JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
2021, VOL. 72, NO. 8, 1879–1897
https://doi.org/10.1080/01605682.2020.1746203

ORIGINAL ARTICLE

A data-driven distributionally robust newsvendor model with a Wasserstein ambiguity set

Sangyoon Lee, Hyunwoo Kim* and Ilkyeong Moon

Seoul National University, Seoul, South Korea

ABSTRACT

In this paper, we derive a closed-form solution and an explicit characterization of the worst-case distribution for the data-driven distributionally robust newsvendor model with an ambiguity set based on the Wasserstein distance of order $p \in [1, \infty)$. We also consider the risk-averse decision with the Conditional Value-at-Risk (CVaR) objective. For the risk-averse model, we derive a closed-form solution for the $p = 1$ case, and propose a tractable formulation to obtain an optimal order quantity for the $p > 1$ case. We conduct numerical experiments to compare out-of-sample performance and convergence results of the proposed solutions against the solutions with other distributionally robust models. We also analyze the risk-averse solutions compared to the risk-neutral solutions.

ARTICLE HISTORY
Received 25 February 2019
Accepted 18 March 2020

KEYWORDS
Newsvendor Model; distributionally robust optimization; Wasserstein distance

1. Introduction

The newsvendor problem is a well-known problem to decide an order quantity considering the trade-off between the risks of *overage* and *underage* under demand uncertainty. It is a building block of many operations management problems, such as inventory control, pricing, supply chain contracts, and retail management. In the classical setting of the newsvendor problem, complete knowledge of demand distribution is assumed. In reality, however, the demand distribution is often impossible to be known precisely. To tackle this difficulty, a decision maker considers an *ambiguity set* which is a set of candidate distributions that may contain the unknown true distribution, and minimizes the worst-case expected cost over the ambiguity set. This approach, called *distributionally robust optimization* (DRO), is widely applied to many operations management problems. Various types of ambiguity sets are proposed with several different prior information about demand distribution. In some practical cases, however, historical data is the only information that can be obtained. Therefore, how to construct an ambiguity set with historical data and optimize over the constructed ambiguity set is important to successful operations. In this study, we consider the data-driven distributionally robust newsvendor model with a Wasserstein ambiguity set.

The distributionally robust newsvendor model dates back to Scarf (1958), who considered the distributionally robust order quantity with an ambiguity set that contained all distributions with known first and second moments. Gallego and Moon (1993) extended Scarf's basic results to various ways with the same ambiguity set. With the development of DRO with moment-based ambiguity sets (Bertsimas et al., 2010; Delage & Ye, 2010; Goh & Sim, 2010; Wiesemann et al., 2014), several extensions were proposed to consider various objective functions and to construct new ambiguity sets considering the shape of distribution, e.g. symmetry/asymmetry and unimodality/multimodality (Hanasusanto et al., 2015; Natarajan et al., 2018; Perakis & Roels, 2008; Saghafian & Tomlin, 2016; Yue et al., 2006; Zhu et al., 2013). In most cases, the distributionally robust newsvendor models with moment constraints are tractable, and in some cases closed-form solutions and explicit characterizations of the worst-case distributions are available. However, the assumption that a decision maker has certain information about moments proves to be unrealistic for many operations management problems. For example, historical data of a newly introduced product is not enough to estimate moments. Moreover, decisions based on an inaccurate estimation of moments can lead to highly suboptimal solutions. Even if the decision maker has the exact moment information, the moment-based ambiguity set is constructed with only moment information and other prior information such as the shape of distribution is abandoned. Another shortcoming of the moment-based ambiguity set is that resulting decisions are sometimes overly conservative due to the unrealistic worst-case distribution (Wang et al., 2016).

CONTACT Ilkyeong Moon ikmoon@snu.ac.kr Department of Industrial Engineering, Seoul National University, Seoul 08826, South Korea
*Current affiliation: LG CNS Co., Ltd. Entrue Consulting Optimization Group, Seoul, South Korea. hwkim0211@lgcns.com.

An important alternative is DRO with an ambiguity set which contains probability distributions close to the reference distribution in terms of a statistical distance. A statistical distance measures the distance between two probability distributions and several studies used various statistical distances for constructing ambiguity sets, such as ϕ-divergences (Bayraksan & Love, 2015; Ben-Tal et al., 2013; Jiang & Guan, 2016; Sun & Xu, 2016) and the Wasserstein distance (Esfahani & Kuhn, 2018; Pichler & Xu, 2017; Wozabal, 2012; Zhao & Guan, 2018). However, as Gao and Kleywegt (2016) pointed out, in some cases, an ambiguity set with ϕ-divergences fails to include distributions that a decision maker wishes to include. For instance, consider historical data generated from a normal distribution and a ϕ-divergence ambiguity set based on the discrete empirical distribution. The ϕ-divergence ambiguity set does not contain the data-generating normal distribution, because probability distributions in the ambiguity set are absolutely continuous with respect to the empirical distribution, i.e. the ambiguity set includes only discrete distributions with the same support of the empirical distribution. Gao and Kleywegt (2016) also pointed out that the ϕ-divergence does not consider the closeness between two points in the support, thus leading to the inclusion of overly conservative or pessimistic distributions.

DRO with an ambiguity set based on the Wasserstein distance not only alleviates the problems mentioned above, but also has several useful properties. The Wasserstein distance captures closeness between two points, which leads to the realistic measurement of distance between two distributions. Furthermore, the Wasserstein ambiguity set contains both discrete and continuous distributions, because the Wasserstein distance between discrete and continuous distributions can be defined (cf. ϕ-divergence). In addition, the Wasserstein ambiguity set with the empirical distribution contains the data-generating distribution with probabilistic guarantees (Fournier & Guillin, 2015). Desirable properties, e.g. finite sample guarantee, asymptotic consistency, and tractability (Bertsimas et al., 2018; Esfahani & Kuhn, 2018; Fournier & Guillin, 2015), are proved for the Wasserstein order $p = 1$.

Although some researches have considered the newsvendor models with a Wasserstein ambiguity set, they are used as examples to emphasize theoretical results of DRO, and the models are limited to discrete and bounded support (Gao & Kleywegt, 2016) or the Wasserstein order $p = 1$ (Esfahani & Kuhn, 2018). To the best of our knowledge, a closed-form solution and an explicit characterization of the worst-case distribution are not studied in the general setting such as the continuous and unbounded support, and the higher Wasserstein order ($p>1$). In this study, we study the distributionally robust newsvendor model with a Wasserstein ambiguity set on support with nonnegative real numbers $\Xi = [0, \infty)$ and the Wasserstein order $p \in [1, \infty)$. We consider the reference distribution for an ambiguity set as the empirical distribution.

The main contributions of this paper are as follows:

- The closed-form expressions of an optimal order quantity and the worst-case distribution for the risk-neutral newsvendor problem are derived with the general support and the higher Wasserstein order.
- For the risk-averse decision, we also consider the Conditional Value-at-Risk (CVaR) objective for the newsvendor model. We derive a closed-form solution for the $p = 1$ case and propose a tractable formulation to obtain an optimal order quantity for the $p > 1$ case.

The rest of this paper is organized as follows. In Section 2, we introduce the definition of the Wasserstein distance and present a strong duality result for the data-driven DRO with a Wasserstein ambiguity set. In Section 3, we derive the closed-form solution of an optimal order quantity and the worst-case distribution for the distributionally robust newsvendor model. We also consider the risk-averse model and analyze the model in Section 4. Section 5 provides numerical experiments based on the theoretical results, and we conclude the paper in Section 6.

2. Distributionally robust optimization with the Wasserstein distance

In this section, we introduce the definition of the Wasserstein distance and discuss properties of the Wasserstein ambiguity set in the optimization perspective. We adopt the strong duality result of data-driven DRO with the Wasserstein distance and related definitions from the result of Gao and Kleywegt (2016).

Let (Ξ, d) be a separable complete metric space (Polish space) and $\mathcal{B}(\Xi)$ be the Borel σ-algebra. Let $\mathcal{P}(\Xi)$ denote a set of Borel measures defined on $(\Xi, \mathcal{B}(\Xi))$. Let $\mathcal{P}_p(\Xi)$ for $p \in [1, \infty)$ denote a set of probability measures with a finite moment of order p for any $x_0 \in \Xi$, i.e. $\mathcal{P}_p(\Xi) := \{\mu \in \mathcal{P}(\Xi) : \int_\Xi d(x_0, x)^p \mu(dx) < \infty\}$.

Definition 2.1. (Wasserstein distance). The Wasserstein distance of order p between two probability measures $\mu, \nu \in \mathcal{P}_p(\Xi)$ is defined as

$$W_p(\mu,\nu) := \left(\min_{\gamma\in\Gamma(\mu,\nu)} \left\{ \int_{\Xi\times\Xi} d(\xi,\zeta)^p d\gamma(\xi,\zeta) \right\} \right)^{\frac{1}{p}},$$

where $\Gamma(\mu,\nu)$ denotes a set of all probability measures on $\mathcal{P}(\Xi\times\Xi)$ with marginals μ and ν.

We use metric d for the definition of the Wasserstein distance as $d(\xi,\zeta) := |\xi-\zeta|$ throughout this study. The Wasserstein distance is motivated by the optimal transport theory whose foundation is rooted back to the Monge's problem. We refer the reader to Villani (2009) for further details. The meaning of the Wasserstein distance is the optimal transport cost of moving mass from μ to ν. Therefore, the Wasserstein distance represents a distance between two different distributions considering the distance between two points in Ξ with respect to the metric d. The Wasserstein distance has been attracting attention in recent studies because of several good properties. First, convergence with respect to the Wasserstein distance implies weak convergence. Second, DRO with a Wasserstein ambiguity set can incorporate the data-driven setting and overcome the absolutely continuous support issue of the ϕ-divergence. The Wasserstein distance has been actively applied in areas other than optimization, e.g. machine learning literature such as a GAN (Generative Adversarial Network) (Arjovsky et al., 2017), an auto-encoder (Tolstikhin et al., 2017), and regularization (Shafieezadeh-Abadeh et al., 2019).

First, we discuss the properties of the Wasserstein distance. The following properties are summarized from various literature (Pflug & Pichler, 2014; Pichler & Xu, 2017; Villani, 2009).

Proposition 2.2. (Properties of the Wasserstein distance). *Let (Ξ,d) be a Polish space. The following hold.*

- *W_p is finite on $\mathcal{P}_p(\Xi)$, so W_p is a distance function on $\mathcal{P}_p(\Xi)$. $\mathcal{P}_p(\Xi)$ equipped with a distance function W_p is a metric space.*
- *W_p metrizes the weak convergence in $\mathcal{P}_p(\Xi)$, i.e., μ_k converges to μ weakly if and only if $W_p(\mu_k,\mu)\to 0$.*
- *The Wasserstein distance is monotone, i.e., if $p_1\le p_2$, then $W_{p_1}(\mu,\nu)\le W_{p_2}(\mu,\nu)$.*
- *W_p is p-convex, i.e., for any $\mu_1,\mu_2\in\mathcal{P}_p(\Xi)$ and $\lambda\in[0,1]$,*

$$W_p(\nu,(1-\lambda)\mu_1+\lambda\mu_2)^p \le (1-\lambda)W_p(\nu,\mu_1)^p + \lambda W_p(\nu,\mu_2)^p.$$

- *A metric space $(\mathcal{P}_p(\Xi),W_p)$ with the Wasserstein distance is a Polish space.*

We define an ambiguity set based on the Wasserstein distance and discuss the properties of the Wasserstein ambiguity set based on Proposition 2.1.

Definition 2.3. (Wasserstein ambiguity set). Let ν be the reference distribution. The Wasserstein ambiguity set $\mathcal{M}$ is defined as

$$\mathcal{M} := \{\mu\in\mathcal{P}_p(\Xi) : W_p(\mu,\nu)\le\theta\}.$$

The Wasserstein radius θ determines the size of the Wasserstein ball. To ensure the certain probability of a Wasserstein ambiguity set containing unknown true distribution, the size of the radius is chosen appropriately (Esfahani & Kuhn, 2018; Fournier & Guillin, 2015).

Since $(\mathcal{P}_p(\Xi),W_p)$ is a metric space, we can define a closed ball $\mathcal{B}_\theta(\nu)$ centered at $\nu\in\mathcal{P}_p(\Xi)$ with radius θ as follows:

$$\mathcal{B}_\theta(\nu) := \{\mu\in\mathcal{P}_p(\Xi) : W_p(\mu,\nu)\le\theta\}$$

Then, $\mathcal{B}_\theta(\nu)$ coincides with Definition 2.2, that is, a Wasserstein ambiguity set is a closed ball in a metric space $(\mathcal{P}_p(\Xi),W_p)$. The following proposition reveals the useful properties of the Wasserstein ambiguity set.

Proposition 2.4. *Let ν be the empirical distribution. Then, the Wasserstein ambiguity set $\mathcal{M}$ is closed, convex, and weakly compact.*

Proof. $\mathcal{M}$ is a closed ball in $(\mathcal{P}_p(\Xi),W_p)$, so $\mathcal{M}$ is closed. To show $\mathcal{M}$ is convex, we use p-convexity of W_p from Proposition *2.1*. For any $\mu_1,\mu_2\in\mathcal{P}_p(\Xi)$ and $\lambda\in[0,1]$, $(1-\lambda)\mu_1+\lambda\mu_2\in\mathcal{P}_p(\Xi)$ and

$$\begin{aligned} W_p(\nu,(1-\lambda)\mu_1+\lambda\mu_2)^p &\le (1-\lambda)W_p(\nu,\mu_1)^p \\ &+ \lambda W_p(\nu,\mu_2)^p \le (1-\lambda)\theta^p+\lambda\theta^p = \theta^p. \end{aligned}$$

The first inequality holds by p-convexity of W_p and the second inequality holds by the definition of $\mathcal{M}$. Therefore, $W_p(\nu,(1-\lambda)\mu_1+\lambda\mu_2)\le\theta$ and $(1-\lambda)\mu_1+\lambda\mu_2\in\mathcal{M}$, which proves that $\mathcal{M}$ is convex. For the compactness, $\mathcal{M}$ is weakly compact by Banach-Alaoglu theorem (Section 3.15 in Rudin (1991)) and tightness of empirical distribution ν. □

The properties of the Wasserstein ambiguity set will be used in various analyses. Using the Wasserstein ambiguity set, DRO with cost function $\Psi : X\times\Xi\to\mathbb{R}$ is expressed as follows:

$$\inf_{x\in X}\sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu[\Psi(x,\xi)]. \tag{1}$$

From the study of Gao and Kleywegt (2016), we adopt the strong duality result for data-driven DRO with the Wasserstein distance of order p when the reference distribution is the empirical distribution.

