e., $\mathbb{P}_1 = \mathbb{P}_2$, if $\mathbb{P}_1(B) = \mathbb{P}_2(B)$ for all $B \in \mathcal{F}$. For a probability measure $\mathbb{P}$, the variance of a random variable Y under $\mathbb{P}$ is denoted as $\mathrm{Var}_{\mathbb{P}}(Y)$, and the covariance of two random variables Y_1 and Y_2 is denoted as $\mathrm{Cov}_{\mathbb{P}}(Y_1, Y_2)$. For a sequence of random variables $\{Y_n\}_{n\in\mathbb{N}}$ and a sequence of positive real numbers $\{a_n\}_{n\in\mathbb{N}}$, we write $Y_n = o_{\mathbb{P}}(a_n)$ if $|Y_n|/a_n \to 0$ in probability.

2. Hierarchical Property of the TRO Ambiguity Set

In this section, we analyze properties of $\mathcal{P}'_{N,\theta}$ and the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$. To facilitate the discussion, we introduce the notion of star-shapedness of the set of distributions $\mathcal{P}_N$ in Definition 2.1 and hierarchical properties of $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ in Definition 2.2.

Definition 2.1 (Star-Shaped Set)**.** The set of distributions $\mathcal{P}_N$ is called *star-shaped* if there exists $\mathbb{M} \in \mathcal{P}_N$ such that

$$(1-\alpha)\mathbb{M} + \alpha\mathbb{P} \in \mathcal{P}_N, \quad \forall \alpha \in [0,1],\, \mathbb{P} \in \mathcal{P}_N. \tag{6}$$

Any $\mathbb{M} \in \mathcal{P}_N$ satisfying condition (6) is called a *star center* of $\mathcal{P}_N$.

Definition 2.2 (Hierarchical Properties)**.** The sequence $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the *hierarchical property* if $\mathcal{P}'_{N,\theta}$ is non-decreasing in θ, i.e., $\mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$ for any $0 \le \theta_1 < \theta_2 \le 1$. The sequence $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the *strict hierarchical property* if $\mathcal{P}'_{N,\theta}$ is increasing in θ, i.e., $\mathcal{P}'_{N,\theta_1} \subset \mathcal{P}'_{N,\theta_2}$ for any $0 \le \theta_1 < \theta_2 \le 1$.

The hierarchical properties indicate that the size of the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ increases with θ, i.e., $\mathcal{P}'_{N,\theta}$ contains more distributions with a larger θ. In other words, with a larger θ, our TRO model hedges against a larger set of distributions. In Theorem 1, we provide necessary and sufficient conditions for the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ to satisfy these properties.

Theorem 1. *The following assertions hold.*

(i) The sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the hierarchical property if and only if $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$.

(ii) The sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the strict hierarchical property if and only if $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$ and $\mathcal{P}_N \neq \{\widehat{\mathbb{P}}_N\}$.

Theorem 1 establishes that constructing the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ using a star-shaped parameter $\mathcal{P}_N$ with a star center $\widehat{\mathbb{P}}_N$ is necessary and sufficient for the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ to satisfy the hierarchical property. Part (i) shows that for a general star-shaped shape parameter $\mathcal{P}_N$, the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ is non-decreasing in θ, i.e., $\mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$ whenever $\theta_1 \le \theta_2$, indicating that the objective function of the trade-off model (4) is non-decreasing in θ. Figure 1 illustrates the relationship between the sets $\{\widehat{\mathbb{P}}_N\}$, $\mathcal{P}'_{N,\theta_1}$, $\mathcal{P}'_{N,\theta_2}$, and $\mathcal{P}_N$ with $0 < \theta_1 < \theta_2 < 1$ as suggested by part (ii) of Theorem 1. Specifically, this figure shows how the

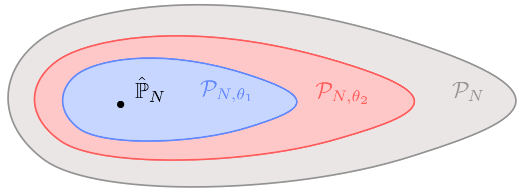

Figure 1: Illustration of the strict hierarchical property of the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid 0 < \theta_1 < \theta_2 < 1\}$.

TRO ambiguity set $\mathcal{P}'_{N,\theta}$ enlarges with θ. This, in turn, implies that the TRO model is more conservative when we pick a larger θ. Note that if $\mathcal{P}'_{N,\theta}$ is constructed using a star-shaped $\mathcal{P}_N$ with a star center $\mathbb{M} \neq \widehat{\mathbb{P}}$, then $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ does not satisfy the hierarchical properties; see Appendix B for an example.

Theorem 1 requires the shape parameter $\mathcal{P}_N$ to be star-shaped with a star center $\widehat{\mathbb{P}}_N$. As is well known, if $\mathcal{P}_N$ is convex and $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, then $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$. Many celebrated (data-driven) ambiguity sets $\mathcal{P}_N$ are convex and contain the empirical distribution; see Examples 1–4 below. Thus, they are star-shaped with a star center $\widehat{\mathbb{P}}_N$. On the other hand, a star-shaped set is not necessarily convex; hence, the hierarchical properties hold for general shape parameters. For example, if $\mathcal{P}_N$ is the union of a collection of convex ambiguity sets $\{\mathcal{P}_{N,\gamma}\}_{\gamma\in\Gamma}$ and $\widehat{\mathbb{P}}_N \in \mathcal{P}_{N,\gamma}$ for all $\gamma \in \Gamma$, then $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$ but not generally convex. Note that in some settings, one may want to consider the union of multiple (convex) ambiguity sets as the shape parameters. For example, the available data may not always follow a structured pattern and, thus, a known shape parameter. Also, the decision-makers may be unable to articulate their preference for the set of distributions to protect against. In such settings, considering a union of two or more ambiguity sets may better represent or approximate the true distribution of the available data. In Propositions 1 and 2, we derive necessary and sufficient conditions for general moment- and distance-based ambiguity sets, respectively, to be star-shaped with a star center $\widehat{\mathbb{P}}_N$.

Proposition 1. *Consider the moment-based ambiguity set* $\mathcal{P}_N = \big\{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathbb{E}_{\mathbb{P}}[\Phi_i(\boldsymbol{\xi})] \in \mathcal{K}_i,\ i \in \{1,\dots,p\}\big\}$, *where* $\Phi_i : \Xi \to \mathbb{R}^{d_i\times d_i}$ *is a matrix-valued function and* $\mathcal{K}_i \subseteq \mathbb{R}^{d_i\times d_i}$ *is a set of matrices for all* $i \in \{1,\dots,p\}$. *Then,* $\mathcal{P}_N$ *is star-shaped with a star center* $\widehat{\mathbb{P}}_N$ *if and only if* $\mathcal{K}_i$ *is star-shaped on* $\mathcal{S}_i := \{\mathbb{E}_{\mathbb{Q}}[\Phi_i(\boldsymbol{\xi})] \mid \mathbb{Q} \in \mathcal{P}_N\} \subseteq \mathbb{R}^{d_i\times d_i}$ *with a star center* $\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] \in \mathcal{S}_i$ *for all* $i \in \{1,\dots,p\}$, *i.e.,* $(1-\alpha)\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] + \alpha\Psi \in \mathcal{K}_i$ *for all* $\alpha \in [0,1]$ *and* $\Psi \in \mathcal{S}_i$.

Example 1. Consider the following moment ambiguity set

$$\mathcal{P}_N = \Big\{\mathbb{P} \in \mathcal{P}(\Xi) \;\Big|\; \mathbb{E}_{\mathbb{P}}[\Phi_i(\boldsymbol{\xi})] = \boldsymbol{\mu}_i^N,\ \forall i \in \{1,\dots,p\},\ \mathbb{E}_{\mathbb{P}}[\Phi_i(\boldsymbol{\xi})] \preceq \boldsymbol{\mu}_i^N,\ \forall i \in \{p+1,\dots,q\}\Big\}, \quad (7)$$

where $\Phi_i : \Xi \to \mathbb{R}^{d_i\times d_i}$ for $i \in \{1,\dots,q\}$ is a symmetric matrix (or a scalar when the dimension d_i is one) with measurable entries and $\mathbb{E}_{\mathbb{P}}[\Phi_i(\boldsymbol{\xi})] \preceq \boldsymbol{\mu}_i^N$ means that $\mathbb{E}_{\mathbb{P}}[\Phi_i(\boldsymbol{\xi})] - \boldsymbol{\mu}_i^N$ is negative semidefinite (Sun and Xu, 2016; Xu et al., 2018). Here, $\boldsymbol{\mu}_i^N$ for $i \in \{1,\ldots,q\}$ is the sample estimate of $\mathbb{E}_{\mathbb{P}^\star}[\Phi_i(\boldsymbol{\xi})]$ based on the given data, i.e., $\boldsymbol{\mu}_i^N = (1/N)\sum_{i=1}^N \Phi(\widehat{\boldsymbol{\xi}}_i)$. The ambiguity set (7) includes several popular ambiguity sets adopted in the literature (Delage and Ye, 2010; So, 2011). It is straightforward to verify that $\mathcal{K}_i := \{\Psi_i \in \mathbb{R}^{d_i\times d_i} \mid \Psi_i = \boldsymbol{\mu}_i^N\}$ for $i \in \{1,\ldots,p\}$ and $\mathcal{K}_i := \{\Psi_i \in \mathbb{R}^{d_i\times d_i} \mid \Psi_i - \boldsymbol{\mu}_i^N \preceq 0\}$ for $i \in \{p+1,\ldots,q\}$ are convex with $\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] \in \mathcal{K}_i$. Thus, $\mathcal{K}_i$ is star-shaped with a star center $\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})]$ for all $i \in \{1,\ldots,q\}$. It follows from Proposition 1 that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$.

Proposition 2. *Let $\mathsf{d} : \mathcal{P}(\Xi)\times\mathcal{P}(\Xi) \to \mathbb{R}_+$ be a statistical distance satisfying $\mathsf{d}(\mathbb{P}_1,\mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$. Consider the distance-based ambiguity set of the form $\mathcal{P}_N(\varepsilon) = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P},\widehat{\mathbb{P}}_N) \le \varepsilon\}$. Then, $\mathcal{P}_N(\varepsilon)$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$ for all $\varepsilon \ge 0$ if and only if d is quasi-convex about $\widehat{\mathbb{P}}_N$ in the first argument, i.e.,*

$$\mathsf{d}\Big((1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q}, \widehat{\mathbb{P}}_N\Big) \le \max\Big\{\mathsf{d}(\widehat{\mathbb{P}}_N,\widehat{\mathbb{P}}_N),\, \mathsf{d}(\mathbb{Q},\widehat{\mathbb{P}}_N)\Big\} = \mathsf{d}(\mathbb{Q},\widehat{\mathbb{P}}_N)$$

for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}(\Xi)$.

Example 2. Consider the following ϕ-divergence ambiguity set

$$\mathcal{P}_N = \left\{\mathbb{P} = \sum_{i=1}^N p_i\delta_{\widehat{\boldsymbol{\xi}}_i} \,\middle|\, \boldsymbol{p} = (p_1,\ldots,p_N) \in \mathbb{R}_+^N,\, \boldsymbol{p}^\top\mathbf{1} = 1,\, \frac{1}{N}\sum_{i=1}^N \phi(Np_i) \le \varepsilon\right\} \tag{8}$$

for some $\varepsilon > 0$ and convex function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\phi(1) = 0$. This contains all the distribution with support $\{\widehat{\boldsymbol{\xi}}_1,\ldots,\widehat{\boldsymbol{\xi}}_N\}$ such that the ϕ-divergence defined by $\mathsf{d}(\mathbb{P},\widehat{\mathbb{P}}_N) = (1/N)\sum_{i=1}^N \phi(Np_i)$ is no greater than the radius ε. It is easy to verify that d is convex in $\mathbb{P}$ (identified with $\boldsymbol{p} \in \mathbb{R}^N$). Thus, it follows from Proposition 2 that ambiguity set (8) is star-shaped with a star center $\widehat{\mathbb{P}}_N$.

Example 3. Consider the following p-Wasserstein ambiguity set

$$\mathcal{P}_N = \big\{\mathbb{P} \in \mathcal{P}(\Xi) \mid W_p(\mathbb{P},\widehat{\mathbb{P}}_N) \le \varepsilon\big\} \tag{9}$$

for some $p \in [1,\infty)$ and $\varepsilon > 0$, where $W_p(\mathbb{P},\widehat{\mathbb{P}}_N)$ is the p-Wasserstein distance between probability measures $\mathbb{P}$ and $\widehat{\mathbb{P}}_N$ (see Villani, 2009; Mohajerin Esfahani and Kuhn, 2018). Since $\mathsf{d} = W_p$ is p-convex in $\mathbb{P}$ (see Lemma 2.10 of Pflug and Pichler, 2014), we have

$$W_p\Big((1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q}, \widehat{\mathbb{P}}_N\Big) \le \Big\{(1-\alpha)W_p^p(\widehat{\mathbb{P}}_N,\widehat{\mathbb{P}}_N) + \alpha W_p^p(\mathbb{Q},\widehat{\mathbb{P}}_N)\Big\}^{\frac{1}{p}} \le \alpha^{\frac{1}{p}}W_p(\mathbb{Q},\widehat{\mathbb{P}}_N) \le W_p(\mathbb{Q},\widehat{\mathbb{P}}_N)$$

for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N$. It follows from Proposition 2 that ambiguity set (9) is star-shaped with a star center $\widehat{\mathbb{P}}_N$.

Finally, in the following example, we show that when the shape parameter $\mathcal{P}_N$ consists only of Dirac measures (as in the classical robust optimization approach), the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the hierarchical property under some mild assumptions.

Example 4. The robust optimization (RO) counterpart of problem (1), defined as $\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi})$ for some uncertainty set $\mathcal{U} \subseteq \Xi$, can be recast as a DRO model with ambiguity set $\mathcal{P}_N = \{\delta_{\boldsymbol{\xi}} \mid \boldsymbol{\xi} \in \mathcal{U}\}$. It is obvious that $\mathcal{P}_N$ contains only Dirac measure and thus, does not contain $\widehat{\mathbb{P}}_N$. However, if for any $\boldsymbol{x} \in \mathcal{X}$, there exists $\boldsymbol{\xi} \in \mathcal{U}$ such that $f(\boldsymbol{x},\boldsymbol{\xi}) \geq f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i)$ for all $i \in \{1,\dots,N\}$, then the RO problem can be written as a DRO problem with ambiguity set

$$\mathcal{P}_N = \operatorname{conv}\left(\widehat{\mathbb{P}}_N \cup \left\{\delta_{\boldsymbol{\xi}} \mid \boldsymbol{\xi} \in \mathcal{U}\right\}\right) \tag{10}$$

(see Appendix C for a proof). By construction, the ambiguity set in (10) is convex with $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$. Hence, it is also star-shaped with a star center $\widehat{\mathbb{P}}_N$, and thus the hierarchical property of $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ follows from Theorem 1.

3. Conservatism of the TRO Model

Recall that the shape parameter $\mathcal{P}_N$ of the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ represents distributional ambiguity, and consequently, it influences the pessimistic view of the objective or conservatism. On the other hand, the size parameter θ controls the level of optimism or, equivalently, the degree of conservatism. In this section, we analyze the conservatism and properties of the optimal value $\widehat{v}_N(\theta)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta)$ of the TRO model for a fixed sample through the lens of quantitative stability analysis. We refer to Appendix E for background results relevant to our analysis.

Let us first introduce some additional notation to lay the foundation for subsequent discussions. For two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, we define the pseudometric $\mathbb{d}(\mathbb{P}_1,\mathbb{P}_2)$ as

$$\mathbb{d}(\mathbb{P}_1,\mathbb{P}_2) = \sup_{x\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}_1}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_2}[f(\boldsymbol{x},\boldsymbol{\xi})]\right| \tag{11}$$

(Liu and Xu, 2013; Römisch, 2003; Sun and Xu, 2016). Using the pseudometric $\mathbb{d}$, we define the distance between a single probability distribution $\mathbb{P}$ and a set $\mathcal{P}$ of distributions as $\mathbb{D}(\mathbb{P},\mathcal{P}) = \inf_{\mathbb{Q}\in\mathcal{P}} \mathbb{d}(\mathbb{P},\mathbb{Q})$. Finally, we define the Hausdorff distance between two sets of probability distributions, $\mathcal{P}_1$ and $\mathcal{P}_2$, as

$$\mathbb{H}(\mathcal{P}_1,\mathcal{P}_2) = \max\left\{\sup_{\mathbb{Q}\in\mathcal{P}_1} \mathbb{D}(\mathbb{Q},\mathcal{P}_2),\ \sup_{\mathbb{Q}\in\mathcal{P}_2} \mathbb{D}(\mathbb{Q},\mathcal{P}_1)\right\}. \tag{12}$$

We make the following technical assumptions to ensure that the TRO model is well-defined.

Assumption 1. The feasible set $\mathcal{X}$ and the function f satisfy the following conditions:

(a) the feasible set $\mathcal{X}$ is compact;

(b) given an ambiguity set $\mathcal{P}_N$, (i) $f(\cdot,\boldsymbol{\xi})$ is Lipschitz continuous on $\mathcal{X}$ for any fixed $\boldsymbol{\xi} \in \Xi$ with Lipschitz modulus bounded by $\kappa(\boldsymbol{\xi})$, where $\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] < \infty$; and (ii) there exists $\boldsymbol{x}_0 \in \mathcal{X}$ such that $\max\left\{\mathbb{E}_{\widehat{\mathbb{P}}_N}\left|f(\boldsymbol{x}_0,\boldsymbol{\xi})\right|, \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\left|f(\boldsymbol{x}_0,\boldsymbol{\xi})\right|\right\} < \infty$.

Assumption 1(a) is a standard assumption in the stochastic optimization literature (Duchi et al., 2021; Shapiro et al., 2014; Van Parys et al., 2021; Zhang et al., 2016). Assumption 1(b) ensures the smoothness of the objective function of (4) and that the objective is finite at some point $\boldsymbol{x}_0 \in \mathcal{X}$ (Gao, 2022; Pichler and Xu, 2018; Sun and Xu, 2016). Essentially, Assumption 1 ensures that

$$C_N := \max\left\{\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]\right|, \sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\mathbb{P}\in\mathcal{P}_N}\left|\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right|\right\} < \infty, \tag{13}$$

implying that our trade-off model (4) is well-defined, i.e., has a finite optimal value and optimal solution for any $\theta \in [0, 1]$. These assumptions hold valid in many applications (e.g., scheduling, inventory control, facility location, etc.).

First, in Theorem 2, we quantify the impact of a perturbation of the trade-off parameter θ on the optimal value and the set of optimal solutions to our TRO model. Specifically, we show the Lipschitz continuity of the optimal value function $\widehat{v}_N(\theta)$ and the stability of the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta)$ to our TRO model.

Theorem 2. *Under Assumption 1, the following assertions hold.*

(i) $|\widehat{v}_N(\theta_1) - \widehat{v}_N(\theta_2)| \leq 2C_N|\theta_1 - \theta_2|$.

(ii) If, in addition, $\sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta_1}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]$ *satisfies the second-order growth condition at* $\widehat{\mathcal{X}}_N(\theta_1)$*, i.e., there exists* $\tau > 0$ *such that*

$$\sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta_1}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \geq \widehat{v}_N(\theta_1) + \tau\left[d\Big(\boldsymbol{x}, \widehat{\mathcal{X}}_N(\theta_1)\Big)\right]^2 \tag{14}$$

for all $\boldsymbol{x} \in \mathcal{X}$*, then* $D\big(\widehat{\mathcal{X}}_N(\theta_2), \widehat{\mathcal{X}}_N(\theta_1)\big) \leq \sqrt{6C_N\tau^{-1}|\theta_1 - \theta_2|}$.

Theorem 2 establishes mechanisms to quantify the difference in $\widehat{v}_N(\theta)$ and $\widehat{\mathcal{X}}_N(\theta)$ (and hence conservatism) incurred by perturbation in θ. In particular, it shows that both the optimal value and the set of optimal solutions change gradually with $\theta \in [0, 1]$; $\widehat{v}_N(\theta)$ is Lipschitz continuous in θ and $\widehat{\mathcal{X}}_N(\theta)$ is Hölder continuous with Hölder exponent $1/2$ under distance D. Moreover, if θ is sufficiently close to zero (resp. one), the optimal value and the set of optimal solutions to our TRO model are close to the SAA (resp. DRO) counterparts. Therefore, in practice, to generate a spectrum of optimal solutions with various extents of conservatism, it suffices to consider multiple disjoint values of θ given the continuity of $\widehat{v}_N(\theta)$ and $\widehat{\mathcal{X}}_N(\theta)$. We remark that the second-order growth condition assumption in (14) is standard in stochastic optimization literature (Liu et al., 2019; Pichler and Xu, 2018; Shapiro, 1994). For example, as discussed in Pichler and Xu (2022), the second-order growth condition (14) holds when the function $f(\boldsymbol{x},\boldsymbol{\xi})$ is $\mu(\boldsymbol{\xi})$-strongly convex in $\boldsymbol{x}$ for each $\boldsymbol{\xi} \in \Xi$ and $\inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\mu(\boldsymbol{\xi})] > 0$.

Note that directly combining SAA and DRO optimal solutions, e.g., via a convex combination after solving each separately, may not yield a feasible solution to the TRO problem (4). This is particularly true in many practical applications where $\mathcal{X}$ is not convex, for example, in problems that involve integer variables such as facility location and scheduling problems. Thus, one needs to solve the TRO model to obtain decisions with different levels of conservatism. In Theorem 3, we leverage the results in Theorem 2 to quantify the difference between our TRO model's optimal value (set of optimal solutions) and the convex combination of optimal values (sets of optimal solutions) to the SAA ($\theta = 0$) and DRO ($\theta = 1$) problems resulting from solving each separately.

Theorem 3. *Under Assumption 1, the following assertion holds.*

(i) $\widehat{v}_N : [0,1] \to \mathbb{R}$ *is concave with*

$$\widehat{v}_N(\theta) = (1-\theta)\cdot \widehat{v}_N(0) + \theta \cdot \widehat{v}_N(1) + \widehat{r}_N(\theta), \tag{15}$$

where $\widehat{r}_N(\theta) \in [0, 4C_N\theta(1-\theta)]$.

(ii) If, in addition, the second order growth rate condition (14) *holds, then*

$$D\Big(\widehat{\mathcal{X}}_N(\theta), (1-\theta)\widehat{\mathcal{X}}_N(0) + \theta\widehat{\mathcal{X}}_N(1)\Big) \le \sqrt{6C_N\tau^{-1}\theta(1-\theta)}\Big(\sqrt{\theta} + \sqrt{1-\theta}\Big),$$

where the set $(1-\theta)\widehat{\mathcal{X}}_N(0) + \theta\widehat{\mathcal{X}}_N(1)$ *contains convex combinations of the SAA and DRO optimal solutions.*

Part (i) of Theorem 3 establishes that the optimal value $\widehat{v}_N(\theta)$ to our TRO model is not less than the convex combination $(1-\theta)\widehat{v}_N(0) + \theta\widehat{v}_N(1)$ of the SAA and DRO optimal values. In addition, if $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, then $\widehat{v}_N(\theta)$ is the minimum of the non-decreasing functions $\big\{g(\theta;\boldsymbol{x}) := \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})] + \theta\big[\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]\big] \,\big|\, \boldsymbol{x} \in \mathcal{X}\big\}$ and thus, is non-decreasing in θ as illustrated in Figure 2. Therefore, by solving our TRO model with different $\theta \in [0,1]$, we can obtain a spectrum of decisions that spans $[\widehat{v}_N(0), \widehat{v}_N(1)]$, representing decisions with different levels of conservatism. Moreover, since $\widehat{v}_N$ is concave, the rate of change in conservatism (i.e., slope of $\widehat{v}_N$) is larger for smaller values of θ. In other words, the increase in conservatism $\widehat{v}_N(\theta+\Delta) - \widehat{v}_N(\theta)$ is more significant when θ is small, where $\Delta > 0$ is the perturbation on θ. To obtain a decision with a specific target level of conservatism, say $\overline{v} = (1-\lambda)\widehat{v}_N(0) + \lambda\widehat{v}_N(1)$ for some $\lambda \in (0,1)$, one should pick a value of θ less than λ (see Figure 2). Part (ii) of Theorem 3 indicates that the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta)$ to our TRO model can be approximated by $\overline{\mathcal{X}}_N(\theta) := (1-\theta)\widehat{\mathcal{X}}_N(0) + \theta\widehat{\mathcal{X}}_N(1)$ only when θ is close to zero or one; however, the difference could be huge for intermediate values of $\theta \in (0,1)$.

4. Finite-Sample Properties

In this section, we investigate finite-sample properties of the TRO model. First, in Section 4.1, we analyze the bias of the optimal value of our TRO model $\widehat{v}_N(\theta)$ as an estimator of the true optimal value $v^\star$. Then, inSection 4.2, we derive the generalization bound for our TRO model.

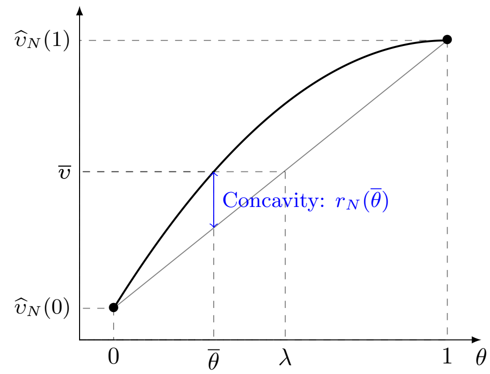

Figure 2: Illustration of $\widehat{\upsilon}_N(\theta)$ when $\mathcal{P}_N$ is convex and $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$. For a given target level of conservatism $\overline{\upsilon} = (1-\lambda)\widehat{\upsilon}_N(0) + \lambda\widehat{\upsilon}_N(1)$ with $\lambda \in (0,1)$, one should pick $\overline{\theta} < \lambda$ in the TRO model.

4.1. Bias Analysis

The optimal value of our TRO model $\widehat{\upsilon}_N(\theta)$ represents an estimator of the optimal value $\upsilon^\star$ of problem (1). In this section, we analyze the bias of $\widehat{\upsilon}_N(\theta)$, i.e., $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta)] - \upsilon^\star$, where $\mathbb{P}^N$ is the joint distribution of $\{\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_N\}$. Recall that when $\theta = 0$, our TRO estimator reduces to the SAA estimator $\widehat{\upsilon}_N(0)$, which suffers from downward bias (Mak et al., 1999; Norkin et al., 1998; Shapiro et al., 2014), i.e.,

$$\mathbb{E}_{\mathbb{P}^N}\left[\widehat{\upsilon}_N(0)\right] = \mathbb{E}_{\mathbb{P}^N}\left[\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i)\right] \le \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^N}\left[\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i)\right] = \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}, \boldsymbol{\xi})] = \upsilon^\star. \tag{16}$$

Although the bias $\upsilon^\star - \mathbb{E}_{\mathbb{P}^N}\left[\widehat{\upsilon}_N(0)\right]$ decreases monotonically with N and goes to zero as $N \to \infty$ (Homem-de Mello and Bayraksan, 2014), it may diminish slowly and could remain significant even for large N (Dentcheva and Lin, 2022). Thus, constructing an accurate estimator of $\upsilon^\star$, and ideally, an unbiased estimator of $\upsilon^\star$ is of practical interest. In this section, we show that our TRO model could produce estimators with a smaller bias than the SAA estimator. Moreover, under mild assumptions, we show that there exists $\theta_N^{\mathrm{u}} \in [0,1]$ such that $\widehat{\upsilon}_N(\theta_N^{\mathrm{u}})$ is an unbiased estimator. Furthermore, we derive the asymptotic convergence rate of θ_N^{u} as $N \to \infty$. In what follows, we say Assumption 1 holds if Assumption 1(a) holds and Assumption 1(b) holds for almost every ambiguity set $\mathcal{P}_N$.

First, in Proposition 3, we derive an upper bound on the bias of the TRO estimator $\widehat{\upsilon}_N(\theta)$.

