Astin Bulletin (2023), **53**, pp. 443–465
doi:10.1017/asb.2023.3

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

Worst-case moments under partial ambiguity

Qihe Tang and Yunshen Yang*

School of Risk and Actuarial Studies, UNSW Business School, UNSW Sydney, Sydney, NSW 2052, Australia
*Corresponding author. E-mail: yunshen.yang@unsw.edu.au

Received: 11 January 2022; **Revised:** 24 November 2022; **Accepted:** 9 January 2023; **First published online:** 13 March 2023

Keywords: Worst-case risk; scenario analysis; partial ambiguity; the Wasserstein distance; bootstrapping

Abstract

The model uncertainty issue is pervasive in virtually all applied fields but especially critical in insurance and finance. To hedge against the uncertainty of the underlying probability distribution, which we refer to as ambiguity, the worst case is often considered in quantifying the underlying risk. However, this worst-case treatment often yields results that are overly conservative. We argue that, in most practical situations, a generic risk is realized from multiple scenarios and the risk in some ordinary scenarios may be subject to negligible ambiguity so that it is safe to trust the reference distributions. Hence, we only need to consider the worst case in the other scenarios where ambiguity is significant. We implement this idea in the study of the worst-case moments of a risk in the hope to alleviate the over-conservativeness issue. Note that the ambiguity in our consideration exists in both the scenario indicator and the risk in the corresponding scenario, leading to a two-fold ambiguity issue. We employ the Wasserstein distance to construct an ambiguity ball. Then, we disentangle the ambiguity along the scenario indicator and the risk in the corresponding scenario, so that we convert the two-fold optimization problem into two one-fold problems. Our main result is a closed-form worst-case moment estimate. Our numerical studies illustrate that the consideration of partial ambiguity indeed greatly alleviates the over-conservativeness issue.

1. Introduction

This work revisits the notorious issue of model uncertainty, which is pervasive in virtually all applied fields using data due to scarcity or low quality of data. This model uncertainty issue is especially critical in insurance and finance for reasons such as the changing environment that introduces multiple layers of uncertainty, increasingly complex modern insurance and financial products, and information asymmetry between borrowers and lenders or between insurers and insureds.

Let us consider the estimation of $E\left[h(X)\right]$, where X is a general non-negative risk variable and $h(\cdot)$ is some non-negative deterministic loss function. As the probability distribution of X, denoted by F_X, is in general unknown, we need to use a reference distribution inferred from available data and sometimes also based on expert opinions. In view of the possible scarcity or low quality of the data as well as the possible bias of the expert, the objective is subject to an ambiguity issue, which may lead to significant deviations from the true value of $E\left[h(X)\right]$ and potentially derail the entire risk analysis.

In the literature, the two terms uncertainty and ambiguity are often used interchangeably. Following Pflug and Wozabal (2007), uncertainty refers to a situation that the model is known but the realizations are unknown while ambiguity refers to a situation that the probability distribution is unknown. This way to distinguish the two terms seems to be consistent with a majority of recent works in this field. Thus, in this paper, we use the term ambiguity to indicate that the underlying probability distribution is unknown.

To tackle the ambiguity issue, the worst-case treatment is often adopted. This treatment constructs an ambiguity set B of possible probability distributions and considers the worst case in this ambiguity set; that is,

$$\sup_{F_X \in B} E\left[h(X)\right]. \tag{1.1}$$

At the core of the worst-case treatment is the construction of such an ambiguity set B, which involves a trade-off issue. On the one hand, this set should be broad enough to cover the true distribution with a prescribed confidence level, but on the other hand, its construction should also be prudent enough to avoid over-conservativeness. While efforts continue to be devoted to methodological innovations, the over-conservativeness issue becomes prominent: When an ambiguity set is constructed to meet a prescribed confidence level, the results provided by the worst-case treatment are often too conservative and the realized confidence level may be unnecessarily too high. This issue represents a great obstacle for the results to be practically useful, but it so far has received very limited discussions.

To alleviate this over-conservativeness issue, we propose a scenario analysis in accordance with the reality that the realization of a generic risk is usually the aggregation of multiple scenarios subject to varying extents of ambiguity. For example, when a company extends to new lines of business, these new lines are likely subject to more ambiguity than the company's existing lines. Similarly, an insurance company faces losses from its unfamiliar territories that are subject to more ambiguity—hence may charge an additional ambiguity premium to those losses—than from its familiar territories. In studies of mortality and morbidity, the population is often described as the composition of different age groups among which we need to take special care of the oldest-old group. Note that scenario analysis as one of classical approaches to risk management is emphasized in some modern regulatory frameworks including Basel III and Swiss Solvency Test.

Suppose that there are two scenarios between which one, called ordinary scenario, has abundant data, and the other, called ambiguous scenario, may suffer from scarcity or low quality of data. Here, we consider only two scenarios for simplicity. It is relatively straightforward to extend the study to more than two scenarios although the formulations will become more complicated. Denoting by Y and Z the risks in the respective scenarios, the generic risk X, which is understood as the result of the aggregation of Y and Z, has the stochastic representation

$$X = (1 - I)Y + IZ,$$

where I is the indicator of the ambiguous scenario, assumed to be independent of the vector (Y, Z). Thus, X is a mixture of Y and Z with weights decided by the indicator I.

It is important to note that this is a process of aggregating the risks Y and Z from the respective scenarios into the generic risk X rather than a process of thinning the generic risk X into different scenarios. During this aggregation, the independence assumption between the scenario indicator I and the vector (Y, Z) comes quite naturally. For example, still consider an insurance company extending its business to a new territory. It would be a common practice in loss models to express its generic loss distribution as a mixture of the loss distributions corresponding to its existing and new territories, respectively, which is equivalent to the above-mentioned independence assumption.

The ordinary scenario, due to the abundant data, may be subject to negligible ambiguity and hence it is safe to trust the empirical distribution when making inferences about the corresponding risk Y. However, the ambiguous scenario may suffer from scarcity or low quality of the data, and for this reason, the corresponding risk Z is subject to significant ambiguity. In such a situation, apparently it becomes unwise to still apply the worst-case treatment to the generic risk in the usual way. Instead, we propose to simply treat the risk Y in the ordinary scenario as subject to no ambiguity and fully in accordance with the reference distribution and to apply the worst-case treatment only to the ambiguous scenario where ambiguity is significant.

Note that the ambiguity in our consideration is twofold, adhering to both the ambiguous scenario indicator I and the corresponding risk Z. Thus, we need to first construct an ambiguity set of possible

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

probability distributions of the pair (I, Z), denoted by (F_I, F_Z), and then consider the worst case in this ambiguity set. With a slight abuse of notation, we still denote by B this ambiguity set. The problem adopted to the current context is reformulated as

$$\sup_{F_X:\ (F_I,F_Z)\in B} E\left[h(X)\right]. \tag{1.2}$$

Following a recent research trend in distributionally robust optimization, we construct the ambiguity set B in terms of the Wasserstein distance and centered at the empirical distribution of the pair (I, Z).

In order to link the problem (1.2) to an established strong duality, we disentangle the total ambiguity described by the set B in (1.2) along the ambiguous scenario indicator I and the corresponding risk Z. Based on this disentanglement, we convert the two-fold optimization problem into two one-fold problems.

For a power loss function $h(x) = x^p$ with $p \geq 1$, we apply the established strong duality to solve the worst-case estimation problem of $E\left[Z^p\right]$ and eventually obtain the worst-case moment $E\left[X^p\right]$ in (1.2); see Theorem 3.1. We use a power loss function here for simplicity, but extensions to other loss functions are possible; see Remark 3.1. Based on Theorem 3.1, we conduct numerical studies to illustrate that the consideration of partial ambiguity is indeed a promising solution to alleviate the over-conservativeness issue.

To summarize, we study the worst-case moments of a risk whose distribution is subject to ambiguity. To alleviate the over-conservativeness issue, we trace the realization of the risk from two scenarios among which one is an ordinary scenario subject to negligible ambiguity, so that we only need to apply the worst-case treatment to the other ambiguous scenario. We construct a Wasserstein ball to describe the two-fold ambiguity in both the ambiguous scenario indicator and the risk in the corresponding ambiguous scenario, and then we derive the worst-case moment estimates both analytically and numerically. Our numerical studies illustrate that the consideration of partial ambiguity, which is our main contribution of this paper, indeed greatly alleviates the over-conservativeness issue.

We would like to point out that many insurance and financial products, such as traditional reinsurance, insurance-linked securities, catastrophe bonds, industry loss warranties, contingent convertible bonds, and credit default swaps, are designed mainly to hedge risks in extreme scenarios where issues such as scarcity and low quality of data are often critical. Our consideration of partial ambiguity becomes particularly relevant to such products.

The rest of the paper is organized as follows: Section 1 ends with a brief literature review; Section 2 formulates our worst-case estimation problem under partial ambiguity; Section 3 presents our main results; Section 4 conducts numerical studies to illustrate the benefit of our consideration of partial ambiguity; Section 5 makes some concluding remarks; finally, the Appendix collects all proofs.

1.1. A brief literature review

The worst-case treatment as formulated in (1.1) often appears in the field of decision-making under ambiguity. Typically, the decision-maker solves a minimax problem to find decisions that minimize the worst-case risk. See, for example, Scarf (1958), Popescu (2007), and Delage and Ye (2010), among others. The worst-case treatment is popular for applications in insurance, finance, and risk management. See, for example, Artzner *et al*. (1999), Hürlimann (2002), Embrechts *et al*. (2005), Kaas *et al*. (2009), Wang *et al*. (2013), Bernard *et al*. (2017, 2020), Li *et al*. (2018), and Cornilly *et al*. (2018), among many others. Our work naturally follows these strands of research.

Many of the works cited above employ a characteristics-based approach to constructing an ambiguity set; that is, the set over which the worst-case treatment takes place is constructed based on certain distributional characteristics such as marginal distributions, moments, and a dependence structure. A disadvantage of characteristics-based approaches is that they often yield overly conservative estimates because the constructed ambiguity set contains all kinds of distributions including pathological ones as long as they possess the required characteristics. For this reason, a majority of recent works have

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

shifted towards divergence-based approaches following which an ambiguity set is constructed to be a ball centered at a reference distribution with a radius specified in terms of a certain statistical divergence.

Ben-Tal *et al*. (2013) advocate to use ϕ-divergences to construct ambiguity balls for which the authors come up with a natural interpretation in terms of statistical confidence sets. See also Glasserman and Xu (2014), Bayraksan and Love (2015), Gupta (2019), Lam (2019), and Rahimian *et al*. (2019). An advantage, among others, of using ϕ-divergences is that many of them enable to conduct goodness-of-fit tests. However, some popular ϕ-divergences such as the Kullback–Leibler divergence and the chi-square divergence have an obvious problem that all distributions in the ball have to be absolutely continuous with respect to the reference distribution, which implies that they have to be discrete when, as usual, an empirical distribution is selected as the reference distribution. See Wozabal (2012) and Gao and Kleywegt (2022) for related discussions.