Theorem 2.5. (Strong duality for data-driven DRO, Gao and Kleywegt (2016)). *Let ν be the empirical distribution with historical data $\{\hat{\xi}^1,\ldots,\hat{\xi}^N\}$, i.e.*

$\nu = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{\xi}^i}$ *where* δ *denotes the Dirac measure. Then, the strong dual of (1) is*

$$\inf_{x\in X,\,\lambda\geq 0}\left\{\lambda\theta^p + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}\left[\Psi(x,\xi) - \lambda d(\xi,\hat{\xi}^i)^p\right]\right\}.$$

To avoid the trivial case of the dual formulation in Theorem 2.4, the inner supremum should be finite. To guarantee finiteness of the inner supremum, we need the following definition.

Definition 2.6. (Growth rate, Gao and Kleywegt (2016)). The growth rate κ of Ψ is defined as

$$\kappa := \inf\left\{\lambda \geq 0 : \int_{\Xi}\sup_{\xi\in\Xi}\left[\Psi(x,\xi) - \lambda d(\xi,\zeta)^p\right]\nu(d\zeta) < \infty\right\}.$$

The growth rate κ is the minimum value of the dual variable λ, which makes the inner supremum finite (see also the definition of steepness of the objective function in Theorem 6.3 and Proposition 6.5 of Esfahani and Kuhn (2018)). The dual variable λ should be greater than or equal to the growth rate of the function Ψ, otherwise, the dual formulation becomes infeasible. We utilize the above results to analyze the distributionally robust newsvendor model with a Wasserstein ambiguity set.

3. Risk-neutral newsvendor model

In this section, we consider the data-driven distributionally robust newsvendor model with a risk-neutral decision maker. We consider the general Wasserstein order $p \in [1,\infty)$ and derive the closed-form solution of the optimal order quantity. Then, we characterize the worst-case distribution with perturbations from historical data.

In the newsvendor model, the decision maker sells a single product for a single period. The decision maker decides the order quantity before the random demand $\xi \in \Xi$ is observed. After the demand is realized, the *overage cost* h per unit of unsold goods or the *underage cost* b per unit of shortage is imposed. The objective of the decision maker is to minimize the expected total cost. The newsvendor model can be expressed as follows:

$$\min_{x\geq 0}\mathbb{E}_{\mu}\left[h(x-\xi)^+ + b(\xi-x)^+\right]$$

where $X^+ := \max\{X, 0\}$. In the classical newsvendor problem with the known demand distribution μ, the optimal order quantity is well known as the critical ratio, i.e. the $\frac{b}{h+b}$ quantile of the demand distribution.

However, in practice, the demand distribution is restricted to be known precisely. Although complete knowledge of demand distribution is restricted, historical data can be obtained. Therefore, we propose the data-driven distributionally robust newsvendor model. We consider the support of the demand distribution is $\Xi = [0,\infty)$, and we assume that without loss of generality N historical data $\{\hat{\xi}^1,\ldots,\hat{\xi}^N\}$ is sorted in nondecreasing order, that is, $\hat{\xi}^1 \leq \ldots \leq \hat{\xi}^N$. Using historical data, we define the empirical distribution $\nu = \frac{1}{N}\sum_{i=1}^{N}\delta_{\hat{\xi}^i}$, which is used for constructing a Wasserstein ambiguity set. Then, the data-driven distributionally robust risk-neutral newsvendor model with a Wasserstein ambiguity set is expressed as follows:

$$\min_{x\geq 0}\sup_{\mu\in\mathcal{P}_p(\Xi)}\{\mathbb{E}_{\mu}\left[h(x-\xi)^+ + b(\xi-x)^+\right] : W_p(\mu,\nu)\leq\theta\}. \tag{2}$$

Using Theorem 2.5 based on the empirical distribution ν and $d(\xi,\zeta) = |\xi-\zeta|$, the dual reformulation of (2) can be expressed as follows:

$$\min_{x\geq 0}\inf_{\lambda\geq 0}\left\{\lambda\theta^p + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}\left[h(x-\xi)^+ + b(\xi-x)^+ - \lambda|\xi-\hat{\xi}^i|^p\right]\right\}. \tag{3}$$

To obtain meaningful and simple analysis, especially for the analysis of the inner supremum and derivation of closed-form solutions, we impose a weak restriction on *overage* and *underage* costs.

Assumption 3.1. *The underage cost is greater than or equal to the overage cost, i.e.* $b \geq h$.

The assumption is needed for further analysis, e.g. analysis of the inner supremum in (3), the partition of sample points based on x and λ, and feasibility issues in Section 3.2. In addition, the assumption has real-world meaning: the decision maker considers the underage situation to be more important than the overage situation. In practice, the shortage is more important in many cases because it results not only in a penalty cost but also in loss of goodwill or trust, which may be costly for the decision maker. Therefore, the underage cost is greater than the overage cost for many real-world situations which means that the decision maker tends to order more than the mean of random demand in the newsvendor problem.

We consider two cases: $p=1$ and $p>1$. The reason we divide the cases is that the analysis is easier for the Wasserstein order $p=1$ case. Due to the Kantorovich-Rubinstein duality, calculation of the Wasserstein distance for $p=1$ is much more tractable. The inner supremum of the dual formulation is first order, which makes the derivation of the explicit form of the inner supremum simple. Even

with the difficulty of analysis in the $p>1$ case, the differentiation can be used for the explicit expression of the inner supremum.

3.1. Wasserstein order $p=1$

First, we consider the Wasserstein order $p=1$ for the dual formulation (3). We note that the growth rate of the newsvendor cost function (Definition 2.5) is b. Therefore, dual variable λ should be greater than or equal to b, i.e. the constraint $\lambda \geq b$ should be added.

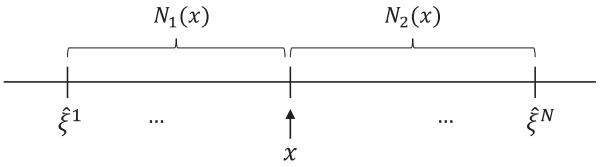

Figure 1. Definitions of $N_1(x)$ and $N_2(x)$.

To obtain the closed-form solution, we need to characterize equivalent expressions of the inner supremum. For given (x,λ), let $f_i(\xi) := h(x-\xi)^+ + b(\xi-x)^+ - \lambda|\xi-\hat{\xi}^i|$ for $i=1,\ldots,N$. To analyze the supremum of f_i, we define $N_1(x) := \{1 \leq i \leq N : \hat{\xi}^i < x\}$ and $N_2(x) := \{1 \leq i \leq N : \hat{\xi}^i \geq x\}$ such that N data points are divided into two sets based on x (Figure 1).

Under Assumption 3.1, the supremum of f_i can be derived by simple analysis using strong duality results. One specific instance of f_i is presented in Figure 2. There are four cases according to whether $x<\hat{\xi}^i$ or $x \geq \hat{\xi}^i$, and whether $\lambda = b$ or $\lambda > b$. We can see that the supremum of f_i is attained at $\xi = \hat{\xi}^i$ for all four cases. This result can be extended to all instances of f_i, that is, the supremum of f_i is attained at $\xi = \hat{\xi}^i$, which leads to $\sup_{\xi\in\Xi} f_i(\xi) = h(x-\hat{\xi}^i)$ for $i \in N_1(x)$ and $\sup_{\xi\in\Xi} f_i(\xi) = b(\hat{\xi}^i - x)$ for $i \in N_2(x)$.

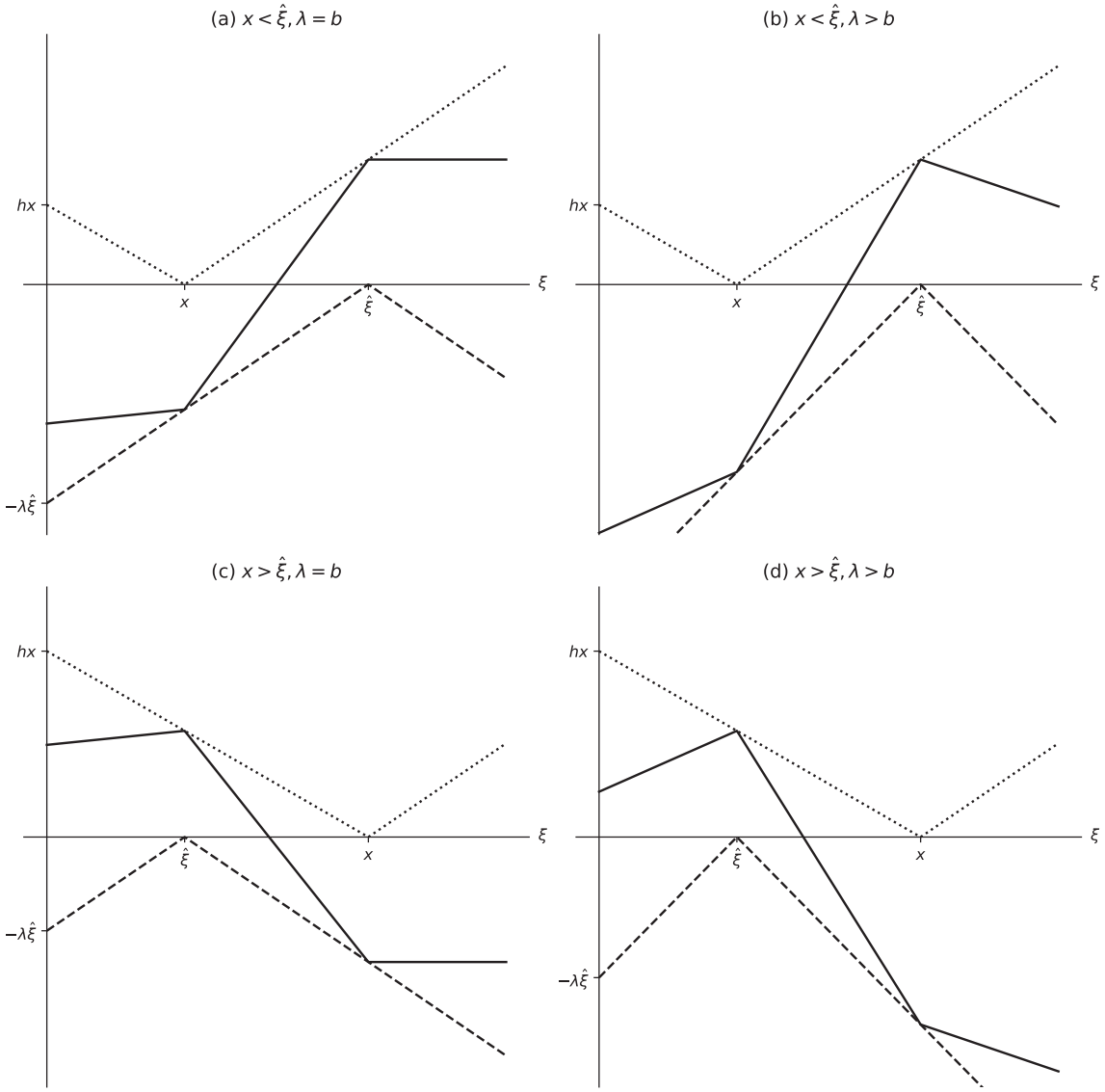

Figure 2. Four cases for $f_i(\xi)$. Dotted lines represent $h(x-\xi)^+ + b(\xi-x)^+$ and $-\lambda|\xi-\hat{\xi}^i|$, respectively, and the solid line represents $f_i(\xi)$.

This result shows that the inner supremum of dual formulation is independent of dual variable λ, which disconnects the linkage between x and λ. Therefore, we obtain the following equivalent expression of the objective function of dual formulation (3).

$$\begin{aligned}&\min_{x\geq 0,\,\lambda\geq b}\lambda\theta+\frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}\left[h(x-\xi)^+ + b(\xi-x)^+ - \lambda|\xi-\hat{\xi}^i|\right]\\ &=\min_{x\geq 0,\,\lambda\geq b}\lambda\theta+\frac{1}{N}\left(\sum_{i\in N_1(x)}h(x-\hat{\xi}^i)+\sum_{i\in N_2(x)}b(\hat{\xi}^i-x)\right)\\ &=b\theta+\min_{x\geq 0}\frac{1}{N}\left(\sum_{i\in N_1(x)}h(x-\hat{\xi}^i)+\sum_{i\in N_2(x)}b(\hat{\xi}^i-x)\right)\end{aligned}$$

Since x and λ are separated, the optimal dual variable $\lambda^*=b$ is obtained. Now, we derive the closed-form solution based on the critical ratio $\frac{b}{h+b}$ in Theorem 3.2.

Theorem 3.2. *Suppose the underage cost is greater than or equal to the overage cost, i.e. $b\geq h$. If $i^*\in\{1,\ldots,N\}$ satisfies $\frac{i^*-1}{N}<\frac{b}{h+b}\leq\frac{i^*}{N}$, then $\hat{\xi}^{i^*}$ is an optimal order quantity and the optimal cost is*

$$b\theta+\frac{1}{N}\left(\sum_{k=1}^{i^*-1}h(\hat{\xi}^{i^*}-\hat{\xi}^k)+\sum_{k=i^*}^{N}b(\hat{\xi}^k-\hat{\xi}^{i^*})\right).$$

Proof. Define region i as $(\hat{\xi}^i,\hat{\xi}^{i+1}]$ for $i=0,\ldots,N$ where $\hat{\xi}^0:=0$ and $\hat{\xi}^{N+1}:=\infty$. For $x\in(\hat{\xi}^i,\hat{\xi}^{i+1}]$, $|N_1(x)|=i$ and $|N_2(x)|=N-i$. Let $g(x):=\frac{1}{N}(\sum_{i\in N_1(x)}h(x-\hat{\xi}^i)+\sum_{i\in N_2(x)}b(\hat{\xi}^i-x))$. Suppose $h|N_1(x)|<b|N_2(x)|$, then $g(x)$ is nonincreasing as $x\to\hat{\xi}^{i+1}$. Suppose $h|N_1(x)|\geq b|N_2(x)|$, then $g(x)$ is nonincreasing as $x\to\hat{\xi}^i$.