Proposition 3. *Under Assumption 1, we have*

$$\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta)] - \upsilon^\star \le \theta\Big\{\mathbb{E}_{\mathbb{P}^N}\left[\widehat{\upsilon}_N(1)\right] - \mathbb{E}_{\mathbb{P}^N}\left[\widehat{\upsilon}_N(0)\right]\Big\} + R_N(\theta), \tag{17}$$

where $R_N(\theta) \in [0, 4\overline{C}_N\theta(1-\theta)]$ *and* $\overline{C}_N := \mathbb{E}_{\mathbb{P}^\star}(C_N) < \infty$, *where* C_N *is as defined in* (13). *If* $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$ *almost surely, then the upper bound in* (17) *is non-negative.*

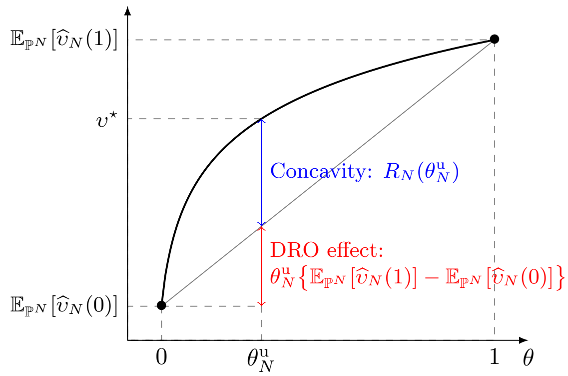

Figure 3: Illustration of the SAA bias reduction effect. The solid curve represents the concave function $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta)]$.

Proposition 3 shows that the bias of the TRO estimator $\widehat{\upsilon}_N(\theta)$ is upper bounded by two terms in (17). The first term $\theta\{\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(1)] - \mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)]\}$ is due to the DRO objective component in the TRO model, which is increasing with θ and is zero when $\theta = 0$. The second term $R_N(\theta)$ is a consequence of the concavity of $\widehat{\upsilon}_N$ (see Theorem 3). Thanks to the concavity, $R(\theta)$ is non-negative when $\theta \in (0, 1)$. Thus, Proposition 3 suggests that the bias of $\widehat{\upsilon}_N(\theta)$ may not be a downward bias as that of the SAA estimator. Indeed, in the next theorem, we show that $\widehat{\upsilon}_N(\theta)$ is an unbiased estimator of $\upsilon^\star$ for some choice of θ under a mild assumption.

Theorem 4. *Suppose that Assumption 1 holds and $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(1)] \geq \upsilon^\star$. Then, there exists $\theta_N^{\mathrm{u}} \in [0, 1]$ such that $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta_N^{\mathrm{u}})] = \upsilon^\star$, i.e., $\widehat{\upsilon}_N(\theta_N^{\mathrm{u}})$ is an unbiased estimator of $\upsilon^\star$.*

Corollary 1. *Under the same assumptions as in Theorem 4, there exists $\theta_N^u \in [0, 1]$ such that $|\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta)] - \upsilon^\star| \leq |\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)] - \upsilon^\star|$ for all $\theta \in [0, \theta_N^u]$, i.e., the bias of $\widehat{\upsilon}_N(\theta)$ is not greater than the bias of the SAA estimator $\widehat{\upsilon}_N(0)$.*

We illustrate the results of Theorem 4 and Corollary 1 in Figure 3. First, in line with Theorem 4, this figure shows that when $\upsilon^\star$ is between $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)]$ and $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(1)]$, one can find $\theta_N^{\mathrm{u}} \in [0, 1]$ for which $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta_N^{\mathrm{u}})] = \upsilon^\star$. Here, θ_N^{u} is a constant that depends only on the sample size N. Second, any choice of $\theta \in [0, \theta_N^{\mathrm{u}}]$ leads to an estimator $\widehat{\upsilon}_N(\theta)$ that has a smaller bias than the SAA estimator as indicated by Corollary 1. Finally, this figure shows that $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta)]$ is the sum of three terms (see Proposition 3) (a) the expected value of the SAA estimator $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)]$; (b) the DRO effect $\theta\{\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(1)] - \mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)]\}$; and (c) the concavity effect $R_N(\theta)$. Specifically, the DRO effect adjusts $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)]$ up to the linear combination of the two endpoints of $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(\theta)]$, while the concavity effect accounts for the remaining difference. In practice, finding θ_N^{u} that leads to the unbiased estimator $\widehat{\upsilon}_N(\theta_N^{\mathrm{u}})$ as shown in Theorem 4 is difficult. However, Corollary 1 suggests that the bias of $\widehat{\upsilon}_N(\theta_N^{\mathrm{u}})$ is smaller than that of the SAA estimator for sufficiently small θ. Therefore, we could always choose a small θ in our TRO model to produce an estimator with a smaller bias.

Remark 1. In Theorem 4, we assume that $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] \geq v^\star$. This assumption will likely hold when the ambiguity set $\mathcal{P}_N$ is rich enough such that it contains a wide range of distributions. For example, if the (data-driven) ambiguity set $\mathcal{P}_{N,\alpha}$ satisfies $\mathbb{P}^N(\mathbb{P}^\star \in \mathcal{P}_N) \geq 1-\alpha$ for some $\alpha \in (0,1)$ (see, e.g., Delage and Ye, 2010; Mohajerin Esfahani and Kuhn, 2018), the assumption would probably hold by choosing $\mathcal{P}_{N,\alpha}$ with α close to zero.

Note that under Assumption 1, if $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] > v^\star$, then θ_N^{u} is unique by concavity of $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)]$. Therefore, without loss of generality, we assume that θ_N^{u} is unique in the following discussions. In the case when $\Theta_N := \{\theta \in [0,1] \mid \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)] = v^\star\}$ is not a singleton, we consider θ_N^{u} as one of the elements in Θ_N. Next, we analyze the asymptotic behavior of θ_N^{u}. In Theorem 5, we prove the convergence of θ_N^{u} as $N \to \infty$ and derive the rate of convergence under mild assumptions.

Theorem 5. *Suppose that Assumption 1 and the following hold.*

(a) The samples $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$ are i.i.d. following the true distribution $\mathbb{P}^\star$.

(b) There exists $\boldsymbol{x}_0 \in \mathcal{X}$ such that $\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}_0, \boldsymbol{\xi})^2] < \infty$.

(c) There exists a measurable function $\kappa : \boldsymbol{\xi} \to \mathbb{R}_+$ such that $\mathbb{E}_{\mathbb{P}^\star}[\kappa(\boldsymbol{\xi})^2] < \infty$ and

$$\left| f(\boldsymbol{x}_1, \boldsymbol{\xi}) - f(\boldsymbol{x}_2, \boldsymbol{\xi}) \right| \leq \kappa(\boldsymbol{\xi}) \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$$

for any $\{\boldsymbol{x}_1, \boldsymbol{x}_2\} \subseteq \mathcal{X}$ and $\mathbb{P}^\star$-almost everywhere $\boldsymbol{\xi}$.

If $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] \geq v^\star$ for all $N \in \mathbb{N}$ and $\inf_{N\in\mathbb{N}} \{\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] - \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)]\} > 0$, then $\theta_N^u = o(\sqrt{\log\log N}/\sqrt{N})$.

Theorem 5 shows that θ_N^{u} converges to zero at a rate of $o(\sqrt{\log\log N}/\sqrt{N})$. This is not surprising since when the size parameter θ converges to zero, the objective function of our TRO model reduces to the SAA objective, and the bias of the SAA estimator $|\mathbb{E}_{\mathbb{P}^N}[\widehat{v}(0)] - v^\star|$ is of order $o(\sqrt{\log\log N}/\sqrt{N})$ (Banholzer et al., 2022). Thus, to construct an unbiased estimator $\widehat{v}_N(\theta_N)$ of $v^\star$, one could use the convergence order as a criterion for choosing the size parameter θ_N. Next, under a stronger assumption widely adopted in the stochastic optimization literature (see, e.g., Homem-de Mello and Bayraksan, 2014; Kim et al., 2015; Shapiro et al., 2014), we can derive an improved convergence rate for θ_N^{u} in Theorem 6.

Theorem 6. *In addition to the assumptions in Theorem 5, suppose that the sequence $\{X_N := \sqrt{N}(\widehat{v}_N(0) - v^\star)\}_{N\in\mathbb{N}}$ is asymptotic uniformly integrable, i.e., $\lim_{M\to\infty} \limsup_{N\to\infty} \mathbb{E}\big(|X_N| \mathbf{1}(|X_N| > M)\big) = 0$. If $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] \geq v^\star$ for all $N \in \mathbb{N}$ and $\inf_{N\in\mathbb{N}} \{\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] - \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)]\} > 0$, then*

1. *$\theta_N^u = o(1/\sqrt{N})$ when the optimal solution to (1) is unique;*

2. *$\theta_N^u = O(1/\sqrt{N})$ when there are multiple optimal solutions to (1).*

We close this section by highlighting that some recent studies in the stochastic optimization literature have also proposed approaches to construct estimators of the true optimal value $v^\star$ with smaller bias than the SAA estimator, see, e.g., Blanchet et al. (2019); Dentcheva and Lin (2022). Other studies focused on obtaining confidence intervals for $v^\star$ (see, e.g., Bertsimas et al., 2018; Duchi et al., 2021; Lam, 2022). Our proposed TRO model and results in this section add to the first stream of literature on alternative approaches to constructing estimators with smaller biases.

4.2. Generalization Bound

In this section, we analyze the generalization error (also known as the out-of-sample disappointment) of the TRO estimator. This error quantifies how much the model generalizes in the (unseen) out-of-sample setting. Specifically, let $\boldsymbol{x}_N(\theta) \in \widehat{\mathcal{X}}_N(\theta)$ be an optimal solution to the TRO model. The generalization error (GE) of the TRO model is given by $\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}_N(\theta), \boldsymbol{\xi})] - \sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}_N(\theta), \boldsymbol{\xi})]$, i.e., the difference between the actual expected cost associated with implementing $\boldsymbol{x}_N(\theta)$ under the true unknown distribution $\mathbb{P}^\star$ and the cost estimated by the TRO model. A positive GE indicates that the TRO model is overly optimistic. As in the literature, we quantify the GE of our TRO model via the following probability:

$$\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] - \sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\} > \delta\right) \tag{18}$$

for some $\delta > 0$ (Duchi and Namkoong, 2021; Gao, 2022; Liu et al., 2023b). Note that (18) provides an upper bound on the probability that the GE is greater than δ. Thus, our goal is to derive an upper bound on (18) that decays rapidly (e.g., exponentially) when N increases. To this end, in Theorem 7, we first derive an upper bound on (18).

Theorem 7. *Assume that the generalization error of the SAA model satisfies*

$$\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\} > \delta\right) \le \alpha_{N,1} \tag{19}$$

and the generalization error of the DRO model with ambiguity set $\mathcal{P}_N$ satisfies

$$\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] - \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\} > \delta\right) \le \alpha_{N,2} \tag{20}$$

for some constants $\alpha_{N,1} = \alpha_{N,1}(\delta) > 0$ and $\alpha_{N,2} = \alpha_{N,2}(\delta) > 0$ (which may depend on the complexity of the function class $\mathcal{H} = \{f(\boldsymbol{x},\cdot) \mid \boldsymbol{x}\in\mathcal{X}\}$). Then, we have

$$\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] - \sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\} > \delta\right) \le \alpha_{N,1} + \alpha_{N,2}. \tag{21}$$

Theorem 7 shows that the probability (18) (i.e., GE) is upper bounded by the sum of the upper bounds on the GE of the SAA and DRO models. This is reasonable because the objective of our

TRO model is a convex combination of the SAA and DRO objectives. The generalization bound (19) for the SAA model is well-known in the literature. Specifically, under some mild assumptions, there exist constants $C_1 > 0$ and $\beta_1 = \beta_1(\delta) > 0$ such that (19) holds with $\alpha_{N,1} = C_1 \exp\{-N\beta_1\}$ (see, e.g., Theorem 7.73 of Shapiro et al., 2014). Similarly, the generalization bound (20) for the DRO model has been widely studied in the literature under various choices of the shape parameter $\mathcal{P}_N$. For example, for specific moment-based ambiguity sets (Delage and Ye, 2010) and distance-based ambiguity sets (Duchi and Namkoong, 2021; Liu et al., 2023b; Mohajerin Esfahani and Kuhn, 2018; Van Parys et al., 2021), there exist constants $C_2 > 0$ and $\beta_2 = \beta_2(\delta) > 0$ such that (20) holds with $\alpha_{N,2} = C_2 \exp\{-N\beta_2\}$. Hence, if we construct the TRO ambiguity set using one of these shape parameters $\mathcal{P}_N$, we can apply Theorem 7 to obtain an exponentially decaying bound on the probability (18)

$$\begin{aligned}\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] - \sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta}}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\} > \delta\right) &\le C_1\exp\{-N\beta_1\} + C_2\exp\{-N\beta_2\}\\ &\le C\exp\{-N\beta\},\end{aligned}$$

where $C = C_1 + C_2$ and $\beta = \min\{\beta_1, \beta_2\}$. This shows that the generalization error of our TRO model has an exponentially decaying tail when the TRO ambiguity set is constructed using some specific choice of the shape parameter.

5. Asymptotic Properties

In this section, we analyze the asymptotic properties of our TRO model. Specifically, in Section 5.1, we show the almost sure convergence of the optimal value $\widehat{v}_N(\theta_N)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta_N)$ of the TRO problem (4) to their true counterparts when $N \to \infty$. In Section 5.2, we derive the asymptotic distribution of $\widehat{v}_N(\theta_N)$ when $N \to \infty$.

5.1. Asymptotic Convergence

In this section, we show the almost sure convergence of the optimal value $\widehat{v}_N(\theta_N)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta_N)$ of our TRO model respectively to the true optimal value $v^\star$ and set of optimal solutions $\mathcal{X}^*$ of problem (1) when $N \to \infty$ (Theorem 8). Our proof idea of these convergence properties leverages the fact that the objective function of our TRO model has two components: $\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]$ and $\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]$. Hence, to establish the desired convergence results, we want to ensure that $\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]$ converges to $\mathbb{E}_{\widehat{\mathbb{P}}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]$ and $\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]$ does not diverge (to infinity) as $N \to \infty$ (Lemma 1). First, we adopt the following standard assumption that guarantees the uniform convergence of $\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]$ over $\boldsymbol{x}\in\mathcal{X}$. In Example 5, we provide two sufficient conditions commonly employed in the literature for this assumption to hold.

Assumption 2. The function class $\mathcal{H} := \{f(\boldsymbol{x},\cdot) : \Xi \to \mathbb{R} \mid \boldsymbol{x}\in\mathcal{X}\}$ is $\mathbb{P}^\star$-Glivenko-Cantelli, i.e.,

$$\|\widehat{\mathbb{P}}_N - \mathbb{P}^\star\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N h(\widehat{\boldsymbol{\xi}}_i) - \mathbb{E}_{\mathbb{P}^\star}[h(\boldsymbol{\xi})]\right| \to 0$$

almost surely as $N \to \infty$.

Example 5. Suppose the samples $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$ are i.i.d. generated from the true distribution $\mathbb{P}^\star$. The following are two sufficient conditions for $\mathcal{H}$ to be $\mathbb{P}^\star$-Glivenko-Cantelli.

(a) $\mathcal{X}$ is compact; $f(\cdot, \boldsymbol{\xi})$ is continuous for $\mathbb{P}^\star$-almost every $\boldsymbol{\xi} \in \Xi$; $f(\boldsymbol{x}, \boldsymbol{\xi})$ is dominated by an integrable function for all $\boldsymbol{x} \in \mathcal{X}$ (see Theorem 7.53 of Shapiro et al., 2014).

(b) $\mathcal{X}$ is compact; $f(\cdot, \boldsymbol{\xi})$ is equicontinuous, i.e., for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(\boldsymbol{x}, \boldsymbol{\xi}) - f(\boldsymbol{x}', \boldsymbol{\xi})| \le \varepsilon$ for all $\boldsymbol{\xi} \in \Xi$ and $\boldsymbol{x}' \in \mathcal{X}$ with $\|\boldsymbol{x} - \boldsymbol{x}'\| \le \delta$ (see Lemma EC.5 of Bertsimas and Kallus, 2020).

The second step toward deriving the desired convergence results is to establish a finite upper bound on the DRO component $\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})]$. Note that under Assumption 1, we have

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}, \boldsymbol{\xi})| &\le \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}, \boldsymbol{\xi}) - f(\boldsymbol{x}_0, \boldsymbol{\xi})| + \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}_0, \boldsymbol{\xi})| \\
&\le \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] \cdot \operatorname{diam}(\mathcal{X}) + \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}_0, \boldsymbol{\xi})|, \qquad (22)
\end{aligned}$$

where $\operatorname{diam}(\mathcal{X}) = \sup_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\| < \infty$. From (22), we observe that the choice of ambiguity set $\mathcal{P}_N$ plays a critical role in the finiteness of $\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})]$. In particular, $\mathcal{P}_N$ should be carefully chosen to ensure that $\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})]$ and $\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}_0, \boldsymbol{\xi})|$ in (22) are finite for all $N \in \mathbb{N}$. To this end, we make the following assumptions on the objective function and/or the sequence of ambiguity sets $\{\mathcal{P}_N\}_{N \in \mathbb{N}}$.

Assumption 3. Either the objective function is uniformly bounded, i.e., $\sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{\xi} \in \Xi} |f(\boldsymbol{x}, \boldsymbol{\xi})| < \infty$, or the sequence of ambiguity sets $\{\mathcal{P}_N\}_{N \in \mathbb{N}}$ satisfies one of the following conditions.

(a) There exists $\widehat{\mathcal{P}} \subseteq \mathcal{P}(\Xi)$ (independent of the data) such that $\mathcal{P}_N \subset \widehat{\mathcal{P}}$ almost surely for sufficiently large N with $\sup_{\mathbb{P} \in \widehat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}_0, \boldsymbol{\xi})| < \infty$ and $\sup_{\mathbb{P} \in \widehat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} |\kappa(\boldsymbol{\xi})| < \infty$.

(b) There exists $\widehat{\mathcal{P}} \subseteq \mathcal{P}(\Xi)$ (independent of the data) such that $\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) \to 0$ almost surely as $N \to \infty$ with $\sup_{\mathbb{P} \in \widehat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} |f(\boldsymbol{x}_0, \boldsymbol{\xi})| < \infty$ and $\sup_{\mathbb{P} \in \widehat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} |\kappa(\boldsymbol{\xi})| < \infty$.

(c) For ambiguity sets $\mathcal{P}_N$ containing only distributions with support on the i.i.d. sample $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$, which can be identified as a vector $\boldsymbol{p} \in \mathbb{R}^N$, we have $\limsup_{N \to \infty} \sup_{\mathbb{P} \in \mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_\infty < \infty$ almost surely with $\mathbb{E}_{\mathbb{P}^\star} |f(\boldsymbol{x}_0, \boldsymbol{\xi})| < \infty$ and $\mathbb{E}_{\mathbb{P}^\star} |\kappa(\boldsymbol{\xi})| < \infty$.

The uniform boundedness assumption on f has been widely adopted in the literature and holds valid in various real applications, for example, when $f(\boldsymbol{x}, \boldsymbol{\xi})$ represents the cost of action $\boldsymbol{x}$ under scenario $\boldsymbol{\xi}$, and both the sets $\mathcal{X}$ and Ξ are compact. Next, in Examples 6–10, we provide sequences $\{\mathcal{P}_N\}_{N \in \mathbb{N}}$ that satisfy Assumptions 3(a)–(c).

Example 6. If $\mathcal{P}_N = \widehat{\mathcal{P}}$ for some fixed $\widehat{\mathcal{P}} \subseteq \mathcal{P}(\Xi)$, then Assumption 3(a) holds immediately. For example, in robust optimization models, $\widehat{\mathcal{P}}$ may contain Dirac measures on a set of scenarios $\Xi' \subseteq \Xi$, i.e., $\widehat{\mathcal{P}} = \{\delta_{\boldsymbol{\xi}'} \mid \boldsymbol{\xi}' \in \Xi'\}$.

Example 7. Consider the moment-based ambiguity set of the form (7) with dimension $d = 1$. If μ_i^N for $i \in \{1, \ldots, q\}$ are the sample estimates, then $\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) \to 0$ almost surely under mild conditions, where $\widehat{\mathcal{P}}$ is the moment-based ambiguity set with true moments μ_i (see, e.g., Sun and Xu, 2016); thus, Assumption 3(b) holds.

Example 8. Consider the distance-based ambiguity set $\mathcal{P}_N = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P}, \widehat{\mathbb{P}}_N) \leq r\}$ for some radius $r > 0$. Here, d is a statistical distance satisfying (i) $\mathsf{d}(\mathbb{P}, \mathbb{P}) = 0$ for any $\mathbb{P} \in \mathcal{P}(\Xi)$, (ii) $\mathsf{d}(\mathbb{P}_1, \mathbb{P}_2) = \mathsf{d}(\mathbb{P}_2, \mathbb{P}_1)$ for any $\{\mathbb{P}_1, \mathbb{P}_2\} \subseteq \mathcal{P}(\Xi)$, (iii) $\mathsf{d}(\mathbb{P}_1, \mathbb{P}_2) \leq \mathsf{d}(\mathbb{P}_1, \mathbb{P}_3) + \mathsf{d}(\mathbb{P}_3, \mathbb{P}_2)$ for any $\{\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3\} \subseteq \mathcal{P}(\Xi)$, and (iv) d is convex in the first argument. These properties hold, for example, if d is the Wasserstein metric. Let $\widehat{\mathcal{P}} = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P}, \mathbb{P}^\star) \leq r\}$. If $\sup_{\mathbb{P}_1 \in \mathcal{P}(\Xi), \mathbb{P}_2 \in \mathcal{P}(\Xi)} \mathbb{d}(\mathbb{P}_1, \mathbb{P}_2) < \infty$ and $\mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star) \to 0$ almost surely, then $\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) \to 0$ almost surely as $N \to \infty$ (see Appendix D for a proof); thus, Assumption 3(b) holds.

Example 9. Consider the ϕ-divergence ambiguity set $\mathcal{P}_N$ in (8) with radius r/N, where we note that $\mathcal{P}_N$ only consists of distributions with support $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$. Under some smoothness conditions on ϕ such as differentiability, Lemma 13 in Duchi et al. (2021) shows that $\sup_{N \in \mathbb{N}} \sup_{\mathbb{P} \in \mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_2 \leq \sqrt{rC_\phi}$ almost surely for some constant C_ϕ that depends on ϕ only. Since $\|\cdot\|_\infty \leq \|\cdot\|_2$, our Assumption 3(c) holds.

Example 10. Consider the total variation ambiguity set $\mathcal{P}_N$ by setting $\phi(t)$ as the the non-differentiable function $|t-1|$ in (8) with radius r/N, i.e., $\mathcal{P}_N = \big\{\mathbb{P} = \sum_{i=1}^N p_i \delta_{\widehat{\boldsymbol{\xi}}_i} \mid \boldsymbol{p} \in \mathbb{R}_+^N, \boldsymbol{p}^\top \mathbf{1} = 1, \|\boldsymbol{p} - N^{-1}\mathbf{1}\|_1 \leq r/N\big\}$. From the definition of $\mathcal{P}_N$, we immediately have $\sup_{\mathbb{P} \in \mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_1 \leq r$ almost surely for all $N \in \mathbb{N}$, implying that Assumption 3(c) holds.

In Lemma 1, we show that $\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_\mathbb{P}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ is upper bounded for sufficiently large N under our Assumptions 1 and 3.

Lemma 1. *Under Assumptions 1 and 3, we have* $\limsup_{N \to \infty} \sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_\mathbb{P}|f(\boldsymbol{x}, \boldsymbol{\xi})| \leq M$ *almost surely for some constant M.*

With Lemma 1, we are ready to prove the asymptotic convergence of our TRO model. Specifically, in Theorem 8, we show that, for any sequence $\{\theta_N\}_{N \in \mathbb{N}}$ converging to zero, the optimal value $\widehat{\upsilon}_N(\theta_N)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta_N)$ of our TRO model converges almost surely to the true optimal value $\upsilon^\star$ and the set of optimal solutions $\mathcal{X}^\star$ to (1), respectively.

Theorem 8. *In addition to Assumptions 1–3, suppose that $\theta_N = o(1)$. Then, almost surely, as $N \to \infty$, we have (i) $\widehat{\upsilon}_N(\theta_N) \to \upsilon^\star$, and (ii) $D\big(\widehat{\mathcal{X}}_N(\theta_N), \mathcal{X}^\star\big) \to 0$.*

5.2. Asymptotic Distribution

We now derive the asymptotic distribution of $\widehat{\upsilon}_N(\theta_N)$. We use the following additional notation in our derivations. Let $L^2(\mathbb{P}^\star)$ be the set of $\mathbb{P}^\star$-square-integrable functions equipped with the norm

$\|h\|_{L^2(\mathbb{P}^\star)} = \sqrt{\mathbb{E}_{\mathbb{P}^\star}[h(\boldsymbol{\xi})^2]}$ for $h \in L^2(\mathbb{P}^\star)$. For a class of functions $\mathcal{H} \subseteq L^2(\mathbb{P}^\star)$, let $\ell^\infty(\mathcal{H}) = \{g : \mathcal{H} \to \mathbb{R} \mid \|g\|_{\mathcal{H}} < \infty\}$ be the space of bounded functions, where $\|g\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} |g(h)|$ is the sup-norm of $g \in \ell^\infty(\mathcal{H})$. Following the convention in empirical process theory, we use the shorthand notation $\mathbb{P}(h) := \mathbb{E}_{\mathbb{P}}(h)$ for $h \in \mathcal{H}$. We adopt the following standard assumption on the complexity of the objective function class (see, e.g., Eichhorn and Römisch, 2007; Lam, 2019, 2022).

Assumption 4. The function class $\mathcal{H} := \{f(\boldsymbol{x}, \cdot) : \Xi \to \mathbb{R} \mid \boldsymbol{x} \in \mathcal{X}\}$ is $\mathbb{P}^\star$-Donsker, i.e., $\sqrt{N}(\widehat{\mathbb{P}}_N - \mathbb{P}^\star) \Rightarrow \mathbb{G}$ in $\ell^\infty(\mathcal{H})$, where "$\Rightarrow$" denotes weak convergence and $\mathbb{G}$ is a tight Gaussian process indexed by $\mathcal{H}$ with mean zero and covariance function $\mathrm{Cov}(\mathbb{G}(h_1), \mathbb{G}(h_2)) = \mathrm{Cov}_{\mathbb{P}^\star}(h_1(\boldsymbol{\xi}), h_2(\boldsymbol{\xi}))$.

In Assumption 4, the measures $\widehat{\mathbb{P}}_N$ and $\mathbb{P}^\star$ are considered as elements in $\ell^\infty(\mathcal{H})$. Example 11 provides a sufficient condition for Assumption 4 to hold.

Example 11. Suppose that $\{\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \dots\}$ is i.i.d. following $\mathbb{P}^\star$, $\mathcal{X}$ is compact, $\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}, \boldsymbol{\xi})] < \infty$, and $\mathrm{Var}_{\mathbb{P}^\star}(f(\boldsymbol{x}, \boldsymbol{\xi})) < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$. If there exists a measurable function $\kappa : \boldsymbol{\xi} \to \mathbb{R}_+$ such that $f(\cdot, \boldsymbol{\xi})$ is Lipschitz with modulus $\kappa(\boldsymbol{\xi})$ almost surely with $\mathbb{E}_{\mathbb{P}^\star}[\kappa(\boldsymbol{\xi})^2] < \infty$, then $\mathcal{H}$ is $\mathbb{P}^\star$-Donsker (Lam, 2019). This sufficient condition is widely adopted in the literature to derive asymptotic distribution in stochastic optimization problems (Guigues et al., 2018; Shapiro et al., 2014).

A key step toward deriving the asymptotic distribution of $\widehat{v}_N(\theta_N)$ is to show that $\mathbb{S}_N := \sqrt{N}\big[(1-\theta_N)\widehat{\mathbb{P}}_N + \theta_N \mathbb{P}_N - \mathbb{P}^\star\big]$ converges weakly to some tight Gaussian process $\mathbb{G}'$ in $\ell^\infty(\mathcal{H})$ for any $\mathbb{P}_N \in \mathcal{P}_N$. Here, $(1-\theta_N)\widehat{\mathbb{P}}_N + \theta_N \mathbb{P}_N$ is a probability measure in the TRO ambiguity set. Lemma 2 establishes the desired convergence under some mild assumptions.