To avoid this absolute continuity restriction on the ambiguity ball, one may turn to the Wasserstein distance, which has attracted a great deal of attention from scholars who follow divergence-based approaches. Among many works along this direction, we name several that are closely related to our current study: Wozabal (2012, 2014), Mohajerin Esfahani and Kuhn (2018), Blanchet and Murthy (2019), Pesenti and Jaimungal (2020), and Gao and Kleywegt (2022). Using the Wasserstein distance has clear advantages over other ϕ-divergences. With the absolute continuity restriction lifted, the constructed ambiguity set allows for both discrete and continuous distributions. Moreover, due to its non-parametric nature, the resulting worst-case estimation is immune to distribution types and can largely avoid the model misspecification issue.

Note that (1.1) is a semi-infinite optimization problem, and so is (1.2), which cannot be solved computationally in a straightforward way. It is necessary to further convert such an optimization problem into a finite dimensional problem. In this regard, under the Wasserstein distance, Gao and Kleywegt (2022) establish a strong duality result, which plays a key role in solving our worst-case estimation problem (1.2). Independently, Blanchet and Murthy (2019) obtain a similar result, which from various aspects is more general than that of Gao and Kleywegt (2022). Related results under simpler settings can be found in Mohajerin Esfahani and Kuhn (2018) and Zhao and Guan (2018).

In order to alleviate the over-conservativeness issue, we propose a scenario analysis following which we focus on ambiguous scenarios only, leading to a partial ambiguity issue. In behavioral finance, Payzan-LeNestour and Woodford (2022) raise an outlier blindness issue that "*people are hampered in their perception of outcomes that they expect to seldom encounter, and view the magnitude of such outcomes as less extreme than they really are*." This also motivates our consideration of partial ambiguity.

We notice that the exact phrase of partial ambiguity has appeared in the study of Ellsberg's two-urn paradox; see, for example, Chew *et al*. (2017) and Berger (2021). In their context, partial ambiguity refers to the situation that some prior knowledge limits the possible compositions in an urn of 100 red and black balls, while in our context it refers to the situation that abundant information about the risk in ordinary scenarios allows us to pin down ambiguity to the other scenarios. In the two contexts, partial ambiguity has essentially the same meaning though presented quite differently.

Recently, Lam and Mottet (2017) evaluate performance measures over a tail region where there are very limited data or even no data. They adopt a new approach based on the geometric premise of tail convexity, a feature shared by virtually all practically useful models, and then consider all possible tail densities. Although their work and ours share a similar flavor of pinning down ambiguity due to data scarcity, essential difference exists in that they conduct optimization over all possible tail densities fulfilling several postulated distributional constraints (thus, they follow a characteristics-based approach), while we resort to a divergence-based approach.

Often being a critical issue in insurance practice, model uncertainty has received increasing attention in the insurance literature. The following works, among others, have investigated or touched this issue in various insurance contexts: Cairns (2000), Chen and Su (2009), Peters *et al*. (2009), Zhao and Zhu (2011), Landsman and Tsanakas (2012), Robert and Therond (2014), Liu and Wang (2017), Fujii *et al*.

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

(2017), and Jiang *et al*. (2020). More applications to practical problems in insurance have been proposed during recent years but are still sporadic; see Wozabal (2014), Pflug *et al*. (2017), Lam and Mottet (2017), Asimit *et al*. (2017, 2019), Ghossoub (2019a, 2019b), Blanchet *et al*. (2019), Birghila and Pflug (2019), and Birghila *et al*. (2020), among others. We hope that our work can draw attention from actuaries and financial analysts to incorporate the consideration of partial ambiguity when addressing the ambiguity issue in practical problems.

2. Formulation of the problem

2.1. Notational conventions

For a general probability space $(\Omega, \mathcal{F}, P)$, denote by $L^p(\Omega, \mathcal{F}, P)$ the Lebesgue space with exponent $p \geq 1$, namely, the space of random variables with finite pth moments. For a measurable set $A \subset \mathbb{R}^d$, denote by $\mathcal{P} \Leftarrow A)$ the set of probability distributions supported on A. Denote by $\|\cdot\|_p$ the usual p norm in $\mathbb{R}^d$ and by $\|\cdot\|_{L^p}$ the L^p norm in $L^p(\Omega, \mathcal{F}, P)$. Thus, for a vector $\mathbf{x} \in \mathbb{R}^d$ we have $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{\frac{1}{p}}$ while for a random variable $X \in L^p(\Omega, \mathcal{F}, P)$ we have $\|X\|_{L^p} = (E\,|X|^p)^{\frac{1}{p}}$. For $x, y \in \mathbb{R}$, we write $x \vee y = \max\{x, y\}$, $x \wedge y = \min\{x, y\}$, $x_+ = x \vee 0$, and denote by δ_x a Dirac measure that assigns mass 1 at point x. For a random variable ξ, we denote its distribution by F_ξ and its quantile function at level $q \in [0, 1]$ by

$$F_\xi^{-1}(q) = \inf\left\{x \in \mathbb{R} : F_\xi(x) \geq q\right\} = \sup\left\{x \in \mathbb{R} : F_\xi(x) < q\right\},$$

where $\inf \emptyset$ is the right endpoint of F_ξ and $\sup \emptyset$ is the left endpoint of F_ξ. Throughout the paper, we will tacitly follow these conventions without extra explanations.

2.2. A general formulation

Consider the estimation of $E\,[h(X)]$, where X is a non-negative risk variable defined on the probability space $(\Omega, \mathcal{F}, P)$ and $h(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a deterministic measurable loss function such that the expectation involved is finite.

Suppose that, as described before, X is the result of the aggregation of the risks in an ordinary scenario and an ambiguous scenario, labelled as 0 and 1, respectively. Introduce a Bernoulli variable I, called scenario indicator, with distribution

$$F_I = (1 - q)\delta_0 + q\delta_1,$$

where $q = P(I = 1) \in [0, 1]$ denotes the probability that scenario 1 is picked. The risk in scenario 0, denoted by Y, is subject to negligible ambiguity and for this reason we will trust its reference distribution. However, the risk in scenario 1, denoted by Z, is subject to significant ambiguity so that the worst-case treatment needs to be applied when estimating the risk. In this way, the generic risk X takes the stochastic representation

$$X = (1 - I)Y + IZ. \tag{2.1}$$

Assuming independence between I and (Y, Z), the expectation $E\,[h(X)]$ is decomposed into

$$E\,[h(X)] = P(I = 0)E\,[h(Y)] + P(I = 1)E\,[h(Z)]\,. \tag{2.2}$$

Remark 2.1 *It is important to keep in mind the following:*

(a) *To derive a worst-case estimate for $E\,[h(X)]$ starting from (2.2), it suffices to take into account the ambiguity adheres to F_I and F_Z since we fully trust the reference distribution of Y. This leads to a two-fold ambiguity issue, which will be at the core of our analysis.*

(b) *The scenario indicator I and the risk Z in the ambiguous scenario are in different scales and play distinct roles in the worst-case estimation, leading to an asymmetry issue.*

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

To address the asymmetry issue between I and Z in Remark 2.1(b), we scale down the risk Z by a constant $s > 0$; that is, we instead consider the scaled risk $\tilde{Z} = \frac{Z}{s}$. The decomposition (2.2) now becomes

$$E\left[h(X)\right] = (1-q)E\left[h(Y)\right] + qE\left[h\left(s\tilde{Z}\right)\right].$$

For the moment, the scale parameter s is simply treated as exogenous when tackling the twofold ambiguity issue. Later in Subsection 4.2, we will establish a mechanism to endogenize the scale parameter s so that in the worst-case estimation of $E\left[h(X)\right]$ the two folds play comparable roles in a certain sense.

Further, we introduce a vector $\mathbf{V} = \left(I, \tilde{Z}\right)$ whose distribution is

$$F_{\mathbf{V}} = F_I \times F_{\tilde{Z}} \in \mathcal{P}\left(\{0,1\} \times \mathbb{R}_+\right).$$

Following the divergence-based approach, we consider the true distribution $F_{\mathbf{V}}$ to be within a ball centered at its empirical version

$$\hat{F}_I \times \hat{F}_{\tilde{Z}} \quad \text{with} \quad \begin{cases} \hat{F}_I = (1-q_n)\delta_0 + q_n\delta_1, \\ \hat{F}_{\tilde{Z}} = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{z}_i}, \end{cases}$$

where $n = n_0 + N$ is the total number of observations counting both $\{y_1, \ldots, y_{n_0}\}$ of Y from the ordinary scenario 0 and $\{z_1, \ldots, z_N\}$ of Z from the ambiguous scenario 1, $\tilde{z}_i = \frac{z_i}{s}$ for $i = 1, \ldots, N$ denote the scaled realizations of the risk Z, and $q_n = \frac{N}{n}$ is the empirical estimate for q. This ball for $F_{\mathbf{V}}$ will be described by $B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right)$, where $r > 0$ represents the radius.

To conclude, we need to solve the worst-case estimation problem

$$\sup_{F_X:\ F_I \times F_{\tilde{Z}} \in B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right)} E\left[h(X)\right] = \sup_{F_I \times F_{\tilde{Z}} \in B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right)} \left\{(1-q)E\left[h(Y)\right] + qE\left[h\left(s\tilde{Z}\right)\right]\right\}. \tag{2.3}$$

2.3. Under the Wasserstein distance

To construct the ambiguity ball in (1.1), we use the following Wasserstein distance of order $p \geq 1$. For two multivariate distributions F_1 and F_2, their Wasserstein distance is defined to be

$$W(F_1, F_2) = \inf_{\Pi:\Pi_{\mathbf{V}_1} = F_1,\ \Pi_{\mathbf{V}_2} = F_2} \left(E_\Pi\left[d(\mathbf{V}_1, \mathbf{V}_2)^p\right]\right)^{\frac{1}{p}}, \tag{2.4}$$

where $\mathbf{V}_1$ and $\mathbf{V}_2$ are two vectors distributed by F_1 and F_2, respectively, $d(\cdot,\cdot)$ denotes a certain distance, Π denotes a joint distribution of $(\mathbf{V}_1, \mathbf{V}_2)$ with marginal distributions $\Pi_{\mathbf{V}_1} = F_1$ and $\Pi_{\mathbf{V}_2} = F_2$.

The definition of the Wasserstein distance (2.4) originates from optimal transportation problems. According to Villani (2009), each joint distribution Π is interpreted as a plan of transporting goods between producers and consumers, whose spatial distributions are described by F_1 and F_2, respectively, the quantity $d(\mathbf{v}_1, \mathbf{v}_2)^p$ is interpreted as the cost for transporting one unit of goods from $\mathbf{v}_1$ to $\mathbf{v}_2$, and consequently the quantity $W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p$ is interpreted as the optimal total transportation cost.

In most existing studies, the distance $d(\cdot,\cdot)$ is specified as the p norm; that is, for two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of the same dimension,

$$d(\mathbf{v}_1, \mathbf{v}_2) = \left\|\mathbf{v}_1 - \mathbf{v}_2\right\|_p.$$

We will simply follow this specification. Nevertheless, we would like to point out that depending on the situation it may become crucial to employ other distances to address some practical considerations. Then for the one-dimensional case, it is well known that the Wasserstein distance between two distributions $F_1, F_2 \in \mathcal{P}(\mathbb{R})$ takes an explicit expression in terms of their quantile functions:

$$W(F_1, F_2) = \left(\int_0^1 \left|F_1^{-1}(u) - F_2^{-1}(u)\right|^p du\right)^{\frac{1}{p}}; \tag{2.5}$$

see, for example, Panaretos and Zemel (2019).