There exists i^* such that $h(i^*-1)<b(N-i^*+1)$ and $h(i^*)\geq b(N-i^*)$. Then it is optimal to order $\hat{\xi}^{i^*}$. With this order quantity, the optimal cost is $b\theta+g(\hat{\xi}^{i^*})$. □

Remark 1. (Sample average approximation). The optimal order quantity derived in Theorem 3.2 is the $\frac{b}{h+b}$ quantile of the empirical distribution ν, i.e. $x^*=\inf\{q:\nu([0,q])\geq\frac{b}{b+h}\}$. Under Assumption 3.1, we derive that the distributionally robust solution with respect to the Wasserstein ambiguity set is equivalent to the optimal solution of the data-driven newsvendor model or sample average approximation (SAA) solution. This result coincides with Remark 6.7 in Esfahani and Kuhn (2018).

The worst-case distribution in $\mathcal{M}$ is an optimal solution of the inner optimization of (1). It is important to analyze the closed-form solution of the worst-case distribution because the structure of the distribution affects the conservativeness of the DRO solution. Therefore, the existence conditions and structure of the worst-case distribution in the general distributionally robust optimization with a Wasserstein ambiguity set are studied. We refer to Gao and Kleywegt (2016) and Esfahani and Kuhn (2018) for more details. We focus on the newsvendor case and propose the explicit characterization of the worst-case distribution based on historical data. The following worst-case distribution in $\mathcal{P}_1(\Xi)$ is the optimal solution of the inner optimization of (2).

Proposition 3.3. (Worst-case distribution for p = 1). *For each $x\geq 0$, let*

$$\mu^*(x):=\frac{1}{N}\sum_{i\in N_1(x)}\delta_{\hat{\xi}^i}+\frac{1}{N}\sum_{i\in N_2(x)}\delta_{(\hat{\xi}^i+\frac{N\theta}{|N_2(x)|})}.$$

Then, $\mu^*(x)$ is the worst-case distribution for a given x.

Proof. To check if $\mu^*(x)$ is the maximizer of the inner maximization of (2) for each x, we have to prove that $\mu^*(x)$ is a feasible distribution in the Wasserstein ambiguity set and satisfies strong duality. First, we show $\mu^*(x)$ satisfies strong duality.

$$\begin{aligned}&\mathbb{E}_{\mu^*(x)}\left[h(x-\xi)^+ + b(\xi-x)^+\right]\\ &=b\theta+\sum_{i\in N_1(x)}h(x-\hat{\xi}^i)+\sum_{i\in N_2(x)}b(\hat{\xi}^i-x)\\ &=\max_{\mu\in\mathcal{P}(\Xi)}\{\mathbb{E}_\mu\left[h(x-\xi)^+ + b(\xi-x)^+\right]:W_1(\mu,\nu)\leq\theta\}.\end{aligned}$$

The first equality holds by the characterization of $\mu^*(x)$ and the second equality holds by strong duality. Hence, $\mu^*(x)$ satisfies strong duality.

To verify $\mu^*(x)$ is a feasible distribution, let $\zeta^i=\hat{\xi}^i$ for $i\in N_1(x)$ and $\zeta^i=\hat{\xi}^i+\frac{N\theta}{|N_2(x)|}$ for $i\in N_2(x)$. Then, by the definition of the Wasserstein distance,

$$\begin{aligned}&W_1(\mu^*(x),\nu)\\ &=\min\left\{\frac{1}{N}\sum_{i=1}^{N}|\hat{\xi}^i-\hat{\zeta}^{\sigma(i)}|:\sigma\in\Pi_N\right\}\leq\frac{1}{N}\sum_{i=1}^{N}|\hat{\xi}^i-\hat{\zeta}^i|\\ &=\frac{1}{N}\left\{\sum_{i\in N_1(x)}|\hat{\xi}^i-\hat{\xi}^i|+\sum_{i\in N_2(x)}\left|\hat{\xi}^i-\hat{\xi}^i-\frac{N\theta}{|N_2(x)|}\right|\right\}=\theta,\end{aligned}$$

where Π_N is all permutations of $\{1,\ldots,N\}$. The inequality holds by letting $\sigma(i)=i$. Then, $\mu^*(x)$ is a feasible distribution and inside the Wasserstein ambiguity set. □

We note that the worst-case distribution is constructed by historical data itself for data in $N_1(x)$ and the perturbation of the data in $N_2(x)$. The perturbation depends on historical data and the Wasserstein radius. We also note that the worst-case

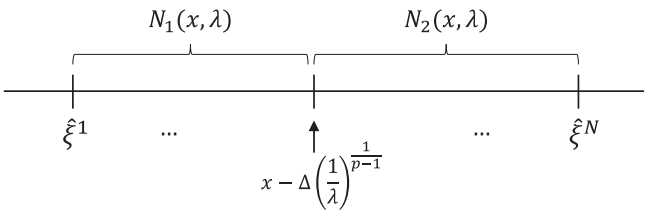

Figure 3. Definitions of $N_1(x,\lambda)$ and $N_2(x,\lambda)$.

distribution is not unique. The worst-case distribution implies that the supremum of (2) is indeed a maximum.

3.2. Wasserstein order $p > 1$

For the $p>1$ case, the analysis is more difficult than the $p=1$ case, but the Wasserstein distance of greater order is stronger by monotonicity (Proposition 2.1) and reflects geometric properties better (Villani, 2009). When $p>1$, the growth rate is 0, which leads to $\lambda \geq 0$.

The derivation of the closed-form solution (Theorem 3.7) consists of several stages. To express the inner supremum of (3) in the explicit form, we partition data based on a specific point determined by x and λ. Based on the partition, the inner supremum can be characterized in explicit form, which is essential for the analysis of the dual formulation (Proposition 3.4). The explicit characterization of the inner supremum leads to a concise expression of the objective function of the dual formulation. We derive the optimality condition based on the structure of the objective function (Lemma 3.6). Using the optimality condition, the objective function is separable in the order of data and dual variable λ, which leads to the derivation of the closed-form solution.

First, we derive the equivalent expression of the inner supremum similar to the $p=1$ case. We consider the partition of historical data to weaken the dependence of x and λ, which leads to the explicit characterization of the inner supremum. For given (x,λ), we define $f_i(\xi) := h(x-\xi)^+ + b(\xi-x)^+ - \lambda|\xi-\hat{\xi}^i|^p$ for $i=1,\ldots,N$ to analyze the inner supremum of (3). By dividing the intervals according to the positions of (x,λ) and $\hat{\xi}^i$ and analyzing the cases, the maximum of $f_i(\xi)$ is attained at $\xi_r^i := \hat{\xi}^i + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}}$ or $\xi_l^i := \hat{\xi}^i - \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}}$ based on $f_i(\xi_r^i)$ and $f_i(\xi_l^i)$. The comparison of two values $b(\hat{\xi}^i - x) + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)$ and $h(x-\hat{\xi}^i) + \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right)$ leads to the following definitions. For given (x,λ), we define $N_1(x,\lambda) := \{1 \leq i \leq N : x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} > \hat{\xi}^i\}$ and $N_2(x,\lambda) := \{1 \leq i \leq N : x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} \leq \hat{\xi}^i\}$, where $\Delta := \frac{1}{h+b}\left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\left(b^{\frac{p}{p-1}} - h^{\frac{p}{p-1}}\right)$ (Figure 3). Under Assumption 3.1, $\Delta \geq 0$ and $\Delta = 0$ when $b=h$. Then, for $i \in N_1(x,\lambda)$, $f_i(\xi_l^i) = h(x-\hat{\xi}^i) + \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right)$, and for $i \in N_2(x,\lambda)$, $f_i(\xi_r^i) = b(\hat{\xi}^i - x) + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)$.

There are two interpretations of $N_1(x,\lambda)$ and $N_2(x,\lambda)$. First, by the definitions, $i \in N_1(x,\lambda)$ represents the samples whose values are less than $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}}$ and $i \in N_2(x,\lambda)$ represents the samples whose values are greater than or equal to $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}}$. Second, the inequality $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} > \hat{\xi}^i$ is equivalent to $h(x-\hat{\xi}^i) + \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right) > b(\hat{\xi}^i - x) + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)$. On the contrary, the inequality $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} \leq \hat{\xi}^i$ is equivalent to $h(x-\hat{\xi}^i) + \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right) \leq b(\hat{\xi}^i - x) + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)$. With the second interpretation of $N_1(x,\lambda)$ and $N_2(x,\lambda)$, we obtain the explicit expression of the inner supremum.

Proposition 3.4. *Under Assumption 3.1, if $\xi_l^i \geq 0$ for $i \in N_1(x,\lambda)$,*

$$\sup_{\xi \in \Xi} f_i(\xi) = f_i(\xi_l^i) = h(x-\hat{\xi}^i) + \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right),$$

and for $i \in N_2(x,\lambda)$,

$$\sup_{\xi \in \Xi} f_i(\xi) = f_i(\xi_r^i) = b(\hat{\xi}^i - x) + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right).$$

In the proposition, ξ_l^i should be nonnegative for $i \in N_1(x,\lambda)$ to attain the maximum of $f_i(\xi)$ at $\xi = \xi_l^i$. The following assumption guarantees that $\xi_l^i \geq 0$ when a dual variable λ is chosen properly.

Assumption 3.5. *For all $i=1,\ldots,N$, i-th data is greater than or equal to the Wasserstein radius, i.e. $\hat{\xi}^i \geq \theta$.*

Under Assumption 3.5 with properly chosen λ, we will prove that $\xi_l^i \geq 0$ for $i \in N_1(x,\lambda)$ in the proof of Theorem 3.7. Therefore, we can obtain the

maximum of $f_i(\xi)$ at $\xi = \xi_l^i$ for $i \in N_1(x,\lambda)$. Even though the above assumption is needed for the technical reason, the assumption holds in most practical instances. The size of the Wasserstein radius to guarantee the probability that a Wasserstein ambiguity set contains the unknown true distribution is $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ (Esfahani & Kuhn, 2018; Fournier & Guillin, 2015). Hence, except for the extremely small demand, Assumption 3.5 holds in most cases.

The next step is to express the objective function of dual formulation with the explicit form of the inner supremum. By the definitions of $N_1(x,\lambda)$ and $N_2(x,\lambda)$ and Proposition 3.4, the objective function of dual formulation (3) can be expressed as follows:

$$
\begin{aligned}
&\lambda\theta^p + \frac{1}{N}\left\{\sum_{i\in N_1}\left(h(x-\hat{\xi}^i) + \left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right)\right) + \sum_{i\in N_2}\left(b(\hat{\xi}^i - x) + \left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)\right)\right\} \\
&= \lambda\theta^p + \frac{1}{N}\left\{\sum_{i\in N_1} h\left(x - \Delta(\lambda)^{-\frac{1}{p-1}} - \hat{\xi}^i\right) + \sum_{i\in N_2} b\left(\hat{\xi}^i - x + \Delta(\lambda)^{-\frac{1}{p-1}}\right)\right\} \\
&+ \frac{1}{N}\left\{\sum_{i\in N_1} h\Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} - \sum_{i\in N_2} b\Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} + \sum_{i\in N_1}\left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right) + \sum_{i\in N_2}\left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)\right\} \\
&= \lambda\theta^p \mathrm{I} + \mathrm{II} + \left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}}
\end{aligned}
$$

where $\mathrm{I} := \frac{1}{N}\left\{\sum_{i\in N_1} h\left(x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} - \hat{\xi}^i\right) + \sum_{i\in N_2} b\left(\hat{\xi}^i - x + \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}}\right)\right\}$ and $\mathrm{II} := \left(\frac{|N_1|}{N} h\Delta - \frac{|N_2|}{N} b\Delta + \frac{|N_1|}{N}\left(\frac{h}{p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right) + \frac{|N_2|}{N}\left(\frac{b}{p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right)\right)$. We suppress the dependence of x and λ on N_1 and N_2 for notational brevity. The first equality holds by adjusting the term $\Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}}$ for x to make a similar structure of the objective function for the $p=1$ case. The objective function consists of three parts: $\lambda\theta^p$, the data-driven newsvendor cost based on $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}}$, and the remainder. Now, we derive the optimality condition based on the structure of the objective function.

Lemma 3.6. *Under Assumptions 3.1 and 3.5, the optimal solution* (x^*, λ^*) *satisfies* $x^* - \Delta\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}} = \hat{\xi}^i$ *for some* $i = 1, \ldots, N$.