Lemma 2. *Let $\{\mathbb{P}_N\}_{N\in\mathbb{N}}$ be any sequence of probability measure $\mathbb{P}_N \in \mathcal{P}_N$. In addition to Assumptions 1 and 3, suppose that*

(a) there exists a square-integrable envelope H of the function class $\mathcal{H}$, i.e., $h(\boldsymbol{\xi}) \le H(\boldsymbol{\xi})$ for all $h \in \mathcal{H}$ with $\mathbb{E}_{\mathbb{P}^\star}[H(\boldsymbol{\xi})^2] < \infty$, and

(b) $\theta_N = o(N^{-1/2})$.

Then, as $N \to \infty$, the process $\mathbb{S}_N := \sqrt{N}[(1-\theta_N)\widehat{\mathbb{P}}_N + \theta_N \mathbb{P}_N - \mathbb{P}^\star] \Rightarrow \mathbb{G}'$ in $\ell^\infty(\mathcal{H})$, where $\mathbb{G}'$ is a tight Gaussian process indexed by $\mathcal{H}$ with mean zero and covariance function $\mathrm{Cov}(\mathbb{G}'(h_1), \mathbb{G}'(h_2)) = \mathrm{Cov}_{\mathbb{P}^\star}(h_1(\boldsymbol{\xi}), h_2(\boldsymbol{\xi}))$.

With Lemma 2, we are ready to derive the asymptotic distribution of $\widehat{v}_N(\theta_N)$. Specifically, in Theorem 9, we show that $\sqrt{N}(\widehat{v}_N(\theta_N) - v^\star)$ converges to the infimum of some Gaussian process indexed by $\mathcal{X}$. For notational simplicity, we write $\widehat{v}_N = \widehat{v}_N(\theta_N)$.

Theorem 9. *In addition to assumptions in Lemma 2, suppose that Assumption 4 holds and there exists a worst-case distribution $\mathbb{P}_N^\star \in \arg\max_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ such that $\mathbb{P}_N^\star \in \mathcal{P}_N$ for any $\boldsymbol{x} \in \mathcal{X}$. Then, as $N \to \infty$,* *(i)* $\sqrt{N}(\widehat{v}_N - v^\star) \Rightarrow \inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})$*, where* $\mathbb{G}$ *is a tight Gaussian process indexed by* $\mathcal{X}$ *with mean zero and covariance function* $\mathrm{Cov}(\mathbb{G}(\boldsymbol{x}_1)), \mathbb{G}(\boldsymbol{x}_2)) = \mathrm{Cov}_{\mathbb{P}^\star}(f(\boldsymbol{x}_1,\boldsymbol{\xi}), f(\boldsymbol{x}_2,\boldsymbol{\xi}))$*;*

(ii) $\widehat{v}_N - v^\star = \inf_{\boldsymbol{x}\in\mathcal{X}^\star} \Big\{(1-\theta_N)\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})] + \theta_N \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]\Big\} + o_{\mathbb{P}^\star}(N^{-1/2})$*.*

In particular, $\widehat{v}_N = \inf_{\boldsymbol{x}\in\mathcal{X}^\star} \Big\{(1-\theta_N)\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})] + \theta_N \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\Big\} + o_{\mathbb{P}^\star}(N^{-1/2})$*.*

It follows from Theorem 9 that when set of optimal solutions to (1) is a singleton, say $\mathcal{X}^\star = \{\boldsymbol{x}^\star\}$, $\sqrt{N}(\widehat{v}_N - v^\star) \Rightarrow N\big(0, \mathrm{Var}_{\mathbb{P}^\star}(f(\boldsymbol{x}^\star,\boldsymbol{\xi}))\big)$, i.e., a normal distribution with mean zero and variance $\mathrm{Var}_{\mathbb{P}^\star}(f(\boldsymbol{x}^\star,\boldsymbol{\xi}))$.

Our asymptotic convergence results hold for TRO models with TRO ambiguity sets constructed using general shape parameters $\mathcal{P}_N$, such as moment- and distance-based ambiguity sets. This differs from results in the existing literature focusing on a specific ambiguity set. In the special case where the shape parameter is chosen as a distance-based ambiguity set $\mathcal{P}_{N,r_N} = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P},\widehat{\mathbb{P}}_N) \le r_N\}$, we can recover the asymptotics of the optimal value of classical distance-based DRO models (see, e.g., Blanchet et al., 2021; Blanchet and Shapiro, 2023). Specifically, in Theorem 10, we derive the asymptotic distribution of $\widehat{v}_N$ under three different convergence rates of the size parameter θ_N and the radius r_N in the shape parameter $\mathcal{P}_{N,r_N}$.

Theorem 10. *Let* $\mathbb{G}$ *be a tight Gaussian process indexed by* $\mathcal{X}$ *with mean zero and covariance function* $\mathrm{Cov}(\mathbb{G}(\boldsymbol{x}_1)), \mathbb{G}(\boldsymbol{x}_2)) = \mathrm{Cov}_{\mathbb{P}^\star}(f(\boldsymbol{x}_1,\boldsymbol{\xi}), f(\boldsymbol{x}_2,\boldsymbol{\xi}))$*. Also, let* $\varepsilon_N(\boldsymbol{x})$ *represent any term that converges to zero in probability uniformly over* $\boldsymbol{x} \in \mathcal{X}$*, i.e.,* $\sup_{\boldsymbol{x}\in\mathcal{X}} |\varepsilon_N(\boldsymbol{x})| = o_{\mathbb{P}^\star}(1)$*. In addition to Assumptions 1 and 4, suppose we construct the TRO ambiguity set using a distance-based shape parameter* $\mathcal{P}_{N,r_N} = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P},\widehat{\mathbb{P}}_N) \le r_N\}$ *and that* $\widehat{v}_N(1)$ *exhibits the following expansion*

$$\sup_{\mathbb{P}\in\mathcal{P}_{N,r_N}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] = \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})] + r_N^\gamma g_N(\boldsymbol{x}) + r_N^\gamma \varepsilon_N(\boldsymbol{x}) \tag{23}$$

for some $\gamma > 0$*, where* $g_N(\boldsymbol{x})$ *satisfies* $g_N(\boldsymbol{x}) = h(\boldsymbol{x}) + \varepsilon_N(\boldsymbol{x})$ *for some continuous deterministic process* $h(\boldsymbol{x})$*. Then, the following assertions hold.*

(i) If $\theta_N r_N^\gamma = o(N^{-1/2})$*, then* $\sqrt{N}(\widehat{v}_N - v^\star) \Rightarrow \inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})$*.*

(ii) If $\theta_N r_N^\gamma = N^{-1/2}$*, then* $\sqrt{N}(\widehat{v}_N - v^\star) \Rightarrow \inf_{\boldsymbol{x}\in\mathcal{X}^\star}\{\mathbb{G}(\boldsymbol{x}) + h(\boldsymbol{x})\}$*.*

(iii) If $o(\theta_N r_N^\gamma) = N^{-1/2}$*, then* $\theta_N^{-1} r_N^{-\gamma}(\widehat{v}_N - v^\star) \Rightarrow \inf_{\boldsymbol{x}\in\mathcal{X}^\star} h(\boldsymbol{x})$*.*

Theorem (10) demonstrates that when the TRO ambiguity is constructed using a distance-based shape parameter $\mathcal{P}_{N,r}$ such that $\widehat{v}_N(1) = \sup_{\mathbb{P}\in\mathcal{P}_{N,r_N}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]$ satisfies (23), the asymptotic distribution of $\widehat{v}_N$ depends on the convergence rate of $\theta_N r_N^\gamma$. In particular, when $r_N = r$ is fixed, Theorem 10 implies that $\widehat{v}_N$ converges to three different distributions depending on the convergence rate of the size parameter θ_N. This resembles the asymptotics of distance-based DRO models, where different convergence rates of the radius r_N lead to different asymptotic distributions of the DRO optimal value (see, e.g., Blanchet et al., 2021; Blanchet and Shapiro, 2023). These results indicate that the size parameter, θ_N, of the TRO model with TRO ambiguity set constructed using the shape parameter $\mathcal{P}_{N,r}$ and the radius r_N in classical distance-based DRO models play a similar role in controlling the asymptotic distribution of the model's optimal value. Finally, we highlight that the expansion (23) is commonly used in the DRO literature to derive asymptotics of the DRO optimal value and is satisfied by popular distance-based ambiguity sets such as ϕ-divergence and Wasserstein ambiguity sets (Blanchet and Shapiro, 2023).

6. Numerical Examples

In this section, we illustrate our theoretical results in the context of two stylized optimization problems: an inventory control problem (Section 6.1) and a portfolio optimization problem (Section 6.2).

6.1. Inventory Control

Consider the classical inventory control problem, where the decision-maker needs to decide the order quantity x of a single item before observing the random demand $\xi \in \mathbb{R}$ (Gallego and Moon, 1993). If x units are ordered, then $\min\{x, \xi\}$ units are sold and $(x - \xi)_+$ units are salvaged. Let c be the per-unit ordering cost, p be the per-unit selling price, and h be the per-unit salvage value. The goal is to minimize the expected ordering cost minus the expected profit and salvage value:

$$\underset{x \geq 0}{\text{minimize}} \quad \mathbb{E}_{\mathbb{P}^\star}[f(x, \xi)] = \mathbb{E}_{\mathbb{P}^\star}[cx - p\min\{x, \xi\} - h(x - \xi)_+],$$

where $\mathbb{P}^\star$ is the true distribution of ξ. Writing $f(x, \xi) = (c-h)x + (p-h)\big[(\xi - x)_+ - \xi\big]$ (see Gallego and Moon, 1993), we formulate the following TRO model for this problem

$$\underset{x \geq 0}{\text{minimize}} \quad (c-h)x + (p-h)\left\{(1-\theta)\frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i\big] + \theta \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+ - \xi\big]\right\}, \tag{24}$$

where $\{\widehat{\xi}_i\}_{i=1}^N$ is the set of samples. Note that $f(x, \xi)$ is Lipschitz in x with Lipschitz constant $c + p - 2h$ (independent of ξ). Moreover, one can impose a practical upper bound on x so that the feasible set $\mathcal{X}$ is compact, and thus, Assumption 1 holds.

We consider the following shape parameters (ambiguity sets) $\mathcal{P}_N$ in our experiment: (a) the mean-variance ambiguity set (Gallego and Moon, 1993), (b) the 1-Wasserstein ambiguity set (Mohajerin Esfahani and Kuhn, 2018), (c) the empirical Burg-entropy divergence ball (Lam, 2019), (d) a set of Dirac measures on points that lie in the confidence interval of ξ (Fabozzi et al., 2007). Table 1 summarizes these sets and the parameter settings we used for each. It is easy to verify that ambiguity sets (a)–(d) satisfy Assumption 3 (see Appendix F.2 for a proof for ambiguity set (d)). In Appendix F.1, we provide tractable reformulations of the TRO model (24) under each set. We adopt similar parameter settings as in Gotoh et al. (2021). Specifically, we set $p = 30$, $c = 2$, and $h = 1$. For illustrative purposes, we use the exponential distribution with mean 50 as the true

Table 1: Shape parameters for the inventory control problem. *Notation:* $\widehat{\mu}_N$ is the sample mean; $\widehat{\sigma}_N^2$ is the sample variance; $\Delta_N \subseteq \mathbb{R}^N$ is the probability simplex; $t_{\nu,\alpha/2}$ is the upper $(1-\alpha)/2$-th quantile of a Student's t-distribution with degree of freedom ν.

	Ambiguity Set $\mathcal{P}_N$	Parameters
(a) Mean-Variance	$\mathcal{P}_N = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}) \mid \mathbb{E}_{\mathbb{P}}(\xi) = \mu,\ \mathrm{Var}_{\mathbb{P}}(\xi) = \sigma^2\}$	$(\mu, \sigma^2) = (\widehat{\mu}_N, \widehat{\sigma}_N^2)$
(b) 1-Wasserstein	$\mathcal{P}_N = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}) \mid W_1(\mathbb{P}, \widehat{\mathbb{P}}_N) \le r\}$	$r = 100$
(c) Burg-Divergence	$\mathcal{P}_N = \{\boldsymbol{p} \in \Delta_N \mid -\frac{1}{N}\sum_{i=1}^N \log(Np_i) \le \frac{r}{N}\}$	$r = 10$
(d) Confidence Interval	$\mathcal{P}_N = \{\delta_\xi \mid \lvert\xi - \widehat{\mu}_N\rvert \le t_{N-1,\alpha/2}\widehat{\sigma}_N/\sqrt{N}\}$	$\alpha = 0.95$

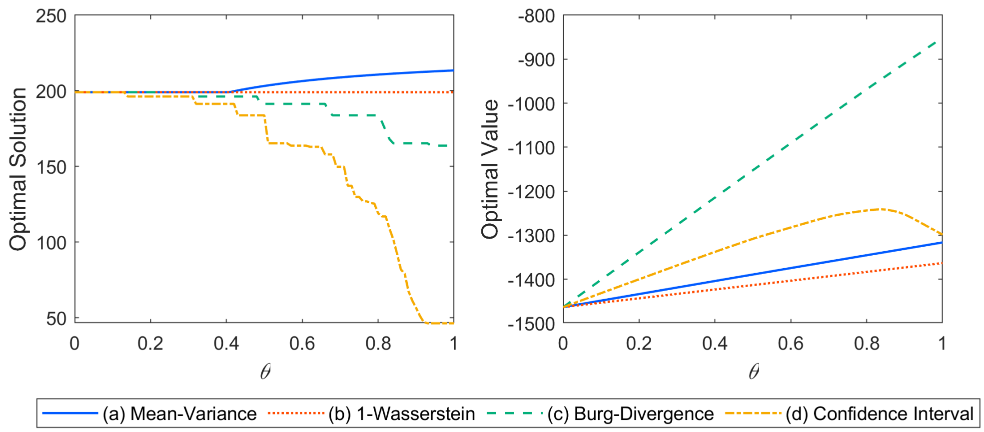

Figure 4: Optimal solution and optimal value for different values of θ in the inventory control problem.

distribution of the demand ξ for generating the data. Under this distribution, we can compute the true optimal value $\upsilon^\star = 1,232$ (Gallego and Moon, 1993).

First, to demonstrate the effect of θ on the optimal solution and value, we solve our TRO model with different values of $\theta \in \{0, 0.01, 0.02, \ldots, 1\}$ and $N = 100$. Figure 4 illustrates the resulting optimal solutions and objective values. Clearly, the optimal value function is concave under shape parameters (a)–(d), which is consistent with Theorem 3. The optimal value function is increasing under sets (a)–(c). increasing on the entire interval [0,1] under set (d). This is because sets (a)–(c) are star-shaped with a star center $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, and thus, the sequence of TRO ambiguity sets constructed using sets (a)–(c) satisfy the hierarchical properties in Theorem 1. In contrast, the optimal value function is first increasing on $\theta \in [0, 0.85]$ and then decreasing. This is because by the construction of the set (d), $\widehat{\mathbb{P}}_N \notin \mathcal{P}_N$, and thus, $\widehat{\mathbb{P}}_N$ is not a star center of $\mathcal{P}_N$. It follows that the sequence of TRO ambiguity sets constructed using set (d) does not satisfy the hierarchical property. We also observe that different choices of the shape parameter $\mathcal{P}_N$ result in different spectra of optimal solutions and values. For example, the TRO model with shape parameter (a) suggests ordering more under a larger θ and suggests ordering less when using shape parameters (c) and (d).

Let us now investigate the bias and standard deviation of the TRO estimator $\widehat{\upsilon}_N(\theta)$. We

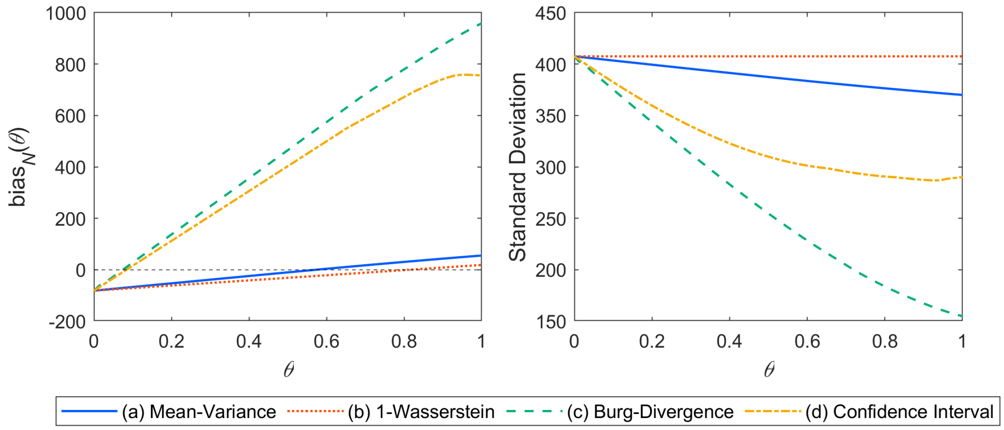

Figure 5: Bias and standard deviation of $\widehat{v}_N(\theta)$ for different values of θ with $N = 10$ in the inventory control problem.

estimate these quantities as follows. First, for each $\theta \in \{0, 0.01, 0.02, \ldots, 1\}$, we solve our TRO model with $R = 1,000$ samples, each consisting of $N = 10$ scenarios. Let $\{\widehat{v}_{N,r}(\theta)\}_{r=1}^R$ denote the resulting optimal values of the TRO model from the R replications under each θ. We compute the mean of $\{\widehat{v}_{N,r}(\theta)\}_{r=1}^R$ as $\texttt{mean}_N(\theta) = R^{-1}\sum_{r=1}^R \widehat{v}_{N,r}(\theta)$. Then, we estimate the bias as $\texttt{bias}_N(\theta) = \texttt{mean}_N(\theta) - v^\star$ and the standard deviation as $\texttt{std}_N(\theta) = R^{-1}\sum_{r=1}^R[\widehat{v}_{N,r}(\theta) - \texttt{mean}_N(\theta)]^2$. Figure 5 presents the estimated bias and standard deviation. As expected, the SAA estimator $\widehat{v}_N(0)$ exhibits a downward bias. In contrast, the TRO estimator $\widehat{v}_N(\theta)$ is an unbiased estimator for some θ. Moreover, when θ is sufficiently small, the absolute bias of $\widehat{v}_N(\theta)$ is smaller than that of the SAA estimator $\widehat{v}_N(0)$. Note that the absolute bias of $\widehat{v}_N(\theta)$ varies across different shape parameters employed in the model. For example, when using sets (a)–(b), the absolute bias of $\widehat{v}_N(\theta)$ is smaller than that of the SAA estimator for $\theta \in (0, 1]$. In contrast, when using sets (c)–(d), the absolute bias of $\widehat{v}_N(\theta)$ is smaller for $\theta \in (0, 0.12]$. These results are consistent with Theorem 4 and Corollary 1. Finally, we observe that the standard deviation of $\widehat{v}_N(\theta)$ decreases when θ increases from zero, demonstrating that the TRO estimator's variability reduces with θ.

Finally, we demonstrate the asymptotic properties of $\widehat{v}_N(\theta_N)$. First, to show the asymptotic convergence of $\widehat{v}_N(\theta_N)$, we solve the TRO model with $N \in \{10, 50, 100, 500, 1000\}$ and $\theta_N = 10/N = o(N^{-1/2})$ (consistent with the rate suggested in Lemma 2) and compute the absolute difference $|\widehat{v}_N(\theta_N) - v^\star|$ between the optimal value of the TRO model $\widehat{v}_N(\theta_N)$ and the true optimal value $v^\star$. Table 2 presents $|\widehat{v}_N(\theta_N) - v^\star|$ for $N \in \{10, 50, 100, 500, 1000\}$. It is clear that $|\widehat{v}_N(\theta_N) - v^\star|$ decreases with N, illustrating the asymptotic convergence of $\widehat{v}_N(\theta_N)$ to $v^\star$ as shown in Theorem 8. Next, to demonstrate the asymptotic distribution of $\widehat{v}_N(\theta_N)$, we solve the TRO model with $R = 1,000$ sets of data, each consisting of N scenarios, to obtain the optimal values $\{\widehat{v}_{N,r}(\theta_N)\}_{r=1}^R$. We then compute the Kolmogorov–Smirnov (KS) statistic that quantifies the difference between the standard normal distribution and the empirical distribution of $\sqrt{N}\big[\widehat{v}_N(\theta) - v^\star\big]/V^\star$ based on the samples $\{\widehat{v}_{N,r}(\theta_N)\}_{r=1}^R$, where $V^\star = \mathrm{Var}_{\mathbb{P}^\star}\big(f(\boldsymbol{x}^\star, \boldsymbol{\xi})\big)$ and $\boldsymbol{x}^\star$ is the (unique) true op-

Table 2: Absolute difference between $\widehat{v}_N(\theta_N)$ and $v^\star$ for different N in the inventory control problem.

	$N = 10$	$N = 50$	$N = 100$	$N = 500$	$N = 1000$
(a) Mean-Variance	150.31	96.11	140.18	24.42	2.79
(b) 1-Wasserstein	264.01	102.63	142.20	24.35	2.76
(c) Burg-Divergence	1004.17	29.20	95.42	21.28	2.03
(d) Confidence Interval	998.49	35.05	123.69	24.41	3.34

Table 3: Kolmogorov–Smirnov statistics for $\sqrt{N}\big[\widehat{v}_N(\theta) - v^\star\big]/V^\star$ with $V^\star = \mathrm{Var}_{\mathbb{P}^\star}\big(f(\boldsymbol{x}^\star,\boldsymbol{\xi})\big)$ for different N in the inventory control problem.

	$N = 10$	$N = 100$	$N = 1000$
(a) Mean-Variance	0.1121	0.0666	0.0312
(b) 1-Wasserstein	0.0681	0.0651	0.0318
(c) Burg-Divergence	0.9071	0.1796	0.0365
(d) Confidence Interval	0.7422	0.1103	0.0269

timal solution. Table 3 presents the KS statistics for $N \in \{10, 100, 1000\}$. The KS statistic converges to zero when N increases, suggesting that $\sqrt{N}\big[\widehat{v}_N(\theta) - v^\star\big]/V^\star$ converges weakly to the standard normal distribution. These observations are consistent with Theorem 9. These results also emphasize that the asymptotic properties of $\widehat{v}_N(\theta_N)$ hold for for general shape parameters.

6.2. *Portfolio Optimization*

Consider the classical mean-risk portfolio optimization problem, where there are n trading assets, and an investor needs to decide the proportion $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ of the total investment amount allocated to each trading asset before observing the random return $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$. Given the portfolio $\boldsymbol{x}$, the corresponding portfolio loss is $-\boldsymbol{x}^\top\boldsymbol{\xi}$. The goal is to minimize the sum of the expected portfolio loss and the portfolio risk measured by conditional value-at-risk (CVaR):

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \quad \beta\, \mathbb{E}_{\mathbb{P}^\star}(-\boldsymbol{x}^\top\boldsymbol{\xi}) + (1-\beta)\, \mathbb{P}^\star\text{-}\,\mathrm{CVaR}_\alpha(-\boldsymbol{x}^\top\boldsymbol{\xi}), \tag{25}$$

where $\mathbb{P}^\star$ is the true distribution of $\boldsymbol{\xi}$, $\beta \in (0,1)$, $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \mathbf{1}^\top\boldsymbol{x} = 1,\, \boldsymbol{x} \geq 0\}$, and

$$\mathbb{P}\text{-}\,\mathrm{CVaR}_\alpha(-\boldsymbol{x}^\top\boldsymbol{\xi}) = \min_t \left\{ t + \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}}\big[(-\boldsymbol{x}^\top\boldsymbol{\xi} - t)_+\big] \right\} \tag{26}$$

for some $\alpha \in (0,1)$ (Rockafellar and Uryasev, 2000). Using (26), we can reformulate (25) into

$$\underset{\boldsymbol{x}\in\mathcal{X},\, t\in\mathbb{R}}{\text{minimize}} \quad \mathbb{E}_{\mathbb{P}^\star}\left[(1-\beta)t + \beta(-\boldsymbol{x}^\top\boldsymbol{\xi}) + \frac{1-\beta}{1-\alpha}(-\boldsymbol{x}^\top\boldsymbol{\xi} - t)_+\right] = \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}, t, \boldsymbol{\xi})], \tag{27}$$

where $f(\boldsymbol{x}, t, \boldsymbol{\xi}) = (1-\beta)t + \beta(-\boldsymbol{x}^\top\boldsymbol{\xi}) + [(1-\beta)/(1-\alpha)](-\boldsymbol{x}^\top\boldsymbol{\xi} - t)^+$. Using (27), we formulate the following TRO model for this problem

$$\underset{\boldsymbol{x}\in\mathcal{X},\, t\in\mathbb{R}}{\text{minimize}} \quad (1-\theta)\,\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}, t, \widehat{\boldsymbol{\xi}}_i) + \theta \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, t, \boldsymbol{\xi})], \tag{28}$$

Table 4: Shape parameters for the portfolio optimization problem. *Notation:* $\widehat{\boldsymbol{\mu}}_N$ is the sample mean; $\widehat{\boldsymbol{\Sigma}}_N$ is the sample covariance matrix; $\Delta_N \subseteq \mathbb{R}^N$ is the probability simplex.

	Ambiguity Set $\mathcal{P}_N$	Parameters
(a) Mean-Variance	$\mathcal{P}_N = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^n) \mid \mathbb{E}_{\mathbb{P}}(\boldsymbol{\xi}) = \boldsymbol{\mu},\, \mathrm{Var}_{\mathbb{P}}(\boldsymbol{\xi}) = \boldsymbol{\Sigma}\}$	$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\widehat{\boldsymbol{\mu}}_N, \widehat{\boldsymbol{\Sigma}}_N)$
(b) 1-Wasserstein	$\mathcal{P}_N = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^n) \mid W_1(\mathbb{P}, \widehat{\mathbb{P}}_N) \le r\}$	$r = 0.1$
(c) Total Variation	$\mathcal{P}_N = \{\boldsymbol{p} \in \Delta_N \mid \lVert \boldsymbol{p} - \frac{1}{N}\mathbf{1} \rVert_1 \le \frac{r}{N}\}$	$r = 100$

where $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$ is the set of samples.

Note that for any $\boldsymbol{x} \in \mathcal{X}$, a minimizer of (26) over $t \in \mathbb{R}$ is the value-at-risk $\mathbb{P}^\star$- $\mathrm{VaR}_\alpha(-\boldsymbol{x}^\top \boldsymbol{\xi}) = \inf\{t \in \mathbb{R} \mid \mathbb{P}^\star(-\boldsymbol{x}^\top \boldsymbol{\xi} \le t) \ge 1 - \alpha\}$ (Shapiro et al., 2014), which is finite whenever $\mathbb{E}_{\mathbb{P}^\star}|\xi_j| < \infty$ for all $j \in \{1, \dots, n\}$. Thus, we can impose a large upper bound on $|t|$ so that the feasible set is compact. Moreover, $f(\boldsymbol{x}, t, \boldsymbol{\xi})$ is Lipschitz in $(\boldsymbol{x}, t)$ with Lipschitz constant depending on $\lVert \boldsymbol{\xi} \rVert_1$ (see Appendix F.4). Thus, Assumption 1 holds if $\sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}} \lVert \boldsymbol{\xi} \rVert_1 < \infty$. This is true for the shape parameters (ambiguity sets) $\mathcal{P}_N$ used in our experiment: (a) the mean-variance ambiguity set, (b) the 1-Wassestein ambiguity set with ℓ_1-norm on $\mathbb{R}^n$, and (c) a ϕ-divergence ball based on total variational distance (Huang et al., 2021). Table 4 summarizes these sets and the parameter settings we used for each. It is straightforward to verify that ambiguity sets (a)–(c) satisfy Assumption 3. In Appendix F.3, we provide tractable reformulations of the TRO model (28) under each set. For illustrative purposes, we consider four major indices (S&P 500, DAX, HSI, and FTSE 1000). As in Yam et al. (2016), we use the multivariate normal distribution as the true distribution of the (monthly) return $\boldsymbol{\xi}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\mu} = \begin{pmatrix} 0.06116 \\ 0.109547 \\ 0.090358 \\ 0.040923 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 0.018632 & 0.020056 & 0.020646 & 0.015213 \\ 0.020056 & 0.034507 & 0.027412 & 0.020652 \\ 0.020646 & 0.027412 & 0.048680 & 0.021663 \\ 0.015213 & 0.020652 & 0.021663 & 0.018791 \end{pmatrix}.$$

These quantities are estimated based on ten-year historical data. Under this distribution, we can compute the true optimal value $v^\star = 0.0719$, where we set $\beta = 0.5$ and $\alpha = 0.95$.