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Corresponding to our setting in Subsection 2.2, $\mathbf{V}_i = \left(I_i, \tilde{Z}_i\right)$ and $F_{\mathbf{V}_i} = F_{I_i} \times F_{\tilde{Z}_i}$ for each $i = 1, 2$, Π denotes a joint distribution of $(\mathbf{V}_1, \mathbf{V}_2) = \left(I_1, \tilde{Z}_1, I_2, \tilde{Z}_2\right)$, and $d(\cdot, \cdot)$ is specified as the p norm. To simplify the notation, we introduce the following sets:

$$\begin{aligned} S &= \left\{\Pi \in \mathcal{P}\left((\{0, 1\} \times \mathbb{R}_+)^2\right) : \Pi_{\mathbf{V}_1} = F_{\mathbf{V}_1}, \Pi_{\mathbf{V}_2} = F_{\mathbf{V}_2}\right\}, \\ S_I &= \left\{\mu \in \mathcal{P}\left(\{0, 1\}^2\right) : \mu_{I_1} = F_{I_1}, \mu_{I_2} = F_{I_2}\right\}, \\ S_{\tilde{Z}} &= \left\{\nu \in \mathcal{P}\left(\mathbb{R}_+^2\right) : \nu_{\tilde{Z}_1} = F_{\tilde{Z}_1}, \nu_{\tilde{Z}_2} = F_{\tilde{Z}_2}\right\}. \end{aligned}$$

Then $S = S_I \times S_{\tilde{Z}}$ and

$$W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2}) = \inf_{\Pi \in S} \left(E_\Pi\left[\|\mathbf{V}_1 - \mathbf{V}_2\|_p^p\right]\right)^{\frac{1}{p}}, \tag{2.6}$$

so that

$$W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p = \inf_{\Pi \in S} \left\{E_\Pi\left[|I_1 - I_2|^p\right] + E_\Pi\left[\left|\tilde{Z}_1 - \tilde{Z}_2\right|^p\right]\right\}.$$

Intuitively, we can take $\inf_{\Pi \in S}$ onto the two terms $E_\Pi\left[|I_1 - I_2|^p\right]$ and $E_\Pi\left[\left|\tilde{Z}_1 - \tilde{Z}_2\right|^p\right]$ separately, giving

$$W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p = W(F_{I_1}, F_{I_2})^p + W\left(F_{\tilde{Z}_1}, F_{\tilde{Z}_2}\right)^p.$$

It turns out that this intuition is correct, as proved by Lemma A.1 in a general context. Moreover, by (2.5), we can easily verify the following two identities:

- With $q_i = P(I_i = 1)$ for $i = 1, 2$,
$$W(F_{I_1}, F_{I_2})^p = \int_0^1 \left|F_{I_1}^{-1}(u) - F_{I_2}^{-1}(u)\right|^p du = |q_1 - q_2|;$$
- With $s > 0$ and $\tilde{Z}_i = \frac{Z_i}{s}$ for $i = 1, 2$,
$$W\left(F_{\tilde{Z}_1}, F_{\tilde{Z}_2}\right)^p = \frac{1}{s^p} W(F_{Z_1}, F_{Z_2})^p.$$

It follows that

$$W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p = |q_1 - q_2| + \frac{1}{s^p} W(F_{Z_1}, F_{Z_2})^p.$$

Formally, we construct the ambiguity ball in our worst-case estimation problem (2.3) as

$$\begin{aligned} B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right) &= \left\{F_I \times F_{\tilde{Z}} : W\left(F_I \times F_{\tilde{Z}}, \hat{F}_I \times \hat{F}_{\tilde{Z}}\right) \le r\right\} \\ &= \left\{(q, F_Z) : q \in [0, 1], |q - q_n| + \frac{1}{s^p} W\left(F_Z, \hat{F}_Z\right)^p \le r^p\right\}. \end{aligned}$$

Thus, our worst-case estimation problem (2.3) can be reformulated as

$$\begin{aligned} &\sup_{F_X:\ F_I \times F_{\tilde{Z}} \in B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right)} E\left[h(X)\right] \\ &= \sup_{(q, F_Z):\ q \in [0,1], |q - q_n| + \frac{1}{s^p} W\left(F_Z, \hat{F}_Z\right)^p \le r^p} \left\{(1 - q)E\left[h(Y)\right] + qE\left[h(Z)\right]\right\}. \end{aligned} \tag{2.7}$$

Remark 2.2 *The next task is to optimally allocate the given amount of ambiguity between the probability q and the distribution F_Z. This involves a trade-off issue given that the total amount of ambiguity as reflected by the radius r is fixed. For example, increasing q from its empirical estimate q_n influences the value of $(1 - q)E\left[h(Y)\right] + qE\left[h(Z)\right]$ in an intricate way: Directly, it decreases the first term and*

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

increases the second term, while indirectly, the deviation of q from q_n consumes part of the ambiguity and thus reduces the range for F_Z in optimizing $E\left[h(Z)\right]$. Also keep in mind two extreme cases in which the ambiguity is maximally allocated to the probability q and the distribution F_Z, respectively. In each case, the given amount of ambiguity is likely used up by one argument, causing the other argument to stick to its empirical estimate. Moreover, we note that, to address the asymmetry issue in Remark 2.1(b), we have used a scale parameter s, taken as exogenous for the moment, in constructing the worst-case estimation problem in (2.7).

Further, we factorize the worst-case estimation problem (2.7) into two layers as

$$\begin{aligned}
&\sup_{F_X:\ F_I\times F_{\tilde{Z}}\in B_r\left(\hat{F}_I\times\hat{F}_{\tilde{Z}}\right)} E\left[h(X)\right]\\
&= \sup_{q\in[0,1]:\ |q-q_n|\le r^p}\left\{(1-q)E\left[h(Y)\right]+q\sup_{F_Z:\ W\left(F_Z,\hat{F}_Z\right)^p\le s^p(r^p-|q-q_n|)} E\left[h(Z)\right]\right\}\\
&= \sup_{q\in[q_n^-,q_n^+]}\left\{(1-q)E\left[h(Y)\right]+q\sup_{F_Z\in B_\epsilon\left(\hat{F}_Z\right)} E\left[h(Z)\right]\right\}, \qquad (2.8)
\end{aligned}$$

where we have used the following notation:

$$[q_n^-,q_n^+]=[q_n-r^p,q_n+r^p]\cap[0,1], \qquad (2.9)$$

$$\epsilon=\epsilon(q)=s\left(r^p-|q-q_n|\right)^{\frac{1}{p}}\in[0,rs]. \qquad (2.10)$$

This way, we have successfully disentangled the ambiguity along the two folds q and F_Z and hence converted the worst-case estimation problem to a two-stage optimization problem. We need to first solve the inner optimization, which is a standard worst-case estimation problem. Then it will become straightforward to solve the outer optimization in (2.8), eventually completing the worst-case problem (2.3).

Remark 2.3 *Observe the range (2.9) for q, which comes from the ingenuous belief of $q\in[0,1]$. In practice, however, we may have a prior belief in q reflected by a restricted range $q\in[\underline{q},\bar{q}]\subseteq[0,1]$, which can be, for example, a confidence interval based on a rich dataset or an expert opinion. In this case, we may utilize this information to reduce the range for q in the hope to further help alleviate the over-conservativeness issue. Formally, we have the worst-case estimation problem*

$$\sup_{F_X:\ F_I\times F_{\tilde{Z}}\in\tilde{B}_r\left(\hat{F}_I\times\hat{F}_{\tilde{Z}}\right)} E\left[h(X)\right]$$

over the modified ball

$$\tilde{B}_r\left(\hat{F}_I\times\hat{F}_{\tilde{Z}}\right)=B_r\left(\hat{F}_I\times\hat{F}_{\tilde{Z}}\right)\cap\left\{q\in[\underline{q},\bar{q}]\right\}.$$

Then by going along the same lines as (2.8), we can still convert the worst-case estimation problem to a two-stage optimization problem in which the outer optimization is over

$$q\in[q_n-r^p,q_n+r^p]\cap[\underline{q},\bar{q}].$$

Thus, it becomes straightforward to address this modification to potentially refine the study, but we will omit it to keep the paper short.

3. The main result

Many existing studies in the literature consider a concave loss function $h(\cdot)$. However, the case with a convex loss function is arguably more relevant to some applications. A potential challenge with a convex loss function is that it grows too fast, causing the worst-case estimate to easily explode. For simplicity, we specify $h(\cdot)$ as a power loss function

$$h(x) = x^p, \qquad p \geq 1, \tag{3.1}$$

so that our work amounts to providing the worst-case pth moment of a generic risk X.

For ease of reference, we recollect here some essential steps in the worst-case estimation of $E\left[X^p\right]$. Following the traditional divergence-based approaches, the worst-case estimation of $E\left[X^p\right]$ is constructed as

$$\sup_{F_X \in B_r\left(\hat{F}_X\right)} E\left[X^p\right], \qquad p \geq 1, \tag{3.2}$$

which is conducted over the Wasserstein ball $B_r\left(\hat{F}_X\right)$, $r > 0$, centered at the empirical distribution $\hat{F}_X$ based on the whole dataset of the generic risk X. When constructing the Wasserstein ball, following the mainstream in related research we choose the order p of the Wasserstein distance (2.6) to be the same as the order p of the power function (3.1). Then, the corresponding worst-case estimation problem (3.2) is bounded according to Lemma A.2.

In our scenario analysis, the generic risk X, as described by (2.1), results from the aggregation of the risks in an ordinary scenario 0 and an ambiguous scenario 1, and we aim to estimate

$$E\left[X^p\right] = (1-q)E\left[Y^p\right] + qE\left[\left(s\tilde{Z}\right)^p\right],$$

where $q = P(I=1)$, $\tilde{Z} = \frac{Z}{s}$, and $s > 0$ is a scale parameter. Following (2.8), our worst-case estimation problem becomes

$$\sup_{F_X:\ F_I \times F_{\tilde{Z}} \in B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right)} E\left[X^p\right] \tag{3.3}$$

$$= \sup_{q \in \left[q_n^-, q_n^+\right]} \left\{(1-q)E\left[Y^p\right] + q \sup_{F_Z \in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right]\right\}, \tag{3.4}$$

with $\epsilon = s\left(r^p - |q - q_n|\right)^{\frac{1}{p}}$ given in (2.10).

As the first step to solve (3.3), we work on the inner optimization in (3.4),

$$\sup_{F_Z \in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right], \tag{3.5}$$

and the result is shown in Proposition 3.1 below, the proof of which is postponed to the Appendix. To avoid triviality, we only consider a proper sample $\{z_1, \ldots, z_N\}$ of Z in the sense that not all sample points are 0. Nevertheless, in case all sample points $\{z_1, \ldots, z_N\}$ of Z are 0, the aimed results are still valid: In this case, $\hat{F}_Z$ is degenerate at 0, $\sup_{F_Z \in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right] = \epsilon^p$ is attained at a distribution F_Z^* degenerate at ϵ, and $W\left(F_Z^*, \hat{F}_Z\right) = \epsilon$.