Proof. We prove the lemma by contradiction. When we suppose the lemma does not hold, there exist i and an optimal solution (x^*, λ^*) such that $\hat{\xi}^i < x^* - \Delta\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}} < \hat{\xi}^{i+1}$. If we fix λ^*, then the first and third parts of the objective function, $\lambda^*\theta^p + \mathrm{II}\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}}$, are fixed. To see the change of the objective function value as x^* changes, we increase x^* as $x^* - \Delta\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}}$ remains in the interval $(\hat{\xi}^i, \hat{\xi}^{i+1})$. Then the cost change of the second part, I, is $\frac{i}{N}h - \frac{N-i}{N}b$. By increasing or decreasing x, we can change the total cost downward, which contradicts the assumption that (x^*, λ^*) is optimal. □

Therefore, we can express the optimal solution (x^*, λ^*) as $x^* - \Delta\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}} = \hat{\xi}^i$ for some $i = 1, \ldots, N$ and the corresponding cost $\mathrm{I} = \frac{1}{N}\left\{\sum_{k=1}^{i-1} h(\hat{\xi}^i - \hat{\xi}^k) + \sum_{k=i}^{N} b(\hat{\xi}^k - \hat{\xi}^i)\right\}$. By the above lemma, we can assume that $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} = \hat{\xi}^{i+1}$ for some i. Then $|N_1(x,\lambda)| = i$ and $|N_2(x,\lambda)| = N - i$. To see the independence of x and λ on II, we arrange the equation.

$$
\begin{aligned}
\mathrm{II} &= \frac{i}{N}h\Delta - \frac{N-i}{N}b\Delta + \frac{i}{N}\left(\frac{h}{p}\right)^{\frac{1}{p-1}} h\left(\frac{p-1}{p}\right) \\
&\quad + \frac{N-i}{N}\left(\frac{b}{p}\right)^{\frac{1}{p-1}} b\left(\frac{p-1}{p}\right) \\
&= \frac{i}{N}\frac{h}{h+b}\left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\left(b^{\frac{p}{p-1}} - h^{\frac{p}{p-1}}\right) \\
&\quad - \frac{N-i}{N}\frac{b}{h+b}\left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\left(b^{\frac{p}{p-1}} - h^{\frac{p}{p-1}}\right) \qquad (4) \\
&\quad + \frac{i}{N}\left(\frac{1}{p}\right)^{\frac{1}{p-1}} h^{\frac{p}{p-1}}\left(\frac{p-1}{p}\right) + \frac{N-i}{N}\left(\frac{1}{p}\right)^{\frac{1}{p-1}} b^{\frac{p}{p-1}}\left(\frac{p-1}{p}\right) \qquad (5) \\
&= \left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\left(\frac{1}{h+b}\right)\left(b^{\frac{p}{p-1}}h + h^{\frac{p}{p-1}}b\right) \qquad (6)
\end{aligned}
$$

The second equality holds by the substituting $\Delta := \frac{1}{h+b}\left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\left(b^{\frac{p}{p-1}} - h^{\frac{p}{p-1}}\right)$ back to the original equation. Then, Equality (6) shows that II does not

depend on x and λ. Let $\Lambda := \frac{1}{h+b}\left(b^{\frac{p}{p-1}}h + h^{\frac{p}{p-1}}b\right) \geq 0$. Then $\text{II} = \left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\Lambda$ and the objective function of (3) can be expressed as follows:

$$\vartheta(i+1,\lambda) := \lambda\theta^p + \left(\frac{1}{p}\right)^{\frac{1}{p-1}}\left(\frac{p-1}{p}\right)\Lambda\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} + \frac{1}{N}\left(\sum_{k=1}^{i} h(\hat{\xi}^{i+1} - \hat{\xi}^k) + \sum_{k=i+1}^{N} b(\hat{\xi}^k - \hat{\xi}^{i+1})\right) \tag{7}$$

Using the fact $x - \Delta\left(\frac{1}{\lambda}\right)^{\frac{1}{p-1}} = \hat{\xi}^{i+1}$, the objective function (7) can be expressed as the function of $i+1$ and λ. By the above analysis, $\vartheta(i+1,\lambda)$ is separable in $i+1$ and λ, which leads to an optimal solution.

Theorem 3.7. *Under Assumptions 3.1 and 3.5, an optimal order quantity is* $x^* = \hat{\xi}^{i^*} + \Delta p^{\frac{1}{p-1}}\theta\left(\frac{1}{\Lambda}\right)^{\frac{1}{p}}$, *and an optimal dual variable is* $\lambda^* = \frac{1}{p\theta^{p-1}}(\Lambda)^{\frac{p-1}{p}}$ *where* $i^* \in \{1,\ldots,N\}$ *satisfies* $\frac{i^*-1}{N} < \frac{b}{h+b} \leq \frac{i^*}{N}$. *The optimal cost is*

$$\theta\Lambda^{\frac{p-1}{p}} + \frac{1}{N}\left(\sum_{k=1}^{i^*-1} h(\hat{\xi}^{i^*} - \hat{\xi}^k) + \sum_{k=i^*}^{N} b(\hat{\xi}^k - \hat{\xi}^{i^*})\right).$$

Proof. Under the assumptions, Lemma 3.6 holds and the objective function is expressed as (7). First, we derive the optimal dual variable λ^*. Since $\vartheta(i,\lambda)$ is a separable function of i and λ, the optimal dual variable λ^* can be derived by partial differentiation in λ.

$$\frac{\partial\vartheta}{\partial\lambda} = \theta^p - \left(\frac{1}{p}\right)^{\frac{p}{p-1}}\Lambda\left(\frac{1}{\lambda}\right)^{\frac{p}{p-1}} = 0$$

Then, $\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}} = p^{\frac{1}{p-1}}\theta\left(\frac{1}{\Lambda}\right)^{\frac{1}{p}}$ and $\lambda^* = \frac{1}{p\theta^{p-1}}\Lambda^{\frac{p-1}{p}}$. To obtain the optimal i^*, we express the objective function using λ^*.

$$\begin{aligned}\vartheta(i+1,\lambda^*) &= \frac{1}{p\theta^{p-1}}\Lambda^{\frac{p-1}{p}}\theta^p + \frac{(p-1)\theta}{p}\Lambda^{\frac{p-1}{p}} \\ &\quad + \frac{1}{N}\left(\sum_{k=1}^{i} h(\hat{\xi}^{i+1} - \hat{\xi}^k) + \sum_{k=i+1}^{N} b(\hat{\xi}^k - \hat{\xi}^{i+1})\right) \\ &= \theta\Lambda^{\frac{p-1}{p}} + \frac{1}{N}\left(\sum_{k=1}^{i} h(\hat{\xi}^{i+1} - \hat{\xi}^k) + \sum_{k=i+1}^{N} b(\hat{\xi}^k - \hat{\xi}^{i+1})\right)\end{aligned}$$

To minimize $\vartheta(i+1,\lambda^*)$ in terms of $i+1$, we refer the analysis of Theorem 3.2. There exists i^* such that $\frac{i^*-1}{N} < \frac{b}{h+b} \leq \frac{i^*}{N}$ and i^* minimizes $\vartheta(i,\lambda^*)$, that is, $x^* - \Delta\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}} = \hat{\xi}^{i^*}$. Then, $x^* = \hat{\xi}^{i^*} + \Delta\left(\frac{1}{\lambda^*}\right)^{\frac{1}{p-1}} = \hat{\xi}^{i^*} + \Delta p^{\frac{1}{p-1}}\theta\left(\frac{1}{\Lambda}\right)^{\frac{1}{p}}$. Therefore, the optimal solution of (3) is (x^*,λ^*) and the optimal cost is $\theta\Lambda^{\frac{p-1}{p}} + \frac{1}{N}\left(\sum_{k=1}^{i^*-1} h(\hat{\xi}^{i^*} - \hat{\xi}^k) + \sum_{k=i^*}^{N} b(\hat{\xi}^k - \hat{\xi}^{i^*})\right)$.

To see that the optimal dual variable λ^* under Assumption 3.5 satisfies $\xi_l^i = \hat{\xi}^i - \left(\frac{h}{\lambda^* p}\right)^{\frac{1}{p-1}} \geq 0$ for all $i \in N_1$ and Proposition 3.4 holds, we analyze the following inequality.

$$\begin{aligned}&\hat{\xi}^i \geq \left(\frac{h}{\lambda^* p}\right)^{\frac{1}{p-1}} \\ &\iff \lambda^*(\hat{\xi}^i)^{p-1} \geq \frac{h}{p} \\ &\iff \left(\frac{\hat{\xi}^i}{\theta}\right)^{p-1}\left(\left(\frac{1}{h+b}\right)\left(b^{\frac{p}{p-1}}h + h^{\frac{p}{p-1}}b\right)\right)^{\frac{p-1}{p}} \geq h \\ &\iff \left(\frac{\hat{\xi}^i}{\theta}\right)^{p}\left(\frac{1}{h+b}\right)\left(b^{\frac{p}{p-1}}h + h^{\frac{p}{p-1}}b\right) \geq h^{\frac{p}{p-1}}\end{aligned}$$

If $\left(\frac{\hat{\xi}^i}{\theta}\right)^p \geq 1$, then λ^* satisfies $\xi_l^i \geq 0$, because $\left(b^{\frac{p}{p-1}}h + h^{\frac{p}{p-1}}b\right) \geq (h+b)h^{\frac{p}{p-1}}$. In short, if Assumption 3.5 holds, then the optimal dual variable λ^* satisfies $\xi_l^i \geq 0$ for $i \in N_1$. □

Remark 2. If $b = h$, then $\Delta = 0$ and $x^* = \hat{\xi}^{i^*}$. In this case, the optimal order quantity is equal to the SAA solution the same as the $p = 1$ case. The optimal cost is $\theta b^{\frac{p}{p-1}} + \frac{1}{N}\left(\sum_{k=1}^{i^*-1} h(\hat{\xi}^{i^*} - \hat{\xi}^k) + \sum_{k=i^*}^{N} b(\hat{\xi}^k - \hat{\xi}^{i^*})\right)$. The only difference is $\theta b^{\frac{p}{p-1}}$ compared to θb in the $p = 1$ case.

For the $p > 1$ case, the optimal order quantity is the sum of the SAA solution and $\Delta p^{\frac{1}{p-1}}\theta\left(\frac{1}{\Lambda}\right)^{\frac{1}{p}}$, where the second part is determined by parameters h, b, and p. If $b > h$, the optimal order quantity x^* is greater than the SAA solution. The first derivative of $\Delta p^{\frac{1}{p-1}}\theta\left(\frac{1}{\Lambda}\right)^{\frac{1}{p}}$ with respect to $p \geq 2$ is negative, so the optimal order quantity is decreasing in $p \geq 2$ with a fixed set of parameters. The first derivative goes to 0 as p increases to infinity, and $\Delta p^{\frac{1}{p-1}}\theta\left(\frac{1}{\Lambda}\right)^{\frac{1}{p}}$ goes to $\frac{b-h}{b+h}\theta$ as p increases to infinity. The optimal cost is decreasing in $p > 1$, because $\theta\Lambda^{\frac{p-1}{p}}$ is decreasing in $p > 1$. This result is explained by the monotone property of the Wasserstein distance. According to Proposition 2.1, $W_{p_1} \leq W_{p_2}$ if $p_1 \leq p_2$, and the ambiguity set becomes smaller as the Wasserstein order p increases. Therefore, the newsvendor model with a higher order is less conservative and the optimal cost is smaller.

We now characterize the worst-case distribution based on $N+1$ points perturbed from historical data. The structure of the worst-case distribution is similar to the $p = 1$ case except for the split of mass for $\hat{\xi}^{i^*}$. The worst-case distribution has mass $\frac{p_0}{N}$ and $\frac{1-p_0}{N}$ at two points perturbed from $\hat{\xi}^{i^*}$, respectively,

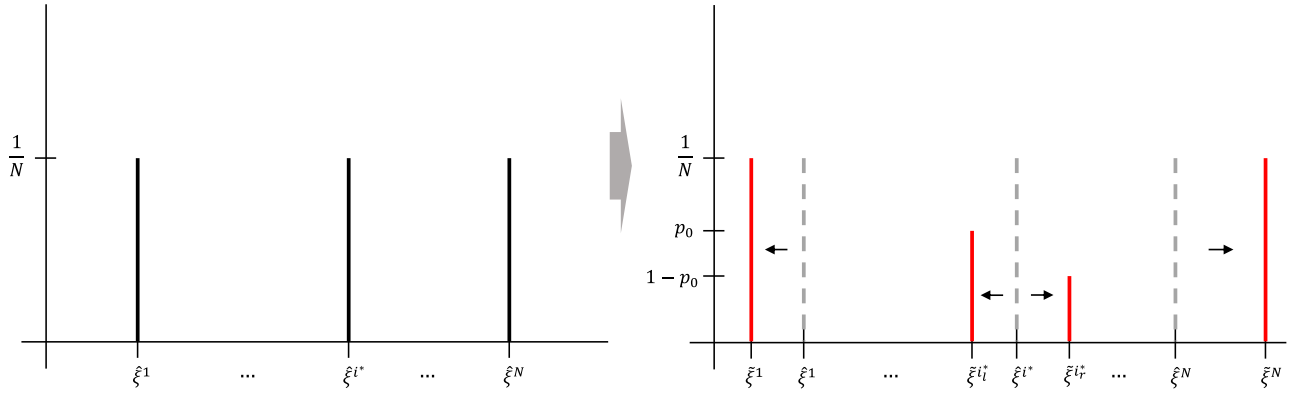

Figure 4. Worst-case distribution perturbed from the empirical distribution.

and mass $\frac{1}{N}$ at $N-1$ points perturbed from historical data (Figure 4).

Proposition 3.8. (Worst-case distribution for $p>1$). *For given optimal order quantity and dual variable* (x^*,λ^*), *let* $\tilde{\xi}^i := \hat{\xi}^i - \left(\frac{h}{\lambda^* p}\right)^{\frac{1}{p-1}} = \hat{\xi}^i - h^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}$ *for* $i \in N_1(x^*,\lambda^*)$ *and* $\tilde{\xi}^i := \hat{\xi}^i + \left(\frac{b}{\lambda^* p}\right)^{\frac{1}{p-1}} = \hat{\xi}^i + b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}$ *for* $i \in N_2(x^*,\lambda^*)\setminus\{i^*\}$. *Let* $\tilde{\xi}^{i_l^*} := \hat{\xi}^{i^*} - h^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}$ *and* $\tilde{\xi}^{i_r^*} := \hat{\xi}^{i^*} + b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}$ *be two points perturbed from* $\hat{\xi}^{i^*}$. *Then,* $\mu^*(x^*) := \frac{1}{N}\sum\delta_{\tilde{\xi}^i} + \frac{p_0}{N}\delta_{\tilde{\xi}^{i_l^*}} + \frac{1-p_0}{N}\delta_{\tilde{\xi}^{i_r^*}}$ *is the worst-case distribution where* $p_0 \in [0,1]$ *satisfies* $\frac{|N_1(x^*,\lambda^*)|+p_0}{N} = \frac{b}{h+b}$.

Proof. We prove that $\mu^*(x^*)$ is a feasible distribution and satisfies the strong duality similar to the $p=1$ case. For notational brevity, we suppress dependence on x and λ on N_1 and N_2. First, we will show that points perturbed from data points satisfy the desirable ordering based on x^*, even separation of N_1 and N_2 is based on $x^* - \Delta(\frac{1}{\lambda^*})^{\frac{1}{p-1}}$, that is, $\tilde{\xi}^i < x^*$ for $i \in N_1$ and i_l^*, and $\tilde{\xi}^i \geq x^*$ for $i \in N_2\setminus\{i^*\}$ and i_r^*. This ordering seems unintuitive when we consider definitions of N_1 and N_2, but the ordering is important for the evaluation of the objective function value. By the definitions of N_1, $\tilde{\xi}^i$ and $\tilde{\xi}^{i_l^*}$, $\tilde{\xi}^i < \hat{\xi}^i < x^*$ for $i \in N_1$ and $\tilde{\xi}^{i_l^*} < \hat{\xi}^{i^*} < x^*$. Since $b^{\frac{1}{p-1}} \geq \Delta p^{\frac{1}{p-1}} = \frac{1}{h+b}\frac{p-1}{p}\left(b^{\frac{p}{p-1}} - h^{\frac{p}{p-1}}\right)$, $\tilde{\xi}^{i_r^*} = \hat{\xi}^{i^*} + b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}} \geq \hat{\xi}^{i^*} + \Delta p^{\frac{p}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}} = x^*$. Then, $\tilde{\xi}^{i_r^*} \leq \tilde{\xi}^i$ for $i \in N_2\setminus\{i^*\}$, that is, $\tilde{\xi}^i \geq x^*$ for $i \in N_2\setminus\{i^*\}$ and i_r^*.