Let us first analyze the optimal value and solution to the TRO model (28) for this problem. Figure 6 illustrates the optimal value for different values of θ. Clearly, the optimal value function is concave, which is consistent with Theorem 3. Also, the optimal value function is increasing. This is because sets (a)–(c) are star-shaped with a star center $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, and thus, the sequence of the TRO ambiguity sets constructed using these sets satisfies the hierarchical properties in Theorem 1. Figure 7 illustrates the optimal proportion invested in each asset obtained from solving the TRO model with different values of $\theta \in \{0, 0.01, 0.02, \dots, 1\}$ and $N = 500$. We again observe that different choices of the shape parameter $\mathcal{P}_N$ result in different spectra of optimal solutions. The TRO model with shape parameter (b) suggests investing equally in each asset under larger θ; in

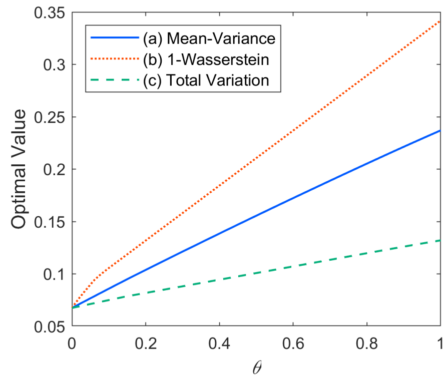

Figure 6: Optimal value for different values of θ in the portfolio optimization problem.

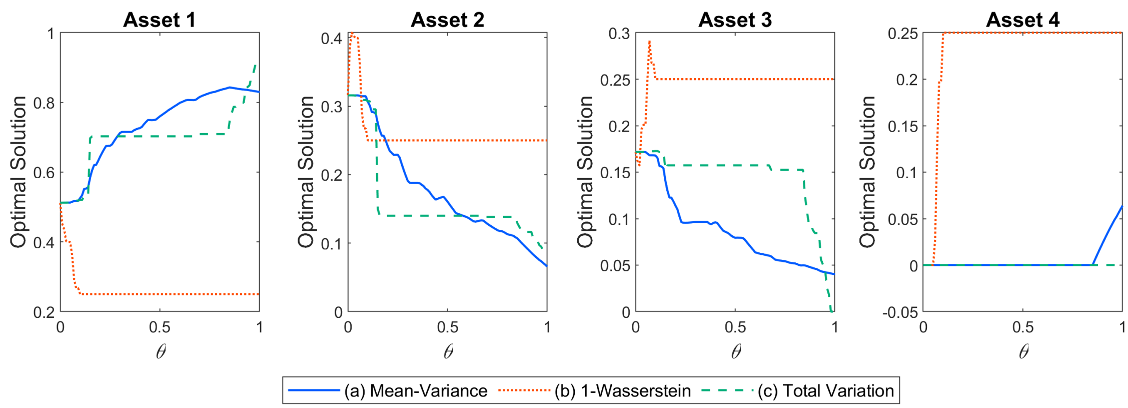

Figure 7: Optimal solution for different values of θ in the portfolio optimization problem.

particular, the optimal proportion invested in each asset converges to 1/4 for sufficiently large θ. In contrast, the TRO model with shape parameters (a) and (c) suggests investing more in assets with smaller variances under a larger θ. For example, the optimal proportion invested in asset 1, which has the smallest variance among the four assets, increases with θ. The optimal proportion invested in asset 4, which has a similar variance but a smaller return compared with asset 1, remains zero for most values of θ. Finally, the optimal proportion invested in assets 2 and 3, which have higher returns and variances compared with asset 1, decreases under a larger θ.

Next, we analyze the bias and standard deviation of the TRO estimator $\widehat{\upsilon}_N(\theta)$ presented in Figure 8. We estimate these quantities as discussed in Section 6.1 with $N = 10$. Again, the SAA estimator $\widehat{\upsilon}_N(0)$ exhibits a downward bias. We also observe that our TRO estimator $\widehat{\upsilon}_N(\theta)$ is an unbiased estimator for some θ. Moreover, the absolute bias of $\widehat{\upsilon}_N(\theta)$ with sufficiently small $\theta > 0$ is smaller than that of $\widehat{\upsilon}_N(0)$. For example, when using set (c), the absolute bias of $\widehat{\upsilon}_N(\theta)$ is smaller for $\theta \in (0, 1]$. These results are consistent with Theorem 4 and Corollary 1. In contrast to the inventory optimization problem, where the standard deviation of $\widehat{\upsilon}_N(\theta)$ decreases with θ, we observe that the standard deviation may increase or decrease, depending on the choice of the shape parameter $\mathcal{P}_N$. With shape parameter (c), the TRO estimator $\widehat{\upsilon}_N(\theta)$ has a larger standard

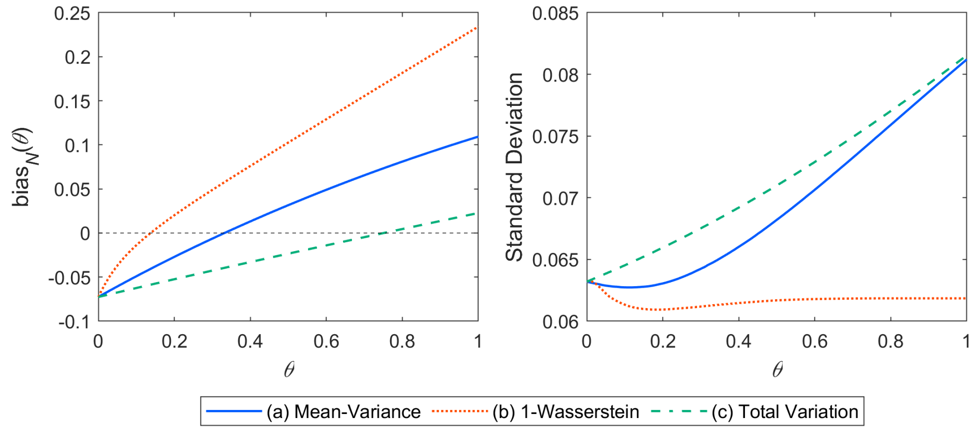

Figure 8: Bias and standard deviation of $\widehat{v}_N(\theta)$ for different values of θ with $N = 10$ in the portfolio optimization problem.

Table 5: Absolute difference between $\widehat{v}_N(\theta_N)$ and $v^\star$ for different N in the portfolio optimization problem.

	$N = 10$	$N = 50$	$N = 100$	$N = 500$	$N = 1000$
(a) Mean-Variance	0.0953	0.0391	0.0239	0.0059	0.0002
(b) 1-Wasserstein	0.2792	0.0117	0.0422	0.0010	0.0035
(c) Total Variation	0.0699	0.0442	0.0204	0.0070	0.0007

deviation compared with the SAA estimator $\widehat{v}_N(0)$ for $\theta \in (0, 1]$. In contrast, with shape parameter (b), $\widehat{v}_N(\theta)$ has a smaller standard deviation than $\widehat{v}_N(0)$ for $\theta \in (0, 1]$. Finally, with shape parameter (a), the standard deviation of $\widehat{v}_N(\theta)$ is smaller than that of $\widehat{v}_N(0)$ for $\theta \in (0, 0.21]$. These results show that for this problem, our TRO model could produce estimators with a smaller bias when the TRO ambiguity set is constructed using shape parameters (a)–(c) and a smaller standard deviation when constructed using shape parameters (a) for small θ and (b) for $\theta \in (0, 1]$.

Finally, we demonstrate the asymptotic properties of $\widehat{v}_N(\theta_N)$. First, we compute the absolute difference $|\widehat{v}_N(\theta_N) - v^\star|$ between the optimal value of the TRO model $\widehat{v}_N(\theta_N)$ and the true optimal value $v^\star$ as discussed in Section 6.1. Table 5 presents $|\widehat{v}_N(\theta_N) - v^\star|$ for $N \in \{10, 50, 100, 500, 1000\}$. Clearly, $|\widehat{v}_N(\theta_N) - v^\star|$ decreases with N, illustrating the asymptotic convergence of $\widehat{v}_N(\theta_N)$ to $v^\star$ as shown in Theorem 8. Second, to demonstrate the asymptotic distribution of $\widehat{v}_N(\theta_N)$, we compute the KS statistic between the standard normal distribution and the empirical distribution of $\sqrt{N}\big[\widehat{v}_N(\theta) - v^\star\big]/V^\star$ as discussed in Section 6.1. Table 6 presents the KS statistics for $N \in \{10, 100, 1000\}$. The KS statistic converges to zero when N increases, suggesting that $\sqrt{N}\big[\widehat{v}_N(\theta) - v^\star\big]/V^\star$ converges weakly to the standard normal distribution. These observations are consistent with Theorem 9.

7. Conclusion

In this paper, we propose and analyze a new TRO approach for modeling uncertainty in optimization problems that serves as a middle ground between the optimistic approach that adopts

Table 6: Kolmogorov–Smirnov statistics for $\sqrt{N}\big[\widehat{v}_N(\theta) - v^\star\big]/V^\star$ with $V^\star = \mathrm{Var}_{\mathbb{P}^\star}\left(f(\boldsymbol{x}^\star, \boldsymbol{\xi})\right)$ for different N in the portfolio optimization problem.

	$N = 10$	$N = 100$	$N = 1000$
(a) Mean-Variance	0.5040	0.1725	0.0683
(b) 1-Wasserstein	0.9275	0.5103	0.2674
(c) Total Variation	0.1144	0.0995	0.0319

a distributional belief and the pessimistic approach that protects against distributional ambiguity. We equip the TRO model with a TRO ambiguity set $\mathcal{P}'_{N,\theta}$ characterized by a size parameter θ controlling the level of optimism and a shape parameter $\mathcal{P}_N$ representing distributional ambiguity, which could be any ambiguity set satisfying mild assumptions. Our theoretical investigations and results include the following. First, we derive necessary and sufficient conditions for $\mathcal{P}'_{N,\theta}$ to satisfy the hierarchical property. Second, we analyze the conservatism of the TRO model and quantify the difference in the optimal value and the set of optimal solutions (and hence conservatism) incurred by perturbation in θ. Moreover, we show how using the TRO model, one can obtain a spectrum of optimal solutions, ranging from optimistic to conservative solutions. Third, we show that our TRO model could produce an unbiased estimator of the true optimal value. Additionally, we derive the generalization bound for the TRO model, demonstrating that the generalization error has an exponentially decaying tail under specific choices of the shape parameter. Finally, we prove the almost sure convergence of the optimal value and the set of optimal solutions of the TRO model to their true counterparts. In addition, we derive the asymptotic distribution of the optimal value to the TRO model. These properties hold for TRO models with TRO ambiguity sets constructed using general shape parameters, such as moment- and distance-based ambiguity sets. We numerically demonstrate these theoretical results using an inventory control problem and a portfolio optimization problem.

Our work opens avenues for further research in various directions. These include extending and identifying properties of the TRO approach for risk-averse settings. Moreover, developing computationally efficient algorithms aimed at obtaining the complete spectrum of optimal solutions of the TRO model would be of significant value in many application domains. Another interesting area is investigating the characteristics of the spectrum of TRO optimal solutions of the TRO model under different shape parameters.

Appendix A. Proofs

Appendix A.1. Proof of Theorem 1

Proof. Proof. First, we prove part (i). Suppose that $\mathcal{P}_N$ is star-shaped and $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$ is a star center of $\mathcal{P}_N$. We show that $\{\mathcal{P}_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the hierarchical property. For any $0 \le \theta_1 < \theta_2 \le 1$ and $\mathbb{Q} \in \mathcal{P}_N$, we have

$$\begin{aligned}(1-\theta_1)\,\widehat{\mathbb{P}}_N + \theta_1\,\mathbb{Q} &= (1-\theta_1)\,\widehat{\mathbb{P}}_N + \theta_2\,\widehat{\mathbb{P}}_N - \theta_2\,\widehat{\mathbb{P}}_N + \theta_1\,\mathbb{Q} \\ &= (1-\theta_2)\,\widehat{\mathbb{P}}_N + (\theta_2-\theta_1)\,\widehat{\mathbb{P}}_N + \theta_1\,\mathbb{Q} \\ &= (1-\theta_2)\widehat{\mathbb{P}}_N + \theta_2\left[\left(1-\frac{\theta_1}{\theta_2}\right)\widehat{\mathbb{P}}_N + \frac{\theta_1}{\theta_2}\mathbb{Q}\right], \end{aligned} \tag{A.1}$$

where $\theta_1/\theta_2 \in [0,1)$. Since $\mathcal{P}_N$ is star-shaped and $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$ is a star center of $\mathcal{P}_N$, the measure $(1-\theta_1/\theta_2)\widehat{\mathbb{P}}_N + (\theta_1/\theta_2)\mathbb{Q}$ in the second term of (A.1) belongs to $\mathcal{P}_N$. It follows from the definition of $\mathcal{P}'_{N,\theta_2}$ that $(1-\theta_1)\,\widehat{\mathbb{P}}_N + \theta_1\,\mathbb{Q} \in \mathcal{P}'_{N,\theta_2}$ for any $\mathbb{Q} \in \mathcal{P}_N$, and thus, $\mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$. This shows that $\{\mathcal{P}_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the hierarchical property.

Now, suppose that $\{\mathcal{P}_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the hierarchical property, i.e., $\mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$ for all $0 \le \theta_1 < \theta_2 \le 1$. We show that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$. For any $\mathbb{Q} \in \mathcal{P}_N$, by (A.1), we have that

$$(1-\theta_1)\,\widehat{\mathbb{P}}_N + \theta_1\,\mathbb{Q} = (1-\theta_2)\widehat{\mathbb{P}}_N + \theta_2\left[\left(1-\frac{\theta_1}{\theta_2}\right)\widehat{\mathbb{P}}_N + \frac{\theta_1}{\theta_2}\mathbb{Q}\right] \in \mathcal{P}'_{N,\theta_2} \tag{A.2}$$

since $(1-\theta_1)\,\widehat{\mathbb{P}}_N + \theta_1\,\mathbb{Q} \in \mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$. By definition of $\mathcal{P}'_{N,\theta_2}$, the inclusion in (A.2) implies that $(1-\theta_1/\theta_2)\widehat{\mathbb{P}}_N + (\theta_1/\theta_2)\mathbb{Q} \in \mathcal{P}_N$. Since $\theta_1/\theta_2 \in [0,1)$ is arbitrary, we have $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} \in \mathcal{P}_N$ for all $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N$. (Note that when $\alpha = 1$, we have $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} = \mathbb{Q} \in \mathcal{P}_N$.) This shows that $\mathcal{P}_N$ is star-shaped and $\widehat{\mathbb{P}}_N$ is a star center.

Next, we prove part (ii). Suppose that $\{\mathcal{P}_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the strict hierarchical property, i.e., $\mathcal{P}'_{N,\theta}$ is increasing in θ. We show that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$ and $\mathcal{P}_N \neq \{\widehat{\mathbb{P}}_N\}$. From part (i), we have that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$. Suppose, on the contrary, that $\mathcal{P}_N = \{\mathbb{P}_N\}$. Then, we have $\mathbb{P}'_{N,\theta} = \{\widehat{\mathbb{P}}_N\}$ for all $\theta \in [0,1]$, contradicting that $\mathcal{P}'_{N,\theta}$ is increasing in θ. This shows that $\mathcal{P}_N \neq \{\mathbb{P}_N\}$.

Now, suppose that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$ and $\mathcal{P}_N \neq \{\widehat{\mathbb{P}}_N\}$. We show that $\{\mathcal{P}_{N,\theta} \mid \theta \in [0,1]\}$ satisfies the strict hierarchical property. From part (i), we have $\mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$, for any $0 \le \theta_1 < \theta_2 \le 1$. To show the strict inclusion $\mathcal{P}'_{N,\theta_1} \subset \mathcal{P}'_{N,\theta_2}$, we need to show that there exists a probability measure $\mathbb{M}$ such that $\mathbb{M} \in \mathcal{P}'_{N,\theta_2}$ but $\mathbb{M} \notin \mathcal{P}'_{N,\theta_1}$. We first claim that this is equivalent to

$$\mathcal{P}_N \setminus \left\{\left(1-\frac{\theta_1}{\theta_2}\right)\widehat{\mathbb{P}}_N + \frac{\theta_1}{\theta_2}\,\mathbb{Q}\,\middle|\,\mathbb{Q} \in \mathcal{P}_N\right\} \neq \emptyset. \tag{A.3}$$

Indeed, (A.3) is equivalent to the existence of $\mathbb{Q}' \in \mathcal{P}_N$ such that $\mathbb{Q}' \neq (1-\theta_1/\theta_2)\widehat{\mathbb{P}}_N + (\theta_1/\theta_2)\mathbb{Q}$ for all $\mathbb{Q} \in \mathcal{P}_N$. Note that $\mathbb{Q}' \neq (1-\theta_1/\theta_2)\widehat{\mathbb{P}}_N + (\theta_1/\theta_2)\mathbb{Q}$ for all $\mathbb{Q} \in \mathcal{P}_N$ is equivalent to $\theta_2\mathbb{Q}' + (1-\theta_2)\widehat{\mathbb{P}}_N \neq \theta_1\mathbb{Q} + (1-\theta_1)\widehat{\mathbb{P}}_N$ for all $\mathbb{Q} \in \mathcal{P}_N$. Thus, we have that $\theta_2\mathbb{Q}' + (1-\theta_2)\widehat{\mathbb{P}}_N \in \mathcal{P}'_{N,\theta_2}$ cannot be expressed as $\theta_1\mathbb{Q} + (1-\theta_1)\widehat{\mathbb{P}}_N$ for any $\mathbb{Q} \in \mathcal{P}_N$, i.e., $\theta_2\mathbb{Q}' + (1-\theta_2)\widehat{\mathbb{P}}_N \notin \mathcal{P}'_{N,\theta_1}$. This completes the proof of the claim.

Next, suppose, for the sake of contradiction, that condition (A.3) does not hold, i.e.,

$$\mathcal{P}_N \subseteq \left\{ \left(1 - \frac{\theta_1}{\theta_2}\right) \widehat{\mathbb{P}}_N + \frac{\theta_1}{\theta_2}\,\mathbb{Q} \,\middle|\, \mathbb{Q} \in \mathcal{P}_N \right\}. \tag{A.4}$$

We claim that (A.4) is equivalent to $\mathcal{P}_N = \{\widehat{\mathbb{P}}_N\}$ is a singleton, which contradicts with $\mathcal{P}_N \neq \{\widehat{\mathbb{P}}_N\}$. Consider an arbitrary probability measure $\mathbb{M} \in \mathcal{P}_N$. By (A.4), we can write $\mathbb{M} = (1-\theta_1/\theta_2)\widehat{\mathbb{P}}_N + (\theta_1/\theta_2)\mathbb{Q}_1$ for some $\mathbb{Q}_1 \in \mathcal{P}_N$. Moreover, note that for any $i \in \mathbb{N}$ and probability measure $\mathbb{Q}_i \in \mathcal{P}_N$, by (A.4), we can write $\mathbb{Q}_i = (1-\theta_1/\theta_2)\widehat{\mathbb{P}}_N + (\theta_1/\theta_2)\mathbb{Q}_{i+1}$ for some $\mathbb{Q}_{i+1} \in \mathcal{P}_N$. Using this recursion, we have

$$\begin{aligned}
\mathbb{M} = \left(1 - \frac{\theta_1}{\theta_2}\right) \widehat{\mathbb{P}}_N + \frac{\theta_1}{\theta_2}\,\mathbb{Q}_1 &= \left(1 - \frac{\theta_1}{\theta_2}\right) \widehat{\mathbb{P}}_N \cdot \left(1 + \frac{\theta_1}{\theta_2}\right) + \left(\frac{\theta_1}{\theta_2}\right)^2 \mathbb{Q}_2 \\
&= \cdots \\
&= \left(1 - \frac{\theta_1}{\theta_2}\right) \widehat{\mathbb{P}}_N \cdot \sum_{i=0}^{n-1} \left(\frac{\theta_1}{\theta_2}\right)^i + \left(\frac{\theta_1}{\theta_2}\right)^n \mathbb{Q}_n
\end{aligned}$$

for any $n \in \mathbb{N}$. Since $\theta_1 < \theta_2$, $\sum_{i=0}^{n-1}(\theta_1/\theta_2)^i \to 1/(1-\theta_1/\theta_2)$ and $(\theta_1/\theta_2)^n \to 0$ as $n \to \infty$. Thus, for any $\varepsilon > 0$, there exists $n' \in \mathbb{N}$ such that $|(1-\theta_1/\theta_2)\sum_{i=0}^{n-1}(\theta_1/\theta_2)^i - 1| < \varepsilon/2$ and $|(\theta_1/\theta_2)^n| < \varepsilon/2$ for all $n > n'$. Therefore, for any $B \in \mathcal{B}$, we have that

$$\begin{aligned}
|\mathbb{M}(B) - \widehat{\mathbb{P}}_N(B)| &= \left| \left(1 - \frac{\theta_1}{\theta_2}\right) \widehat{\mathbb{P}}_N(B) \cdot \sum_{i=0}^{n-1} \left(\frac{\theta_1}{\theta_2}\right)^i + \left(\frac{\theta_1}{\theta_2}\right)^n \mathbb{Q}_n(B) - \widehat{\mathbb{P}}_N(B) \right| \\
&\leq \left| \left(1 - \frac{\theta_1}{\theta_2}\right) \sum_{i=0}^{n-1} \left(\frac{\theta_1}{\theta_2}\right)^i - 1 \right| \cdot \widehat{\mathbb{P}}_N(B) + \left| \left(\frac{\theta_1}{\theta_2}\right)^n \right| \cdot \mathbb{Q}_n(B) \\
&\leq \frac{\varepsilon}{2}\,\widehat{\mathbb{P}}_N(B) + \frac{\varepsilon}{2}\,\mathbb{Q}_n(B) \\
&\leq \varepsilon
\end{aligned}$$

for all $n > n'$, where the first inequality follows from triangular inequality. Since $\varepsilon > 0$ is arbitrary, it follows that $\mathbb{M} = \widehat{\mathbb{P}}_N$ for any $\mathbb{M} \in \mathcal{P}_N$, implying that $\mathcal{P}_N = \{\widehat{\mathbb{P}}_N\}$. This contradicts the assumption that $\mathcal{P}_N \neq \{\widehat{\mathbb{P}}_N\}$ and completes the proof. □

Appendix A.2. Proof of Proposition 1

Proof. Proof. Suppose that $\mathcal{K}_i$ is star-shaped on $\mathcal{S}_i$ with a star center $\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})]$ for all $i \in \{1,\ldots,p\}$. We show that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$. Indeed, for all $i \in \{1,\ldots,p\}$, by the star-shapedness of $\mathcal{K}_i$, we have $(1-\alpha)\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] + \alpha\mathbb{E}_{\mathbb{Q}}[\Phi_i(\boldsymbol{\xi})] = \mathbb{E}_{(1-\alpha)\widehat{\mathbb{P}}_N+\alpha\mathbb{Q}}[\Phi_i(\boldsymbol{\xi})] \in \mathcal{K}_i$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N$. It follows from the definition of $\mathcal{P}_N$ that $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} \in \mathcal{P}_N$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N$, i.e., $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$.

Now, suppose that $\mathcal{P}_N$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$. We show that $\mathcal{K}_i$ is star-shaped on $\mathcal{S}_i := \{\mathbb{E}_{\mathbb{Q}}[\Phi_i(\boldsymbol{\xi})] \mid \mathbb{Q} \in \mathcal{P}_N\} \subseteq \mathbb{R}^{d_i \times d_i}$ with a star center $\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] \in \mathcal{S}_i$ for all $i \in \{1,\ldots,p\}$. Since $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} \in \mathcal{P}_N$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N$, we have $\mathbb{E}_{(1-\alpha)\widehat{\mathbb{P}}_N+\alpha\mathbb{Q}}[\Phi_i(\boldsymbol{\xi})] = (1-\alpha)\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] + \alpha\mathbb{E}_{\mathbb{Q}}[\Phi_i(\boldsymbol{\xi})] \in \mathcal{K}_i$ for all $i \in \{1,\ldots,p\}$. This, in turn, implies that $(1-\alpha)\mathbb{E}_{\widehat{\mathbb{P}}_N}[\Phi_i(\boldsymbol{\xi})] + \alpha\Psi \in \mathcal{K}_i$ for any $\alpha \in [0,1]$ and $\Psi \in \mathcal{S}_i$. This completes the proof. □

Appendix A.3. Proof of Proposition 2

Proof. Proof. Suppose that d is quasi-convex about $\widehat{\mathbb{P}}_N$ in the first argument. We show that $\mathcal{P}_N(\varepsilon)$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$ for all $\varepsilon \geq 0$. Indeed, for any given $\varepsilon \geq 0$, by quasi-convexity, we have $\mathsf{d}\big((1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q}, \widehat{\mathbb{P}}_N\big) \leq \mathsf{d}(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N(\varepsilon)$. It follows from the definition of $\mathcal{P}_N(\varepsilon)$ that $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} \in \mathcal{P}_N(\varepsilon)$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N(\varepsilon)$, i.e., $\mathcal{P}_N(\varepsilon)$ is star-shaped with a star-center $\widehat{\mathbb{P}}_N$.