Proposition 3.1 *Consider the worst-case estimation problem (3.5) in which $p \geq 1$, $\epsilon > 0$, and the Wasserstein distance is specified as (2.6) with the same order p. A proper sample $\{z_1, \ldots, z_N\}$ of Z is given. Then*

$$\sup_{F_Z \in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right] = \left(\epsilon + \|Z_N\|_{L^p}\right)^p, \tag{3.6}$$

where Z_N denotes a random variable uniformly distributed over the sample $\{z_1, \ldots, z_N\}$ and thus $\|Z_N\|_{L^p} = \left(\frac{1}{N}\sum_{i=1}^N z_i^p\right)^{\frac{1}{p}}$.

Clearly, the supremum (3.6) is attained at

$$F_Z^* = \frac{1}{N}\sum_{i=1}^{N}\delta_{z_i^*},$$

where $z_i^* = \left(1 + \frac{\epsilon}{\|Z_N\|_{L^p}}\right) z_i$ for $i = 1, \ldots, N$. Moreover, it is easy to check that $W\left(F_Z^*, \hat{F}_Z\right) = \epsilon$, which means that the worst-case distribution F_Z^* consumes all the ambiguity ϵ and lies on the boundary of $B_\epsilon\left(\hat{F}_Z\right)$. To see this, recall the alternative expression (2.5) of the Wasserstein distance for the univariate case. We have

$$\begin{aligned} W\left(F_Z^*, \hat{F}_Z\right) &= \left(\int_0^1 \left|F_Z^{*-1}(u) - \hat{F}_Z^{-1}(u)\right|^p du\right)^{\frac{1}{p}} \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}\left|z_i^* - z_i\right|^p\right)^{\frac{1}{p}} \\ &= \epsilon. \end{aligned}$$

Remark 3.1 *Although we focus on a non-negative risk variable X and a power loss function, our method is applicable to a real-valued risk variable X and a general loss function h:$\mathbb{R} \to \mathbb{R}$ as long as $h(\cdot)$ is upper semi-continuous and satisfies $h(x) = O(|x|^p)$ for some $p \geq 1$ as $|x| \to \infty$. Indeed, we can still arrive at the two-stage optimization (2.8) by disentangling the ambiguity along the two folds. Furthermore, following the proof of Proposition 3.1, we can always convert the inner optimization in (2.8) to a one-dimensional convex optimization problem to make it numerically tractable but an explicit expression may be available only for certain special loss functions. One such example is the loss function*

$$h(x) = (x - d)_+^p, \qquad d > 0,\ p \geq 1,$$

for which $E[h(X)]$ corresponds to the pth moment of the insurance payoff in a stop-loss contract. For this example, we can achieve an explicit expression as

$$\sup_{F_Z \in B_\epsilon\left(\hat{F}_Z\right)} E\left[(Z - d)_+^p\right] = \left(\epsilon + \|(Z_N - d)_+\|_{L^p}\right)^p.$$

However, it does not seem to be easy to establish a unified result for general convex loss functions.

Under the help of Proposition 3.1, by completing the outer optimization in (3.4), we will eventually solve the worst-case estimation problem (3.3). To this end, introduce

$$C_0 = \|Y\|_{L^p} \qquad \text{and} \qquad sC_1 = \|Z_N\|_{L^p},$$

where the distribution of Y is fully in accordance with the empirical distribution from the sample $\{y_1, \ldots, y_{n_0}\}$ in scenario 0. There are two cases:

$$C_0 \leq C_1 \qquad \text{and} \qquad C_0 > C_1.$$

The case $C_0 \leq C_1$ means that, in the L^p norm, the risk realization in the ambiguous scenario 1 tends to be larger than that in the ordinary scenario 0, which may be the case if, for example, the ambiguous scenario 1 is more catastrophic. For this case, we are able to obtain an analytical solution to the worst-case estimation problem (3.3).

Theorem 3.1 *Consider the worst-case estimation problem (3.3). In addition to the conditions in Proposition 3.1, assume that $C_0 \leq C_1$. Then the worst-case pth moment of X is*

$$\sup_{F_X:\ F_I \times F_{\tilde{Z}} \in B_r\left(\hat{F}_I \times \hat{F}_{\tilde{Z}}\right)} E[X^p] = \begin{cases} (1 - q_n)C_0^p + q_n(rs + C_1)^p, & \text{when } r \leq r_*, \\ (1 - q_*)C_0^p + q_*(\epsilon_* + C_1)^p, & \text{when } r > r_*. \end{cases} \tag{3.7}$$

In (3.7):

- $r_* = \bar{r} \vee 0$, *with* $\bar{r}$ *the unique solution* r *to*
$$-C_0^p + (rs + C_1)^{p-1} \left(rs(1 - q_n r^{-p}) + C_1\right) = 0; \tag{3.8}$$

- $q_* = \bar{q}_n \wedge q_n^+$, *with* q_n^+ *defined in (2.9) and with* $\bar{q}_n$ *the unique solution* q *to*
$$-C_0^p + \left(s(r^p + q_n - q)^{\frac{1}{p}} + C_1\right)^p - qs\left(s(r^p + q_n - q)^{\frac{1}{p}} + C_1\right)^{p-1} (r^p + q_n - q)^{\frac{1}{p}} = 0; \tag{3.9}$$

- $\epsilon_* = \epsilon(q_*) = s(r^p + q_n - q_*)^{\frac{1}{p}}$ *as defined in (2.10).*

The proof of Theorem 3.1 is postponed to the Appendix, from which it is easy to see that the supremum of (3.7) is always attainable. Let us observe Theorem 3.1 for the cases $p > 1$ and $p = 1$ separately. We will see that it echoes Remark 2.2 to a certain extent.

For $p > 1$, it is easy to see that $r_* > 0$ as Equation (3.8) yields a unique solution $\bar{r} \in \left(0, q_n^{\frac{1}{p}}\right)$. Actually, the left-hand side of (3.8), as a continuous and increasing function in r, diverges to $-\infty$ as $r \downarrow 0$, while it takes a positive value at $r = q_n^{\frac{1}{p}}$ since $C_0 \le C_1$. The first piece in (3.7) shows that, when r is relatively small such that $0 < r \le r_* = \bar{r}$, the optimization procedure requires that the ambiguity be fully allocated to F_Z to raise $E[Z^p]$, and subsequently the range for q boils down to the singleton $\{q_n\}$. However, when r is relatively large such that $r > r_* = \bar{r}$, the total amount of ambiguity is allocated to both q and F_Z according to the second piece in (3.7). Precisely, part of the ambiguity is allocated to q to shift it from q_n to $q_* \in (q_n, q_n + r^p)$, and the remaining ambiguity, as quantified by $\epsilon_* \in (0, rs)$ after scaling by s, is allocated to F_Z to raise $E[Z^p]$.

For $p = 1$, it is possible that Equation (3.8) yields a negative solution depending on the value of q_n. For this case, $r_* = 0$, and thus only the second piece in (3.7) is relevant, which indicates that the ambiguity is not fully allocated to F_Z. However, it is possible that $q_* = q_n + r$ and hence $\epsilon_* = 0$, which means that the optimization procedure fully allocates the ambiguity to q to shift it from q_n to $q_* = q_n + r$ and subsequently squeezes the range for F_Z to the singleton $\left\{\hat{F}_Z\right\}$.

Remark 3.2 *The other case* $C_0 > C_1$ *means that, in the* L^p *norm, the risk realization in the ambiguous scenario 1 is smaller than that in the ordinary scenario 0, which may be the case when, for example, an insurance company extends with prudence to a new line of business in which losses incurred are subject to more ambiguity but not necessarily larger than in its ordinary business. Unfortunately, the proof of Theorem 3.1 does not cover this case: The auxiliary function* $f_1(q)$ *in (A12) does not exhibit a clear convexity feature on* $[q_n^-, q_n]$*, and hence it becomes troublesome to achieve an analytical solution. Nevertheless, under the help of Proposition 3.1, the worst-case estimation problem (3.3) is numerically tractable and we will instead seek numerical solutions in Section 4.*

4. Numerical studies

This section is devoted to numerical studies to illustrate the benefit of our new approach via (3.3) giving consideration to partial ambiguity compared with the traditional approach via (3.2).

4.1. Synthetic data

To assess the performance of the traditional and new approaches, we will generate synthetic data from a known distribution, and then use this dataset in both approaches but pretend that we do not know the true distribution. Such an idea of synthetic data has often been used to facilitate similar numerical studies; see, for example, Lam and Mottet (2017).

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Precisely, given the distributions F_I, F_Y, and F_Z, the synthetic data of X can be generated in the following steps: First, we generate a uniform random sample of U and define $I = 1_{(U \geq 1-q)}$; Second, if $U < 1-q$, then $I = 0$ and we generate a random sample of Y, while if $U \geq 1-q$, then $I = 1$ and we generate a random sample of Z; Third, putting these into (2.1), we obtain a random sample of X.

In our numerical experiments, we will use the following distributions to model Y and Z:

- We call ξ a folded normal variable with parameters $(m, v^2) \in (-\infty, \infty) \times (0, \infty)$ if $\xi = |\eta|$, where η is normally distributed with mean m and variance v^2, namely, $\eta \sim \mathcal{N}(m, v^2)$;
- We call ξ a beta variable with parameters $(a, b, c) \in (0, \infty)^3$ if ξ has the density

$$g(t) = \frac{1}{B(a,b)c} \left(\frac{t}{c}\right)^{a-1} \left(1 - \frac{t}{c}\right)^{b-1}, \qquad 0 \leq t \leq c,$$

where $B(\cdot, \cdot)$ is the beta function;

- We call ξ a Pareto variable with parameters $(\alpha, \gamma) \in (0, \infty)^2$ if ξ has the density

$$g(t) = \frac{\alpha \gamma^{\alpha}}{t^{\alpha+1}}, \qquad t \geq \gamma.$$

4.2. On the scale parameter

Recall the asymmetry issue mentioned in Remark 2.1(b) between the scenario indicator I and the risk Z in the ambiguous scenario. To address this issue, in Subsection 2.2, we have proposed to introduce a constant $s > 0$ to scale down the risk Z, eventually leading to the worst-case estimation problem (2.7) with the constraint

$$|q - q_n| + \frac{1}{s^p} W\left(F_Z, \hat{F}_Z\right)^p \leq r^p; \tag{4.1}$$

namely, we let q and F_Z compete for the given amount of ambiguity in the worst-case estimation, while the amount of ambiguity allocated to F_Z is further scaled by s. Now we establish a mechanism to endogenize the scale parameter s. To clarify, we do not claim that our mechanism is universally appropriate, but rather we think that, depending on the situation, there should be other potentially better mechanisms to endogenize s.

Recall the two extreme cases mentioned in Remark 2.2 when allocating the ambiguity along the two folds. The first extreme case is to maximally allocate the ambiguity to the probability q, which must be realized at either endpoint of the interval (2.9). There is a subtlety that q may not consume all the given ambiguity, for which case there is still ambiguity left for F_Z according to (2.10). Putting together, we obtain an estimate for $E[X^p]$ as

$$(1 - \tilde{q})C_0^p + \tilde{q}(\tilde{\epsilon} + C_1)^p, \tag{4.2}$$

where $\tilde{q}$ is either q_n^+ or q_n^-, whichever yields a larger value of (4.2), and where $\tilde{\epsilon}$ is defined by

$$\tilde{\epsilon} = \epsilon(\tilde{q}) = s\left(r^p - |\tilde{q} - q_n|\right)^{\frac{1}{p}} \tag{4.3}$$

according to (2.10). The second extreme case is to maximally allocate the ambiguity described by (4.1) to the distribution F_Z to raise the pth moment of Z. As F_Z can always consume all the given ambiguity, which raises the pth moment of Z to $(rs + C_1)^p$ by Proposition 3.1, there is no ambiguity left for q. This gives another estimate for $E[X^p]$ as

$$(1 - q_n)C_0^p + q_n(rs + C_1)^p. \tag{4.4}$$

Our mechanism for determining the scale parameter s is based on the reasoning that, in competing for the given ambiguity to optimize $E[X^p]$, the scenario indicator I and the risk Z in the ambiguous scenario should have equal power. Quantitatively, we interpret this as that the two extreme cases yield

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

the same estimate for $E\left[X^p\right]$. Thus, by equating (4.2) and (4.4), we arrive at the equation

$$(1-\tilde{q})C_0^p+\tilde{q}(\tilde{\epsilon}+C_1)^p=(1-q_n)C_0^p+q_n(rs+C_1)^p, \tag{4.5}$$

which we use to endogenize s.