Second, the objective function value with $\mu^*(x^*)$ satisfies the strong duality. The objective function with x^* and $\mu^*(x^*)$ is expressed as follows:

$$
\begin{aligned}
&\mathbb{E}_{\mu^*(x^*)}\left[h(x^*-\xi)^+ + b(\xi-x^*)^+\right] \\
&= \frac{1}{N}\sum_{i\in N_1} h(x^*-\tilde{\xi}^i) + \frac{1}{N}\sum_{i\in N_2\setminus\{i^*\}} b(\tilde{\xi}^i - x^*) + \frac{p_0}{N}h(x-\tilde{\xi}^{i_l^*}) + \frac{1-p_0}{N}b(\tilde{\xi}^{i_r^*}-x^*) \\
&= \frac{1}{N}\sum_{i\in N_1} h\left(\hat{\xi}^{i^*} + \Delta p^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}} - \hat{\xi}^i + h^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}\right) + \frac{1}{N}\sum_{i\in N_2\setminus\{i^*\}} b\left(\hat{\xi}^i + b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}} - \hat{\xi}^{i^*} - \Delta p^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}\right) \\
&+ \frac{p_0}{N}h\left(\hat{\xi}^{i^*} + \Delta p^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}} - \hat{\xi}^{i^*} + h^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}\right) + \frac{1-p_0}{N}b\left(\hat{\xi}^{i^*} + b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}} - \hat{\xi}^{i^*} - \Delta p^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^{\frac{1}{p}}\right) \\
&= \frac{1}{N}\sum_{i\in N_1} h(\hat{\xi}^{i^*} - \hat{\xi}^i) + \frac{1}{N}\sum_{i\in N_2} b(\hat{\xi}^i - \hat{\xi}^{i^*}) \\
&+ \theta\Lambda^{-\frac{1}{p}}\left\{\frac{|N_1|+p_0}{N}h\Delta p^{\frac{1}{p-1}} - \frac{|N_2|-p_0}{N}b\Delta p^{\frac{1}{p-1}} + \frac{|N_1|+p_0}{N}h^{\frac{p}{p-1}} + \frac{|N_2|-p_0}{N}b^{\frac{p}{p-1}}\right\} \\
&= \frac{1}{N}\sum_{i\in N_1} h(\hat{\xi}^{i^*} - \hat{\xi}^i) + \frac{1}{N}\sum_{i\in N_2} b(\hat{\xi}^i - \hat{\xi}^{i^*}) + \theta\Lambda^{-\frac{1}{p}}\left(\frac{b}{h+b}h^{\frac{p}{p-1}} + \frac{h}{h+b}b^{\frac{p}{p-1}}\right) \\
&= \frac{1}{N}\sum_{i\in N_1} h(\hat{\xi}^{i^*} - \hat{\xi}^i) + \frac{1}{N}\sum_{i\in N_2} b(\hat{\xi}^i - \hat{\xi}^{i^*}) + \theta\Lambda^{\frac{p}{p-1}}
\end{aligned}
$$

The equalities hold by definitions of $\mu^*(x^*), \tilde{\xi}$, and p_0. Therefore, $\mu^*(x^*)$ satisfies the strong duality.

Third, $\mu^*(x^*)$ is a feasible distribution, that is, the Wasserstein distance between $\mu^*(x^*)$ and the empirical distribution ν is less than or equal to the Wasserstein radius θ. By the definition of the Wasserstein distance, $W_p(\mu^*(x^*), \nu)$ can be expressed as follows:

$$W_p(\mu^*(x^*),\nu)^p = \min_{\gamma_{ij}\geq 0} \sum_{i=1}^{N+1}\sum_{j=1}^{N} |\tilde{\xi}^i - \hat{\xi}^j|^p$$

$$\text{s.t.} \sum_{j=1}^{N} \gamma_{ij} = \begin{cases} \frac{1}{N}, & \text{if } i \in N_1 \cup N_2 \setminus \{i_l^*, i_r^*\} \\ \frac{p_0}{N}, & \text{if } i = i_l^* \\ \frac{1-p_0}{N}, & \text{if } i = i_r^* \end{cases}$$

$$\sum_{i=1}^{N+1} \gamma_{ij} = \frac{1}{N} \quad \text{for } j = 1, \ldots, N$$

Then, the following inequality holds.

$$\begin{aligned} & W_p(\mu^*(x^*),\nu)^p \\ & \leq \frac{1}{N}\sum_{i\in N_1} |\tilde{\xi}^i - \hat{\xi}^i|^p + \frac{1}{N}\sum_{i\in N_2\setminus\{i^*\}} |\tilde{\xi}^i - \hat{\xi}^i|^p \\ & + \frac{p_0}{N}|\tilde{\xi}^{i_l^*} - \hat{\xi}^{i^*}|^p + \frac{1-p_0}{N}|\tilde{\xi}^{i_r^*} - \hat{\xi}^{i^*}|^p \\ & = \frac{1}{N}\sum_{i\in N_1}\left(h^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^p\right) + \frac{1}{N}\sum_{i\in N_2\setminus\{i^*\}}\left(b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^p\right) \\ & + \frac{p_0}{N}\left(h^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^p\right) + \frac{1-p_0}{N}\left(b^{\frac{1}{p-1}}\theta(\frac{1}{\Lambda})^p\right) \\ & = \theta^p \frac{1}{\Lambda}\left(\frac{|N_1|+p_0}{N}h^{\frac{p}{p-1}} + \frac{|N_2|-p_0}{N}b^{\frac{p}{p-1}}\right) \\ & = \theta^p \frac{1}{\Lambda}\Lambda = \theta^p \end{aligned}$$

The first inequality holds, because $\gamma_{ij} = \frac{1}{N}$ for $i = j, i \neq i_l^*, i_r^*, \gamma_{i_l^*, i^*} = \frac{p_0}{N}$, and $\gamma_{i_r^*, i^*} = \frac{1-p_0}{N}$ is a feasible solution of γ_{ij}. The second equality holds by the definitions of p_0 and Λ. Therefore, the Wasserstein distance between $\mu^*(x^*)$ and ν is less than or equal to θ and $\mu^*(x^*)$ is a feasible distribution. □

We derived the closed-form expressions of an optimal order quantity and the worst-case distribution for the general Wasserstein order $p \in [1, \infty)$. The closed-form solutions can be applied to various applications such as multistage inventory control, pricing, and retail management. In the next section, we consider the risk measure as an objective for the risk-averse case, rather than an expected cost as the risk-neutral case.

4. Risk-averse newsvendor model

Although the long-run average performance of the risk-neutral solution outperforms that of the risk-averse solution, risk-averse decisions of the first few periods are important in terms of protection against bankruptcy. Moreover, many decision makers are risk-averse in reality. Therefore, we consider a risk-averse model with the Conditional Value-at-Risk (CVaR) objective. CVaR has several strong points. CVaR is a coherent risk measure and preserves convexity of the newsvendor cost function. Optimization with CVaR is much easier by using the following definition.

Definition 4.1. (CVaR, Rockafellar and Uryasev (2000)).

$$\text{CVaR}_\mu^\beta(X) := \inf_{\alpha\in\mathbb{R}} \left\{\alpha + \frac{1}{1-\beta}\mathbb{E}_\mu\left[(X-\alpha)^+\right]\right\}$$

Let $F_\beta(x,\alpha,\mu) := \alpha + \frac{1}{1-\beta}\int (h(x-\xi)^+ + b(\xi-x)^+ - \alpha)^+ \mu(d\xi)$. Then, minimization of the worst-case CVaR of the newsvendor cost function is represented as follows:

$$\begin{aligned} & \min_{x\geq 0} \sup_{\mu\in\mathcal{M}} \text{CVaR}_\mu^\beta(h(x-\xi)^+ + b(\xi-x)^+) \\ & = \min_{x\geq 0} \sup_{\mu\in\mathcal{M}} \inf_{\alpha\in\mathbb{R}} F_\beta(x,\alpha,\mu) \end{aligned} \tag{8}$$

Before we utilize Theorem 2.4 directly, we need to change the order of supremum and infimum with a general version of John von Neumann's minimax theorem.

Lemma 4.2. *(Sion's minimax theorem). Let X be a compact, convex subset of a topological vector space and Y be a convex subset of a topological vector space. Let f be a real-valued function on $X \times Y$ such that, $f(\cdot, y)$ is lower-semicontinuous and quasi-convex on X for each $y \in Y$ and $f(x, \cdot)$ is upper-semicontinuous and quasi-concave on Y for each $x \in X$. Then,*

$$\inf_{x\in X}\sup_{y\in Y} f(x,y) = \sup_{y\in Y}\inf_{x\in X} f(x,y).$$

First, we check that the risk-averse newsvendor model satisfies the conditions of Lemma 4.2 to change the order of operations. $F_\beta(x,\alpha,\mu)$ is convex in α (Rockafellar & Uryasev, 2000), and affine in μ. The Wasserstein ambiguity set $\mathcal{M}$ is convex by Proposition 2.3. For fixed x, minimizer of $\inf_{\alpha\in\mathbb{R}} F_\beta(x,\alpha,\mu)$ is achieved in $[0, hx]$ (Chen et al., 2009; Gotoh & Takano, 2007). It can be easily proved that $F_\beta(x,\alpha,\mu)$ is continuous with respect to α and μ. Then,

$$\begin{aligned} & \sup_{\mu\in\mathcal{M}}\inf_{\alpha\in\mathbb{R}} F_\beta(x,\alpha,\mu) = \sup_{\mu\in\mathcal{M}}\inf_{\alpha\in[0,hx]} F_\beta(x,\alpha,\mu) \\ & = \inf_{\alpha\in[0,hx]}\sup_{\mu\in\mathcal{M}} F_\beta(x,\alpha,\mu) \geq \inf_{\alpha\in\mathbb{R}}\sup_{\mu\in\mathcal{M}} F_\beta(x,\alpha,\mu) \\ & \geq \sup_{\mu\in\mathcal{M}}\inf_{\alpha\in\mathbb{R}} F_\beta(x,\alpha,\mu) \end{aligned}$$

The second equality holds by Lemma 4.2 and the last inequality holds by minimax inequality.

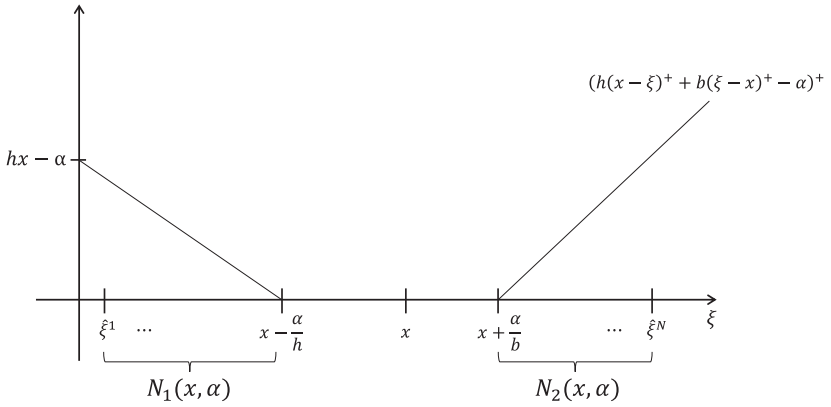

Figure 5. Definitions of $N_1(x,\alpha)$ and $N_2(x,\alpha)$.

By the above result and Theorem 2.5, we obtain

$$\min_{x\geq 0}\sup_{\mu\in\mathcal{M}}\inf_{\alpha\in\mathbb{R}} F_\beta(x,\alpha,\mu) = \min_{x\geq 0}\inf_{\alpha\in\mathbb{R}}\sup_{\mu\in\mathcal{M}} F_\beta(x,\alpha,\mu)$$
$$= \min_{x,\,\lambda\geq 0,\,\alpha\in\mathbb{R}}\left\{\alpha + \frac{1}{1-\beta}\lambda\theta^p + \frac{1}{1-\beta}\frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi} \left[\left(h(x-\xi)^+ + b(\xi-x)^+ - \alpha\right)^+ - \lambda|\xi-\hat{\xi}^i|^p\right]\right\}. \tag{9}$$

We consider two cases similar to the risk-neutral model: $p=1$ and $p>1$. We utilize a similar analysis of (3), which leads to a closed-form solution for the $p=1$ case and a tractable formulation for the $p>1$ case.

4.1. Wasserstein order $p=1$

First, we characterize the explicit form of the inner supremum of (9) for $p=1$. For given (x,λ,α), let $\eta_i(\xi) := (h(x-\xi)^+ + b(\xi-x)^+ - \alpha)^+ - \lambda|\xi-\hat{\xi}^i|$. For the supremum of η_i, we consider the interval of ξ which makes the first part of η_i positive. Otherwise, $\eta_i(\xi)$ would be less than or equal to 0 and the supremum of η_i is 0 when $\xi=\hat{\xi}^i$. We define $N_1(x,\alpha) := \{1\leq i\leq N : \hat{\xi}^i < x-\frac{\alpha}{h}\}$ and $N_2(x,\alpha) := \{1\leq i\leq N : \hat{\xi}^i \geq x+\frac{\alpha}{b}\}$ (Figure 5). Then, $N_1(x,\alpha)\subset N_1(x)$ and $N_2(x,\alpha)\subset N_2(x)$ by the definitions of $N_1(x)$ and $N_2(x)$ in Section 3.1.