Now, suppose that $\mathcal{P}_N(\varepsilon)$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$ for all $\varepsilon \geq 0$. We show that d is quasi-convex about $\widehat{\mathbb{P}}_N$. Since $\mathcal{P}_N(\varepsilon)$ is star-shaped, we have $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} \in \mathcal{P}_N(\varepsilon)$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N(\varepsilon)$. It follows from the definition of $\mathcal{P}_N(\varepsilon)$ that $\mathsf{d}\big((1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q}, \widehat{\mathbb{P}}_N\big) \leq \varepsilon$ for any $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{P}_N(\varepsilon)$. Suppose, for the sake of contradiction, that d is not quasi-convex about $\widehat{\mathbb{P}}_N$ in the first argument. That is, there exist $\alpha \in (0,1)$ and $\mathbb{Q} \in \mathcal{P}(\Xi)$ such that $\mathsf{d}\big((1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q}, \widehat{\mathbb{P}}_N\big) > \mathsf{d}(\mathbb{Q}, \widehat{\mathbb{P}}_N)$. Let $\bar{\varepsilon} := \mathsf{d}(\mathbb{Q}, \widehat{\mathbb{P}}_N) \in [0,\infty)$. Since $\mathcal{P}_N(\bar{\varepsilon})$ is star-shaped with a star center $\widehat{\mathbb{P}}_N$ and $\mathbb{Q} \in \mathcal{P}_N(\bar{\varepsilon})$, we have $(1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q} \in \mathcal{P}_N(\bar{\varepsilon})$. Thus, $\mathsf{d}\big((1-\alpha)\widehat{\mathbb{P}}_N + \alpha\mathbb{Q}, \widehat{\mathbb{P}}_N\big) \leq \bar{\varepsilon} = \mathsf{d}(\mathbb{Q}, \widehat{\mathbb{P}}_N)$, contradicting the assumption that d is not quasi-convex. This completes the proof. □

Appendix A.4. Proof of Theorem 2

Proof. Proof. First, we show that for any $\{\theta_1, \theta_2\} \subset [0,1]$, we have

$$\mathbb{H}(\mathcal{P}'_{N,\theta_1}, \mathcal{P}'_{N,\theta_2}) \leq 2C_N|\theta_1 - \theta_2|. \tag{A.5}$$

For any $\mathbb{P}_1 \in \mathcal{P}'_{N,\theta_1}$ and $\mathbb{P}_2 \in \mathcal{P}'_{N,\theta_2}$, we can write $\mathbb{P}_i = (1-\theta_i)\widehat{\mathbb{P}}_N + \theta_i\mathbb{Q}_i$ for some $\mathbb{Q}_i \in \mathcal{P}_N$ and $i \in \{1,2\}$. Then,

$$\begin{aligned}
\Big|\mathbb{E}_{\mathbb{P}_1}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_2}[f(\boldsymbol{x},\boldsymbol{\xi})]\Big| &= \left|(\theta_2 - \theta_1)\cdot\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i) + \theta_1\cdot\mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x},\boldsymbol{\xi})] - \theta_2\cdot\mathbb{E}_{\mathbb{Q}_2}[f(\boldsymbol{x},\boldsymbol{\xi})]\right| \\
&\leq |\theta_1 - \theta_2|\left|\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i)\right| + \Big|\theta_1\mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x},\boldsymbol{\xi})] - \theta_2\mathbb{E}_{\mathbb{Q}_2}[f(\boldsymbol{x},\boldsymbol{\xi})]\Big| \\
&\leq C_N|\theta_1 - \theta_2| + \Big|\theta_1\mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x},\boldsymbol{\xi})] - \theta_2\mathbb{E}_{\mathbb{Q}_2}[f(\boldsymbol{x},\boldsymbol{\xi})]\Big|, \qquad \text{(A.6)}
\end{aligned}$$

where the last inequality follows from Assumption 1 and the definition of C_N in (13). Note that

$$\begin{aligned}
\sup_{\mathbb{P}_1 \in \mathcal{P}'_{N,\theta_1}} \inf_{\mathbb{P}_2 \in \mathcal{P}'_{N,\theta_2}} \mathbb{d}(\mathbb{P}_1, \mathbb{P}_2) &= \sup_{\mathbb{P}_1 \in \mathcal{P}'_{N,\theta_1}} \inf_{\mathbb{P}_2 \in \mathcal{P}'_{N,\theta_2}} \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| \mathbb{E}_{\mathbb{P}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_2}[f(\boldsymbol{x}, \boldsymbol{\xi})] \Big| \\
&\le C_N |\theta_1 - \theta_2| + \sup_{\mathbb{Q}_1 \in \mathcal{P}_N} \inf_{\mathbb{Q}_2 \in \mathcal{P}_N} \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| \theta_1 \mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \theta_2 \mathbb{E}_{\mathbb{Q}_2}[f(\boldsymbol{x}, \boldsymbol{\xi})] \Big| \qquad \text{(A.7)} \\
&\le C_N |\theta_1 - \theta_2| + \sup_{\mathbb{Q}_1 \in \mathcal{P}_N} \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| \theta_1 \mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \theta_2 \mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] \Big| \qquad \text{(A.8)} \\
&\le C_N |\theta_1 - \theta_2| + \sup_{\mathbb{Q}_1 \in \mathcal{P}_N} \sup_{\boldsymbol{x} \in \mathcal{X}} |\theta_1 - \theta_2| \big| \mathbb{E}_{\mathbb{Q}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] \big| \\
&\le 2C_N |\theta_1 - \theta_2|. \qquad \text{(A.9)}
\end{aligned}$$

Inequality (A.7) follows from (A.6), where we note that the arguments in supremum and infimum are changed to $\mathbb{Q}_1 \in \mathcal{P}_N$ and $\mathbb{Q}_2 \in \mathcal{P}_N$, respectively; inequality (A.8) follows from the fact that choosing $\mathbb{Q}_2 = \mathbb{Q}_1$ for the infimum problem yields an upper bound; inequality (A.9) follows from Assumption 1 and the definition of C_N in (13). Interchanging the role of $\mathbb{P}_1$ and $\mathbb{P}_2$, as well as $\mathcal{P}'_{N,\theta_1}$ and $\mathcal{P}'_{N,\theta_2}$, we can obtain a similar inequality:

$$\sup_{\mathbb{P}_2 \in \mathcal{P}'_{N,\theta_2}} \inf_{\mathbb{P}_1 \in \mathcal{P}'_{N,\theta_1}} \mathbb{d}(\mathbb{P}_1, \mathbb{P}_2) \le 2C_N |\theta_1 - \theta_2|.$$

Therefore, from the definition of $\mathbb{H}$ in (12), we have

$$\mathbb{H}(\mathcal{P}'_{N,\theta_1}, \mathcal{P}'_{N,\theta_2}) = \max \left\{ \sup_{\mathbb{P}_1 \in \mathcal{P}'_{N,\theta_1}} \inf_{\mathbb{P}_2 \in \mathcal{P}'_{N,\theta_2}} \mathbb{d}(\mathbb{P}_1, \mathbb{P}_2), \sup_{\mathbb{P}_2 \in \mathcal{P}'_{N,\theta_2}} \inf_{\mathbb{P}_1 \in \mathcal{P}'_{N,\theta_1}} \mathbb{d}(\mathbb{P}_1, \mathbb{P}_2) \right\} \le 2C_N |\theta_1 - \theta_2|.$$

Now, we prove assertion (i):

$$|\widehat{v}_N(\theta_1) - \widehat{v}_N(\theta_2)| \le \mathbb{H}(\mathcal{P}'_{N,\theta_1}, \mathcal{P}'_{N,\theta_2}) \le 2C_N |\theta_1 - \theta_2|,$$

where the first inequality follows from the quantitative stability analysis; see Proposition 7 in Appendix E.

For (ii), under the second-order growth condition, we have

$$D\Big(\widehat{\mathcal{X}}_N(\theta_2), \widehat{\mathcal{X}}_N(\theta_1)\Big) \le \sqrt{\frac{3}{\tau} \mathbb{H}(\mathcal{P}'_{N,\theta_1}, \mathcal{P}'_{N,\theta_2})} \le \sqrt{\frac{6C_N}{\tau} |\theta_1 - \theta_2|},$$

where the first inequality follows again from the quantitative stability analysis; see Proposition 7 in Appendix E. This completes the proof. □

Appendix A.5. Proof of Theorem 3

Proof. Proof.

First, we prove part (i). To show the concavity of $\widehat{v}_N(\theta)$, we define the function $g(\theta; \boldsymbol{x}) := \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x}, \boldsymbol{\xi})] + \theta \big\{ \sup_{\mathbb{P} \in \mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x}, \boldsymbol{\xi})] \big\}$, which is linear in θ for any $\boldsymbol{x} \in \mathcal{X}$. Since $\widehat{v}_N(\theta)$

is the pointwise minimum of the linear functions $\{g(\theta;\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$, i.e., $\widehat{v}_N(\theta) = \min_{\boldsymbol{x}\in\mathcal{X}} g(\theta;\boldsymbol{x})$, the function $\widehat{v}_N(\theta)$ is concave. Next, to show (15), note that $\theta = (1-\theta)\cdot 0 + \theta \cdot 1$. Thus, concavity of $\widehat{v}_N$ implies

$$\widehat{r}_N(\theta) = \widehat{v}_N(\theta) - \Big[(1-\theta)\cdot\widehat{v}_N(0) + \theta\cdot\widehat{v}_N(1)\Big] \geq 0.$$

Also, by Theorem 2, we have

$$\begin{aligned}
\widehat{r}_N(\theta) &= \widehat{v}_N(\theta) - \Big[(1-\theta)\cdot\widehat{v}_N(0) + \theta\cdot\widehat{v}_N(1)\Big] \\
&\leq (1-\theta)\cdot\Big|\widehat{v}_N(\theta) - \widehat{v}_N(0)\Big| + \theta\cdot\Big|\widehat{v}_N(\theta) - \widehat{v}_N(1)\Big| \\
&\leq (1-\theta)\cdot 2C_N\theta + \theta\cdot 2C_N(1-\theta) \\
&\leq 4C_N\theta(1-\theta).
\end{aligned}$$

This completes the proof of part (i).

Finally, for part (ii), we have

$$\begin{aligned}
& D\Big(\widehat{\mathcal{X}}_N(\theta), (1-\theta)\widehat{\mathcal{X}}_N(0) + \theta\widehat{\mathcal{X}}_N(1)\Big) \\
= & \sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N(\theta)} \inf_{\overline{\boldsymbol{x}}\in(1-\theta)\widehat{\mathcal{X}}_N(0)+\theta\widehat{\mathcal{X}}_N(1)} \|\boldsymbol{x} - \overline{\boldsymbol{x}}\| \\
= & \sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N(\theta)} \inf_{\boldsymbol{x}'\in\widehat{\mathcal{X}}_N(0), \boldsymbol{x}''\in\widehat{\mathcal{X}}_N(1)} \big\|\boldsymbol{x} - \big[(1-\theta)\boldsymbol{x}' + \theta\boldsymbol{x}''\big]\big\| \\
\leq & \sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N(\theta)} \inf_{\boldsymbol{x}'\in\widehat{\mathcal{X}}_N(0), \boldsymbol{x}''\in\widehat{\mathcal{X}}_N(1)} \Big\{(1-\theta)\|\boldsymbol{x} - \boldsymbol{x}'\| + \theta\|\boldsymbol{x} - \boldsymbol{x}''\|\Big\} && \text{(A.10)} \\
\leq & (1-\theta) \sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N(\theta)} \inf_{\boldsymbol{x}'\in\widehat{\mathcal{X}}_N(0)} \|\boldsymbol{x} - \boldsymbol{x}'\| + \theta \sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N(\theta)} \inf_{\boldsymbol{x}''\in\widehat{\mathcal{X}}_N(1)} \|\boldsymbol{x} - \boldsymbol{x}''\| && \text{(A.11)} \\
= & (1-\theta)D\Big(\widehat{\mathcal{X}}_N(\theta), \widehat{\mathcal{X}}_N(0)\Big) + \theta D\Big(\widehat{\mathcal{X}}_N(\theta), \widehat{\mathcal{X}}_N(1)\Big) \\
\leq & (1-\theta)\sqrt{\frac{6C_N\theta}{\tau}} + \theta\sqrt{\frac{6C_N(1-\theta)}{\tau}} && \text{(A.12)} \\
= & \sqrt{\frac{6C_N\theta(1-\theta)}{\tau}}\Big(\sqrt{\theta} + \sqrt{1-\theta}\Big).
\end{aligned}$$

Inequality (A.10) follows from triangle inequality; inequality (A.11) follows from separating the supremum operator to each summand; inequality (A.12) follows from Theorem 2. □

Appendix A.6. Proof of Proposition 3

Proof. Proof. Since $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] \leq v^\star$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)] - v^\star &\leq \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)] - \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] \\
&= \mathbb{E}_{\mathbb{P}^N}[(1-\theta)\cdot\widehat{v}_N(0) + \theta\cdot\widehat{v}_N(1) + \widehat{r}_N(\theta)] - \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] \quad \text{(by (15) in Theorem 3)} \\
&= \theta\Big\{\mathbb{E}_{\mathbb{P}^N}\big[\widehat{v}_N(1)\big] - \mathbb{E}_{\mathbb{P}^N}\big[\widehat{v}_N(0)\big]\Big\} + \mathbb{E}_{\mathbb{P}^N}[\widehat{r}_N(\theta)].
\end{aligned}$$

From Theorem 3, we have $\widehat{r}_N(\theta) \in [0, 4C_N\theta(1-\theta)]$. Thus, $R_N(\theta) := \mathbb{E}_{\mathbb{P}^N}[\widehat{r}_N(\theta)] \in [0, 4\overline{C}_N\theta(1-\theta)]$. Finally, if $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, then $\widehat{v}_N(0) \leq \widehat{v}_N(1)$. It follows that the upper bound on the bias in (17) is non-negative. □

Appendix A.7. Proof of Theorem 4

Proof. Proof. From Theorem 2, we have that $\widehat{v}_N(\theta)$ is Lipschitz continuous with Lipschitz constant $C_N < \infty$ (almost surely). Thus, $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)]$ is also Lipschitz continuous with Lipschitz constant $\overline{C}_N = \mathbb{E}_{\mathbb{P}^N}(C_N) < \infty$. Since $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] \leq v^\star \leq \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)]$, it follows by the intermediate value theorem (see Theorem 4.23 of Rudin et al., 1976) that there exists $\theta_N^{\mathrm{u}} \in [0,1]$ for which $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta_N^{\mathrm{u}})] = v^\star$. This completes the proof. □

Appendix A.8. Proof of Corollary 1

Proof. Proof. By Theorem 3, $\widehat{v}_N(\theta)$ is concave on $[0,1]$, and so is $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)]$. Since $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] \leq \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)]$, we have that $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)]$ is either (a) non-decreasing on $[0,1]$ or (b) first non-decreasing and then non-increasing (see Lemma 1.1.4 of Niculescu and Persson, 2018). By Theorem 4, there exists θ_N^{u} such that $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)] = v^\star$. When it is not unique, take $\theta_N^{\mathrm{u}} = \inf\{\theta \in [0,1] \mid \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)] = v^\star\}$. It follows that $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)]$ is non-decreasing on $[0, \theta_N^{\mathrm{u}}]$ and thus, $|\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta)] - v^\star| \leq |\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] - v^\star|$ for all $\theta \in [0, \theta_N^{\mathrm{u}}]$. □

Appendix A.9. Proof of Theorem 5

Proof. Proof. First, note that θ_N^{u} satisfies $\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta_N^{\mathrm{u}})] = v^\star$, where Theorem 4 ensures the existence of θ_N^{u}. Moreover, Theorem 3 implies that

$$\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(\theta_N^{\mathrm{u}})] = (1-\theta_N^{\mathrm{u}})\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] + \theta_N^{\mathrm{u}}\mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(1)] + R_N(\theta_N^{\mathrm{u}}),$$

where $R_N(\theta_N^{\mathrm{u}}) \in [0, 4C\theta_N^{\mathrm{u}}(1-\theta_N^{\mathrm{u}})]$. Combining these two equations, we have

$$v^\star - \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] = \theta_N^{\mathrm{u}}\Big\{\mathbb{E}_{\mathbb{P}^N}\left[\widehat{v}_N(1)\right] - \mathbb{E}_{\mathbb{P}^N}\left[\widehat{v}_N(0)\right]\Big\} + R_N(\theta_N^{\mathrm{u}}). \tag{A.13}$$

Note that left-hand-side of (A.13) corresponds to the bias of the SAA estimator. Under assumptions (a)–(c), Theorem 6 of Banholzer et al. (2022) gives $v^\star - \mathbb{E}_{\mathbb{P}^N}[\widehat{v}_N(0)] = o(\sqrt{\log\log N}/\sqrt{N})$. For the ease of notation, let $\Delta_N = \mathbb{E}_{\mathbb{P}^N}\left[\widehat{v}_N(1)\right] - \mathbb{E}_{\mathbb{P}^N}\left[\widehat{v}_N(0)\right]$ and $b_N = \sqrt{\log\log N}/\sqrt{N}$. By the assumption that $\inf_{N\in\mathbb{N}} \Delta_N > 0$, we know that $\Delta_N \geq \Delta$ for some $\Delta > 0$. Then,

$$0 \leq \theta_N^{\mathrm{u}}\Delta \leq \theta_N^{\mathrm{u}}\Delta_N + R_N(\theta_N^{\mathrm{u}}) = o(b_N), \tag{A.14}$$

where the second inequality follows from $R_N(\theta_N^{\mathrm{u}}) \geq 0$. In other words, (A.14) implies that

$$0 \leq \left(b_N^{-1}\theta_N^{\mathrm{u}}\right)\Delta \leq o(1),$$

showing that $b_N^{-1}\theta_N^{\mathrm{u}} \to 0$ as $N \to \infty$. This completes the proof. □

Appendix A.10. Proof of Theorem 6

Proof. Proof. Theorem 5.7 of Shapiro et al. (2014) gives the following asymptotic normality of $\widehat{\upsilon}_N(0)$:

$$X_N = \sqrt{N}\Big(\widehat{\upsilon}_N(0) - \upsilon^\star\Big) \Rightarrow \inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x}), \tag{A.15}$$

where "⇒" denotes the convergence in distribution, $\mathcal{X}^\star$ is the set of optimal solutions to (1), and $\mathbb{G}$ is a Gaussian process indexed by $\mathcal{X}$ with mean zero and covariance function $\mathrm{Cov}(\mathbb{G}(\boldsymbol{x}_1), \mathbb{G}(\boldsymbol{x}_2)) = \mathrm{Cov}_{\mathbb{P}^\star}(f(\boldsymbol{x}_1, \boldsymbol{\xi}), f(\boldsymbol{x}_2, \boldsymbol{\xi}))$. Together with the asymptotic uniform integrability of $\{X_N\}_{N\in\mathbb{N}}$, (A.15) implies that

$$\mathbb{E}_{\mathbb{P}^N}\left[\sqrt{N}\Big(\widehat{\upsilon}_N(0) - \upsilon^\star\Big)\right] = \mathbb{E}\Big[\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})\Big] + o(1)$$

(see, e.g., Theorem 2.20 of van der Vaart, 2000). Hence, the bias of the SAA estimator is given by

$$\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)] - \upsilon^\star = \frac{1}{\sqrt{N}}\mathbb{E}\Big[\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})\Big] + o(1/\sqrt{N}). \tag{A.16}$$

If $\mathcal{X}^\star$ is a singleton, then $\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})$ reduces to a normal distribution with mean zero. Therefore, (A.16) implies that the bias of the SAA estimator is of order $o(1/\sqrt{N})$. A similar argument in the proof of Theorem 5 shows that $\theta_N^{\mathrm{u}} = o(1/\sqrt{N})$.

Next, we consider that $\mathcal{X}^\star$ is not a singleton. In the trivial case when $\mathbb{E}_{\mathbb{P}^N}[\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})] = 0$, we immediately have $\theta_N^{\mathrm{u}} = o(1/\sqrt{N})$, which directly implies $\theta_N^{\mathrm{u}} = O(1/\sqrt{N})$. Now, consider the case when $\mathbb{E}_{\mathbb{P}^N}[\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})] < 0$. From (A.13) and (A.16), we have

$$\theta_N^{\mathrm{u}}\Big\{\mathbb{E}_{\mathbb{P}^N}\big[\widehat{\upsilon}_N(1)\big] - \mathbb{E}_{\mathbb{P}^N}\big[\widehat{\upsilon}_N(0)\big]\Big\} + R_N(\theta_N^{\mathrm{u}}) = -\frac{1}{\sqrt{N}}\mathbb{E}\Big[\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})\Big] + o(1/\sqrt{N}), \tag{A.17}$$

where $R_N(\theta_N^{\mathrm{u}}) \in [0, 4C\theta_N^{\mathrm{u}}(1-\theta_N^{\mathrm{u}})]$. Again, for the ease of notation, let $\Delta_N = \mathbb{E}_{\mathbb{P}^N}\big[\widehat{\upsilon}_N(1)\big] - \mathbb{E}_{\mathbb{P}^N}\big[\widehat{\upsilon}_N(0)\big] > 0$. By the assumption that $\inf_{N\in\mathbb{N}} \Delta_N > 0$, we know that $\Delta_N \geq \Delta$ for some $\Delta > 0$. Since $R_N(\theta_N^{\mathrm{u}}) \geq 0$, (A.17) implies

$$0 \leq (\sqrt{N}\theta_N^{\mathrm{u}})\Delta \leq -\mathbb{E}\Big[\inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})\Big] + o(1). \tag{A.18}$$

Thus, (A.18) shows that $\limsup_{N\to\infty} \sqrt{N}\theta_N^{\mathrm{u}}$ is upper bounded, concluding that $\theta_N^{\mathrm{u}} = O(1/\sqrt{N})$. □

Appendix A.11. Proof of Theorem 7

Proof. Proof. We derive the desired generalization bound as follows:

$$
\begin{aligned}
&\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta}}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\}>\delta\right)\\
&=\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{(1-\theta)\cdot\left(\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]\right)+\theta\cdot\left(\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right)\right\}>\delta\right)\\
&\le\mathbb{P}^N\left((1-\theta)\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\}+\theta\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\}>\delta\right)\\
&\le\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\}>\delta\right)+\mathbb{P}^N\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]-\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\}>\delta\right)\qquad\text{(A.19)}\\
&\le\alpha_{N,1}+\alpha_{N,2}.\qquad\text{(A.20)}
\end{aligned}
$$

Here, (A.19) follows from the fact that $\{(1-\theta)A_1+\theta A_2>\delta\}\subseteq\{A_1>\delta\}\cup\{A_2>\delta\}$ for any random variables A_1 and A_2, and (A.20) follows from (19) and (20). □

Appendix A.12. Proof of Lemma 1

Proof. Proof. For brevity, all convergence, equalities and inequalities hold almost surely in this proof. If f is uniformly bounded, i.e., $\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{\xi}\in\Xi}f(\boldsymbol{x},\boldsymbol{\xi})<\infty$, then the desired inequality follows immediately. Now, suppose that Assumption 3(a) holds. We can directly obtain an upper bound from (22):

$$\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x},\boldsymbol{\xi})|\le\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})]\cdot\mathrm{diam}(\mathcal{X})+\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}_0,\boldsymbol{\xi})|:=M<\infty\qquad\text{(A.21)}$$

for all sufficiently large N. Thus, we obtain $\limsup_{N\to\infty}\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x},\boldsymbol{\xi})|\le M$ by taking supremum over $\boldsymbol{x}\in\mathcal{X}$ on both sides of (A.21).

Now, suppose that Assumption 3(b) holds. Then, for any $\boldsymbol{x}\in\mathcal{X}$, we have

$$
\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]&\le\left|\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]-\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\right|+\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\\
&\le\mathbb{H}(\mathcal{P}_N,\widehat{\mathcal{P}})+\left\{\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})]\cdot\mathrm{diam}(\mathcal{X})+\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}_0,\boldsymbol{\xi})|\right\},\qquad\text{(A.22)}
\end{aligned}
$$

where the first term in (A.22) follows from the definition of $\mathbb{H}(\mathcal{P}_N,\widehat{\mathcal{P}})$ and the second term in (A.22) follows from the same argument in (A.21). Since $\mathbb{H}(\mathcal{P}_N,\widehat{\mathcal{P}})\to 0$, (A.22) implies that

$$\limsup_{N\to\infty}\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x},\boldsymbol{\xi})|\le\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})]\cdot\mathrm{diam}(\mathcal{X})+\sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}_0,\boldsymbol{\xi})|<\infty.$$

Finally, suppose that Assumption 3(c) holds. For any $\mathbb{P} \in \mathcal{P}_N$ identified as $\boldsymbol{p} \in \mathbb{R}_+^N$,

$$\begin{aligned}
\left|\mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})]\right| = \left|\sum_{i=1}^N \left(p_i - \frac{1}{N}\right)\kappa(\widehat{\boldsymbol{\xi}}_i)\right| &\le \left\|\boldsymbol{p} - \frac{1}{N}\mathbf{1}\right\|_\infty \cdot \sum_{i=1}^N \kappa(\widehat{\boldsymbol{\xi}}_i) \\
&= \frac{1}{N}\|N\boldsymbol{p} - \mathbf{1}\|_\infty \cdot \sum_{i=1}^N \kappa(\widehat{\boldsymbol{\xi}}_i) \\
&= \|N\boldsymbol{p} - \mathbf{1}\|_\infty \cdot \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})]. \qquad \text{(A.23)}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] &\le \limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \left\{\mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})] + \left|\mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})]\right|\right\} \\
&\le \limsup_{N\to\infty} \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})] + \limsup_{N\to\infty} \left\{\sup_{\mathbb{P}\in\mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_\infty \cdot \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})]\right\} \qquad \text{(A.24)} \\
&\le \limsup_{N\to\infty} \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})] + \left(\limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_\infty\right)\left(\limsup_{N\to\infty} \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})]\right) \\
&= \left(1 + \limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_\infty\right)\left(\limsup_{N\to\infty} \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})]\right),
\end{aligned}$$

where the second term in (A.24) follows from (A.23). Since $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$ are i.i.d. following $\mathbb{P}^\star$, the strong law of large numbers gives $\limsup_{N\to\infty} \mathbb{E}_{\widehat{\mathbb{P}}_N}[\kappa(\boldsymbol{\xi})] = \mathbb{E}_{\mathbb{P}^\star}[\kappa(\boldsymbol{\xi})] < \infty$. Moreover, since $\limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \|N\boldsymbol{p} - \mathbf{1}\|_\infty < \infty$, we have $\limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] < \infty$. Using a similar argument, we can obtain $\limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}_0, \boldsymbol{\xi})| < \infty$. Therefore, we have

$$\begin{aligned}
\limsup_{N\to\infty} \sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}, \boldsymbol{\xi})| &\le \limsup_{N\to\infty} \left\{\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] \cdot \operatorname{diam}(\mathcal{X}) + \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}_0, \boldsymbol{\xi})|\right\} \\
&\le \limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[\kappa(\boldsymbol{\xi})] \cdot \operatorname{diam}(\mathcal{X}) + \limsup_{N\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x}_0, \boldsymbol{\xi})| < \infty.
\end{aligned}$$

This completes the proof. □

Appendix A.13. Proof of Theorem 8

Proof. Proof. For brevity, all convergence, equalities and inequalities hold almost surely in this proof. First, we claim that the following uniform convergence holds:

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\left\{(1-\theta_N) \cdot \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) + \theta_N \cdot \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})]\right\} - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}, \boldsymbol{\xi})]\right| \to 0 \qquad \text{(A.25)}$$

as $N \to \infty$. To prove (A.25), note that

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}} \left| \left\{ (1-\theta_N) \cdot \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) + \theta_N \cdot \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \right\} - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right| \\
&\le (1-\theta_N) \cdot \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right| + \theta_N \cdot \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right|.
\end{aligned} \tag{A.26}$$

Assumption 2 implies that the first term $\sup_{\boldsymbol{x}\in\mathcal{X}} \left| N^{-1} \sum_{i=1}^{N} f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right|$ converges to zero. Next, for the second term, note that

$$\theta_N \cdot \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right| \le \theta_N \left(\sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x},\boldsymbol{\xi})| + \sup_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}|f(\boldsymbol{x},\boldsymbol{\xi})| \right) \tag{A.27}$$

Since $\limsup_{N\to\infty} \sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|f(\boldsymbol{x},\boldsymbol{\xi})| \le M$ by Lemma 1 and $\sup_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}|f(\boldsymbol{x},\boldsymbol{\xi})| < \infty$, the upper bound in (A.27) converges to zero by $\theta_N = o(1)$. Therefore, the upper bound in (A.26) converges to zero, which shows (A.25).

Now, with the use of (A.25), we can prove the desired asymptotic convergence results in a way similar to Theorem 5.3 of Shapiro et al. (2014). For completeness, we also provide the details here. To show assertion (i), note that

$$\begin{aligned}
|\widehat{\upsilon}_N(\theta_N) - \upsilon^\star| &= \left| \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ (1-\theta_N) \cdot \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) + \theta_N \cdot \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \right\} - \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right| \\
&\le \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \left\{ (1-\theta_N) \cdot \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}_i) + \theta_N \cdot \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \right\} - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})] \right|,
\end{aligned}$$

which converges to zero by (A.25). Next, to show assertion (ii), suppose, on the contrary, that $D\big(\widehat{\mathcal{X}}_N(\theta_N), \mathcal{X}^\star\big) \to 0$ does not hold almost surely, i.e., $\mathbb{P}^\infty\big(D\big(\widehat{\mathcal{X}}_N(\theta_N), \mathcal{X}^\star\big) \not\to 0\big) > 0$. Consider a data sequence such that the event $D\big(\widehat{\mathcal{X}}_N(\theta_N), \mathcal{X}^\star\big) \not\to 0$ holds. Then, for some $\epsilon > 0$, there exists a sequence $\{\boldsymbol{x}_{N_j}\}_{j\in\mathbb{N}}$ such that $\boldsymbol{x}_{N_j} \in \widehat{\mathcal{X}}_{N_j}(\theta_{N_j})$ and $d\big(\boldsymbol{x}_{N_j}, \mathcal{X}^\star\big) > \epsilon$. Since $\mathcal{X}$ is compact by Assumption 1, without loss of generality, we can assume that $\boldsymbol{x}_{N_j} \to \overline{\boldsymbol{x}}$ for some $\overline{\boldsymbol{x}} \in \mathcal{X}$. By continuity of the distance function d, we have $d(\overline{\boldsymbol{x}}, \mathcal{X}^\star) = \lim_{j\to\infty} d(\boldsymbol{x}_{N_j}, \mathcal{X}^\star) > \epsilon$. Thus, $\overline{\boldsymbol{x}} \notin \mathcal{X}^\star$, which implies that $\mathbb{E}_{\mathbb{P}^\star}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] > \upsilon^\star$. However, note that

$$\left| \mathbb{E}_{\mathbb{P}^\star}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] - \widehat{\upsilon}_{N_j}(\theta_{N_j}) \right| \le \left| \mathbb{E}_{\mathbb{P}^\star}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}_{N_j},\boldsymbol{\xi})] \right| + \left| \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}_{N_j},\boldsymbol{\xi})] - \widehat{\upsilon}_{N_j}(\theta_{N_j}) \right|,$$

where the first term converges to zero by the continuity of $\mathbb{E}_{\mathbb{P}^\star}[f(\cdot,\boldsymbol{\xi})]$, and the second term converges to zero by (A.25). Therefore, we arrive at $\lim_{j\to\infty} \widehat{\upsilon}_{N_j}(\theta_{N_j}) = \mathbb{E}_{\mathbb{P}^\star}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] > \upsilon^\star$, which contradicts assertion (i) that $\lim_{j\to\infty} \widehat{\upsilon}_{N_j}(\theta_{N_j}) = \upsilon^\star$. □

Appendix A.14. Proof of Lemma 2

Proof. Proof.