Once the dataset is available and the radius r is given, the values of C_0, C_1, and q_n are known, and we can decide $\tilde{q}$ and $\tilde{\epsilon}$ by comparing the values of (4.2) at $\tilde{q}=q_n^+$ and $\tilde{q}=q_n^-$. Then it becomes straightforward to check the existence and uniqueness of the solution s to (4.5). We remark that, for most practical situations where the radius r is modest, q moving from q_n to $\tilde{q}$ is able to consume all the ambiguity, leaving no ambiguity to F_Z. More precisely, if $[q_n-r^p, q_n+r^p]\subset[0,1]$, then $[q_n^-, q_n^+]=[q_n-r^p, q_n+r^p]$ by (2.9), and subsequently, for either $\tilde{q}=q_n^+$ or $\tilde{q}=q_n^-$, we have $\tilde{\epsilon}=0$ by (4.3). In such a situation, Equation (4.5) is simplified to

$$(1-\tilde{q})C_0^p+\tilde{q}C_1^p=(1-q_n)C_0^p+q_n(rs+C_1)^p,$$

which gives a closed-form solution

$$s=\frac{\left((q_n-\tilde{q})C_0^p+\tilde{q}C_1^p\right)^{\frac{1}{p}}-C_1q_n^{\frac{1}{p}}}{rq_n^{\frac{1}{p}}}.$$

4.3. Bootstrapping

To produce the worst-case estimate for $E\left[X^p\right]$ with a desired coverage probability, a key step is to determine an appropriate radius r for the Wasserstein ball. There have already been a few theoretical studies under various distributional assumptions such as light tails, but the corresponding selection of the radius either is too conservative or involves unknown constants; see, for example, Pflug and Wozabal (2007), Fournier and Guillin (2015), and Zhao and Guan (2018), among others. Most of those theoretical studies by far are not immediately applicable in practice, and thus scholars usually resort to numerical approaches such as bootstrapping to calibrate the Wasserstein radius r; see, for example, Mohajerin Esfahani and Kuhn (2018) and Kannan *et al.* (2020) for related discussions.

In our numerical studies, we will prudently calibrate the radius r using bootstrapping. For a given dataset of size n, we first construct resamplings from it. Each resampling yields a training dataset and a validation dataset. Suppose that we already have k resamplings. Given a radius r for the worst-case estimation, we process these k resamplings to examine if this level of r is high enough to guarantee the desired coverage probability for $E\left[X^p\right]$. Then we identify the smallest radius r such that the coverage probability is no less than the desired coverage probability.

Formally, the procedure consists of the following steps:

- For each resampling, with the reference distribution generated from the training dataset and the radius r, we produce a worst-case estimate for $E\left[X^p\right]$.
- With the reference distribution generated from the corresponding validation dataset, we have a direct estimate for $E\left[X^p\right]$.
- Repeat these steps for k resamplings and count the frequency that the worst-case estimate from the training dataset is no less than the direct estimate from the validation dataset.
- We regard this frequency as the coverage probability of the corresponding approach with the radius r. We gradually raise r if the obtained coverage probability is lower than the desired, or gradually reduce r otherwise.
- We redo the steps above for T times.

Following the procedure, the radius r is prudently calibrated based on the available data.

Now we show how to construct the k resamplings in the traditional and new approaches, respectively. In the traditional approach as formulated by (3.2), the generic risk X is subject to ambiguity. To construct a resampling, we simply sample with replacement the total n data points to construct the training dataset,

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

and the validation dataset comprises the rest, roughly

$$\left(1-\frac{1}{n}\right)^n \approx \frac{1}{e}$$

of the original data points that are absent from the training dataset; see Subsection 7.2 of Mohajerin Esfahani and Kuhn (2018). Moreover, when processing this resampling, the reference distributions generated from the training dataset and the validation dataset are selected to be the empirical distributions.

In our new approach as formulated by (3.3), the moment $E[X^p]$ is now decided by (F_I, F_Y, F_Z) among which F_I and F_Z are subject to ambiguity. To reflect this two-fold ambiguity, we construct the training dataset and the validation dataset in each resampling as follows:

- We keep the n_0 data points of Y in scenario 0 unchanged and denote by $\hat{F}_Y$ the empirical distribution, which we treat as the true distribution of Y.
- Regarding the ambiguity of F_Z, we sample with replacement the N data points in scenario 1 to construct the training dataset, which gives the empirical distribution $\hat{F}_Z^{\mathrm{tr}}$. The rest of the data points absent from the training dataset form the validation dataset, which gives the empirical distribution $\hat{F}_Z^{\mathrm{va}}$.
- Regarding the ambiguity of F_I, we sample with replacement the $n = n_0 + N$ mixed data to construct the training dataset and estimate $\hat{q}^{\mathrm{tr}}$ to be the frequency of scenario 1, yielding the Bernoulli distribution $\hat{F}_I^{\mathrm{tr}}$ with $\hat{F}_I^{\mathrm{tr}}(\{1\}) = \hat{q}^{\mathrm{tr}}$. The rest of the data points absent from the training dataset form the validation dataset from which we obtain in the same way the frequency $\hat{q}^{\mathrm{va}}$ and the Bernoulli distribution $\hat{F}_I^{\mathrm{va}}$.
- When processing this resampling, the reference distributions generated from the training dataset and the validation dataset are $\left(\hat{F}_I^{\mathrm{tr}}, \hat{F}_Y, \hat{F}_Z^{\mathrm{tr}}\right)$ and $\left(\hat{F}_I^{\mathrm{va}}, \hat{F}_Y, \hat{F}_Z^{\mathrm{va}}\right)$, respectively.

4.4. Comparison with the central limit theorem approach

In this subsection, following a reviewer's request we use the classical central limit theorem (CLT) to provide an upper confidence bound as another conservative estimate for the pth moment. In the next subsection, we will conduct numerical experiments to examine the advantage of our new approach over the CLT approach.

Recall the decomposition (2.2), which, with $h(x)$ specified to the power function x^p, becomes

$$E[X^p] = P(I=0)E[Y^p] + P(I=1)E[Z^p] = (1-E[I])E[Y^p] + E[W],$$

where W denotes the product IZ^p and the last step holds due to the independence between I and Z. With $\pi = E[X^p]$, $q = E[I]$, and $w = E[W]$, we further rewrite the decomposition as

$$\pi = (1-q)E[Y^p] + w. \tag{4.6}$$

To use (4.6) to estimate π, note that $E[Y^p]$ is known since we fully trust the empirical distribution of Y from scenario 0, but q and w, due to the unknown distributions of I and W, need to be estimated. In view of the simple linear relationship in (4.6), it is customary to estimate π by the same expression with q and w replaced by their sample versions

$$q_n = \frac{1}{n}\sum_{i=1}^n I_i \qquad \text{and} \qquad w_n = \frac{1}{n}\sum_{i=1}^n I_i z_i^p,$$

respectively, yielding

$$\pi_n = (1-q_n)E[Y^p] + w_n.$$

In the expression for w_n, whenever $I_i = 0$ there is no observation of Z but the product $I_i z_i^p$ is understood as 0.

By the central limit theorem, we have

$$\sqrt{n}\left(\begin{pmatrix} q_n \\ w_n \end{pmatrix} - \begin{pmatrix} q \\ w \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\right),$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\Sigma = \left(\sigma_{ij}\right)_{2\times 2}$ is the covariance matrix of (I, W). It follows that

$$\sqrt{n}\left(\pi_n - \pi\right) \xrightarrow{d} \mathcal{N}\left(0, \left(E\left[Y^p\right]\right)^2 \sigma_{11} - 2E\left[Y^p\right]\sigma_{12} + \sigma_{22}\right).$$

Therefore, for a desired coverage probability $\beta \in (0, 1)$, namely, the probability that the estimate is no less than the true value, the CLT estimate for π is

$$\pi_n + \frac{\Phi^{-1}\left(\beta\right)}{\sqrt{n}}\sqrt{\left(E\left[Y^p\right]\right)^2 \hat{\sigma}_{11} - 2E\left[Y^p\right]\hat{\sigma}_{12} + \hat{\sigma}_{22}},$$

where $\Phi\left(\cdot\right)$ is the standard normal distribution function and $\hat{\sigma}_{11}$, $\hat{\sigma}_{12}$, and $\hat{\sigma}_{22}$ are the sample versions of σ_{11}, σ_{12}, and σ_{22}.

We would like to point out that the CLT approach and our new approach focus on different topics and applications. The former provides approximations based on the classical limit theory, while the latter is designed to address the ambiguity issue and the resulting worst-case estimates can be viewed as distributionally robust bounds.

4.5. Numerical results

4.5.1. First, consider the case $C_0 \leq C_1$ in Section 3

We specify a Bernoulli distribution for I with $P(I = 1) = 0.1$, a folded normal distribution for Y with parameters $(m, v^2) = (2, 2^2)$, and a Pareto distribution for Z with parameters $(\alpha, \gamma) = (5, 20)$. Experiments are conducted for the first four moments $p = 1, 2, 3, 4$. It is easy to check that these moments of Y are much smaller than those of Z, and thus the condition $C_0 \leq C_1$ can be easily fulfilled in numerical experiments.

We first generate a dataset of size n but pretend that we do not know the true distributions and set the desired coverage probability to be $\beta = 0.95$. Based on this dataset, we compute the upper bound for $E\left[X^p\right]$ using the traditional and the new approaches adopting the worst-case treatment. In each approach, we calibrate the radius r using bootstrapping with $k = 100$ resamplings and then produce the upper bounds by solving the corresponding worst-case estimation problems. Note that for the new approach, we use the mechanism described in Subsection 4.2 to determine the scale parameter s. Moreover, we also apply the CLT approach proposed in Subsection 4.4 to produce the 0.95 confidence upper bound for $E\left[X^p\right]$. Thus, we employ three approaches, which we call the "traditional", the "new", and the "CLT" approaches to distinguish them.

Repeating the above procedure $T = 100$ times, the realized coverage probability in each approach is estimated to be the frequency of the upper bound no less than the true moment $E\left[X^p\right]$. We specify the size of the dataset to be $n = 200$ and $n = 2000$, and the corresponding results are shown in Tables 1 and 2, respectively. We observe the following. The traditional approach always achieves a full coverage probability, which signals the aforementioned over-conservativeness issue, while the CLT approach always achieves coverage probabilities that are significantly lower than the desired level 0.95, which signals significant underestimation. In contrast, our new approach achieves generally satisfactory coverage probabilities. Actually, only for the fourth moment, the coverage probability from the new approach is slightly below the desired, which indicates that a larger dataset is required for estimating higher moments.

