Discrete Approximation and Quantification in Distributionally Robust Optimization[*]

Yongchao Liu[†] Alois Pichler[‡] Huifu Xu[§]

September 11, 2017

Abstract

Discrete approximation of probability distributions is an important topic in stochastic programming. In this paper, we extend the research on this topic to distributionally robust optimization (DRO), where discretization is driven by either limited availability of empirical data (samples) or a computational need for improving numerical tractability. We start with a one-stage DRO where the ambiguity set is defined by generalized prior moment conditions and quantify the discrepancy between the discretized ambiguity set and the original one by employing the Kantorovich/ Wasserstein metric. The quantification is achieved by establishing a new form of Hoffman's lemma for moment problems under a general class of metrics, namely ζ-structures. We then investigate how the discrepancy propagates to the optimal value in one-stage DRO and discuss further the multistage DRO under nested distance. The technical results lay down a theoretical foundation for various discrete approximation schemes to be applied to solve one stage and multistage distributionally robust optimization problems.

Keywords: Hoffman's lemma, Kantorovich/ Wasserstein metric, discretization of ambiguity set, moment conditions, nested distance

Classification: 90C15, 90C31

1 Introduction

A key step in decision making under uncertainty is to quantify the probability distribution of the underlying uncertain parameters. In some cases we can obtain a sufficiently large

[*] The research is supported by EPSRC grant EP/M003191/1.

[†] School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China. lyc@dlut.edu.cn. The work of this author was carried out in the School of Mathematical Sciences, University of Southampton. His work is supported in part by EPSRC grant EP/M003191/1 and NSFC #11571056.

[‡] Faculty of Mathematics, Chemnitz, University of Technology, Germany. alois.pichler@mathematik.tu-chemnitz.de

[§] School of Mathematical Sciences, University of Southampton, SO17 1BJ, Southampton, UK, H.Xu@soton.ac.uk number of samples or empirical data and use them to construct an approximation of the true distribution. An advantageous side effect of the latter is that it leads to discretization of a stochastic programming problem which is an indispensable step for numerical solutions in many cases. The well-known sample average approximation method in stochastic programming is fundamentally based on this (cf. Shapiro [27]). A significant drawback of this approach is that when the sample size is large, solving the resulting optimization problem may be difficult particularly in a multistage decision making process (cf. Pflug and Pichler [17]). On the other hand, in his recent monograph, Savage [24] points out that the number of discretization points must not be too small to ensure sufficiently good approximations.

In other cases such as signal processing of mobile ad hoc networks, the number of samples is relatively small and consequently an additional measure may have to be taken to hedge the risk of inadequate information for approximating the true probability distribution. These are indeed some main motivations behind scenario models in stochastic optimization and distributionally robust optimization (see also the recent monograph by Pflug and Pichler [18] and references therein).

Over the past decade, effectively quantifying uncertainty and addressing the trade-off between using less information for approximating the true probability distribution such as samples and securing specified confidence of the resulting approximate optimal decision has been a challenging research topic in data-driven optimization problems, either because there is limited number of available samples or it is more desirable to use fewer samples to increase numerical tractability of the resulting optimization problem as we discussed before.

In this paper, we extend this important topic of research to distributionally robust optimization. To explain the idea, we consider the following one-stage distributionally robust minimization problem

$$\min_{x\in X} \sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\xi(\omega))], \tag{1}$$

where x is decision vector taking values from a closed set X of $\mathbb{R}^n$, $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is a continuous cost function, $\xi : \Omega \to \Xi \subset \mathbb{R}^k$ is a vector of random variables defined on a measurable space $(\Omega, \mathcal{F})$, the expectation in (1) is with respect to $\omega \in \Omega$, $\mathcal{P}$ is a set of probability measures defined by a set of generalized moment conditions

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\Omega) : \ \mathbb{E}_P[\Psi(\xi(\omega))] \in \mathcal{K} \ \right\}, \tag{2}$$

where Ψ is a random mapping consisting of vectors and/ or matrices with measurable random components, the mathematical expectation of Ψ is taken w.r.t. each component of Ψ, $\mathscr{P}(\Omega)$ denotes the set of all probability measures on Ω and $\mathcal{K}$ is a closed convex cone in the Cartesian product of some finite dimensional vector and/ or matrix spaces.