Recall that, following the convention in empirical process theory, we view $\mathbb{P} \in \mathcal{P}(\Xi)$ as an element in $\ell^\infty(\mathcal{H})$ defined as $\mathbb{P}(h) = \mathbb{E}_{\mathbb{P}}(h)$ for $h \in \mathcal{H}$. We divide the proof of the desired weak convergence $\mathbb{S}_N \Rightarrow \mathbb{G}'$ into two steps.

Step 1. We first show that $\big(\mathbb{S}_N(h_1), \ldots, \mathbb{S}_N(h_k)\big) \Rightarrow \big(\mathbb{G}'(h_1), \ldots, \mathbb{G}'(h_k)\big)$ for any finite subset $\{h_1, \ldots, h_k\} \subset \mathcal{H}$. To prove this, we write $\mathbb{S}_N = \sqrt{N}(1-\theta_N)(\widehat{\mathbb{P}}_N - \mathbb{P}^\star) + \sqrt{N}\theta_N(\mathbb{P}_N - \mathbb{P}^\star)$. Note that $\sqrt{N}(\widehat{\mathbb{P}}_N - \mathbb{P}^\star)(h_1, \ldots, h_k) \Rightarrow \big(\mathbb{G}'(h_1), \ldots, \mathbb{G}'(h_k)\big)$ by Assumption 4. Also, from the assumption that $\theta_N = o(N^{-1/2})$, we have $(1-\theta_N) \to 1$. Hence, for the first term in $\mathbb{S}_N$, we have $(1-\theta_N) \cdot \sqrt{N}(\widehat{\mathbb{P}}_N - \mathbb{P}^\star)(h_1, \ldots, h_k) \Rightarrow \big(\mathbb{G}'(h_1), \ldots, \mathbb{G}'(h_k)\big)$ by Slutsky's Theorem (see, e.g., Example 1.4.7 of van der Vaart and Wellner, 1996). Next, for the second term in $\mathbb{S}_N$, Lemma 1 implies that $\limsup_{N\to\infty} \sup_{h\in\mathcal{H}} \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|h| \le M$ almost surely for some constant M. Therefore, there exists constant M' such that $\limsup_{N\to\infty} |\mathbb{E}_{\mathbb{P}_N}(h) - \mathbb{E}_{\mathbb{P}^\star}(h)| \le \limsup_{N\to\infty} \sup_{h\in\mathcal{H}} \sup_{\mathbb{P}\in\mathcal{P}_N} |\mathbb{E}_{\mathbb{P}}(h) - \mathbb{E}_{\mathbb{P}^\star}(h)| \le M'$ for any $\mathbb{P}_N \in \mathcal{P}_N$ and $h \in \mathcal{H}$ (since $\mathbb{E}_{\mathbb{P}^\star}|h| < \infty$). By assumption (b), $\sqrt{N}\theta_N$ converges to zero as $N \to \infty$ and hence, $\sqrt{N}\theta_N(\mathbb{P}_N - \mathbb{P}^\star)(h)$ also converges to zero almost surely. This implies that $\sqrt{N}\theta_N(\mathbb{P}_N - \mathbb{P}^\star)(h_1, \ldots, h_k) \to (0, \ldots, 0) \in \mathbb{R}^k$ almost surely. Therefore, provoking Slutsky's Theorem again, we obtain the desired convergence, i.e., $\big(\mathbb{S}_N(h_1), \ldots, \mathbb{S}_N(h_k)\big) \Rightarrow \big(\mathbb{G}'(h_1), \ldots, \mathbb{G}'(h_k)\big)$. This completes step 1.

Step 2. We show that $\mathbb{S}_N$ is asymptotically tight. Note that for any $\varepsilon > 0$,

$$\limsup_{\delta\to 0} \limsup_{N\to\infty} \mathbb{P}^N\left(\sup_{\|h-h'\|_{L^2(\mathbb{P}^\star)}<\delta} \big|\mathbb{S}_N(h-h')\big| \ge \varepsilon \right) \tag{A.28}$$

$$\le \limsup_{\delta\to 0} \limsup_{N\to\infty} \mathbb{P}^N\left(\sup_{\|h-h'\|_{L^2(\mathbb{P}^\star)}<\delta} \big|(1-\theta_N)\sqrt{N}(\widehat{\mathbb{P}}_N - \mathbb{P}^\star)(h-h')\big| \ge \frac{\varepsilon}{2} \right) \tag{A.29}$$

$$+ \limsup_{\delta\to 0} \limsup_{N\to\infty} \mathbb{P}^N\left(\sup_{\|h-h'\|_{L^2(\mathbb{P}^\star)}<\delta} \big|\theta_N\sqrt{N}(\mathbb{P}_N - \mathbb{P}^\star)(h-h')\big| \ge \frac{\varepsilon}{2} \right). \tag{A.30}$$

Since $\sqrt{N}(1-\theta_N)(\widehat{\mathbb{P}}_N - \mathbb{P}^\star) \Rightarrow \mathbb{G}'$, the sequence $\{\sqrt{N}(1-\theta_N)(\widehat{\mathbb{P}}_N - \mathbb{P}^\star)\}_{N\in\mathbb{N}}$ is asymptotically tight. Therefore, (A.29) equals zero by Theorem 1.5.7 of van der Vaart and Wellner (1996). Now, we show that (A.30) also vanishes. It suffices to show that the sequence $\{\sqrt{N}\theta_N(\mathbb{P}_N - \mathbb{P}^\star)\}_{N\in\mathbb{N}}$ is asymptotically tight. Note that

$$\mathbb{Q}_N(h) := \sqrt{N}\theta_N(\mathbb{P}_N - \mathbb{P}^\star)(h) = \sqrt{N}\theta_N[\mathbb{E}_{\mathbb{P}_N}(h) - \mathbb{E}_{\mathbb{P}^\star}(h)].$$

By assumption (a), the metric space $(\mathcal{H}, \|\cdot\|_{L^2(\mathbb{P}^\star)})$ is totally bounded. Also, as shown earlier (in step 1), $\mathbb{Q}_N(h)$ converges to zero almost surely. It follows by Theorem 1.5.4 of van der Vaart and Wellner (1996) that $\mathbb{Q}_N(h)$ is asymptotically tight for any $h \in \mathcal{H}$. Next, we claim that $\mathbb{Q}_N$ is asymptotically uniform equicontinuous in probability, i.e., for any $\varepsilon > 0$, and $\eta > 0$, there exists

$\delta > 0$ such that

$$\limsup_{N\to\infty} \mathbb{P}^N\left(\sup_{\|h-h'\|_{L^2(\mathbb{P}^\star)}<\delta} |\mathbb{Q}_N(h-h')| > \varepsilon\right) < \eta.$$

For notational simplicity, we write $\|h-h'\| = \|h-h'\|_{L^2(\mathbb{P}^\star)}$. Note that

$$\begin{aligned}
\sup_{\|h-h'\|<\delta} \left|\mathbb{Q}_N(h-h')\right| &= \sup_{\|h-h'\|<\delta} \left|\sqrt{N}\theta_N(\mathbb{P}_N - \mathbb{P}^\star)(h-h')\right| \\
&\le \sqrt{N}\theta_N \sup_{\|h-h'\|<\delta} \left\{\left|\mathbb{P}_N(h-h')\right| + \left|\mathbb{P}^\star(h-h')\right|\right\} \\
&\le \sqrt{N}\theta_N \sup_{\|h-h'\|<\delta} \left|\mathbb{P}_N(h-h')\right| + \sqrt{N}\theta_N \sup_{\|h-h'\|<\delta} \left|\mathbb{P}^\star(h-h')\right|.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}^N\left(\sup_{\|h-h'\|_{L^2(\mathbb{P}^\star)}<\delta} |\mathbb{Q}_N(h-h')| > \varepsilon\right) \\
&\le \mathbb{P}^N\left(\sqrt{N}\theta_N \sup_{\|h-h'\|<\delta} |\mathbb{P}_N(h-h')| > \frac{\varepsilon}{2}\right) + \mathbb{P}^N\left(\sqrt{N}\theta_N \sup_{\|h-h'\|<\delta} |\mathbb{P}^\star(h-h')| > \frac{\varepsilon}{2}\right) \\
&=: A_N + B_N. \qquad \text{(A.31)}
\end{aligned}$$

Consider the term B_N in (A.31). For any h and h' such that $\|h-h'\| < \delta$, we have $|\mathbb{P}^\star(h-h')| = |\mathbb{E}_{\mathbb{P}^\star}(h-h')| \le \mathbb{E}_{\mathbb{P}}|h-h'| \le \sqrt{\mathbb{E}_{\mathbb{P}^\star}[(h-h')^2]} = \|h-h'\| < \delta$, where the second inequality follows from Cauchy-Schwartz inequality. Also, by assumption (b), we have $\sqrt{N}\theta_N = o(1)$. Therefore,

$$\limsup_{N\to\infty} B_N \le \mathbb{P}^\infty\left(\limsup_{N\to\infty}\left\{\sqrt{N}\theta_N\delta\right\} > \frac{\varepsilon}{2}\right) = 0. \qquad \text{(A.32)}$$

Consider the term A_N in (A.31). Note that for any $\mathbb{P}_N \in \mathcal{P}_N$,

$$\sup_{\|h-h'\|<\delta} \mathbb{E}_{\mathbb{P}_N}|h-h'| \le 2\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbb{P}_N}|h| \le 2\sup_{h\in\mathcal{H}}\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}|h|. \qquad \text{(A.33)}$$

By Lemma 1, (A.33) implies that $\limsup_{N\to\infty}\sup_{\|h-h'\|<\delta} \mathbb{E}_{\mathbb{P}_N}|h-h'| \le M''$ almost surely for some constant M''. Since $\sqrt{N}\theta_N = o(1)$, we have $\limsup_{N\to\infty}\left\{\sqrt{N}\theta_N \sup_{\|h-h'\|<\delta}\mathbb{E}_{\mathbb{P}_N}|h-h'|\right\} = 0$. Therefore, we have

$$\limsup_{N\to\infty} A_N \le \mathbb{P}^\infty\left(\limsup_{N\to\infty}\left\{\sqrt{N}\theta_N \sup_{\|h-h'\|<\delta} \mathbb{E}_{\mathbb{P}_N}|h-h'|\right\} > \frac{\varepsilon}{2}\right) = 0. \qquad \text{(A.34)}$$

Combining (A.32) and (A.34) with (A.31), we have

$$0 \le \limsup_{N\to\infty} \mathbb{P}^N\left(\sup_{\|h-h'\|_{L^2(\mathbb{P}^\star)}<\delta} |\mathbb{Q}_N(h-h')| > \varepsilon\right) \le \limsup_{N\to\infty} A_N + \limsup_{N\to\infty} B_N \le 0. \qquad \text{(A.35)}$$

Since (A.35) holds for any $\delta > 0$, this shows that $\mathbb{Q}_N$ is asymptotically uniform continuous in probability. Since (a) the metric space $(\mathcal{H}, \|\cdot\|_{L^2(\mathbb{P}^\star)})$ is totally bounded, (b) $\mathbb{Q}_N(h)$ is asymptotically tight for any $h \in \mathcal{H}$, and (c) $\mathbb{Q}_N$ is asymptotically uniform continuous in probability, $\mathbb{Q}_N$ is asymptotically tight by Theorem 1.5.7 of van der Vaart and Wellner (1996), implying that (A.30) equals zero. Since both (A.29) and (A.30) equal zero, (A.28) also equals zero, showing that the sequence $\{\mathbb{S}_N\}_{N\in\mathbb{N}}$ is asymptotically tight. This completes step 2.

Combining the two steps, we have (a) $\{\mathbb{S}_N\}$ is asymptotically tight and (b) the marginals $\big(\mathbb{S}_N(h_1),\ldots,\mathbb{S}_N(h_k)\big)$ converge weakly to $\big(\mathbb{G}'(h_1),\ldots,\mathbb{G}'(h_k)\big)$. It follows from Theorem 1.5.4 of van der Vaart and Wellner (1996) that $\mathbb{S}_N \Rightarrow \mathbb{G}'$. □

Appendix A.15. Proof of Theorem 9

Proof. Proof.

By our assumption, there exists $\mathbb{P}_N^\star \in \arg\max_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]$ such that $\mathbb{P}_N^\star \in \mathcal{P}_N$ for any $\boldsymbol{x}\in\mathcal{X}$. Thus, by Lemma 2, we have

$$\sqrt{N}\left[(1-\theta_N)\widehat{\mathbb{P}}_N(\cdot)+\theta_N\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{P}(\cdot)-\mathbb{P}^\star(\cdot)\right]\Rightarrow\mathbb{G}'(\cdot)\ \text{ in }\ \ell^\infty(\mathcal{H}),\tag{A.36}$$

where $\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{P}(\cdot)\in\mathcal{P}_N$ denotes the worst-case distribution of the input function $h\in\mathcal{H}$. Recall that $\mathcal{H}=\{f(\boldsymbol{x},\cdot)\mid \boldsymbol{x}\in\mathcal{X}\}$. Note that the map from $\ell^\infty(\mathcal{H})$ to $\ell^\infty(\mathcal{X})$ given by $g(\cdot)\mapsto g(h(\cdot,\cdot))$ is continuous, where $\ell^\infty(\mathcal{X})$ is the Banach space of bounded functions $\psi:\mathcal{X}\to\mathbb{R}$ equipped with the supremum norm $\|\psi\|=\sup_{\boldsymbol{x}\in\mathcal{X}}|\psi(\boldsymbol{x})|$. By continuous mapping theorem (see Theorem 1.3.6 of van der Vaart and Wellner, 1996), (A.36) implies that

$$\sqrt{N}\left[(1-\theta_N)\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\cdot,\boldsymbol{\xi})]+\theta_N\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\cdot,\boldsymbol{\xi})]-\mathbb{E}_{\mathbb{P}^\star}[f(\cdot,\boldsymbol{\xi})]\right]\Rightarrow\mathbb{G}(\cdot)\ \text{ in }\ \ell^\infty(\mathcal{X}).\tag{A.37}$$

Consider the functional $V:\ell^\infty(\mathcal{X})\to\mathbb{R}$ by $V(\psi)=\inf_{\boldsymbol{x}\in\mathcal{X}}\psi(\boldsymbol{x})$. Since $\mathcal{X}$ is compact by Assumption 1, the Hadamard directional derivative of V at ψ is given by $V'_\psi(\phi)=\inf_{\boldsymbol{x}\in B(\psi)}\phi(\boldsymbol{x})$, where $B(\psi)=\arg\min_{\boldsymbol{x}\in\mathcal{X}}\psi(\boldsymbol{x})$ (see, e.g., Corollary 2.2 of Cárcamo et al., 2020). Thus, together with (A.37), applying the Delta's method (see, e.g., Theorem 2.2 of Cárcamo et al., 2020), we obtain the desired assertions: (i) $\sqrt{N}(\widehat{v}_N-v^\star)\Rightarrow\inf_{\boldsymbol{x}\in\mathcal{X}^\star}\mathbb{G}(\boldsymbol{x})$ and (ii)

$$\widehat{v}_N-v^\star=\inf_{\boldsymbol{x}\in\mathcal{X}^\star}\left\{(1-\theta_N)\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]+\theta_N\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]-\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]\right\}+o_{\mathbb{P}^\star}(N^{-1/2}).\tag{A.38}$$

Finally, since $-\infty<v^\star=\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]$ for any $\boldsymbol{x}\in\mathcal{X}^\star$, (A.38) directly implies $\widehat{v}_N=\inf_{\boldsymbol{x}\in\mathcal{X}^\star}\big\{(1-\theta_N)\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x},\boldsymbol{\xi})]+\theta_N\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\big\}+o_{\mathbb{P}^\star}(N^{-1/2})$. This completes the proof. □

Appendix A.16. Proof of Theorem 10

Proof. Proof. Using (23), we can rewrite the objective function of the TRO model with shape parameter $\mathcal{P}_{N,r_N}$ as follows:

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}'_{N,\theta}}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]&=(1-\theta_N)\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]+\theta_N\sup_{\mathbb{P}\in\mathcal{P}_{N,r_N}}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\\&=\mathbb{E}_{\widehat{\mathbb{P}}_N}[f(\boldsymbol{x},\boldsymbol{\xi})]+\big(\theta_N r_N^\gamma\big)g_N(\boldsymbol{x})+\big(\theta_N r_N^\gamma\big)\varepsilon_N(\boldsymbol{x}),\end{aligned}$$

which resembles the expansion (23) with r_N^γ replaced by $\theta_N r_N^\gamma$. Thus, we can prove the desired assertions by following the same proof techniques of Theorem 1 in Blanchet and Shapiro (2023). $\square$

Appendix B. An Example of a Sequence of TRO Ambiguity Sets

Let $\widehat{\mathbb{P}}_N = \delta_0$, i.e., the Dirac measure on 0, and $\mathcal{P}_N = \{(1-t)\delta_1 + t\delta_e \mid t \in [0,1],\, e \in \{0,2\}\}$. That is, $\mathcal{P}_N$ contains the one-point distributions $\{\delta_0, \delta_1, \delta_2\}$, as well as all two-point distributions with support on either $\{0,1\}$ or $\{1,2\}$. Note that $\mathcal{P}_N$ is star-shaped with a star center δ_1. Indeed, for any $\alpha \in [0,1]$ and $\mathbb{Q} = (1-t)\delta_1 + t\delta_e \in \mathcal{P}_N$, we have

$$(1-\alpha)\delta_1 + \alpha\mathbb{Q} = (1-\alpha)\delta_1 + \alpha\big[(1-t)\delta_1 + t\delta_e\big] = (1-\alpha t)\delta_1 + \alpha t\delta_e \in \mathcal{P}_N$$

since $\alpha t \in [0,1]$. However, $\widehat{\mathbb{P}}_N = \delta_0$ is not a star center. To see this, note that

$$\frac{1}{2}\delta_0 + \frac{1}{2}\left(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2\right) = \frac{1}{2}\delta_0 + \frac{1}{4}\delta_1 + \frac{1}{4}\delta_2 \notin \mathcal{P}_N$$

since it is a three-point distribution.

Now, we show that $\mathcal{P}'_{N,\theta}$ is *not* non-decreasing. In particular, we show that for any $0 < \theta_1 < \theta_2 \le 1$, there exists $\mathbb{M} \in \mathcal{P}'_{N,\theta_1}$ but $\mathbb{M} \notin \mathcal{P}'_{N,\theta_2}$. Indeed, since $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2 \in \mathcal{P}_N$, we construct the measure $\mathbb{M} \in \mathcal{P}'_{N,\theta_1}$ as follows:

$$\mathbb{M} = (1-\theta_1)\delta_0 + \theta_1\left(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2\right). \tag{B.1}$$

We show that $\mathbb{M}$ defined in (B.1) does not belong to $\mathcal{P}'_{N,\theta_2}$. Note that

$$\begin{aligned}\mathbb{M} = (1-\theta_1)\delta_0 + \theta_1\left(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2\right) &= (1-\theta_2)\delta_0 + (\theta_2-\theta_1)\delta_0 + \theta_1\left(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2\right)\\ &= (1-\theta_2)\delta_0 + \theta_2\left\{\left(1-\frac{\theta_1}{\theta_2}\right)\delta_0 + \frac{\theta_1}{\theta_2}\left(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2\right)\right\}.\end{aligned}$$

Since $(1-\frac{\theta_1}{\theta_2})\delta_0 + \frac{\theta_1}{\theta_2}(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2)$ is a three-point distribution that does not belong to $\mathcal{P}_N$, we have $\mathbb{M} \notin \mathcal{P}'_{N,\theta_2}$.

Appendix C. Robust Optimization Ambiguity Set

Proposition 4. *For a fixed $\boldsymbol{x} \in \mathcal{X}$, if there exists $\boldsymbol{\xi}_0 \in \mathcal{U}$ such that $f(\boldsymbol{x},\boldsymbol{\xi}_0) \ge f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i)$ for all $i \in \{1,\dots,N\}$, then $\sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi}) = \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]$, where $\mathcal{P}_N = \operatorname{conv}\big(\widehat{\mathbb{P}}_N \cup \{\delta_{\boldsymbol{\xi}} \mid \boldsymbol{\xi} \in \mathcal{U}\}\big)$.*

Proof. Proof. First, if $\sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi}) = \infty$, then there exists a sequence $\{\boldsymbol{\xi}_j\}_{j\in\mathbb{N}} \subset \mathcal{U}$ such that $f(\boldsymbol{x},\boldsymbol{\xi}_j) \to \infty$ as $j \to \infty$. Since $\boldsymbol{\xi}_j \in \mathcal{U}$, we have $\delta_{\boldsymbol{\xi}_j} \in \mathcal{P}_N$ and $\mathbb{E}_{\delta_{\boldsymbol{\xi}_j}}[f(\boldsymbol{x},\boldsymbol{\xi})] = f(\boldsymbol{x},\boldsymbol{\xi}_j) \to \infty$ as $j \to \infty$. This shows that $\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] = \infty$.

Now, assume that $\sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi}) < \infty$, and thus $f(\boldsymbol{x},\boldsymbol{\xi}) < \infty$ for all $\boldsymbol{\xi} \in \mathcal{U}$. Note that for any $\mathbb{P} \in \mathcal{P}_N$, by definition of $\mathcal{P}_N$, there exists $K \in \mathbb{N}$, $\lambda \in [0,1]$, and $\{\alpha_k\}_{k=1}^K \subseteq [0,1]$ with $\alpha_k \ge 0$ and

$\sum_{k=1}^K \alpha_k = 1$ such that $\mathbb{P} = (1-\lambda)\widehat{\mathbb{P}}_N + \lambda \sum_{k=1}^K \alpha_i \delta_{\bar{\boldsymbol{\xi}}_k}$ for some $\bar{\boldsymbol{\xi}}_k \in \mathcal{U}$, $k \in \{1,\dots,K\}$. Then,

$$
\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \\
&= \sup\left\{ (1-\lambda)\cdot\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i) + \lambda\sum_{k=1}^K \alpha_k f(\boldsymbol{x},\bar{\boldsymbol{\xi}}_k) \;\middle|\; \begin{array}{l} K\in\mathbb{N},\ \lambda\in[0,1],\ \alpha_k\in[0,1], \\ \sum_{k=1}^K \alpha_k = 1,\ \bar{\boldsymbol{\xi}}_k\in\mathcal{U},\ k\in\{1,\dots,K\} \end{array} \right\} \quad \text{(C.1)} \\
&= \sup_{\lambda\in[0,1]} \left\{ (1-\lambda)\cdot\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i) + \lambda\cdot\sup\left\{ \sum_{k=1}^K \alpha_k f(\boldsymbol{x},\bar{\boldsymbol{\xi}}_k) \;\middle|\; \begin{array}{l} K\in\mathbb{N},\ \alpha_k\in[0,1], \sum_{k=1}^K\alpha_k=1, \\ \bar{\boldsymbol{\xi}}_k\in\mathcal{U},\ k\in\{1,\dots,K\} \end{array} \right\} \right\} \\
&= \sup_{\lambda\in[0,1]} \left\{ (1-\lambda)\cdot\frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i) + \lambda\cdot\sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi}) \right\} \quad \text{(C.2)} \\
&= \sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi}). \quad \text{(C.3)}
\end{aligned}
$$

Equality (C.1) follows from $\mathbb{P}\in\mathcal{P}_N$ and the definition of $\mathcal{P}_N$. Equality (C.2) follows from the fact that

$$
\sum_{k=1}^K \alpha_k f(\boldsymbol{x},\bar{\boldsymbol{\xi}}_k) \le \sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi}) \le \sup\left\{ \sum_{k=1}^K \alpha_k f(\boldsymbol{x},\bar{\boldsymbol{\xi}}_k) \;\middle|\; \begin{array}{l} K\in\mathbb{N},\ \alpha_k\in[0,1], \sum_{k=1}^K\alpha_k=1, \\ \bar{\boldsymbol{\xi}}_k\in\mathcal{U},\ k\in\{1,\dots,K\} \end{array} \right\} \quad \text{(C.4)}
$$

for any $k\in\mathbb{N}$, $\alpha_k\in[0,1]$ and $\bar{\boldsymbol{\xi}}_k\in\mathcal{U}$ for $k\in\{1,\dots,K\}$ with $\sum_{k=1}^K\alpha_k = 1$. Here, the first inequality in (C.4) follows from $f(\boldsymbol{x},\bar{\boldsymbol{\xi}}) \le \sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi})$ for all $\bar{\boldsymbol{\xi}}\in\mathcal{U}$ while the second inequality in (C.4) follows from letting $K=1$. Taking supremum over K, $\{\alpha_k\}_{k=1}^K$ and $\{\bar{\boldsymbol{\xi}}_k\}_{k=1}^K$ in (C.4), we obtain the desired equality in (C.2). Finally, equality (C.3) follows from the fact that $\lambda=1$ is optimal to (C.2) since

$$
f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i) \le f(\boldsymbol{x},\boldsymbol{\xi}_0) \le \sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi})
$$

for all $i\in\{1,\dots,N\}$, where the first inequality follows from our assumption that $f(\boldsymbol{x},\boldsymbol{\xi}_0)\ge f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i)$ for all $i\in\{1,\dots,N\}$. This shows that $(1/N)\sum_{i=1}^N f(\boldsymbol{x},\widehat{\boldsymbol{\xi}}_i) \le \sup_{\boldsymbol{\xi}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{\xi})$. □

Appendix D. Convergence of Distance-Based Ambiguity Sets

Proposition 5. *Consider the distance-based ambiguity set* $\mathcal{P}_N = \{\mathbb{P}\in\mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P},\widehat{\mathbb{P}}_N)\le r\}$ *for some radius* $r>0$, *where* d *is any statistical distance satisfying*

- *(i) (normalization)* $\mathsf{d}(\mathbb{P},\mathbb{P}) = 0$ *for any* $\mathbb{P}\in\mathcal{P}(\Xi)$,
- *(ii) (symmetry)* $\mathsf{d}(\mathbb{P}_1,\mathbb{P}_2) = \mathsf{d}(\mathbb{P}_2,\mathbb{P}_1)$ *for any* $\{\mathbb{P}_1,\mathbb{P}_2\}\subseteq\mathcal{P}(\Xi)$,
- *(iii) (triangle inequality)* $\mathsf{d}(\mathbb{P}_1,\mathbb{P}_2)\le\mathsf{d}(\mathbb{P}_1,\mathbb{P}_3)+\mathsf{d}(\mathbb{P}_3,\mathbb{P}_2)$ *for any* $\{\mathbb{P}_1,\mathbb{P}_2,\mathbb{P}_3\}\subseteq\mathcal{P}(\Xi)$, *and*
- *(iv) (convexity)* d *is convex in the first argument.*