Furthermore, we compare the traditional and new approaches that generally guarantee the desired coverage probability. Specifically, we measure the estimation error of each approach by the mean squared

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Table 1. *Performance of the different approaches when $n = 200$, with the desired coverage probability specified to $\beta = 0.95$, I a Bernoulli variable with $P(I = 1) = 0.1$, Y a folded normal variable with parameters $(m, v^2) = (2, 2^2)$, and Z a Pareto variable with parameters $(\alpha, \beta) = (5, 20)$.*

p	Coverage (traditional)	Coverage (new)	Reduction R in MSE (%)	Coverage (CLT)
1	1.00	0.96	68.38	0.79
2	1.00	0.96	69.65	0.89
3	1.00	0.97	72.63	0.85
4	1.00	0.93	82.81	0.76

Table 2. *Performance of the different approaches when $n = 2000$, with the desired coverage probability specified to $\beta = 0.95$, I a Bernoulli variable with $P(I = 1) = 0.1$, Y a folded normal variable with parameters $(m, v^2) = (2, 2^2)$, and Z a Pareto variable with parameters $(\alpha, \beta) = (5, 20)$.*

p	Coverage (traditional)	Coverage (new)	Reduction R in MSE (%)	Coverage (CLT)
1	1.00	0.96	71.87	0.77
2	1.00	0.97	68.54	0.89
3	1.00	0.98	71.27	0.87
4	1.00	0.94	78.90	0.81

error (MSE) of the upper bounds in the $T = 100$ experiments compared with the true moment. To demonstrate the benefit of our new approach, we calculate the reduction ratio in the MSE:

$$R = 1 - \frac{\text{MSE of the new approach}}{\text{MSE of the traditional approach}},$$

and the corresponding results are also shown in Tables 1 and 2. We can observe that the new approach significantly reduces the estimation error compared with the traditional approach.

Note that the implementation of the new approach requires an adequately large dataset. On the one hand, the new approach is based on the assumption of abundant data from the ordinary scenario; on the other hand, it also requires a reasonable amount of data from the ambiguous scenario in order to effectively construct the training dataset and the validation dataset during bootstrapping.

4.5.2. Next, consider the other case $C_0 > C_1$ in Section 3

We specify a Bernoulli distribution for I with $P(I = 1) = 0.1$, a beta distribution for Y with parameters $(a, b, c) = (5, 1, 2)$, and a folded normal distribution for Z with parameters $(m, v^2) = (0, 1^2)$. Experiments are conducted still for the first four moments $p = 1, 2, 3, 4$. It is easy to check that these moments of Y are now larger than those of Z, and thus the condition $C_0 > C_1$ can be fulfilled in numerical experiments.

We conduct the same experiments as before with $k = 100$, $T = 100$, and the desired coverage probability $\beta = 0.95$. The numerical results for the datasets of size $n = 200$ and 2000 are shown in Tables 3 and 4, respectively, from which we have similar observations to those from Tables 1 and 2. In particular, we see that the desired coverage probability 0.95 is achieved only in the traditional and new approaches, between which the new approach, while generally retaining the desired coverage probability, greatly reduces the estimation error for all four moments.

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Table 3. *Performance of the different approaches when $n = 200$, the desired coverage probability specified to $\beta = 0.95$, I a Bernoulli variable with $P(I = 1) = 0.1$, Y a beta variable with parameters $(a, b, c) = (5, 1, 2)$, and Z a folded normal variable with parameters $(m, v^2) = (0, 1^2)$.*

p	Coverage (traditional)	Coverage (new)	Reduction R in MSE (%)	Coverage (CLT)
1	1.00	1.00	53.57	0.85
2	1.00	1.00	66.33	0.79
3	1.00	0.94	79.80	0.77
4	1.00	0.94	85.05	0.75

Table 4. *Performance of the different approaches when $n = 2000$, the desired coverage probability specified to $\beta = 0.95$, I a Bernoulli variable with $P(I = 1) = 0.1$, Y a beta variable with parameters $(a, b, c) = (5, 1, 2)$, and Z a folded normal variable with parameters $(m, v^2) = (0, 1^2)$.*

p	Coverage (traditional)	Coverage (new)	Reduction R in MSE (%)	Coverage (CLT)
1	1.00	1.00	45.47	0.90
2	1.00	1.00	56.22	0.87
3	1.00	0.99	63.79	0.86
4	1.00	0.97	63.05	0.86

5. Concluding remarks

We revisit the worst-case estimation of the moments of a risk X whose distribution is subject to ambiguity. To alleviate the over-conservativeness issue, we consider the risk X as resulting from the aggregation of the risk Y in an ordinary scenario subject to no ambiguity and the risk Z in an ambiguous scenario subject to significant ambiguity. The ambiguity exists in both the scenario indicator and the risk in the ambiguous scenario. We construct a Wasserstein ball to describe this two-fold ambiguity and then we derive worst-case estimates for the moments of X both analytically and numerically.

Several extensions are worthy of pursuit in the future. First, we may consider multiple risk scenarios each of which is subject to a varying extent of ambiguity. With the extents of ambiguity specified, we need to disentangle the total ambiguity along the scenario indicator and the risks from the respective scenarios. Our current work already lends useful hints to this extension. Second, loss functions often involve control variables, which represent, for example, a contract design, an investment strategy, or a risk management rule. Then, we face an additional layer of optimization with respect to the control variables, which, as pointed out by Kuhn *et al.* (2019), may amplify the impact of ambiguity. Third, it is also desirable to consider a general situation involving multiple risk factors rather than multiple risk scenarios. Then we need to deal with the worst-case estimation of the expectation $E\left[h(Y, Z_1, \ldots, Z_d)\right]$ in which h is a general multivariate loss function and $Y, Z_1, \ldots, Z_d$ are risks subject to different extents of ambiguity. Note that in this situation ambiguity also exists in the dependence structure of $(Y, Z_1, \ldots, Z_d)$.

Acknowledgments. The authors would like to thank the three anonymous referees for their constructive comments which have greatly helped us improve the work. The authors feel indebted to Editor Ruodu Wang, who kindly pointed out an asymmetry issue between the scenario indicator I and the risk Z in the decomposition (2.2) and made a nice suggestion on how to address this issue in conducting the worst-case estimation. Qihe Tang acknowledges the financial support by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (DP200101859 and DP220100090), and Yunshen Yang acknowledges the financial support from the Scientia PhD Scholarship and the PhD Teaching Fellowship of UNSW Sydney.

References

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999) Coherent measures of risk. *Mathematical Finance*, **9**(3), 203–228.

Asimit, A.V., Bignozzi, V., Cheung, K.C., Hu, J. and Kim, E.S. (2017) Robust and Pareto optimality of insurance contracts. *European Journal of Operational Research*, **262**(2), 720–732.

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Asimit, A.V., Hu, J. and Xie, Y. (2019) Optimal robust insurance with a finite uncertainty set. *Insurance: Mathematics and Economics*, **87**, 67–81.

Bayraksan, G. and Love, D.K. (2015) Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pp. 1–19. INFORMS.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B. and Rennen, G. (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, **59**(2), 341–357.

Berger, L. (2021) What is partial ambiguity? *Economics & Philosophy*, **38**(2), 206–220.

Bernard, C., Kazzi, R. and Vanduffel, S. (2020) Range Value-at-Risk bounds for unimodal distributions under partial information. *Insurance: Mathematics and Economics*, **94**, 9–24.

Bernard, C., Rüschendorf, L., Vanduffel, S. and Wang, R. (2017) Risk bounds for factor models. *Finance and Stochastics*, **21**(3), 631–659.

Birghila, C., Boonen, T.J. and Ghossoub, M. (2020) Optimal insurance under maxmin expected utility. arXiv preprint arXiv:2010.07383.

Birghila, C. and Pflug, G.C. (2019) Optimal XL-insurance under Wasserstein-type ambiguity. *Insurance: Mathematics and Economics*, **88**, 30–43.

Blanchet, J., Lam, H., Tang, Q. and Yuan, Z. (2019) Robust actuarial risk analysis. *North American Actuarial Journal*, **23**(1), 33–63.

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

Cairns, A.J. (2000) A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, **27**(3), 313–330.

Chen, A. and Su, X. (2009) Knightian uncertainty and insurance regulation decision. *Decisions in Economics and Finance*, **32**(1), 13–33.

Chew, S.H., Miao, B. and Zhong, S. (2017) Partial ambiguity. *Econometrica*, **85**(4), 1239–1260.

Cornilly, D., Rüschendorf, L. and Vanduffel, S. (2018) Upper bounds for strictly concave distortion risk measures on moment spaces. *Insurance: Mathematics and Economics*, **82**, 141–151.

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

Embrechts, P., Höing, A. and Puccetti, G. (2005) Worst VaR scenarios. *Insurance: Mathematics and Economics*, **37**(1), 115–134.

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

Fujii, Y., Iwaki, H. and Osaki, Y. (2017) An economic premium principle under the dual theory of the smooth ambiguity model. *ASTIN Bulletin*, **47**(3), 787–801.

Gao, R. and Kleywegt, A. (in press) Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research.* Available at https://doi.org/10.1287/moor.2022.1275.

Ghossoub, M. (2019a) Budget-constrained optimal insurance without the nonnegativity constraint on indemnities. *Insurance: Mathematics and Economics*, **84**, 22–39.

Ghossoub, M. (2019b) Budget-constrained optimal insurance with belief heterogeneity. *Insurance: Mathematics and Economics*, **89**, 79–91.

Glasserman, P. and Xu, X. (2014) Robust risk measurement and model risk. *Quantitative Finance*, **14**(1), 29–58.

Gupta, V. (2019) Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Science*, **65**(9), 4242–4260.

Hürlimann, W. (2002) Analytical bounds for two value-at-risk functionals. *ASTIN Bulletin*, **32**(2), 235–265.

Jiang, W., Escobar-Anel, M. and Ren, J. (2020). Optimal insurance contracts under distortion risk measures with ambiguity aversion. *ASTIN Bulletin*, **50**(2), 619–646.

Kaas, R., Laeven, R.J. and Nelsen, R.B. (2009) Worst VaR scenarios with given marginals and measures of association. *Insurance: Mathematics and Economics*, **44**(2), 146–158.

Kannan, R., Bayraksan, G. and Luedtke, J.R. (2020) Residuals-based distributionally robust optimization with covariate information. arXiv preprint arXiv:2012.01088.

Kuhn, D., Mohajerin Esfahani, P., Nguyen, V.A. and 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**(4), 1090–1105.

Lam, H. and Mottet, C. (2017) Tail analysis without parametric models: A worst-case perspective. *Operations Research*, **65**(6), 1696–1711.

Landsman, Z. and Tsanakas, A. (2012) Parameter uncertainty in exponential family tail estimation. *ASTIN Bulletin*, **42**(1), 123–152.

Li, L., Shao, H., Wang, R. and Yang, J. (2018) Worst-case range value-at-risk with partial information. *SIAM Journal on Financial Mathematics*, **9**(1), 190–218.

Liu, H. and Wang, R. (2017) Collective risk models with dependence uncertainty. *ASTIN Bulletin*, **47**(2), 361–389.