The DRO model (1) differs from the standard one stage stochastic minimization formulation in that here the true probability distribution of ξ is unknown but there is some partial information revealing that it satisfies the generalized prior moment condition (2). To immunize the risk arising from ambiguity of the true probability distribution, the optimal decision is taken on the basis of the worst probability measure from $\mathcal{P}$, which is called the *ambiguity set*. This kind of robust optimization framework can be traced back to earlier work of Scarf [25], whose primary aim was to address incomplete information on the underlying uncertainty in supply chain and inventory control problems. DRO models have found many applications in operations research, finance and management sciences and the research on DRO has grown rapidly over the past few decade, see Bertsimas and Popescu [4], Delage and Ye [5], Wiesemann et al. [33] and references therein.

A great deal of research in the literature to date is devoted to developing tractable numerical methods for solving DRO by reformulating the inner maximization problem into a semi-infinite programming problem through Lagrange dualization and further as an semi-definite programming problem through $\mathcal{S}$-Lemma or dual method, cf. Zymler et al. [40], Wiesemann et al. [33]. This kind of approach requires the underlying functions in the objective and moments to have some specific structure in terms of the variable ξ and also the support set of ξ to have some structure, see Wiesemann et al. [33] for a comprehensive discussion.

Another important approach pioneered by Pflug and Wozabal [19] is to discretize the ambiguity set of DRO and then solve the discretized mini-max optimization problem directly as a saddle point problem in deterministic optimization. The discretization approach has received increasing attention over the past few years. For instance, Mehrotra and Papp [14] extend the approach to a general class of DRO problems and design a process which generates a *cutting surface* of the inner optimal value at each iterate. Xu et al. [35] observe that the discretization scheme is equivalent to discrete approximation of the semi-infinite constraints of the dualized inner maximization problem and apply the well known cutting plane method (Kelley [12]) to solve the minimax optimization. Under some moderate conditions, they show convergence of the optimal value of the discretized problem to its true counterpart as the discretization refines.

While the convergence result gives some qualitative guarantee for asymptotic consistency of the optimal value, it does not address a quantitative relationship between the sample size and error of optimal value. This paper aims to fill out the gap. The main contributions in this regard can be summarized as follows:

We derive a new form of Hoffman's lemma for the moment problem (2) by showing that the distance between any probability measure and the ambiguity set $\mathcal{P}$ under a generic metric with ζ-structure (cf. Rachev [21, Chapter 4]) is linearly bounded by the residual of the moment system (Theorem 2 below) under the Slater condition. These metrics include the Kantorovich/ Wasserstein metric, the total variation metric and Fortet–Mourier metric. The new Hoffman's lemma (Section 3) complements the existing results established by Sun and Xu [31] and Zhang et al. [36], where the distance of probability measures is characterized by the total variation metric.

We propose a general discretization scheme for the ambiguity set defined by the moment problem (2) where the probability measures in the discretized ambiguity set are supported at a finite subset of Ξ. By exploiting an earlier result due to Pflug and Pichler [18] and the new Hoffman's lemma, we establish a quantitative relationship between the distance of $\mathcal{P}$ and its discretized counterpart under Kantorovich (Wasserstein) metric and the Hausdorff distance between the support sets of the two ambiguity sets of probability measures (Theorem 12, Section 4).

We investigate the relationship between the DRO problem (1) and its discretization in terms of the optimal value and quantify the difference of the latter by the Kantorovich/ Wasserstein distance of $\mathcal{P}$ and its discretized counterpart (Theorem 14). The result lays down a theoretical foundation for a wide range of discretization schemes including data-driven optimization problems where there is limited availability of samples/ empirical data.

A key condition that we impose in deriving the new Hoffman's lemma under ζ-metric is the Slater condition. This implicitly excludes moment problems with equality constraints. While this may be viewed as a significant limitation, we note that a number of interesting moment based ambiguity sets in the literature only involve inequality constraints.