Let $\widehat{\mathcal{P}} = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P}, \mathbb{P}^\star) \leq r\}$*. If* $\Delta := \sup_{\mathbb{P}_1 \in \mathcal{P}(\Xi), \mathbb{P}_2 \in \mathcal{P}(\Xi)} \mathbb{d}(\mathbb{P}_1, \mathbb{P}_2) < \infty$ *and* $\mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star) \to 0$ *almost surely, then* $\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) \to 0$ *almost surely as* $N \to \infty$*.*

Proof. Proof. The idea of the proof follows from the proof of Hoffman's Lemma for moment problems (see Theorem 2 in Liu et al., 2019). First, we claim that for any $\mathbb{P}_1 \in \mathcal{P}_N$, $(1-\rho)\mathbb{P}_1 + \rho\mathbb{P}^\star \in \widehat{\mathcal{P}}$, where $\rho = \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)/\big(r + \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)\big)$. To see this, note that

$$\begin{aligned}
\mathsf{d}\big((1-\rho)\mathbb{P}_1 + \rho\mathbb{P}^\star, \mathbb{P}^\star\big) &\leq (1-\rho)\mathsf{d}(\mathbb{P}_1, \mathbb{P}^\star) \\
&\leq \frac{r}{r + \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)}\big[\mathsf{d}(\mathbb{P}_1, \widehat{\mathbb{P}}_N) + \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)\big] \\
&\leq \frac{r}{r + \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)}\big[r + \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)\big] = r,
\end{aligned} \tag{D.1}$$

where the first inequality follows from properties (iv) and (i), the second inequality follows from the definition of ρ and property (iii). Then, we have

$$\begin{aligned}
\mathbb{D}(\mathbb{P}_1, \widehat{\mathcal{P}}) = \inf_{\mathbb{P} \in \widehat{\mathcal{P}}} \mathbb{d}(\mathbb{P}_1, \mathbb{P}) &\leq \mathbb{d}\big(\mathbb{P}_1, (1-\rho)\mathbb{P}_1 + \rho\mathbb{P}^\star\big) \\
&= \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| \mathbb{E}_{\mathbb{P}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \Big\{(1-\rho)\mathbb{E}_{\mathbb{P}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] + \rho\mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}, \boldsymbol{\xi})]\Big\} \Big| \\
&= \rho \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| \mathbb{E}_{\mathbb{P}_1}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}^\star}[f(\boldsymbol{x}, \boldsymbol{\xi})] \Big| \\
&= \frac{\mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)}{r + \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)} \cdot \mathbb{d}(\mathbb{P}_1, \mathbb{P}^\star) \leq \frac{\Delta}{r} \cdot \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star),
\end{aligned} \tag{D.2}$$

where the last inequality follows from the fact that $\mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star) \geq 0$ and $\mathbb{d}(\mathbb{P}_1, \mathbb{P}^\star) \leq \Delta$. This implies that $\sup_{\mathbb{P}_1 \in \mathcal{P}_N} \mathbb{D}(\mathbb{P}_1, \widehat{\mathcal{P}}) \leq (\Delta/r)\mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)$. Similarly, following a similar argument in (D.1), we can show that for any $\mathbb{P}_2 \in \widehat{\mathcal{P}}$, $(1-\tau)\mathbb{P}_2 + \tau\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, where $\tau = \mathsf{d}(\mathbb{P}^\star, \widehat{\mathbb{P}}_N)/\big(r + \mathsf{d}(\mathbb{P}^\star, \widehat{\mathbb{P}}_N)\big)$. A similar argument as in (D.2) shows that

$$\mathbb{D}(\mathbb{P}_2, \widehat{\mathcal{P}}_N) \leq \frac{\Delta}{r} \cdot \mathsf{d}(\mathbb{P}^\star, \widehat{\mathbb{P}}_N) = \frac{\Delta}{r} \cdot \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star),$$

where the last inequality follows from property (ii). This implies that $\sup_{\mathbb{P}_2 \in \widehat{\mathcal{P}}} \mathbb{D}(\mathbb{P}_2, \widehat{\mathcal{P}}_N) \leq (\Delta/r)\mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star)$. Therefore, we have

$$\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) = \max\left\{ \sup_{\mathbb{P}_1 \in \mathcal{P}_N} \mathbb{D}(\mathbb{P}_1, \widehat{\mathcal{P}}), \sup_{\mathbb{P}_2 \in \widehat{\mathcal{P}}} \mathbb{D}(\mathbb{P}_2, \mathcal{P}_N) \right\} \leq \frac{\Delta}{r} \cdot \mathsf{d}(\widehat{\mathbb{P}}_N, \mathbb{P}^\star) \to 0$$

almost surely as $N \to \infty$. □

Appendix E. Some Quantitative Stability Analysis Results

Consider two DRO models with two different ambiguity sets:

$$\upsilon_i = \inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}_i} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{\xi})]$$

for $i \in \{1, 2\}$, where we assume that υ_i is finite and the set of optimal solutions $\mathcal{X}_i^\star$ is non-empty. In quantitative stability analysis, we analyze how the change in the ambiguity set would affect the optimal value and the set of optimal solutions to the DRO model. In this appendix, we summarize some relevant results in the existing literature, in particular, the upper bounds on the differences between the optimal values and the set of optimal solutions from the two DRO models (see, e.g., Liu and Xu, 2013; Pichler and Xu, 2018; Sun and Xu, 2016). For the sake of completeness, we also provide the proof of these results.

Proposition 6. *The (pointwise) absolute difference between the two objective functions is upper bounded by the Hausdorff distance between the two ambiguity sets, i.e., for any $\boldsymbol{x} \in \mathcal{X}$,*

$$\left| \sup_{\mathbb{Q} \in \mathcal{P}_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \sup_{\mathbb{Q}' \in \mathcal{P}_2} \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})] \right| \leq \mathbb{H}(\mathcal{P}_1, \mathcal{P}_2). \tag{E.1}$$

Proof. Proof. By definition of the pseudometric $\mathbb{d}$ in (11), for any $\mathbb{Q} \in \mathcal{P}_1$ and $\mathbb{Q}' \in \mathcal{P}_2$, we have $\left|\mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})]\right| \leq \mathbb{d}(\mathbb{Q}, \mathbb{Q}')$ for all $\boldsymbol{x} \in \mathcal{X}$. Since

$$\begin{aligned}
\sup_{\mathbb{Q} \in \mathcal{P}_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \sup_{\mathbb{Q}' \in \mathcal{P}_2} \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})] &= \sup_{\mathbb{Q} \in \mathcal{P}_1} \inf_{\mathbb{Q}' \in \mathcal{P}_2} \Big\{ \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})] \Big\} \\
&\leq \sup_{\mathbb{Q} \in \mathcal{P}_1} \inf_{\mathbb{Q}' \in \mathcal{P}_2} \mathbb{d}(\mathbb{Q}, \mathbb{Q}')
\end{aligned}$$

and

$$\begin{aligned}
\sup_{\mathbb{Q}' \in \mathcal{P}_2} \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \sup_{\mathbb{Q} \in \mathcal{P}_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] &= \sup_{\mathbb{Q}' \in \mathcal{P}_2} \inf_{\mathbb{Q} \in \mathcal{P}_1} \Big\{ \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] \Big\} \\
&\leq \sup_{\mathbb{Q}' \in \mathcal{P}_2} \inf_{\mathbb{Q} \in \mathcal{P}_1} \mathbb{d}(\mathbb{Q}, \mathbb{Q}'),
\end{aligned}$$

we obtain

$$\begin{aligned}
\left| \sup_{\mathbb{Q} \in \mathcal{P}_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x}, \boldsymbol{\xi})] - \sup_{\mathbb{Q}' \in \mathcal{P}_2} \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x}, \boldsymbol{\xi})] \right| &\leq \max \left\{ \sup_{\mathbb{Q} \in \mathcal{P}_1} \inf_{\mathbb{Q}' \in \mathcal{P}_2} \mathbb{d}(\mathbb{Q}, \mathbb{Q}'), \sup_{\mathbb{Q}' \in \mathcal{P}_2} \inf_{\mathbb{Q} \in \mathcal{P}_1} \mathbb{d}(\mathbb{Q}, \mathbb{Q}') \right\} \\
&= \mathbb{H}(\mathcal{P}_1, \mathcal{P}_2).
\end{aligned}$$

This completes the proof. □

Proposition 7. *The following assertions hold.*

(i) $|\upsilon_1 - \upsilon_2| \leq \mathbb{H}(\mathcal{P}_1, \mathcal{P}_2)$.

(ii) If, in addition, $\sup_{\mathbb{P} \in \mathcal{P}_1} \mathbb{E}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ *satisfies the second order growth condition at* $\mathcal{X}_1^\star$*, i.e., there exists* $\tau > 0$ *such that*

$$\sup_{\mathbb{P} \in \mathcal{P}_1} \mathbb{E}[f(\boldsymbol{x}, \boldsymbol{\xi})] \geq \upsilon_1 + \tau [d(\boldsymbol{x}, \mathcal{X}_1^\star)]^2$$

for all $\boldsymbol{x} \in \mathcal{X}$*, then*

$$D(\mathcal{X}_2^\star, \mathcal{X}_1^\star) \leq \sqrt{\frac{3}{\tau} \mathbb{H}(\mathcal{P}_1, \mathcal{P}_2)}.$$

Proof. Proof. Part (i) follows directly from

$$\begin{aligned}
|v_1 - v_2| &= \left| \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_1} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_2} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \right| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \sup_{\mathbb{Q}\in\mathcal{P}_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \sup_{\mathbb{Q}'\in\mathcal{P}_2} \mathbb{E}_{\mathbb{Q}'}[f(\boldsymbol{x},\boldsymbol{\xi})] \right| \\
&\leq \mathbb{H}(\mathcal{P}_1,\mathcal{P}_2),
\end{aligned}$$

where the last inequality follows from Proposition 6. For part (ii), suppose, on the contrary, that

$$D(\mathcal{X}_2^\star,\mathcal{X}_1^\star) = \sup_{\boldsymbol{x}\in\mathcal{X}_2^\star} d(\boldsymbol{x},\mathcal{X}_1^\star) > \sqrt{\frac{3}{\tau}\mathbb{H}(\mathcal{P}_1,\mathcal{P}_2)}.$$

Then, there exists $\overline{\boldsymbol{x}} \in \mathcal{X}_2^\star$ such that $d(\overline{\boldsymbol{x}},\mathcal{X}_1^\star) > \sqrt{3\tau^{-1}\mathbb{H}(\mathcal{P}_1,\mathcal{P}_2)}$. The second order growth condition implies that

$$\sup_{\mathbb{P}\in\mathcal{P}_1} \mathbb{E}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] - v_1 \geq \tau\big[d(\overline{\boldsymbol{x}},\mathcal{X}_1^\star)\big]^2 > 3\,\mathbb{H}(\mathcal{P}_1,\mathcal{P}_2). \tag{E.2}$$

However, by Proposition 6 and part (i), we have

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}_1} \mathbb{E}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] - v_1 &\leq \left\{ \sup_{\mathbb{P}\in\mathcal{P}_2} \mathbb{E}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] + \mathbb{H}(\mathcal{P}_1,\mathcal{P}_2) \right\} - \big[v_2 - \mathbb{H}(\mathcal{P}_1,\mathcal{P}_2)\big] \\
&= \sup_{\mathbb{P}\in\mathcal{P}_2} \mathbb{E}[f(\overline{\boldsymbol{x}},\boldsymbol{\xi})] - v_2 + 2\,\mathbb{H}(\mathcal{P}_1,\mathcal{P}_2) \\
&= 2\,\mathbb{H}(\mathcal{P}_1,\mathcal{P}_2),
\end{aligned}$$

where the first equality follows from $\overline{\boldsymbol{x}} \in \mathcal{X}_2^\star$. Thus, this contradicts with (E.2). □

Appendix F. Details of Numerical Experiments

Appendix F.1. Inventory Control – Reformulations

Recall the TRO model (24) under shape parameters (a)–(d) in Table 1. This model cannot be solved directly because of the inner supremum problem $\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+ - \xi\big]$. In this section, we provide tractable reformulation to this inner supremum problem, and thus our TRO model.

First, consider ambiguity set (a) in Table 1. We have

$$\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+ - \xi\big] = \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+\big] - \widehat{\mu}_N = \frac{1}{2}\left[-(x - \widehat{\mu}_N) + \sqrt{\widehat{\sigma}_N^2 + (x - \widehat{\mu}_N)^2} \right] - \widehat{\mu}_N,$$

where the first equality follows from $\mathbb{E}_{\mathbb{P}}(\xi) = \widehat{\mu}_N$ for any $\mathbb{P} \in \mathcal{P}_N$, and the second equality follows from Lemma 2.2 of Chen et al. (2011). Therefore, the TRO model under set (a) is equivalent to

$$\underset{x\geq 0}{\text{minimize}}\ (c-h)x+(p-h)\left\{(1-\theta)\frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i-x)_+-\widehat{\xi}_i\big]+\theta\left\{\frac{1}{2}\left[\sqrt{\widehat{\sigma}_N^2 + (x - \widehat{\mu}_N)^2}-(x-\widehat{\mu}_N)\right]-\widehat{\mu}_N\right\}\right\}.$$

Next, consider ambiguity set (b) in Table 1. Note that $(\xi - x)_+ - \xi = \max\{-x, -\xi\}$ is piecewise linear. By the strong duality result in Theorem 6.3 and Remark 6.6 of Mohajerin Esfahani and Kuhn (2018), we have

$$\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+ - \xi\big] = r + \frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i\big].$$

Therefore, the TRO model under set (b) is equivalent to

$$\underset{x\geq 0}{\text{minimize}} \quad (c-h)x + (p-h)\bigg\{(1-\theta)\frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i\big] + \theta\bigg\{r + \frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i\big]\bigg\}\bigg\}.$$

Now, consider ambiguity set (c) in Table 1. By the strong duality result for ϕ-divergence DRO problems in Bayraksan and Love (2015), we have

$$\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+ - \xi\big] = \inf_{\lambda\geq 0,\,\tau}\left\{\tau + \lambda r - \frac{\lambda}{N}\sum_{i=1}^{N}\log\left(1 - \frac{(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i - \tau}{\lambda}\right)\right\},$$

where $-0\log(1 - s/0) = 0$ for $s \leq 0$ and $-0\log(1-s/0) = \infty$ for $s > 0$. Therefore, the TRO model under set (c) is equivalent to

$$\begin{aligned}\underset{x\geq 0,\,\lambda\geq 0,\,\tau}{\text{minimize}} \quad & (c-h)x + (p-h)\bigg\{(1-\theta)\frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i\big] \\ & + \theta\bigg\{\tau + \lambda r - \frac{\lambda}{N}\sum_{i=1}^{N}\log\left(1 - \frac{(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i - \tau}{\lambda}\right)\bigg\}\bigg\}.\end{aligned}$$

Finally, consider ambiguity set (d) in Table 1. For notational simplicity, let $\gamma := t_{N-1,\alpha/2}\widehat{\sigma}_N/\sqrt{N}$. Then, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}\big[(\xi - x)_+ - \xi\big] = \sup_{\xi:|\xi-\widehat{\mu}_N|\leq\gamma}\big\{(\xi - x)_+ - \xi\big\} = \max\big\{-x, -\widehat{\mu}_N + \gamma\big\},$$

where the last equality follows from the fact that $(\xi - x)_+ - \xi = \max\{-x, -\xi\}$ and the objective function is non-increasing in ξ. Thus, the supremum over $\xi \in [\widehat{\mu}_N - \gamma, \widehat{\mu}_N + \gamma]$ is attained at the lower bound. Therefore, the TRO model under set (d) is equivalent to

$$\underset{x\geq 0}{\text{minimize}} \quad (c-h)x + (p-h)\bigg\{(1-\theta)\frac{1}{N}\sum_{i=1}^{N}\big[(\widehat{\xi}_i - x)_+ - \widehat{\xi}_i\big] + \theta\max\big\{-x, -\widehat{\mu}_N + \gamma\big\}\bigg\}.$$

Appendix F.2. Inventory Control – Ambiguity Set (d)

We show that ambiguity set (d) in Table 1, i.e., $\mathcal{P}_N = \{\delta_\xi \mid |\xi - \widehat{\mu}_N| \leq t_{N-1,\alpha/2}\widehat{\sigma}_N/\sqrt{N}\}$, satisfies Assumption 3(b). Recall that $f(x,\xi) = (\xi - x)_+ - \xi$, which is Lipschitz in ξ with modulus

1. We claim that $\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) \to 0$ as $N \to \infty$ almost surely, where $\widehat{\mathcal{P}} = \{\delta_{\mu^\star}\}$ with $\mu^\star = \mathbb{E}_{\mathbb{P}^\star}(\xi)$. Indeed,

$$\begin{aligned}
\mathbb{H}(\mathcal{P}_N, \widehat{\mathcal{P}}) = \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{D}(\mathbb{P}, \widehat{\mathcal{P}}) &= \sup_{\xi:|\xi-\widehat{\mu}_N|\le t_{N-1,\alpha/2}\widehat{\sigma}_N/\sqrt{N}} \sup_{x\ge 0} \left| f(x,\xi) - f(x,\mu^\star) \right| && \text{(F.1)}\\
&\le \sup_{\xi:|\xi-\widehat{\mu}_N|\le t_{N-1,\alpha/2}\widehat{\sigma}_N/\sqrt{N}} |\xi - \mu^\star| && \text{(F.2)}\\
&= \max\left\{ \left| \widehat{\mu}_N - \mu^\star - t_{N-1,\alpha/2}\frac{\widehat{\sigma}_N}{\sqrt{N}} \right|, \left| \widehat{\mu}_N - \mu^\star + t_{N-1,\alpha/2}\frac{\widehat{\sigma}_N}{\sqrt{N}} \right| \right\} && \text{(F.3)}\\
&\to 0
\end{aligned}$$

almost surely. Here, (F.1) follows from the definitions in (12) and (11), and that $\sup_{\mathbb{P}\in\widehat{\mathcal{P}}} \mathbb{D}(\mathbb{P}, \widehat{\mathcal{P}}) \le \sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{D}(\mathbb{P}, \widehat{\mathcal{P}})$; (F.2) follows from the Lipschitz continuity of $f(x,\xi)$ in ξ; (F.3) follows from the fact that the supremum over a compact interval of the absolute value function $|\xi - \mu^\star|$ is attained either at the lower or the upper bound of the interval. Finally, by the strong law of large numbers, $|\widehat{\mu}_N - \mu^\star| \to 0$ and $|\widehat{\sigma}_N - \sigma^\star| \to 0$ almost surely, where $\sigma^\star = \mathrm{Var}_{\mathbb{P}^\star}(\xi) < \infty$. Also, $t_{N-1,\alpha/2} \to z_{\alpha/2} < \infty$, where $z_{\alpha/2}$ is the upper $(1-\alpha)/2$-th quantile of a standard normal distribution. Thus, we have (F.3) converges to zero almost surely. Thus, ambiguity set (d) satisfies Assumption 3(b).

Appendix F.3. Portfolio Optimization – Reformulations

Recall the TRO model (28) under shape paraemters (a)–(c) in Table 4. Again, this model cannot be solved directly because of the inner supremum problem $\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, t, \boldsymbol{\xi})]$. In this section, we provide tractable reformulation to this inner supremum problem, and thus our TRO model.

First, consider ambiguity set (a) in Table 4. By Lemma 2.2 and Lemma 2.4 of Chen et al. (2011), we have

$$\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, t, \boldsymbol{\xi})] = (1-\beta)t - \beta\boldsymbol{x}^\top \widehat{\mu}_N + \frac{1-\beta}{1-\alpha}\cdot\frac{1}{2}\left[-\boldsymbol{x}^\top \widehat{\boldsymbol{\mu}}_N - t + \sqrt{\boldsymbol{x}^\top \widehat{\boldsymbol{\Sigma}}_N \boldsymbol{x} + (-\boldsymbol{x}^\top \widehat{\boldsymbol{\mu}}_N - t)^2} \right].$$

Therefore, the TRO model under set (a) is equivalent to

$$\begin{aligned}
\underset{\boldsymbol{x}\in\mathcal{X},\, t\in\mathbb{R}}{\text{minimize}} \quad & (1-\theta)\cdot\frac{1}{N}\sum_{i=1}^{N}\left\{ (1-\beta)t + \beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i) + \frac{1-\beta}{1-\alpha}(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i - t)^+ \right\}\\
& + \theta\left\{ (1-\beta)t - \beta\boldsymbol{x}^\top\widehat{\mu}_N + \frac{1-\beta}{1-\alpha}\cdot\frac{1}{2}\left[-\boldsymbol{x}^\top\widehat{\boldsymbol{\mu}}_N - t + \sqrt{\boldsymbol{x}^\top\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{x} + (-\boldsymbol{x}^\top\widehat{\boldsymbol{\mu}}_N - t)^2} \right] \right\}.
\end{aligned}$$

Now, consider ambiguity set (b) in Table 4. By Theorem 6.3 of Mohajerin Esfahani and Kuhn (2018), we have

$$\sup_{\mathbb{P}\in\mathcal{P}_N} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, t, \boldsymbol{\xi})] = (1-\beta)t + r\left(\beta + \frac{1-\beta}{1-\alpha}\right)\|\boldsymbol{x}\|_\infty + \frac{1}{N}\sum_{i=1}^{N}\left[\beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i) + \frac{1-\beta}{1-\alpha}(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i - t)_+ \right].$$

Therefore, the TRO model under set (b) is equivalent to

$$\begin{aligned}\underset{\boldsymbol{x}\in\mathcal{X},\, t\in\mathbb{R}}{\text{minimize}} \quad & (1-\theta)\cdot\frac{1}{N}\sum_{i=1}^{N}\left\{(1-\beta)t+\beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i)+\frac{1-\beta}{1-\alpha}(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i-t)^+\right\} \\ & +\theta\left\{(1-\beta)t+r\left(\beta+\frac{1-\beta}{1-\alpha}\right)\|\boldsymbol{x}\|_\infty+\frac{1}{N}\sum_{i=1}^{N}\left[\beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i)+\frac{1-\beta}{1-\alpha}(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i-t)_+\right]\right\}.\end{aligned}$$

Finally, consider ambiguity set (c). Note that ambiguity set (c) is characterized by the ℓ_1 norm constraint. Thus, as in Lemma 3.1 of Huang et al. (2021), using linear programming duality, $\sup_{\mathbb{P}\in\mathcal{P}_N}\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x},t,\boldsymbol{\xi})]$ is equivalent to

$$\underset{\tau\in\mathbb{R},\,\lambda\in\mathbb{R}}{\text{minimize}} \quad (1-\beta)t+\tau+r\lambda+\frac{1}{N}\sum_{i=1}^{N}(u_i^+-u_i^-) \tag{F.4a}$$

$$\text{subject to} \quad \tau\geq\beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i)+\frac{1-\beta}{1-\alpha}\big(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i-t\big)_+-u_i^++u_i^-, \quad \forall i\in\{1,\ldots,n\}, \tag{F.4b}$$

$$\lambda\geq u_i^++u_i^-, \quad \forall i\in\{1,\ldots,n\}, \tag{F.4c}$$

$$\lambda\geq 0,\ u_i^+\geq 0,\ u_i^-\geq 0, \quad \forall i\in\{1,\ldots,n\}. \tag{F.4d}$$

Therefore, the TRO model under set (c) is equivalent to

$$\begin{aligned}\underset{\boldsymbol{x}\in\mathcal{X},\, t,\, \tau,\, \lambda,\, u^+,\, u^-}{\text{minimize}} \quad & (1-\theta)\cdot\frac{1}{N}\sum_{i=1}^{N}\left\{(1-\beta)t+\beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i)+\frac{1-\beta}{1-\alpha}(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i-t)^+\right\} \\ & \theta\left\{(1-\beta)t+\tau+r\lambda+\frac{1}{N}\sum_{i=1}^{N}(u_i^+-u_i^-)\right\}\end{aligned} \tag{F.5a}$$

$$\text{subject to} \quad \tau\geq\beta(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i)+\frac{1-\beta}{1-\alpha}\big(-\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}_i-t\big)_+-u_i^++u_i^-, \quad \forall i\in\{1,\ldots,n\}, \tag{F.5b}$$

$$\lambda\geq u_i^++u_i^-, \quad \forall i\in\{1,\ldots,n\}, \tag{F.5c}$$

$$\lambda\geq 0,\ u_i^+\geq 0,\ u_i^-\geq 0, \quad \forall i\in\{1,\ldots,n\}. \tag{F.5d}$$

Appendix F.4. Portfolio Optimization – Lipschitz Continuity

In this section, we show that the function $f(\boldsymbol{x},t,\boldsymbol{\xi})=(1-\beta)t+\beta(-\boldsymbol{x}^\top\boldsymbol{\xi})+[(1-\beta)/(1-\alpha)](-\boldsymbol{x}^\top\boldsymbol{\xi}-t)^+$ is Lipschitz continuous in $(\boldsymbol{x},t)$. Indeed, for any $\{(\boldsymbol{x}_1,t_1),(\boldsymbol{x}_2,t_2)\}\subseteq\mathcal{X}\times\mathbb{R}$, we have

$$
\begin{aligned}
&|f(\boldsymbol{x}_1, t_1, \boldsymbol{\xi}) - f(\boldsymbol{x}_2, t_2, \boldsymbol{\xi})| \\
&= \left| \left[(1-\beta)t_1 + \beta(-\boldsymbol{x}_1^\top \boldsymbol{\xi}) + \frac{1-\beta}{1-\alpha}(-\boldsymbol{x}_1^\top \boldsymbol{\xi} - t_1)^+ \right] - \left[(1-\beta)t_2 + \beta(-\boldsymbol{x}_2^\top \boldsymbol{\xi}) + \frac{1-\beta}{1-\alpha}(-\boldsymbol{x}_2^\top \boldsymbol{\xi} - t_2)^+ \right] \right| \\
&\leq (1-\beta)\big|t_1 - t_2\big| + \beta\big|\boldsymbol{x}_1^\top \boldsymbol{\xi} - \boldsymbol{x}_2^\top \boldsymbol{\xi}\big| + \frac{1-\beta}{1-\alpha}\Big|(-\boldsymbol{x}_1^\top \boldsymbol{\xi} - t_1)^+ - (-\boldsymbol{x}_2^\top \boldsymbol{\xi} - t_2)^+\Big| \\
&\leq (1-\beta)\big|t_1 - t_2\big| + \beta\big|\boldsymbol{x}_1^\top \boldsymbol{\xi} - \boldsymbol{x}_2^\top \boldsymbol{\xi}\big| + \frac{1-\beta}{1-\alpha}\Big(\big|t_1 - t_2\big| + \big|\boldsymbol{x}_1^\top \boldsymbol{\xi} - \boldsymbol{x}_2^\top \boldsymbol{\xi}\big|\Big) && \text{(F.6)} \\
&\leq (1-\beta)\left(1 + \frac{1}{1-\alpha}\right)|t_1 - t_2| + \frac{1-\alpha\beta}{1-\alpha}\|\boldsymbol{\xi}\|_1 \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_\infty && \text{(F.7)} \\
&\leq \left[(1-\beta)\left(1 + \frac{1}{1-\alpha}\right) + \frac{1-\alpha\beta}{1-\alpha}\|\boldsymbol{\xi}\|_1\right]\Big(|t_1 - t_2| + \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_\infty\Big) \\
&\leq \left[(1-\beta)\left(1 + \frac{1}{1-\alpha}\right) + \frac{1-\alpha\beta}{1-\alpha}\|\boldsymbol{\xi}\|_1\right]\left\|\begin{pmatrix}\boldsymbol{x}_1 - \boldsymbol{x}_2 \\ t_1 - t_2\end{pmatrix}\right\|_1 =: \kappa(\boldsymbol{\xi})\left\|\begin{pmatrix}\boldsymbol{x}_1 - \boldsymbol{x}_2 \\ t_1 - t_2\end{pmatrix}\right\|_1. && \text{(F.8)}
\end{aligned}
$$

Here, (F.6) follows from the fact that $|(a_1)^+ - (a_2)^+| \leq |a_1 - a_2|$ for any $\{a_1, a_2\} \subset \mathbb{R}$; (F.7) follows from $\big|\boldsymbol{x}_1^\top \boldsymbol{\xi} - \boldsymbol{x}_2^\top \boldsymbol{\xi}\big| \leq \|\boldsymbol{\xi}\|_1 \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_\infty$; the inequality in (F.8) follows from $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_\infty \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_1$.