We also assume that Assumption 3.1 holds, i.e. $b\geq h$. By the analysis similar to that of $f_i(\xi)$ in Section 3.1, for $i\in N_1(x,\alpha)$ the supremum is attained at $\xi=\hat{\xi}^i$ with $\eta_i(\hat{\xi}^i) = h(x-\hat{\xi}^i)-\alpha$, and for $i\in N_2(x,\alpha)$ the supremum is attained at $\xi=\hat{\xi}^i$ with $\eta_i(\hat{\xi}^i) = b(\hat{\xi}-x)-\alpha$. For $i\in N_1(x)\setminus N_1(x,\alpha)$ and $i\in N_2(x)\setminus N_2(x,\alpha)$, the supremum of η_i is 0. Therefore, the linkage between x and λ is disconnected by the analysis of supremum of η_i, which leads to separation of x and λ in the objective function in (9). Using the explicit form of the inner supremum, the CVaR objective (9) can be expressed as follows:

$$\alpha + \frac{1}{1-\beta}\lambda\theta + \frac{1}{1-\beta}\frac{1}{N}\left(\sum_{i\in N_1(x,\alpha)}\left(h(x-\hat{\xi}^i)-\alpha\right) + \sum_{i\in N_2(x,\alpha)}\left(b(\hat{\xi}^i-x)-\alpha\right)\right) \tag{10}$$

Then, $\lambda^* = b$ by the growth rate constraint, and the remaining part can be expressed as follows, which is the same as the data-driven CVaR model where the true distribution is given as the empirical distribution ν, $\mathrm{CVaR}^\beta_\nu(h(x-\xi)^+ + b(\xi-x)^+)$:

$$\min_{x\geq 0,\,\alpha\in\mathbb{R}} \alpha + \frac{1}{1-\beta}\frac{1}{N}\left(\sum_{i\in N_1(x,\alpha)}\left(h(x-\hat{\xi}^i)-\alpha\right) + \sum_{i\in N_2(x,\alpha)}\left(b(\hat{\xi}^i-x)-\alpha\right)\right) \tag{11}$$

The closed-form solution of the CVaR newsvendor model is characterized by Gotoh and Takano (2007). When distribution function F is given, the inverse distribution function of F is defined as $F^{-1}(q) := \inf\{x\in\mathbb{R} : F(x)\geq q\}$. Then, the closed-form solutions are given as follows:

$$x^* = \frac{h}{h+b}F^{-1}\left(\frac{b(1-\beta)}{h+b}\right) + \frac{b}{h+b}F^{-1}\left(\frac{b+h\beta}{h+b}\right)$$
$$\alpha^* = \frac{hb}{h+b}\left(F^{-1}\left(\frac{b+h\beta}{h+b}\right) - F^{-1}\left(\frac{b(1-\beta)}{h+b}\right)\right).$$

If the true distribution F is given as the empirical distribution ν, the inverse function is written as

$$F^{-1}\left(\frac{b(1-\beta)}{h+b}\right) = \inf\left\{x : \nu([0,x]) \geq \frac{b(1-\beta)}{h+b}\right\}$$
$$F^{-1}\left(\frac{b+h\beta}{h+b}\right) = \inf\left\{x : \nu([0,x]) \geq \frac{b+h\beta}{h+b}\right\}.$$

Therefore, by the definition of the empirical distribution ν, there exist i_1 and i_2 such that

$$\frac{i_1-1}{N}<\frac{b(1-\beta)}{h+b}\le\frac{i_1}{N}$$
$$\frac{i_2-1}{N}<\frac{b+h\beta}{h+b}\le\frac{i_2}{N}.$$

Then the closed-form solution of (9) for $p=1$ can be expressed using i_1 and i_2.

Theorem 4.3. *Under Assumption 3.1, the optimal solutions of the worst-case CVaR newsvendor model are*

$$x^*=\frac{h}{h+b}\hat{\xi}^{i_1}+\frac{b}{h+b}\hat{\xi}^{i_2}$$
$$\alpha^*=\frac{hb}{h+b}\left(\hat{\xi}^{i_2}-\hat{\xi}^{i_1}\right)$$
$$\lambda^*=b.$$

The optimal objective function value is

$$\frac{hb}{h+b}(\hat{\xi}^{i_2}-\hat{\xi}^{i_1})+\frac{1}{1-\beta}b\theta$$
$$+\frac{1}{1-\beta}\frac{1}{N}\left\{\sum_{k=1}^{i_1-1}h(\hat{\xi}^{i_1}-\hat{\xi}^{k})+\sum_{k=i_2}^{N}b(\hat{\xi}^{k}-\hat{\xi}^{i_2})\right\}.$$

By the analysis of the inner supremum and separation of x and λ, the optimal order quantity is the same as the data-driven CVaR solution. Theorem 4.3 shows that the equivalence between the SAA solution and the distributionally robust solution with respect to the Wasserstein ambiguity set is extended to the risk-averse case for $p=1$.

Using the closed-form solutions in Theorem 4.3, $x^*-\frac{\alpha^*}{h}=\hat{\xi}^{i_1}$ and $x^*+\frac{\alpha^*}{b}=\hat{\xi}^{i_2}$, which leads to $N_1(x^*,\alpha^*)=\{1,\dots,i_1-1\}$ and $N_2(x^*,\alpha^*)=\{i_2,\dots,N\}$ if N samples are sorted in nondecreasing order. The above expression is useful for deriving the worst-case distribution.

Proposition 4.4. (Worst-case distribution). *For given (x^*,λ^*,α^*), let*

$$\mu^*(x^*,\alpha^*):=\frac{1}{N}\sum_{i\in N\setminus N_2(x^*,\alpha^*)}\delta_{\hat{\xi}^i}+\frac{1}{N}\sum_{i\in N_2(x^*,\alpha^*)}\delta_{\left(\hat{\xi}^i+\frac{N\theta}{|N_2|}\right)}.$$

Then, $\mu^*(x^*,\alpha^*)$ is the worst-case distribution.

Proof. $\mu^*(x^*,\alpha^*)$ satisfies the strong duality and is a feasible distribution. We omit the detailed proof, because it is similar to the proof of Proposition 3.3. □

4.2. Wasserstein order $p>1$

To characterize the explicit form of the inner supremum of (9) for $p>1$, we define $\eta_i(\xi):=(h(x-\xi)^+ + b(\xi-x)^+-\alpha)^+-\lambda|\xi-\hat{\xi}^i|^p$ for given (x,λ,α) and $i=1,\dots,N$. Then, $\sup_{\xi\in\Xi}\eta_i(\xi)\ge 0$ by the definition of η_i. If $h(x-\xi)^+ + b(\xi-x)^+-\alpha\ge 0$, i.e. $\xi<x-\frac{\alpha}{h}$ or $\xi\ge x+\frac{\alpha}{b}$, then $\eta_i(\xi)=h(x-\xi)^+ + b(\xi-x)^+-\alpha+\lambda|\xi-\hat{\xi}^i|^p$. Using the analysis of the inner supremum in Section 3.2, the supremum of η_i may be attained at $\xi_r^i:=\hat{\xi}^i+(\frac{b}{\lambda p})^{\frac{1}{p-1}}$ or $\xi_l^i:=\hat{\xi}^i-(\frac{h}{\lambda p})^{\frac{1}{p-1}}$. Specifically, for $i\in N_1(x,\lambda)$, $\eta_i(\xi_l^i)=h(x-\hat{\xi}^i)-\alpha+(\frac{h}{\lambda p})^{\frac{1}{p-1}}h\left(\frac{p-1}{p}\right)$ should be greater than or equal to 0, that is, $\hat{\xi}^i<x-\frac{\alpha}{h}+(\frac{h}{\lambda p})^{\frac{1}{p-1}}\frac{p-1}{p}$. For $i\in N_2(x,\lambda)$, $\eta_i(\xi_r^i)=b(\hat{\xi}^i-x)-\alpha+(\frac{b}{\lambda p})^{\frac{1}{p-1}}b\left(\frac{p-1}{p}\right)$ should be greater than or equal to 0, that is, $\hat{\xi}^i\ge x+\frac{\alpha}{b}-(\frac{b}{\lambda p})^{\frac{1}{p-1}}\frac{p-1}{p}$.

Using the above result, we need definitions of index sets considering interactions of both (x,λ) and (x,α). We define $N_1(x,\lambda,\alpha):=\{1\le i\le N:\hat{\xi}^i<x-\Delta(\frac{1}{\lambda})^{\frac{1}{p-1}},\hat{\xi}^i<x-\frac{\alpha}{h}+(\frac{h}{\lambda p})^{\frac{1}{p-1}}\frac{p-1}{p}\}$ and $N_2(x,\lambda,\alpha):=\{1\le i\le N:\hat{\xi}^i\ge x-\Delta(\frac{1}{\lambda})^{\frac{1}{p-1}},\hat{\xi}^i\ge x+\frac{\alpha}{b}-(\frac{b}{\lambda p})^{\frac{1}{p-1}}\frac{p-1}{p}\}$. For $i\in N_1(x,\lambda,\alpha)$, the supremum of η_i is attained at ξ_l^i with $\eta_i(\xi_l^i)=h(x-\hat{\xi}^i)-\alpha+(\frac{h}{\lambda p})^{\frac{1}{p-1}}h\left(\frac{p-1}{p}\right)$. For $i\in N_2(x,\lambda,\alpha)$, the supremum of η_i is attained at ξ_r^i with $\eta_i(\xi_r^i)=b(\hat{\xi}^i-x)-\alpha+(\frac{b}{\lambda p})^{\frac{1}{p-1}}b\left(\frac{p-1}{p}\right)$. If $i\notin N_1(x,\lambda,\alpha)$ and $i\notin N_2(x,\lambda,\alpha)$, then supremum of η_i is 0.

Using the explicit form of the inner supremum, we derive the equivalent expression of (9) for $p>1$. We suppress the dependence of (x,λ,α) on N_1 and N_2 for notational convenience.

$$\min_{x,\lambda\ge0,\alpha\in\mathbb{R}}\ \alpha+\frac{1}{1-\beta}\lambda\theta^p+\frac{1}{1-\beta}\frac{1}{N}$$
$$\left(\sum_{i\in N_1}\left(h(x-\hat{\xi}^i)-\alpha+\left(\frac{h}{\lambda p}\right)^{\frac{1}{p-1}}h\frac{p-1}{p}\right)\right.$$
$$\left.+\sum_{i\in N_2}\left(b(\hat{\xi}^i-x)-\alpha+\left(\frac{b}{\lambda p}\right)^{\frac{1}{p-1}}b\frac{p-1}{p}\right)\right).\tag{12}$$

The objective function of (12) without $\frac{1}{1-\beta}\lambda\theta^p$ can be expressed as the CVaR objective. Let $\bar{\xi}^i:=\hat{\xi}^i-(\frac{h}{\lambda p})^{\frac{1}{p-1}}\frac{p-1}{p}$ for $i\in N_1$ and $\bar{\xi}^i:=\hat{\xi}^i+(\frac{b}{\lambda p})^{\frac{1}{p-1}}\frac{p-1}{p}$ for $i\in N_2$. Then, we define

$$\mu(x,\lambda):=\frac{1}{N}\sum_{i\in N_1}\delta_{\bar{\xi}^i}+\frac{1}{N}\sum_{i\in N_2}\delta_{\bar{\xi}^i}+\frac{1}{N}\sum_{i\in N\setminus(N_1\cup N_2)}\delta_{\hat{\xi}^i}.$$

Using the definition of $\mu(x,\lambda)$, (12) can be expressed as follows:

$$\min_{x,\lambda\ge0,\alpha\in\mathbb{R}}\frac{1}{1-\beta}\lambda\theta^p+\mathrm{CVaR}^{\beta}_{\mu(x,\lambda)}\left(h(x-\xi)^+ + b(\xi-x)^+\right)\tag{13}$$

However, there exists dependence between x and λ, which makes further analysis complex. In this case,

it is difficult to derive closed-form solutions, but we can obtain a tractable formulation using the analysis of the inner supremum.

Theorem 4.5. *Under Assumption 3.1, the optimal order quantity of (12) is determined by the following formulation:*

$$\begin{aligned}\min_{x,\lambda\geq 0,y,\alpha\in\mathbb{R}} \quad & \alpha+\frac{1}{1-\beta}\lambda\theta^p+\frac{1}{1-\beta}\frac{1}{N}\sum_{i=1}^{N}y_i \\ \text{s.t.} \quad & y_i \geq h(x-\hat{\xi}^i)-\alpha+\left(\tfrac{h}{\lambda p}\right)^{\frac{1}{p-1}}h\frac{p-1}{p}, \quad \forall i=1,\ldots,N,\\ & y_i \geq b(\hat{\xi}^i-x)-\alpha+\left(\tfrac{b}{\lambda p}\right)^{\frac{1}{p-1}}b\frac{p-1}{p}, \quad \forall i=1,\ldots,N,\\ & y_i \geq 0, \qquad \forall i=1,\ldots,N,\\ & \alpha \leq hx.\end{aligned}$$

For the risk-averse decision, we consider the CVaR objective for the newsvendor model. We derive the closed-form solution for the $p=1$ case, and propose a tractable formulation to obtain the optimal order quantity for the $p>1$ case. In the next section, numerical experiments are conducted to verify the risk-aversion of the CVaR solution using Theorem 4.3 and 4.5.

5. Computational experiments

We conduct numerical experiments to compare the Wasserstein model with other data-driven models such as the ϕ-divergence model (Ben-Tal et al., 2013). We also compare with the Scarf's moment-based model (Scarf, 1958) in terms of convergence of order quantities and total costs to those with the true underlying distribution. Then, we compare the risk-averse solution with the risk-neutral solution in terms of the worst cost with pessimistic demand realization.

5.1. Out-of-sample performance

One important aspect when analyzing a distributionally robust solution is out-of-sample performance, i.e. the average performance of the distributionally robust solution over test samples from the data-generating distribution. We compare the out-of-sample performance of distributionally robust solutions for Wasserstein models and ϕ-divergence models. The ϕ-divergence models are another data-driven distributionally robust approach based on ϕ-divergence, using measures such as KL-divergence and χ^2-distance (Ben-Tal et al., 2013).