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

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Panaretos, V.M. and Zemel, Y. (2019) Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, **6**, 405–431.

Payzan-LeNestour, E. and Woodford, M. (2022) Outlier blindness: A neurobiological foundation for neglect of financial risk. *Journal of Financial Economics*, **143**(3), 1316–1343.

Pesenti, S. and Jaimungal, S. (2020) Portfolio optimisation within a Wasserstein ball. arXiv preprint arXiv:2012.04500.

Peters, G.W., Shevchenko, P.V. and Wüthrich, M.V. (2009) Model uncertainty in claims reserving within Tweedie's compound Poisson models. *ASTIN Bulletin*, **39**(1), 1–33.

Pflug, G.C., Timonina-Farkas, A. and Hochrainer-Stigler, S. (2017) Incorporating model uncertainty into optimal insurance contract design. *Insurance: Mathematics and Economics*, **73**, 68–74.

Pflug, G. and Wozabal, D. (2007) Ambiguity in portfolio selection. *Quantitative Finance*, **7**(4), 435–442.

Popescu, I. (2007) Robust mean-covariance solutions for stochastic optimization. *Operations Research*, **55**(1), 98–112.

Rahimian, H., Bayraksan, G. and Homem-de-Mello, T. (2019) Controlling risk and demand ambiguity in newsvendor models. *European Journal of Operational Research*, **279**(3), 854–868.

Robert, C.Y. and Therond, P.E. (2014) Distortion risk measures, ambiguity aversion and optimal effort. *ASTIN Bulletin*, **44**(2), 277–302.

Scarf, H. (1958) A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production* (eds. K.J. Arrow, S. Karlin and H.E. Scarf), pp. 201–209. Stanford University Press.

Villani, C. (2009) *Optimal Transport: Old and New*. Berlin: Springer.

Wang, R., Peng, L. and Yang, J. (2013) Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics*, **17**(2), 395–417.

Wozabal, D. (2012) A framework for optimization under ambiguity. *Annals of Operations Research*, **193**(1), 21–47.

Wozabal, D. (2014) Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, **62**(6), 1302–1315.

Zhao, C. and Guan, Y. (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, **46**(2), 262–267.

Zhao, L. and Zhu, W. (2011) Ambiguity aversion: A new perspective on insurance pricing. *ASTIN Bulletin*, **41**(1), 157–189.

A. Appendix

Lemma A.1 *For two random pairs $\boldsymbol{V}_1 = (\xi_1, \eta_1)$ and $\boldsymbol{V}_2 = (\xi_2, \eta_2)$, each with independent components so that $F_{\mathbf{V}_i} = F_{\xi_i} \times F_{\eta_i}$ for $i = 1, 2$, we have*

$$W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p = W(F_{\xi_1}, F_{\xi_2})^p + W(F_{\eta_1}, F_{\eta_2})^p, \tag{A1}$$

where $W(\cdot, \cdot)$ is the Wasserstein distance of order $p \geq 1$ defined as in (2.4) with $d(\cdot, \cdot)$ specified as the p norm.

Proof. As in Subsection 2.3, we introduce the following spaces:

$$\begin{aligned} S &= \left\{\Pi \in \mathcal{P}\left(\mathbb{R}^4\right) : \Pi_{\mathbf{V}_1} = F_{\mathbf{V}_1}, \Pi_{\mathbf{V}_2} = F_{\mathbf{V}_2}\right\}, \\ S_\xi &= \left\{\mu \in \mathcal{P}\left(\mathbb{R}^2\right) : \mu_{\xi_1} = F_{\xi_1}, \mu_{\xi_2} = F_{\xi_2}\right\}, \\ S_\eta &= \left\{\nu \in \mathcal{P}\left(\mathbb{R}^2\right) : \nu_{\eta_1} = F_{\eta_1}, \nu_{\eta_2} = F_{\eta_2}\right\}. \end{aligned}$$

Since $F_{\mathbf{V}_i} = F_{\xi_i} \times F_{\eta_i}$ for $i = 1, 2$, we have $S = S_\xi \times S_\eta$. Clearly,

$$W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p = \inf_{\Pi \in S} E_\Pi \left[\|\mathbf{V}_1 - \mathbf{V}_2\|_p^p\right] = \inf_{\Pi \in S} E_\Pi \left[|\xi_1 - \xi_2|^p + |\eta_1 - \eta_2|^p\right]. \tag{A2}$$

On the one hand, due to the super-additivity of the infimum operation, we have

$$\begin{aligned} W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p &\geq \inf_{\Pi \in S} E_\Pi \left[|\xi_1 - \xi_2|^p\right] + \inf_{\Pi \in S} E_\Pi \left[|\eta_1 - \eta_2|^p\right] \\ &= \inf_{\mu \in S_\xi} E_\mu \left[|\xi_1 - \xi_2|^p\right] + \inf_{\nu \in S_\eta} E_\nu \left[|\eta_1 - \eta_2|^p\right] \\ &= W(F_{\xi_1}, F_{\xi_2})^p + W(F_{\eta_1}, F_{\eta_2})^p. \end{aligned}$$

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

On the other hand, for any $\mu \in S_\xi$ and $\nu \in S_\eta$, we simply take the product measure $\Pi_{(\mu,\nu)} = \mu \times \nu$, which defines a joint distribution of the quadruple $(\xi_1, \eta_1, \xi_2, \eta_2)$. Under $\Pi_{(\mu,\nu)}$, for each $i = 1, 2$, the pair $\mathbf{V}_i = (\xi_i, \eta_i)$ follows the distribution $F_{\mathbf{V}_i} = F_{\xi_i} \times F_{\eta_i}$. This verifies $\Pi_{(\mu,\nu)} \in S$. Therefore, by (A2),

$$\begin{aligned} E_\mu\left[|\xi_1-\xi_2|^p\right] + E_\nu\left[|\eta_1-\eta_2|^p\right] &= E_{\Pi_{(\mu,\nu)}}\left[|\xi_1-\xi_2|^p\right] + E_{\Pi_{(\mu,\nu)}}\left[|\eta_1-\eta_2|^p\right] \\ &= E_{\Pi_{(\mu,\nu)}}\left[|\xi_1-\xi_2|^p + |\eta_1-\eta_2|^p\right] \\ &\geq W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p. \end{aligned}$$

Taking infimum of the two terms on the left-hand side over $\mu \in S_\xi$ and $\nu \in S_\eta$, respectively, we obtain

$$W(F_{\xi_1}, F_{\xi_2})^p + W(F_{\eta_1}, F_{\eta_2})^p \geq W(F_{\mathbf{V}_1}, F_{\mathbf{V}_2})^p.$$

This proves (A1). □

Proof of Proposition 3.1. Introduce the space

$$S = \left\{\Pi \in \mathcal{P}\left(\mathbb{R}_+^2\right) : \Pi_Z = F_Z,\ \Pi_{\hat{Z}} = \hat{F}_Z\right\}.$$

Observe that the optimization problem (3.5) is conducted over the Wasserstein ball

$$B_\epsilon\left(\hat{F}_Z\right) = \left\{F_Z : W\left(F_Z, \hat{F}_Z\right) \leq \epsilon\right\}$$

with the Wasserstein distance $W\left(F_Z, \hat{F}_Z\right) = \inf_{\Pi \in S}\left(E_\Pi\left[\left|Z-\hat{Z}\right|^p\right]\right)^{\frac{1}{p}}$. Note that the infimum in defining the Wasserstein distance is always attainable; see, for example, Theorem 4.1 of Villani (2009). Therefore, for each $F_Z \in B_\epsilon\left(\hat{F}_Z\right)$, we can find a joint distribution $\Pi \in S$ such that $E_\Pi\left[\left|Z-\hat{Z}\right|^p\right] \leq \epsilon^p$, while conversely, for each $\Pi \in S$ such that $E_\Pi\left[\left|Z-\hat{Z}\right|^p\right] \leq \epsilon^p$, its marginal distribution $\Pi_Z = F_Z$ belongs to the ambiguity ball $B_\epsilon\left(\hat{F}_Z\right)$. This allows us to rewrite the optimization problem (3.5) in terms of Π, that is,

$$\begin{aligned} &\sup_{\Pi \in S} E\left[Z^p\right] \\ &\quad\text{subject to } E_\Pi\left[\left|Z-\hat{Z}\right|^p\right] \leq \epsilon^p. \end{aligned} \tag{A3}$$

Now we continue on the optimization problem (A3). Given a joint distribution $\Pi \in S$, we denote by F_Z^i the distribution of Z conditioned on $\hat{Z} = z_i$, that is,

$$F_Z^i(dz) = P\left(Z \in dz|\ \hat{Z} = z_i\right), \qquad z \in \mathbb{R}_+,\ i = 1, \ldots, N. \tag{A4}$$

In terms of this conditional distribution, by the law of total probability, we have

$$E\left[Z^p\right] = \frac{1}{N}\sum_{i=1}^N \int_{0-}^\infty z^p F_Z^i(dz)$$

and

$$E_\Pi\left[\left|Z-\hat{Z}\right|^p\right] = \frac{1}{N}\sum_{i=1}^N \int_{0-}^\infty |z-z_i|^p\, F_Z^i(dz). \tag{A5}$$

Note that the construction of the conditional distributions (A4) actually establishes a mapping between a joint distribution $\Pi \in S$ and a set of conditional distributions $\{F_Z^1, \ldots, F_Z^N\}$. Moreover, by (A5), $E_\Pi\left[\left|Z-\hat{Z}\right|^p\right] \leq \epsilon^p$ if and only if

$$\frac{1}{N}\sum_{i=1}^N \int_{0-}^\infty |z-z_i|^p\, F_Z^i(dz) \leq \epsilon^p. \tag{A6}$$

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Thus, we can convert the optimization problem (A3) to

$$\sup_{F_Z^1,\ldots,F_Z^N} \frac{1}{N}\sum_{i=1}^{N}\int_{0-}^{\infty} z^p F_Z^i(dz) \tag{A7}$$
$$\text{subject to (A6).}$$

Introducing the Lagrangian multiplier λ to this new version of the optimization problem (A7), we have

$$\begin{aligned}&\sup_{F_Z\in B_\epsilon(\hat{F}_Z)} E\left[Z^p\right]\\ &= \sup_{F_Z^1,\ldots,F_Z^N}\inf_{\lambda\geq 0}\left\{\frac{1}{N}\sum_{i=1}^{N}\int_{0-}^{\infty} z^p F_Z^i(dz) - \lambda\left(\frac{1}{N}\sum_{i=1}^{N}\int_{0-}^{\infty}\left|z-z_i\right|^p F_Z^i(dz)-\epsilon^p\right)\right\}\\ &= \inf_{\lambda\geq 0}\sup_{F_Z^1,\ldots,F_Z^N}\left\{\frac{1}{N}\sum_{i=1}^{N}\int_{0-}^{\infty} z^p F_Z^i(dz) - \lambda\left(\frac{1}{N}\sum_{i=1}^{N}\int_{0-}^{\infty}\left|z-z_i\right|^p F_Z^i(dz)-\epsilon^p\right)\right\},\end{aligned}$$

where we can switch the order of sup and inf based on an established strong duality in worst-case estimation problems; see, for example, Theorem 1 and Proposition 2 of Gao and Kleywegt (2022). After rearrangement, it follows that

$$\begin{aligned}\sup_{F_Z\in B_\epsilon(\hat{F}_Z)} E\left[Z^p\right] &= \inf_{\lambda\geq 0}\left\{\lambda\epsilon^p - \frac{1}{N}\sum_{i=1}^{N}\inf_{F_Z^i}\int_{0-}^{\infty}\left(\lambda\left|z-z_i\right|^p - z^p\right)F_Z^i(dz)\right\}\\ &= \inf_{\lambda\geq 0}\left\{\lambda\epsilon^p - \frac{1}{N}\sum_{i=1}^{N}\inf_{z\geq 0}\left(\lambda\left|z-z_i\right|^p - z^p\right)\right\},\end{aligned} \tag{A8}$$

where the last step follows from the observation that each F_Z^i can be reduced to a Dirac measure δ_x at any $x\in\mathbb{R}_+$. By now, we have converted the semi-infinite optimization problem (3.5) to a one-dimensional convex minimization problem (A8).