Appendix G. Comparison with Wang et al. (2023)

In this appendix, we provide a detailed comparison between our work and that of Wang et al. (2023).

- *The Proposed Model.* While our TRO model looks similar to Wang et al. (2023)'s model at the outset, our work actually generalizes the idea of Wang et al. (2023). Recall our TRO model in (4), which is equipped with the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ in (5). The TRO ambiguity set (5) is characterized by two parameters: the *shape* parameter $\mathcal{P}_N$ and the *size* parameter θ. The shape parameter $\mathcal{P}_N$ represents distributional ambiguity and could be any (data-driven) ambiguity set satisfying some mild assumptions mentioned in the paper. The size parameter $\theta \in [0, 1]$ controls the level of optimism, i.e., it controls the trade-off between solving the problem under a distributional belief and solving it under ambiguity. By replacing function f in (4) with the loss function h and choosing the shape parameter $\mathcal{P}_N$ as the distance-based ambiguity set $B_{\epsilon_N}(\widehat{\mathbb{P}}_N) = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \Delta(\mathbb{P}, \widehat{\mathbb{P}}_N) \leq \epsilon\}$ (where $\mathcal{P}(\Xi)$ is the set of probability measures on the support Ξ and Δ is a statistical distance), our TRO model reduces to the Bayesian distributionally robust (BDR) optimization model proposed in Wang et al. (2023) (see equation (11) in Wang et al., 2023), i.e., model (4) reduces to

$$
\underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}} \quad \beta_N \max_{\mathbb{P} \in B_{\epsilon_N}(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{P}}[h(\boldsymbol{x}, \boldsymbol{\xi})] + (1 - \beta_N)\mathbb{E}_{\widehat{\mathbb{P}}_N}[h(\boldsymbol{x}, \boldsymbol{\xi})]. \tag{BDR}
$$

Thus, Wang et al. (2023)'s BDR model is a special case of our TRO model. In particular, we emphasize that one can construct the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ using any shape parameter $\mathcal{P}_N$, including general moment- and distance-based ambiguity sets. Hence, our theoretical results are valid for various types of the shape parameter. In contrast, Wang et al. (2023) analyses are limited to the case where $\mathcal{P}_N$ is a distance-based ambiguity set.

- *Hierarchical Properties.* In Section 2, we analyze properties of the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ defined in (5) and the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$. These were not analyzed in Wang et al. (2023). We first formally introduce the notion of *hierarchical properties* of the sequence of the TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$; see Definition 2.2. The hierarchical properties indicate that the size of the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ increases with θ, i.e., $\mathcal{P}'_{N,\theta}$ contains more distributions with a larger θ. This implies that the TRO model is more conservative when we pick a larger θ. Then, in Theorem 1, we provide necessary and sufficient conditions for the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ to satisfy these properties. Specifically, Theorem 1 establishes that constructing the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ using a star-shaped shape parameter $\mathcal{P}_N$ with a star center $\widehat{\mathbb{P}}_N$ is necessary and sufficient for the sequence of TRO ambiguity sets $\{\mathcal{P}'_{N,\theta} \mid \theta \in [0,1]\}$ to satisfy the hierarchical property. Part (i) shows that for a general star-shaped shape parameter $\mathcal{P}_N$, the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ is non-decreasing in θ, i.e., $\mathcal{P}'_{N,\theta_1} \subseteq \mathcal{P}'_{N,\theta_2}$ whenever $\theta_1 \leq \theta_2$, indicating that the objective function of the trade-off model (4) is non-decreasing in θ. Part (ii) illustrates the relationship between the sets $\{\widehat{\mathbb{P}}_N\}$, $\mathcal{P}'_{N,\theta_1}$, $\mathcal{P}'_{N,\theta_2}$, and $\mathcal{P}_N$ with $0 < \theta_1 < \theta_2 < 1$. Specifically, part (ii) shows how the TRO ambiguity set $\mathcal{P}'_{N,\theta}$ enlarges with θ. This, in turn, implies that the TRO model is more conservative when we pick a larger θ. These important new results establish the connection between the specific choice of θ and the conservatism of the TRO model. However, they were not analyzed in Wang et al. (2023). In addition, in Proposition 2, we derive necessary and sufficient conditions for a general distance-based ambiguity set, i.e., the one adopted in the BDR model, to be star-shaped with a star center $\widehat{\mathbb{P}}_N$. Thus, our results provide necessary and sufficient conditions under which the sequence of ambiguity sets corresponding to the BDR model satisfies the hierarchical property. This was not studied in Wang et al. (2023). Moreover, as mentioned in the first point, our findings can be applied to TRO models with a general shape parameter. For instance, in Proposition 1, we establish the corresponding conditions for a general moment-based ambiguity set to be star-shaped, with a star center represented by $\widehat{\mathbb{P}}_N$, a novel contribution not explored in previous studies.

- *Analysis of the Conservatism.* In Section 3, we investigate the conservatism and properties of the optimal value $\widehat{\upsilon}_N(\theta)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta)$ of the TRO model through the lens of quantitative stability analysis. First, in Theorem 2, we establish mechanisms to quantify the difference in $\widehat{\upsilon}_N(\theta)$ and $\widehat{\mathcal{X}}_N(\theta)$ (and hence conservatism) incurred by perturbation in θ. In particular, it shows that $\widehat{\upsilon}_N(\theta)$ is Lipschitz continuous in θ and $\widehat{\mathcal{X}}_N(\theta)$ is Hölder continuous with Hölder exponent $1/2$ under distance D. This shows that both the optimal value and the set of optimal solutions change gradually with $\theta \in [0,1]$. Then, in Theorem 3, we show that naively combining SAA and DRO optimal solutions, e.g., via a convex combination after solving each separately, may not yield a feasible solution to the TRO problem (4). This is particularly true in applications where $\mathcal{X}$ is not convex (e.g., problems involving integer variables such as facility location and scheduling problems). Specifically, part (i) of Theorem 3 establishes that the optimal value $\widehat{\upsilon}_N(\theta)$ to our TRO model is not less than the convex combination $(1-\theta)\widehat{\upsilon}_N(0) + \theta\widehat{\upsilon}_N(1)$ of the SAA and DRO optimal values. In addition, if $\widehat{\mathbb{P}}_N \in \mathcal{P}_N$, Theorem 3 implies that $\widehat{\upsilon}_N$ is non-decreasing in θ as illustrated in Figure 2. Part (ii) of Theorem 3 indicates that the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta)$ to our TRO model can be approximated by $\overline{\mathcal{X}}_N(\theta) := (1-\theta)\widehat{\mathcal{X}}_N(0) + \theta\widehat{\mathcal{X}}_N(1)$ only when θ is close to zero or one; however, the difference could be huge for intermediate values of $\theta \in (0,1)$. These important investigations and results are new. In particular, Wang et al. (2023) only suspected that their BDR model is likely to be less conservative than the DRO model without providing any theoretical analysis. Indeed, we could apply our results to show that by changing β_N in Wang et al. (2023)'s BDR model, one can obtain a spectrum of optimal solutions, ranging from optimistic to conservative solutions.

- *Finite-Sample Properties.* As discussed in the first paragraph of Section 4.1, the optimal value of our TRO model $\widehat{\upsilon}_N(\theta)$ represents an estimator of the true optimal value of the stochastic optimization problem $\upsilon^\star = \inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}\big[f(\boldsymbol{x},\boldsymbol{\xi})\big]$. Analyzing the bias of an estimator to the true optimal value $\upsilon^\star$ is common in the related literature; see, e.g., Blanchet et al. (2019); Dentcheva and Lin (2022). Also, it is well known that the SAA estimator $\widehat{\upsilon}_N(0)$ is a downward biased estimator of $\upsilon^\star$; see (16). Thus, we and Wang et al. (2023) analyze the bias of the TRO and BDR estimators, respectively. Specifically, we and Wang et al. (2023) show that there exists $\theta^{\mathrm{u}}_N \in [0,1]$ in our TRO model and $\beta^{\mathrm{u}}_N \in [0,1]$ in the BDR model such that the optimal values of the TRO model and the BDR model are unbiased estimators of $\upsilon^\star$. However, in Section 4.1, we provide a more detailed investigation of the bias of the more general model, the TRO model, as well as new results. First, in Proposition 3, we derive an upper bound on the bias of the TRO estimator $\widehat{\upsilon}_N(\theta)$. It suggests that the bias of $\widehat{\upsilon}_N(\theta)$ may not be a downward bias as that of the SAA estimator. Second, in Corollary 1, we show that for sufficiently small θ, the TRO estimator $\widehat{\upsilon}_N(\theta)$ has a smaller bias than the SAA estimator, which was not discussed in Wang et al. (2023). Third, we show that $\mathbb{E}_{\mathbb{P}_N}[\widehat{\upsilon}_N(\theta)]$ can be decomposed as the sum of three terms: (a) the expected value of the SAA estimator $\mathbb{E}_{\mathbb{P}^N}[\widehat{\upsilon}_N(0)]$, (b) the DRO effect $\theta\big\{\mathbb{E}_{\mathbb{P}^N}\big[\widehat{\upsilon}_N(1)\big] - \mathbb{E}_{\mathbb{P}^N}\big[\widehat{\upsilon}_N(0)\big]\big\}$, and (c) the concavity effect $R_N(\theta)$ (see Figure 3). Finally, as pointed out by Wang et al. (2023), parameter β^{u}_N is typically hard to obtain. Similarly, θ^{u}_N is hard to obtain. However, in Theorems 5 and 6, we analyze the asymptotic behavior of θ^{u}_N. In particular, we prove the convergence of θ^{u}_N as $N \to \infty$ and derive its convergence rate. Wang et al. (2023) did not conduct such analyses. In Section 4.2, we derive the generalization bound for our TRO model in Theorem 7 based on that of the SAA and DRO models. In particular, we show that the probability (18) (i.e., the generalization error) is upper bounded by the sum of the probabilities $\alpha_{N,1}$ in (19) from the SAA model and $\alpha_{N,2}$ in (20) from the DRO model. This indeed corrects the generalization bound derived in Theorem 3.5 of Wang et al. (2023), where they did not take the sum of the two probability bounds from the SAA and DRO models; see Appendix C.6 of Wang et al. (2023). In addition, we show that for specific choices of the shape parameter $\mathcal{P}_N$, such as popular distance-based ambiguity sets as the one employed in Wang et al. (2023)'s BDR model, the generalization error exhibits an exponentially decaying tail. This important and attractive finite-sample property was not mentioned in Wang et al. (2023).

- *Asymptotic Convergence.* In Section 5, we show the almost sure convergence of the optimal value $\widehat{v}_N(\theta_N)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta_N)$ of the TRO model to their true counterparts when $N \to \infty$, and we derive the asymptotic distribution of $\widehat{v}_N(\theta_N)$ when $N \to \infty$. Our asymptotic convergence results hold for TRO models with TRO ambiguity sets constructed using general shape parameters $\mathcal{P}_N$, such as moment- and distance-based ambiguity sets. This differs from results in the existing literature focusing on a specific ambiguity set, including Wang et al. (2023). Note that the asymptotic convergence and distribution are two basic asymptotic properties that are commonly analyzed in the existing literature for data-driven optimization models (Blanchet and Shapiro, 2023; Kuhn et al., 2019; Shapiro et al., 2014). Hence, Wang et al. (2023) analyzed these asymptotic properties of their BDR model. However, the following are some differences between our analyses and those of Wang et al. (2023).
  - First, Wang et al. (2023) assumed the DRO objective $\sup_{\mathbb{P}\in B_{\epsilon_N}(\widehat{\mathbb{P}}_N)} \mathbb{E}[h(\boldsymbol{x},\boldsymbol{\xi})]$ is $\mathbb{P}^\star$-bounded and attainable for $\boldsymbol{x} \in \mathcal{X}$, where we recall that $\mathbb{P}^\star$ is the true distribution; see assumption C1 of Theorem 3.3 in Wang et al. (2023). To justify this assumption, Wang et al. (2023) provided one example based on the Wasserstein ambiguity set; see Appendix C.2 in Wang et al. (2023). Our convergence analyses also require the DRO objective to be upper-bounded for sufficiently large N. However, we adopt a more general set of assumptions under which the desired boundedness condition holds. Specifically, we impose assumptions on the objective function or the sequence of the ambiguity sets $\{\mathcal{P}_N\}_{N\in\mathbb{N}}$; see Assumption 3. Examples 6–10 provide a wide range of settings under which Assumption 3 holds. These examples include the case where $\mathcal{P}_N$ is constructed based on the Wasserstein ambiguity sets, as well as other distance-based and moment-based ambiguity sets. In Lemma 1, we formally prove that under Assumption 3, the DRO objective is *asymptotically* bounded, which is weaker than assumption C1 adopted by Wang et al. (2023) (which requires the DRO objective to be bounded for all $N \in \mathbb{N}$).
  - Second, in Theorem 8, we prove the almost-sure convergence of our TRO model. Specifically, we show that the optimal value $\widehat{v}_N(\theta_N)$ and the set of optimal solutions $\widehat{\mathcal{X}}_N(\theta_N)$ of our TRO

model converges almost surely to the true optimal value $v^\star$ and the set of optimal solutions $\mathcal{X}^\star$ to (1), respectively. In contrast, in Theorem 3.3 of Wang et al. (2023), they only provided the convergence of the optimal value and the set of optimal solutions of the BDR model in probability, which is weaker than our almost-sure convergence. We would like to highlight that the asymptotic convergence holds for TRO models with TRO ambiguity sets constructed using general shape parameters, such as moment-based ambiguity sets. This differs from the existing convergence results established for data-driven DRO models, which mainly employ distance-based ambiguity sets.

– Third, in Theorem 3.3 of Wang et al. (2023), they derived the asymptotic normality of the optimal value of the BDR model. To show this, they assumed that the optimal solution $\widehat{\boldsymbol{x}}_{b,N}$ to the BDR model converges (in probability) to an optimal solution $\boldsymbol{x}_0$ to the stochastic optimization problem under the true distribution $\mathbb{P}^\star$. This assumption is not common in relevant literature when deriving the asymptotic distribution of the optimal value of data-driven optimization models (Blanchet and Shapiro, 2023; Guigues et al., 2018; Shapiro et al., 2014). In contrast, in Theorem 9, we derive the asymptotic distribution of $\widehat{v}_N(\theta_N)$ without imposing such an assumption. Specifically, we apply the Delta's method (see, e.g., Cárcamo et al., 2020) to derive the following asymptotic distribution of the optimal value $\widehat{v}_N = \widehat{v}_N(\theta_N)$ of the TRO model: $\sqrt{N}(\widehat{v}_N - v^\star) \Rightarrow \inf_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{G}(\boldsymbol{x})$, where $\mathbb{G}$ is a tight Gaussian process indexed by $\mathcal{X}$ with mean zero and covariance function $\mathrm{Cov}(\mathbb{G}(\boldsymbol{x}_1)), \mathbb{G}(\boldsymbol{x}_2)) = \mathrm{Cov}_{\mathbb{P}^\star}(f(\boldsymbol{x}_1, \boldsymbol{\xi}), f(\boldsymbol{x}_2, \boldsymbol{\xi}))$. Thus, our results on the asymptotic distribution of the optimal value are different from those of Wang et al. (2023).

– Finally, for the special case when the shape parameter is chosen as some popular distance-based ambiguity set $\mathcal{P}_{N,r_N} = \{\mathbb{P} \in \mathcal{P}(\Xi) \mid \mathsf{d}(\mathbb{P}, \widehat{\mathbb{P}}_N) \le r_N\}$ as in Wang et al. (2023)'s BDR model, we can recover the asymptotics of the optimal value in classical distance-based DRO models. Specifically, in Theorem 10, we derive the asymptotic distribution of the optimal value $\widehat{v}_N(\theta_N)$ of our TRO model under different convergence rates of the size parameter θ_N and the radius r_N in the shape parameter $\mathcal{P}_{N,r_N}$. This generalizes the asymptotic convergence of the optimal solution derived in Theorem 3.3 of Wang et al. (2023). In particular, the asymptotic distribution results in Theorem 3.3 of Wang et al. (2023) essentially correspond to the case (i) in Theorem 10 only. Cases (ii) and (iii) in Theorem 10 were not investigated in Wang et al. (2023). Moreover, we also investigate the connection between the size parameter θ_N in our TRO model (with TRO ambiguity set constructed using the shape parameter $\mathcal{P}_{N,r}$) and the radius r_N in classical distance-based DRO models. This was not analyzed in Wang et al. (2023).

References

Banholzer, D., Fliege, J., Werner, R., 2022. On rates of convergence for sample average approximations in the almost sure sense and in mean. Mathematical Programming 191, 307–345.

Bayraksan, G., Love, D.K., 2015. Data-driven stochastic programming using phi-divergences, in: The Operations Research Revolution. INFORMS, pp. 1–19.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G., 2013. Robust solutions of optimization problems affected by uncertain probabilities. Management Science 59, 341–357.

Bertsimas, D., Gupta, V., Kallus, N., 2018. Robust sample average approximation. Mathematical Programming 171, 217–282.

Bertsimas, D., Kallus, N., 2020. From predictive to prescriptive analytics. Management Science 66, 1025–1044.

Birge, J.R., Louveaux, F., 2011. Introduction to Stochastic Programming. Springer Science & Business Media.

Blanchet, J., Murthy, K., Nguyen, V.A., 2021. Statistical analysis of Wasserstein distributionally robust estimators, in: Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications. INFORMS, pp. 227–254.

Blanchet, J., Shapiro, A., 2023. Statistical limit theorems in distributionally robust optimization, in: 2023 Winter Simulation Conference (WSC), IEEE. pp. 31–45.

Blanchet, J.H., Glynn, P.W., Pei, Y., 2019. Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. arXiv preprint arXiv:1904.09929 .

Cárcamo, J., Cuevas, A., Rodríguez, L.A., 2020. Directional differentiability for supremum-type functionals: Statistical applications. Bernoulli 26, 2143–2175.

Chen, L., He, S., Zhang, S., 2011. Tight bounds for some risk measures, with applications to robust portfolio selection. Operations Research 59, 847–865.

Chen, Z., Ma, W., Ji, B., 2023. Data-driven approximation of distributionally robust chance constraints using Bayesian credible intervals. arXiv preprint arXiv:2306.12735 .

Das, B., Dhara, A., Natarajan, K., 2021. On the heavy-tail behavior of the distributionally robust newsvendor. Operations Research 69, 1077–1099.

Delage, E., Ye, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research 58, 595–612.

Dentcheva, D., Lin, Y., 2022. Bias reduction in sample-based optimization. SIAM Journal on Optimization 32, 130–151.

Ding, K.w., Huang, N.j., Wang, L., 2023. Globalized distributionally robust optimization problems under the moment-based framework. Optimization , 1–39.

Duchi, J.C., Glynn, P.W., Namkoong, H., 2021. Statistics of robust optimization: A generalized empirical likelihood approach. Mathematics of Operations Research 46, 946–969.

Duchi, J.C., Namkoong, H., 2021. Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics 49, 1378–1406.

Eichhorn, A., Römisch, W., 2007. Stochastic integer programming: Limit theorems and confidence intervals. Mathematics of Operations Research 32, 118–135.

Fabozzi, F.J., Focardi, S.M., Kolm, P.N., Pachamanova, D.A., 2007. Robust Portfolio Optimization and Management. John Wiley & Sons.

Gallego, G., Moon, I., 1993. The distribution free newsboy problem: Review and extensions. Journal of the Operational Research Society 44, 825–834.

Gao, R., 2022. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. Operations Research .

Gao, R., Kleywegt, A., 2022. Distributionally robust stochastic optimization with Wasserstein distance. Mathematics of Operations Research .

Goh, J., Sim, M., 2010. Distributionally robust optimization and its tractable approximations. Operations Research 58, 902–917.

Gotoh, J.y., Kim, M.J., Lim, A.E., 2021. Calibration of distributionally robust empirical optimization models. Operations Research 69, 1630–1650.

Guigues, V., Krätschmer, V., Shapiro, A., 2018. A central limit theorem and hypotheses testing for risk-averse stochastic programs. SIAM Journal on Optimization 28, 1337–1366.

Hanasusanto, G.A., Roitch, V., Kuhn, D., Wiesemann, W., 2015. A distributionally robust perspective on uncertainty quantification and chance constrained programming. Mathematical Programming 151, 35–62.

Huang, R., Qu, S., Yang, X., Xu, F., Xu, Z., Zhou, W., 2021. Sparse portfolio selection with uncertain probability distribution. Applied Intelligence 51, 6665–6684.

Iyengar, G., Lam, H., Wang, T., 2022. Hedging complexity in generalization via a parametric distributionally robust optimization framework. arXiv preprint arXiv:2212.01518 .

Kakouris, I., Rustem, B., 2014. Robust portfolio optimization with copulas. European Journal of Operational Research 235, 28–37.

Kim, S., Pasupathy, R., Henderson, S.G., 2015. A guide to sample average approximation, in: Handbook of Simulation Optimization. Springer, pp. 207–243.

de Klerk, E., Kuhn, D., Postek, K., 2020. Distributionally robust optimization with polynomial densities: Theory, models and algorithms. Mathematical Programming 181, 265–296.

Kuhn, D., Mohajerin Esfahani, P., Nguyen, V.A., Shafieezadeh-Abadeh, S., 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning, in: Operations Research & Management Science in the Age of Analytics. Informs, pp. 130–166.

Lam, H., 2019. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. Operations Research 67, 1090–1105.

Lam, H., 2021. On the impossibility of statistically improving empirical optimization: A second-order stochastic dominance perspective. arXiv preprint arXiv:2105.13419 .

Lam, H., 2022. A cheap bootstrap method for fast inference. arXiv preprint arXiv:2202.00090 .

Lam, H., Liu, Z., Zhang, X., 2021. Orthounimodal distributionally robust optimization: Representation, computation and multivariate extreme event applications. arXiv preprint arXiv:2111.07894 .

Li, B., Jiang, R., Mathieu, J.L., 2019. Ambiguous risk constraints with moment and unimodality information. Mathematical Programming 173, 151–192.

Li, Y., Xing, W., 2023. Globalized distributionally robust optimization based on samples. Journal of Global Optimization , 1–30.

Liu, F., Chen, Z., Wang, S., 2023a. Globalized distributionally robust counterpart. INFORMS Journal on Computing .

Liu, J., Wu, J., Li, B., Cui, P., 2022. Distributionally robust optimization with data geometry, in: 36th Conference on Neural Information Processing Systems (NeurIPS).

Liu, Y., Pichler, A., Xu, H., 2019. Discrete approximation and quantification in distributionally robust optimization. Mathematics of Operations Research 44, 19–37.

Liu, Y., Xu, H., 2013. Stability analysis of stochastic programs with second order dominance constraints. Mathematical Programming 142, 435–460.

Liu, Z., Van Parys, B.P., Lam, H., 2023b. Smoothed f-divergence distributionally robust optimization: Exponential rate efficiency and complexity-free calibration. arXiv preprint arXiv:2306.14041 .

Long, D.Z., Qi, J., Zhang, A., 2023. Supermodularity in two-stage distributionally robust optimization. Management Science .

Mak, W.K., Morton, D.P., Wood, R.K., 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. Operations Research Letters 24, 47–56.

Homem-de Mello, T., Bayraksan, G., 2014. Monte Carlo sampling-based methods for stochastic optimization. Surveys in Operations Research and Management Science 19, 56–85.

Michel, P., Hashimoto, T., Neubig, G., 2021. Modeling the second player in distributionally robust optimization, in: Proceedings of the International Conference on Learning Representations (ICLR).

Michel, P., Hashimoto, T., Neubig, G., 2022. Distributionally robust models with parametric likelihood ratios, in: Proceedings of the International Conference on Learning Representations (ICLR).

Mohajerin Esfahani, P., Kuhn, D., 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming 171, 115–166.

Niculescu, C.P., Persson, L.E., 2018. Convex Functions and Their Applications: A Contemporary Approach. Second ed., Springer.

Nie, J., Yang, L., Zhong, S., Zhou, G., 2023. Distributionally robust optimization with moment ambiguity sets. Journal of Scientific Computing 94, 12.

Norkin, V.I., Pflug, G.C., Ruszczyński, A., 1998. A branch and bound method for stochastic global optimization. Mathematical Programming 83, 425–450.

Pflug, G.C., Pichler, A., 2014. Multistage Stochastic Optimization. volume 1104. Springer.

Pichler, A., Xu, H., 2018. Quantitative stability analysis for minimax distributionally robust risk optimization. Mathematical Programming , 1–31.

Pichler, A., Xu, H., 2022. Quantitative stability analysis for minimax distributionally robust risk optimization. Mathematical Programming 191, 47–77.

Postek, K., Ben-Tal, A., Den Hertog, D., Melenberg, B., 2018. Robust optimization with ambiguous stochastic constraints under mean and dispersion information. Operations Research 66, 814–833.

Postek, K., Romeijnders, W., Den Hertog, D., van der Vlerk, M.H., 2019. An approximation framework for two-stage ambiguous stochastic integer programs under mean-mad information. European Journal of Operational Research 274, 432–444.

Rahimian, H., Mehrotra, S., 2022. Frameworks and results in distributionally robust optimization. Open Journal of Mathematical Optimization 3, 1–85.

Rockafellar, R.T., Uryasev, S., 2000. Optimization of conditional value-at-risk. Journal of Risk 2, 21–42.

Römisch, W., 2003. Stability of stochastic programming problems. Handbooks in Operations Research and Management Science 10, 483–554.

Rudin, W., et al., 1976. Principles of Mathematical Analysis. volume 3. McGraw-hill New York.

Shapiro, A., 1994. Quantitative stability in stochastic programming. Mathematical Programming 67, 99–108.

Shapiro, A., Dentcheva, D., Ruszczyński, A., 2014. Lectures on Stochastic Programming: Modeling and Theory. Second ed., SIAM.

Shehadeh, K.S., Tucker, E.L., 2022. Stochastic optimization models for location and inventory prepositioning of disaster relief supplies. Transportation Research Part C: Emerging Technologies 144, 103871.

Smith, J.E., Winkler, R.L., 2006. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. Management Science 52, 311–322.

So, A.M.C., 2011. Moment inequalities for sums of random matrices and their applications in optimization. Mathematical Programming 130, 125–151.

Sun, H., Xu, H., 2016. Convergence analysis for distributionally robust optimization and equilibrium problems. Mathematics of Operations Research 41, 377–401.

Sun, H., Zhang, D., Wu, S.Y., Chen, L., 2022. A modified exchange algorithm for distributional robust optimization and applications in risk management. International Transactions in Operational Research 29, 130–157.

van der Vaart, A.W., 2000. Asymptotic Statistics. volume 3. Cambridge university press.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications in Statistics. Springer.

Van Parys, B.P., Goulart, P.J., Kuhn, D., 2016. Generalized gauss inequalities via semidefinite programming. Mathematical Programming 156, 271–302.

Van Parys, B.P., Mohajerin Esfahani, P., Kuhn, D., 2021. From data to decisions: Distributionally robust optimization is optimal. Management Science 67, 3387–3402.

Villani, C., 2009. Optimal Transport: Old and New. volume 338. Springer.

Wang, J., Gao, R., Xie, Y., 2021. Sinkhorn distributionally robust optimization. arXiv preprint arXiv:2109.11926 .

Wang, S., Wang, H., Honorio, J., 2023. Learning against distributional uncertainty: On the trade-off between robustness and specificity. arXiv preprint arXiv:2301.13565 .

Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. Operations Research 62, 1358–1376.

Xie, W., 2021. On distributionally robust chance constrained programs with Wasserstein distance. Mathematical Programming 186, 115–155.

Xu, H., Liu, Y., Sun, H., 2018. Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods. Mathematical Programming 169, 489–529.

Yam, S.C.P., Yang, H., Yuen, F.L., 2016. Optimal asset allocation: Risk and information uncertainty. European Journal of Operational Research 251, 554–561.

Zhang, J., Xu, H., Zhang, L., 2016. Quantitative stability analysis for distributionally robust optimization with moment constraints. SIAM Journal on Optimization 26, 1855–1882.

Zhao, C., Guan, Y., 2015. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Available on Optimization Online 2, 1–40.