Let $\mu=(\mu_1,\ldots,\mu_N)$ and $\nu=(\nu_1,\ldots,\nu_N)$ be probability distributions defined on N points. The ϕ-divergence model is given as follows:

$$\min_{x\geq 0}\ \sup_{\mu\in\mathcal{M}_\phi(\nu)}\ \mathbb{E}_\mu\left[h(x-\xi)^+ + b(\xi-x)^+\right], \tag{14}$$

where $\mathcal{M}_\phi(\nu):=\{\mu\in\mathcal{P}(\Xi): I_\phi(\mu,\nu)\leq\rho\}$ and $I_\phi(\mu,\nu)=\sum_{i=1}^{N}\nu_i\phi\left(\frac{\mu_i}{\nu_i}\right)$, i.e. the ϕ-divergence ambiguity set based on the empirical distribution ν. ϕ-divergence is only defined between the empirical distribution and distributions that are absolutely continuous to the empirical distribution. Hence, the ϕ-divergence ambiguity set consists of distributions whose supports are the same as the empirical distribution. From the strong duality of the ϕ-divergence model (Ben-Tal et al., 2013), the newsvendor model with the ϕ-divergence ambiguity set is expressed as follows:

$$\begin{aligned}\min_{x\geq 0,\lambda\geq 0,\eta}\ & \eta+\rho\lambda \\ & +\lambda\sum_{i=1}^{N}\frac{1}{N}\phi^*\left(\frac{h(x-\hat{\xi}^i)^+ + b(\hat{\xi}^i-x)^+-\eta}{\lambda}\right)\end{aligned} \tag{15}$$

where ϕ^* is the conjugate of ϕ. In this experiment, we use KL-divergence and χ^2-distance for ϕ-divergence. The ϕ-divergence function of KL-divergence is $\phi(t)=t\log t-t+1$ and the conjugate function is $\phi^*(s)=e^s-1$. The ϕ-divergence function of χ^2-distance is $\phi(t)=\frac{1}{t}(t-1)^2$ and the conjugate function is $\phi^*(s)=2-2\sqrt{1-s}$ for $s\leq 1$. For both cases, the resulting dual formulations are tractable (Ben-Tal et al., 2013).

We will compare the empirical out-of-sample performance of distributionally robust solutions. We set $h=1$ and $b\in\{1,3,9,19\}$ to set the resulting critical ratio as $\{0.50,0.75,0.90,0.95\}$, respectively. We generate $N\in\{50,500\}$ samples from a normal distribution with two different parameters: mean $m=100$ and standard deviation $s\in\{20,40\}$ for different coefficients of variation $\mathrm{CV}=\frac{s}{m}$ as $\{0.2,0.4\}$. We let the Wasserstein radius $\theta=1$, which is large enough to guarantee probability bounds (Esfahani & Kuhn, 2018; Fournier & Guillin, 2015). We let the radius of ϕ-divergence models $\rho=0.5$. Let p-Wasserstein denote the Wasserstein newsvendor models of the Wasserstein order p for $p=1$, 2. $p=1$ and $p=2$ are the widely used orders for the Wasserstein distance. The Wasserstein and ϕ-divergence solutions are derived based on the N historical data. Then, the demand d is realized and the total cost is calculated based on the order quantities and the realized demand as $\mathrm{TC}(x^*,d)=h(x^*-d)^+ + b(d-x^*)^+$. We generate 500 demand realizations for d to evaluate the average of the realized total cost $\mathrm{TC}(x^*,d)$ from the same normal distribution. The average of the realized total cost represents the empirical out-of-sample performance of distributionally robust solutions. We conduct 100 iterations to generate data sets and calculate the average and maximum of simulated costs when the optimal solutions are implemented.

Table 1. Empirical out-of-sample performance of optimal order quantities for Wasserstein and ϕ-divergence models when CV = 0.2 and CV = 0.4.

			1-Wasserstein			2-Wasserstein			KL-divergence			χ^2-distance		
CV	b	N	x_{avg}	c_{avg}	c_{max}	x_{avg}	c_{avg}	c_{max}	x_{avg}	c_{avg}	c_{max}	x_{avg}	c_{avg}	c_{max}
0.2	1	50	98.91	16.18	17.85	98.91	16.18	17.85	98.91	16.18	17.85	99.89	16.60	20.81
	1	500	99.77	15.93	17.47	99.77	15.93	17.47	99.77	15.93	17.47	100.81	16.43	19.16
	3	50	113.08	25.82	30.34	113.66	25.80	30.05	120.06	26.99	33.75	122.22	28.43	44.44
	3	500	113.31	25.40	28.63	113.89	25.40	28.53	121.74	27.30	29.89	131.74	33.95	56.02
	9	50	124.17	36.07	44.32	125.50	35.90	42.45	136.12	39.89	58.76	135.56	39.66	58.84
	9	500	125.64	35.09	39.22	126.98	35.16	38.99	145.39	46.35	64.02	150.23	50.74	79.70
	19	50	132.02	42.59	54.00	134.09	42.43	51.32	140.85	45.43	64.65	139.97	45.10	63.97
	19	500	132.80	41.39	47.27	134.86	41.54	47.17	155.51	55.98	86.27	156.39	56.80	87.61
0.4	1	50	97.81	32.36	35.71	97.81	32.36	35.71	98.28	32.33	35.71	99.83	33.04	40.20
	1	500	99.54	31.86	34.94	99.54	31.86	34.94	100.16	31.88	35.05	101.61	32.85	38.32
	3	50	126.17	51.64	60.68	126.74	51.62	60.39	140.47	54.04	67.75	141.85	55.24	80.42
	3	500	126.63	50.80	57.26	127.20	50.80	57.16	143.61	54.65	59.84	163.47	67.91	112.04
	9	50	148.34	72.15	88.63	149.67	71.94	86.75	172.24	79.79	117.49	169.45	78.10	112.70
	9	500	151.29	70.19	78.43	152.62	70.22	78.16	191.20	93.11	143.84	200.46	101.48	159.41
	19	50	164.04	85.18	108.00	166.11	84.92	105.15	181.71	90.86	129.31	179.58	89.78	127.18
	19	500	165.59	82.78	94.54	167.66	82.82	94.39	210.99	111.95	173.48	212.79	113.60	175.21

x_{avg} denotes the average of optimal order quantities of each model. c_{avg} and c_{max} denote the average and maximum of empirical out-of-sample performances over 100 iterations. The results are summarized in Table 1.

The optimal order quantities of the ϕ-divergence models are larger than those of the Wasserstein models for most cases, and the gaps between order quantities increase as b increases. Therefore, the gaps between average simulated costs are considerable, which reflects the better out-of-sample performances of the Wasserstein solutions compared to those of the ϕ-divergence solutions. The optimal order quantities for the ϕ-divergence models are sensitive to the coefficient of variation, that is, the decision-maker considering the ϕ-divergence model over-orders when the variance is large. Another reason for the gaps is that the ϕ-divergence ambiguity set cannot contain normal distributions that are data-generating distributions and the resulting ambiguity set becomes unrealistic. In contrast, the Wasserstein ambiguity set contains the unknown true distribution with a certain probability (Esfahani & Kuhn, 2018). In summary, the optimal order quantities for the Wasserstein models have better out-of-sample performances than those for the ϕ-divergence models.

5.2. Convergence property

One important property of Wasserstein DRO is convergence property, i.e. as the sample size N increases, the distributionally robust solution converges to the true optimal solution with the complete knowledge of the probability distribution. According to the result of Theorem 3.2, for $p = 1$, an optimal order quantity with the Wasserstein ambiguity set is equivalent to the SAA solution. If we choose the Wasserstein radius proportional to $\frac{1}{\sqrt{N}}$ for probability guarantee (Fournier & Guillin, 2015), then the ambiguity set shrinks to the empirical distribution as the sample size N goes to infinity. We choose the Wasserstein radius $\theta_N = \frac{10}{\sqrt{N}}$ to control the conservativeness of the Wasserstein models. Then, the optimal solution of the Wasserstein newsvendor model converges to the optimal solution with true distribution as the sample size N increases, even if the decision maker does not know the true distribution. Furthermore, with θ_N, the objective function value of the Wasserstein models converges to the optimal cost with true distribution as N goes to infinity.

In this experiment, we compare the convergence property of the Wasserstein solutions compared to the moment-based DRO model, specifically Scarf's model. Scarf's model is given by

$$\min_{x\geq 0} \sup_{\mu\in\mathcal{M}(m,s)} \mathbb{E}_\mu\left[h(x-\xi)^+ + b(\xi-x)^+\right], \qquad (16)$$

where $\mathcal{M}(m,s)$ denotes the moment-based ambiguity set with known first and second moments, i.e. a set of probability distributions with mean m and standard deviation s. In our setting, the decision maker knows only about historical data. Hence, m and s are estimated by sample mean and sample standard deviation, respectively. Scarf's solution is $x_{\text{Scarf}} = m + \frac{s}{2}\tau$, where $\tau = \sqrt{\frac{b}{h}} - \sqrt{\frac{h}{b}}$ and the optimal objective function value is $s\sqrt{bh}$.

In the data-driven approach, the decision maker can update the order quantity based on the realized data. For example, we consider the repeated setting of the newsvendor problem, that is, a decision maker decides the order quantity repeatedly through the planning horizon. The decision maker decides x_N based on N samples and demand is realized after the decision is implemented. The realized demand can be used to make a decision x_{N+1} with $N+1$ samples. Likewise, order quantities are updated with

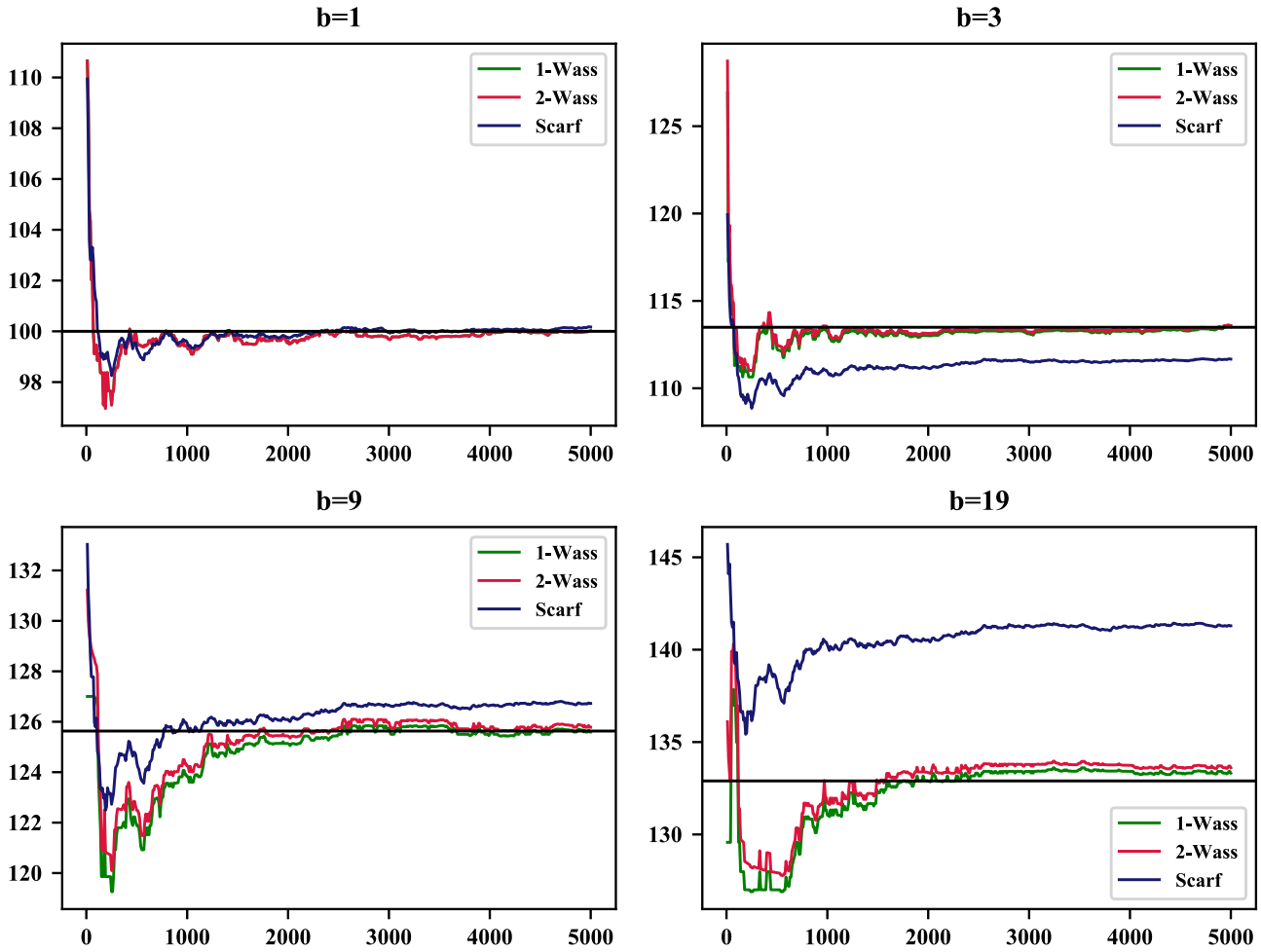

Figure 6. Convergence of optimal order quantities to true optimal order quantities as sample size (N) increases.

realized demand as sample size N increases. We conduct numerical experiments by updating order decisions based on the realized demand data until $N = 5,000$. We generated demand samples from a normal distribution of mean 100 and standard deviation 20, and the experimental setting is similar to that of Section 5.1. The updated order quantities and objective function values are shown in Figures 6 and 7, respectively. The black horizontal line represents the optimal solution when the underlying true distribution is known.

When $b = 1$, the Wasserstein solution is equal to the median and Scarf's solution is equal to the sample mean, because $\tau = 0$. Therefore, Scarf's solution behaves similarly to that of the Wasserstein model, because samples are generated from a normal distribution and mean and median of a normal distribution are equal. In other cases ($b = 3$, 9, 19), the Wasserstein solutions converge to the true optimal solution, whereas the Scarf's solutions do not. Moreover, the Wasserstein radius θ_N decreases as N increases, which leads to the convergence of objective function values of the Wasserstein model. This convergence result reflects one of the important advantages of the Wasserstein model.

5.3. Risk-aversion of the CVaR solution

In some cases, it is important to reduce the risk of extremely large costs caused by the pessimistic demand realization. To verify such a risk-aversion property of the CVaR model, the following experiment is conducted. Risk-neutral and CVaR solutions are derived based on the N historical data generated from a normal distribution of mean 100 and standard deviation 20. Then, the demand d is realized and the total cost is calculated based on the order quantities and the realized demand as $\mathrm{TC}(x^*, d) = h(x^* - d)^+ + b(d - x^*)^+$. The experimental setting is the same as that described in Section 5.1 such that we generate 500 demand realizations for d to evaluate the realized total cost from the same normal distribution. We let $\beta \in \{0.2, 0.5, 0.9\}$ for the CVaR coefficient.