Having established the desired results for one stage stochastic programing problems, we ask ourselves whether the established results can be extended to multistage setting. As far as we are concerned, the research on multistage distributionally robust optimization (MDRO) seems to be still in its infancy with only a few papers appearing on some specific topics. For instance, Analui and Pflug [1] consider a MDRO model for multistage stochastic programs where the data and information structure of the baseline model is a tree and the "ambiguity neighborhoods" around this tree is defined through nested distance. By reformulating the MDRO as a deterministic minimax saddle point problem, they propose a numerical method for solving the latter and apply the MDRO model to stochastic production/ inventory control problem with weekly ordering. Xin et al. [34] propose a distributionally robust optimization model for multistage news vendor problems where there is an ambiguity in the distribution of uncertain demand at each stage. They investigate time consistency of the decision making process. Shapiro [28] formulates the MDRO associated with risk measure and discuss conditions for time consistency of such formulations of stochastic problems. Iyengar [10] and Nilim and Ghaoui [15] study the distributionally robust Markov chain where the ambiguity set is defined through Cartesian product of independent marginal sets where time consistency follows. Wiesemann et al. [32] study the distributionally robust Markov decision processes with a new class of ambiguity sets, which contains the above Cartesian product ambiguity sets as a special case.

In Section 5, we consider a class of multistage distributionally robust minimization problems with ambiguity at each stage being constructed by conditional prior moments. We concentrate on investigating the difference between a discretized distributionally robust minimization problem and its true counterpart in terms of optimal values (Theorem 19, Section 5). A key difference between Theorem 19 and Theorem 14 is that the error bound in the former is derived through nested distance of two processes rather than ambiguity sets and this prevents us from obtaining Hoffman-type error bound in the multistage setting.

Notation. Throughout the paper, we use the following notation. We use $\mathcal{S}^n$, $\mathcal{S}^n_+$ and $\mathcal{S}^n_-$ to denote the space of symmetric matrices, the cone of positive semidefinite matrices and the cone of negative semidefinite matrices in $\mathbb{R}^{n\times n}$. $\mathbb{R}^n_+$ denotes the cone of vectors with non-negative components in $\mathbb{R}^n$. $\langle\cdot,\cdot\rangle$ denotes a bilinear representation of the expected value, $\|\cdot\|$ denote the 2-norm for a vector or the Frobenius norm for a matrix.

2 Metrics of probability measures

Let Ω be a sample space set and $\mathcal{F}$ be the associated sigma algebra. Let $\mathscr{P}(\Omega)$ be the set of all probability measures over the measurable space $(\Omega,\mathcal{F})$. We consider a vector valued measurable function ξ mapping from Ω to $\Xi\subset\mathbb{R}^k$. Let $\mathscr{B}$ be the Borel sigma algebra in $\mathbb{R}^k\cap\Xi$ and $P\in\mathscr{P}(\Omega)$. For each set $A\in\mathscr{B}$, let $P^\xi(A):=P(\xi^{-1}(A))$ denote the image measure, which is also known as push-forward measure. Consequently we may focus on $\mathscr{P}(\Xi)$, the set of all probability measures defined on the measurable space $(\Xi,\mathscr{B})$ with support set contained in Ξ.

In probability theory, various metrics have been introduced to quantify the distance/ difference between two probability measures; see Athreya and Lahiri [3], Gibbs and Su [7]. Here we adopt the metrics with ζ-structure which subsume a number of interesting metrics.