To solve (A8), we first consider the case $p>1$. We separate $\lambda\geq 0$ into two regions: $0\leq\lambda\leq 1$ and $\lambda>1$. When $0\leq\lambda\leq 1$, since there exists at least one $z_i>0$ for which $\lim_{z\to\infty}\left(\lambda\left|z-z_i\right|^p - z^p\right) = -\infty$, we have

$$\lambda\epsilon^p - \frac{1}{N}\sum_{i=1}^{N}\left(\inf_{z\geq 0}\left(\lambda\left|z-z_i\right|^p - z^p\right)\right) = \infty.$$

Therefore, the minimizer of the dual problem (A8) for $p>1$ must be in the region $\lambda>1$. By splitting the range for z into $(z>z_i)$ and $(0\leq z\leq z_i)$, we obtain

$$\sup_{F_Z\in B_\epsilon(\hat{F}_Z)} E\left[Z^p\right] = \min_{\lambda>1}\left\{\lambda\epsilon^p - \frac{1}{N}\sum_{i=1}^{N}\left(\inf_{z>z_i}\left(\lambda(z-z_i)^p - z^p\right)\wedge\inf_{0\leq z\leq z_i}\left(\lambda(z_i-z)^p - z^p\right)\right)\right\}. \tag{A9}$$

Observe the function inside the first inner infimum $\inf_{z>z_i}$ in (A9). By taking derivative $\frac{d}{dz}$ and letting it be zero, we obtain a unique stationary point

$$z_* = \frac{\lambda^{\frac{1}{p-1}} z_i}{\lambda^{\frac{1}{p-1}} - 1},$$

which belongs to the region $z>z_i$ because $\lambda>1$. It is easy to see that this infimum $\inf_{z>z_i}$ is attained at z_*, that is,

$$\inf_{z>z_i}\left(\lambda\left(z-z_i\right)^p - z^p\right) = -z_i^p\left(\frac{\lambda^{\frac{1}{p-1}}}{\lambda^{\frac{1}{p-1}}-1}\right)^{p-1}.$$

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Moreover, the second inner infimum $\inf_{0\le z\le z_i}$ in (A9) is attained at z_i since the inside function is decreasing in $z \le z_i$. Thus, between the two infima in the bracketed part of (A9) the first is smaller, and we have

$$\begin{aligned}\sup_{F_Z\in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right] &= \min_{\lambda>1}\left\{\lambda\epsilon^p + \frac{1}{N}\sum_{i=1}^N\left(\frac{\lambda^{\frac{1}{p-1}}}{\lambda^{\frac{1}{p-1}}-1}\right)^{p-1} z_i^p\right\}\\ &= \min_{\lambda>1}\left\{\lambda\epsilon^p + \left(\frac{\lambda^{\frac{1}{p-1}}}{\lambda^{\frac{1}{p-1}}-1}\right)^{p-1} \|Z_N\|_{L^p}^p\right\}. \qquad \text{(A10)}\end{aligned}$$

We continue to solve the optimization problem (A10). Looking at the inside function of λ in (A10), it is easy to see that the $\min_{\lambda>1}$ is attained at the unique stationary point

$$\lambda_* = \left(1 + \frac{\|Z_N\|_{L^p}}{\epsilon}\right)^{p-1} > 1,$$

and we have

$$\sup_{F_Z\in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right] = \left\{\lambda\epsilon^p + \left(\frac{\lambda^{\frac{1}{p-1}}}{\lambda^{\frac{1}{p-1}}-1}\right)^{p-1} \|Z_N\|_{L^p}^p\right\}\Bigg|_{\lambda=\lambda_*} = \left(\epsilon + \|Z_N\|_{L^p}\right)^p.$$

This proves (3.6) for $p > 1$.

The case $p = 1$ for (A8) can be dealt with in the same way, as follows:

$$\begin{aligned}\sup_{F_Z\in B_\epsilon\left(\hat{F}_Z\right)} E\left[Z^p\right] &= \min_{\lambda\ge 0}\left\{\lambda\epsilon - \frac{1}{N}\sum_{i=1}^N\left(\inf_{z\ge 0}\left(\lambda\left|z-z_i\right| - z\right)\right)\right\}\\ &= \min_{\lambda\ge 1}\left\{\lambda\epsilon - \frac{1}{N}\sum_{i=1}^N\left(\inf_{z\ge 0}\left(\lambda\left|z-z_i\right| - z\right)\right)\right\}\\ &= \min_{\lambda\ge 1}\left\{\lambda\epsilon + \frac{1}{N}\sum_{i=1}^N z_i\right\}\\ &= \epsilon + \|Z_N\|_{L^1}.\end{aligned}$$

This proves (3.6) for $p = 1$ and hence complete the proof of Proposition 3.1. □

Lemma A.2 *Consider the optimization problem (3.2) in which $\hat{F}_X$ is the empirical distribution of a dataset $\{x_1, \ldots, x_n\}$. The supremum (3.2) is finite if and only if the order of the Wasserstein distance embedded in this worst-case estimation problem is no less than the order p of the power function.*

Proof. Denoted by p' the order of the Wasserstein distance embedded in (3.2). Following the beginning part of the proof of Proposition 3.1, we can convert the optimization problem (3.2) to

$$\sup_{F_X\in B_r\left(\hat{F}_X\right)} E\left[X^p\right] = \inf_{\lambda\ge 0}\left\{\lambda r^{p'} - \frac{1}{n}\sum_{i=1}^n \inf_{x\ge 0}\left(\lambda\left|x-x_i\right|^{p'} - x^p\right)\right\}.$$

The right-hand side is finite if and only if there exists some $\lambda \ge 0$ such that

$$\lambda r^{p'} - \frac{1}{n}\sum_{i=1}^n \inf_{x\ge 0}\left(\lambda\left|x-x_i\right|^{p'} - x^p\right) < \infty,$$

if and only if there exists some $\lambda \ge 0$ such that

$$\inf_{x\ge 0}\left(\lambda\left|x-x_i\right|^{p'} - x^p\right) > -\infty \qquad \text{for all } i = 1, \ldots, n,$$

if and only if $p' \ge p$. □

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press

Proof of Theorem 3.1. In view of Proposition 3.1, to solve the worst-case estimation problem (3.3), it remains to complete the outer optimization problem in (3.4). Simply plugging into (3.4) the expression (3.6) with $\epsilon = s\,(r^p - |q-q_n|)^{\frac{1}{p}}$ given in (2.10), we obtain

$$\begin{aligned}&\sup_{q\in[q_n^-,q_n^+]} \left\{(1-q)E\left[Y^p\right] + q\left(s\,(r^p - |q-q_n|)^{\frac{1}{p}} + \|Z_N\|_p\right)^p\right\}\\ =\; &\sup_{q\in[q_n^-,q_n]} f_1(q) \vee \sup_{q\in[q_n,q_n^+]} f_2(q),\end{aligned} \tag{A11}$$

where $f_1(q)$ and $f_2(q)$ are two auxiliary functions defined by

$$\begin{aligned}f_1(q) &= (1-q)C_0^p + q\left(s\,(r^p-(q_n-q))^{\frac{1}{p}} + C_1\right)^p,\\ f_2(q) &= (1-q)C_0^p + q\left(s\,(r^p-(q-q_n))^{\frac{1}{p}} + C_1\right)^p.\end{aligned} \tag{A12}$$

Given $C_0 \le C_1$, it is easy to deal with the first supremum of $f_1(q)$. Observe that the bracketed part $\left(s\,(r^p-(q_n-q))^{\frac{1}{p}} + C_1\right)$ in $f_1(q)$ is monotonically increasing in q and larger than C_0 over $q \in [q_n^-, q_n]$. Thus, the supremum of $f_1(q)$ is attained at the largest possible value $q = q_n$. This reduces the optimization problem (A11) to the second supremum.

To determine the supremum of $f_2(q)$ over $q \in [q_n, q_n^+]$, we observe the following derivatives:

$$\begin{aligned}\frac{d}{dq}f_2(q) &= -C_0^p + \left(s(r^p+q_n-q)^{\frac{1}{p}} + C_1\right)^p\\ &\quad -qs\left(s(r^p+q_n-q)^{\frac{1}{p}} + C_1\right)^{p-1}(r^p+q_n-q)^{\frac{1-p}{p}};\\ \frac{d^2}{dq^2}f_2(q) &= \frac{s}{p}\left(s(r^p+q_n-q)^{\frac{1}{p}} + C_1\right)^{p-2}(r^p+q_n-q)^{\frac{1}{p}-2}\\ &\quad \times\left(-2ps(r^p+q_n-q)^{\frac{1}{p}+1} + C_1\left((1+p)q - 2p(q_n+r^p)\right)\right).\end{aligned}$$

Note that, in the expression for $\frac{d^2}{dq^2}$ above, the last bracketed term after C_1 satisfies

$$(1+p)q - 2p(q_n+r^p) \le (1+p)(q_n+r^p) - 2p(q_n+r^p) \le 0.$$

Thus, $\frac{d^2}{dq^2}f_2(q) \le 0$ and $f_2(q)$ is concave over $q \in [q_n, q_n^+]$. We now look at the derivative

$$\left.\frac{d}{dq}f_2(q)\right|_{q=q_n} = -C_0^p + (rs+C_1)^{p-1}\left(rs(1-q_nr^{-p}) + C_1\right),$$

whose sign depends on the value of r. As it is continuous and strictly increasing in r, we define $\bar{r}$ as the unique solution to $\left.\frac{d}{dq}f_2(q)\right|_{q=q_n} = 0$, namely, Equation (3.8), and separately conclude the two pieces in (3.7) as follows:

- When $0 < r < r_* = \bar{r} \vee 0$, we have $\left.\frac{d}{dq}f_2(q)\right|_{q=q_n} < 0$, which, combined with the concavity of $f_2(q)$ over the region $q \in [q_n, q_n^+]$, implies that $f_2(q)$ is decreasing over this region. Thus, the supremum is attained at $q = q_n$, giving the first expression in (3.7).
- When $r \ge r_*$, we have $\left.\frac{d}{dq}f_2(q)\right|_{q=q_n} \ge 0$. This implies that the supremum is attained at either the unique solution $\bar{q}_n$ to $\frac{d}{dq}f_2(q) = 0$, namely, Equation (3.9), or q_n^+, whichever is smaller, giving the second expression in (3.7).

This completes the proof of Theorem 3.1. □

https://doi.org/10.1017/asb.2023.3 Published online by Cambridge University Press