The optimal solution x^* and the average and maximum of realized total costs ($\mathrm{TC}_{\mathrm{avg}}, \mathrm{TC}_{\mathrm{max}}$) are summarized in Tables 2 and 3 for $p = 1$ and $p = 2$, respectively. The CVaR model becomes more risk averse as β increases. Hence, the optimal order quantity tends to increase as β increases, but there are exceptions because of the data-driven setting. The closed-form solution of the CVaR model

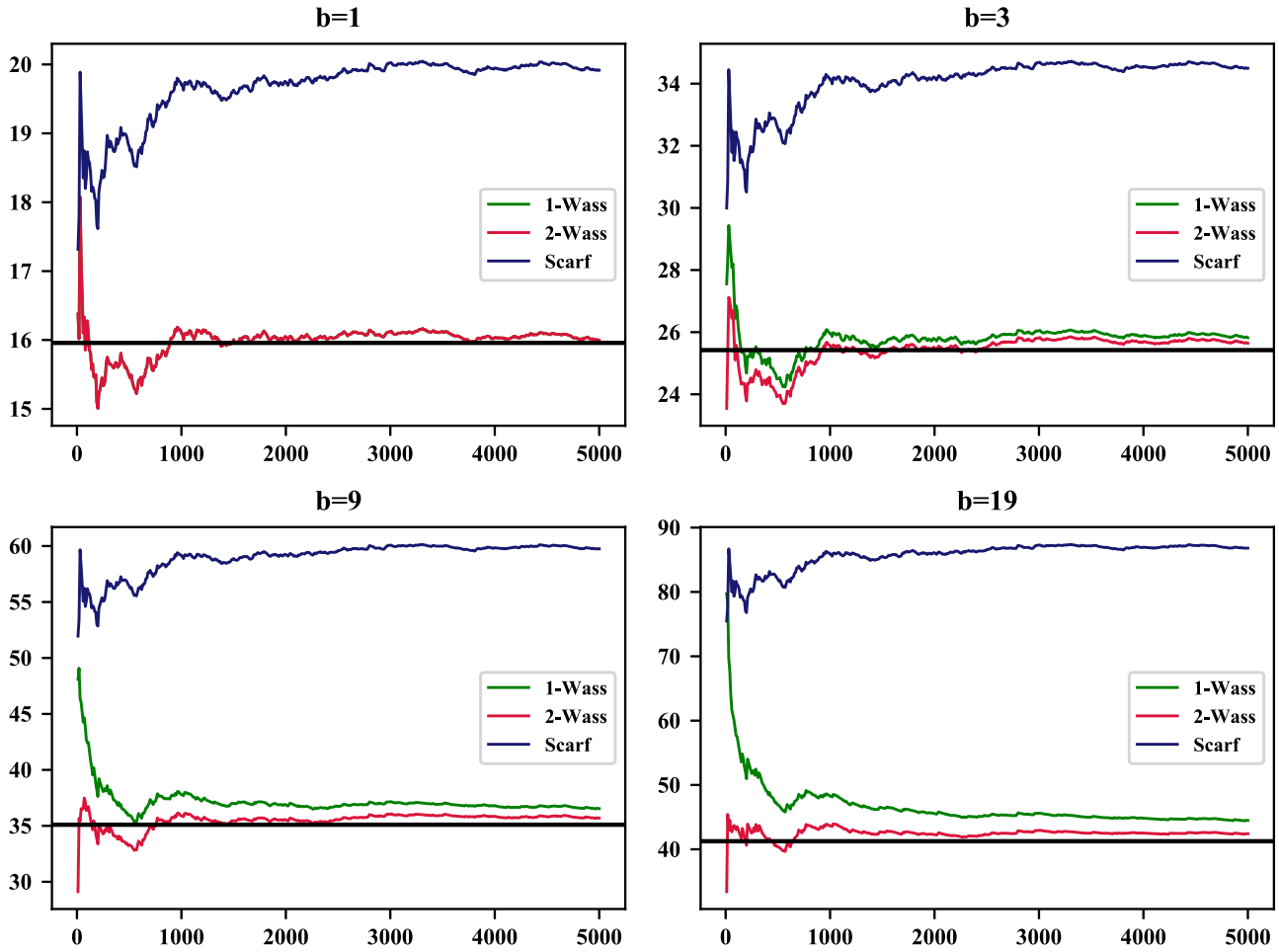

Figure 7. Convergence of objective function values to true optimal cost as sample size (N) increases.

Table 2. Optimal solutions and simulation results for risk-neutral and CVaR models with different values of β when the Wasserstein $p = 1$.

		risk-neutral			CVaR: $\beta = 0.2$			CVaR: $\beta = 0.5$			CVaR: $\beta = 0.9$		
b	N	x^*	TC_{avg}	TC_{max}	x^*	TC_{avg}	TC_{max}	x^*	TC_{avg}	TC_{max}	x^*	TC_{avg}	TC_{max}
1	50	96.44	15.82	66.89	97.64	15.68	65.69	99.87	15.56	66.20	100.02	15.56	66.35
1	500	99.42	15.57	65.75	99.66	15.56	65.99	99.21	15.57	65.54	99.12	15.58	65.44
3	50	111.22	25.14	156.32	110.71	25.18	157.85	114.42	25.27	146.74	117.09	25.79	138.72
3	500	112.23	25.11	153.30	112.60	25.12	152.18	115.98	25.54	142.06	118.41	26.20	134.76
9	50	121.91	35.57	372.76	123.95	35.21	354.40	127.47	35.27	322.70	129.03	35.58	308.74
9	500	125.92	35.13	336.68	126.66	35.18	330.01	128.23	35.40	315.88	130.76	36.15	293.10
19	50	130.98	41.03	614.63	129.99	41.28	633.40	129.73	41.37	638.36	132.01	40.92	595.13
19	500	131.73	40.95	600.33	131.28	40.99	608.91	132.74	40.88	581.18	139.61	43.47	450.73

Table 3. Optimal solutions and simulation results for risk-neutral and CVaR models with different values of β when the Wasserstein $p = 2$.

		risk-neutral			CVaR: $\beta = 0.2$			CVaR: $\beta = 0.5$			CVaR: $\beta = 0.9$		
b	N	x^*	TC_{avg}	TC_{max}	x^*	TC_{avg}	TC_{max}	x^*	TC_{avg}	TC_{max}	x^*	TC_{avg}	TC_{max}
1	50	96.44	15.82	66.89	97.79	15.67	65.54	99.87	15.56	66.20	100.02	15.56	66.35
1	500	99.42	15.57	65.75	99.66	15.56	65.99	99.46	15.57	65.78	99.20	15.58	65.53
3	50	111.80	25.11	154.59	111.36	25.13	155.92	115.23	25.40	144.30	118.92	26.37	133.24
3	500	112.81	25.13	151.57	113.25	25.16	150.24	116.79	25.72	139.61	120.23	26.89	129.30
9	50	123.25	35.29	360.76	125.44	35.13	341.00	129.35	35.67	305.79	133.24	37.30	270.80
9	500	127.25	35.24	324.68	128.15	35.39	316.60	130.27	35.96	297.58	136.26	39.25	243.61
19	50	133.05	40.89	575.40	132.30	40.89	589.54	132.65	40.88	582.89	138.54	42.85	471.08
19	500	133.80	40.97	561.10	133.89	40.98	559.39	135.66	41.53	525.76	146.13	47.91	326.80

depends on historical data, so the order quantity may decrease as β increases, especially for small values of b. The most noticeable result is that, in most cases, the maximum cost of the risk-averse model is less than that of the risk-neutral model, and the difference becomes significant when b and

β are large. It means that the CVaR model prefers less risk of extremely large loss, even the average cost would be increased. When b is large, the order quantity of the risk-averse model is larger than that of the risk-neutral model to decrease the extremely large underage costs. Therefore, more cautious and concerned decision makers would prefer the risk-averse model.

6. Conclusions

In this paper, we considered a Wasserstein ambiguity set for the data-driven distributionally robust newsvendor model. To incorporate a wide range of random demand and the Wasserstein distance, we considered continuous and unbounded support $\Xi = [0, \infty)$ and the general Wasserstein order $p \in [1, \infty)$. We derived the closed-form expressions of the optimal order quantity and the worst-case distribution for the risk-neutral model. Esfahani and Kuhn (2018) also discussed the equivalence between a closed-form solution and the SAA solution when $p = 1$, but we extended the closed-form analysis to the general $p > 1$. We analyzed the structure of an optimal order quantity based on the closed-form expressions, which is characterized by the sum of the SAA solution and the value determined by the parameters. We also considered the risk-averse model with the CVaR objective, and derived the closed-form solution for the $p = 1$ case, and proposed a tractable formulation to obtain the optimal order quantity for the $p > 1$ case. We conducted numerical experiments to verify the out-of-sample performance of distributionally robust solutions and the convergence results of the Wasserstein models. The Wasserstein solutions showed better out-of-sample performance and convergence properties, which is an important advantage when applied to practical circumstances. The risk aversion of the CVaR model was analyzed in terms of the possibility of extremely large costs caused by the pessimistic realization.

This research can serve as a building block for applications of the newsvendor model. There are several extensions of the newsvendor model, such as multi-item setting, risk-averse models with various risk measures, and pricing. Furthermore, the closed-form of a newsvendor order quantity can be applied to various applications, for example, inventory management, supply chain contract, and many other operations management problems. For further research, the Wasserstein ambiguity set with the bounding of the shape of distributions, e.g. symmetry or unimodality, could be considered to incorporate prior information on distributions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT and Future Planning (MSIT) [NRF-2019R1A2C2084616].

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

Bayraksan, G. & Love, D. K. (2015). Data-driven stochastic programming using phi-divergences. D. Aleman & A. Thiele (Eds.), *The operations research revolution, Tutorials in Operations Research* (pp. 1–19). Catonsville, MD: INFORMS. doi:10.1287/educ.2015.0134.

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*(2), 341–357. doi:10.1287/mnsc.1120.1641

Bertsimas, D., Doan, X. V., Natarajan, K., & Teo, C.-P. (2010). Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, *35*(3), 580–602. doi:10.1287/moor.1100.0445

Bertsimas, D., Gupta, V., & Kallus, N. (2018). Robust sample average approximation. *Mathematical Programming*, *171*(1-2), 217–282. doi:10.1007/s10107-017-1174-z

Chen, Y., Xu, M., & Zhang, Z. G. (2009). A risk-averse newsvendor model under the CVaR criterion. *Operations Research*, *57*(4), 1040–1044. doi:10.1287/opre.1080.0603

Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, *58*(3), 595–612. doi:10.1287/opre.1090.0741

Esfahani, P. M., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, *171*(1-2), 115–166. doi:10.1007/s10107-017-1172-1

Fournier, N., & Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, *162*(3-4), 707–738. doi:10.1007/s00440-014-0583-7

Gallego, G., & Moon, I. (1993). The distribution free newsboy problem: Review and extensions. *The Journal of the Operational Research Society*, *44*(8), 825–834. doi:10.2307/2583894

Gao, R., & Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

Goh, J., & Sim, M. (2010). Distributionally robust optimization and its tractable approximations. *Operations Research*, *58*(4-part-1), 902–917. doi:10.1287/opre.1090.0795

Gotoh, J-y., & Takano, Y. (2007). Newsvendor solutions via conditional value-at-risk minimization. *European Journal of Operational Research*, *179*(1), 80–96. doi:10.1016/j.ejor.2006.03.022

Hanasusanto, G. A., Kuhn, D., Wallace, S. W., & Zymler, S. (2015). Distributionally robust multi-item

newsvendor problems with multimodal demand distributions. *Mathematical Programming*, *152*(1-2), 1–32. doi:10.1007/s10107-014-0776-y

Jiang, R., & Guan, Y. (2016). Data-driven chance constrained stochastic program. *Mathematical Programming*, *158*(1-2), 291–327. doi:10.1007/s10107-015-0929-7

Natarajan, K., Sim, M., & Uichanco, J. (2018). Asymmetry and ambiguity in newsvendor models. *Management Science*, *64*(7), 3146–3167. doi:10.1287/mnsc.2017.2773

Perakis, G., & Roels, G. (2008). Regret in the newsvendor model with partial information. *Operations Research*, *56*(1), 188–203. doi:10.1287/opre.1070.0486

Pflug, G. C., & Pichler, A. (2014). *Multistage stochastic optimization*. Springer.

Pichler, A., & Xu, H. (2017). Quantitative stability analysis for minimax distributionally robust risk optimization. *Mathematical Programming*, 1–31. doi:10.1007/s10107-018-1347-4

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, *2*(3), 21–42. doi:10.21314/JOR.2000.038

Rudin, W. (1991). *Functional analysis*. McGrawHill Inc.

Saghafian, S., & Tomlin, B. (2016). The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research*, *64*(1), 167–185. doi:10.1287/opre.2015.1454

Scarf, H. (1958). A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*.

Shafieezadeh-Abadeh, S., Kuhn, D., & Esfahani, P. M. (2019). Regularization via mass transportation. *Journal of Machine Learning Research*, *20*(103), 1–68.

Sun, H., & Xu, H. (2016). Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, *41*(2), 377–401. doi:10.1287/moor.2015.0732

Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2017). Wasserstein Auto-Encoders. *arXiv preprint arXiv:1711.01558*.

Villani, C. (2009). *Optimal transport: old and new*. Springer Science & Business Media.

Wang, Z., Glynn, P. W., & Ye, Y. (2016). Likelihood robust optimization for data-driven problems. *Computational Management Science*, *13*(2), 241–261. doi:10.1007/s10287-015-0240-3

Wiesemann, W., Kuhn, D., & Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, *62*(6), 1358–1376. doi:10.1287/opre.2014.1314

Wozabal, D. (2012). A framework for optimization under ambiguity. *Annals of Operations Research*, *193*(1), 21–47. doi:10.1007/s10479-010-0812-0

Yue, J., Chen, B., & Wang, M.-C. (2006). Expected value of distribution information for the newsvendor problem. *Operations Research*, *54*(6), 1128–1136. doi:10.1287/opre.1060.0318

Zhao, C., & Guan, Y. (2018). Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, *46*(2), 262–267. doi:10.1016/j.orl.2018.01.011

Zhu, Z., Zhang, J., & Ye, Y. (2013). Newsvendor optimization with limited distribution information. *Optimization Methods and Software*, *28*(3), 640–667. doi:10.1080/10556788.2013.768994