Let P, $Q\in\mathscr{P}(\Xi)$ and $\mathscr{G}$ be a family of real-valued bounded measurable functions on Ξ and define

$$\mathsf{dl}_{\mathscr{G}}(P,Q) := \sup_{g\in\mathscr{G}} \left|\mathbb{E}_P[g(\xi)]-\mathbb{E}_Q[g(\xi)]\right|. \tag{3}$$

The distance defined as such is called a metric with ζ-structure which covers a wide range of metrics in probability theory including the total variation metric, Kantorovich/ Wasserstein metric, bounded Lipschitz metric and some other metrics; see Gibbs and Su [7], Rachev [21] or Zolotarev [39] and references therein. Specifically, if

$$\mathscr{G} := \left\{ g:\Xi\to\mathbb{R} |\ g \text{ is } \mathscr{B} \text{ measurable}, \ \sup_{\xi\in\Xi}|g(\xi)|\leq 1\right\},$$

then $\mathsf{dl}_{\mathscr{G}}(P,Q)$ reduces to the *total variation metric*, denoted by dl_{TV}.[1]

If g is restricted further to be Lipschitz continuous with modulus bounded by 1, i.e.,

$$\mathscr{G} = \left\{ g:\sup_{\xi\in\Xi}|g(\xi)|\leq 1, \ g \text{ is Lipschitz continuous with Lipschtiz constant } L_1(g)\leq 1\right\}, \tag{4}$$

[1] Note that in some references such as Gibbs and Su [7], the total variation metric is defined by the maximal difference of two probability measures over all measurable sets of Ξ which is equal to $2\mathsf{dl}_{TV}$.

where $L_1(g) := \sup\{|g(u) - g(v)|/d(u,v) : u \neq v\}$, then the resulting metric is known as *bounded Lipschitz metric*, denoted by dl_{BL}. If the boundedness of g is lifted in (4), that is,

$$\mathscr{G} = \{g : g \text{ is Lipschitz continuous and Lipschtiz modulus } L_1(g) \leq 1\}, \tag{5}$$

then we arrive at Kantorovich/ Wasserstein metric, denoted by dl_K. If we relax the Lipschitz continuity in (5), that is,

$$\mathscr{G} = \{g : g \text{ is Lipschitz continuous and } L_q(g) \leq 1\},$$

with

$$L_q(g) := \inf\{L : |g(u) - g(v)| \leq L\|u - v\| \max(1, \|u\|^{q-1}, \|v\|^{q-1})\}, \qquad u, v \in \Xi,$$

then we obtain *Fortet–Mourier metric*, denoted by dl_{FM}. If

$$\mathscr{G} = \left\{g : g(\cdot) := \mathbb{1}_{(-\infty,t]}(\cdot),\ t \in \mathbb{R}^k\right\},$$

where

$$\mathbb{1}_{(-\infty,t]}(\xi) := \begin{cases} 1, & \text{if } \xi \in (-\infty, t] \\ 0, & \text{otherwise,} \end{cases}$$

then we obtain *uniform (Kolmogorov) metric*, denoted by dl_U. It is well-known that $\mathsf{dl}_{TV}(P,Q) \in [0,1]$ and when Ξ is bounded $\mathsf{dl}_K(P,Q) \in [0, \mathrm{diam}(\Xi)]$, see Gibbs and Su [7]. Moreover, it follows by Zhao and Guan [37, Lemmas 1–4], that

$$\mathsf{dl}_{BL}(P,Q) \leq \max\{\mathsf{dl}_K(P,Q),\ \mathsf{dl}_{TV}(P,Q)\}, \tag{6}$$

$$\mathsf{dl}_{FM}(P,Q) \leq \max\{1,\ \mathrm{diam}(\Xi)^{q-1}\} \cdot \mathsf{dl}_K(P,Q) \tag{7}$$

and $\mathsf{dl}_U \leq \frac{1}{2}\mathsf{dl}_{TV}(P,Q)$.

Based on the ζ-metric, we can define the distance from a point to a set, deviation from one set to another and the Hausdorff distance between two sets in the space of probability measures $\mathscr{P}(\Xi)$. Specifically, for subset $\mathcal{C}$ and $\mathcal{C}'$ of $\mathscr{P}(\Xi)$ set

$$\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}) := \inf_{P \in \mathcal{C}} \mathsf{dl}_{\mathscr{G}}(Q, P), \tag{8}$$

$$\mathbb{D}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}) := \sup_{Q \in \mathcal{C}'} \mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}) \tag{9}$$

and

$$\mathbb{H}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}) := \max\left\{\mathbb{D}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}),\ \mathbb{D}(\mathcal{C}, \mathcal{C}'; \mathsf{dl}_{\mathscr{G}})\right\}. \tag{10}$$

Here $\mathbb{H}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}})$ defines the Hausdorff distance between $\mathcal{C}'$ and $\mathcal{C}$ under the ζ-metric $\mathsf{dl}_{\mathscr{G}}$ in the space of $\mathscr{P}(\Xi)$. It is easy to observe that $\mathbb{H}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}) = 0$ implies $\mathbb{D}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}) = 0$ and

$$\inf_{Q \in \mathcal{C}} \sup_{g \in \mathscr{G}} |\mathbb{E}_P[g(\xi)] - \mathbb{E}_Q[g(\xi)]| = 0$$

for any $P \in \mathcal{C}$. Recall that $\{P_N\}$ is said to converge to $P \in \mathscr{P}$ *weakly* if

$$\lim_{N\to\infty} \int_\Xi h(\xi)P_N(d\xi) = \int_\Xi h(\xi)P(d\xi)$$

for each bounded and continuous function $h : \Xi \to \mathbb{R}$. An important property of the Kantorovich/ Wasserstein metric is that it metrizes weak convergence of probability measures, that is, a sequence of probability measures $\{P_N\}$ converges to P weakly if and only if $\mathsf{dl}_K(P_N, P) \to 0$ (cf. Gibbs and Su [7]).

3 Discrete approximation of the ambiguity set

In some practical instances, we might have some additional partial information other than samples about the true probability distribution P in a decision making problem. Moment information is one of them. We consider a set of probability distributions defined by (2), where the unknown true probability distribution P is characterized by the moments of the reference mapping Ψ. In general $\mathcal{P}$ is a set and we know the true probability distribution lies in the set.

Recall that for each given $P \in \mathscr{P}(\Omega)$, the random variable ξ induces the image probability measure (denoted by P^ξ) on Ξ such that $\mathbb{E}_P[\Psi(\xi(\omega))] = \mathbb{E}_{P^\xi}[\Psi(\xi)]$. We may thus view the ambiguity set $\mathcal{P}$ as being defined on Ξ, i.e.,

$$\mathcal{C} := \{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\Psi(\xi)] \in \mathcal{K}\}, \tag{11}$$

where $\mathscr{P}(\Xi)$ denotes the set of all probability measures on the measurable space $(\Xi, \mathscr{B})$ with Borel sigma-algebra.

The moment condition (11) is considered by Zhang et al. [36]. It covers a wide range of moment conditions in the literature of distributionally robust optimization by choosing a specific structure for $\mathcal{K}$, see Zhang et al. [36], Xu et al. [35] for details.

Here we consider a discrete approximation of the ambiguity set $\mathcal{C}$. To streamline the idea of discretization, let $\Xi^N := \{\xi^1, \ldots, \xi^N\} \subset \Xi$ be a set of points in Ξ. These points may be samples of ξ or selected in deterministic manner. We look into the ambiguity set of probability distributions in $\mathscr{P}(\Xi^N)$ satisfying the moment condition

$$\mathcal{C}_N := \left\{P \in \mathscr{P}(\Xi^N) : \ \mathbb{E}_P[\Psi(\xi)] \in \mathcal{K} \ \right\}. \tag{12}$$

This kind of discretization was considered in [35]. Our focus here is to quantify the difference between $\mathcal{C}_N$ and $\mathcal{C}$. It is easy to observe that any probability measure in $\mathscr{P}(\Xi^N)$ can be presented as $P(\cdot) := \sum_{i=1}^N p_i\, \delta_{\xi^i}(\cdot)$, where $\delta_\xi(\cdot)$ denotes the Dirac probability measure located at ξ and the moment condition reduces to

$$\mathcal{C}_N := \left\{ P = \sum_{j=1}^N p_j \delta_{\xi^j} \in \mathscr{P}(\Xi^N) : \sum_{j=1}^N p_j \Psi(\xi^j) \in \mathcal{K} \right\}.$$

Thus $\mathcal{C}_N \subset \mathcal{C}$. What is unclear is the difference between $\mathcal{C}_N$ and $\mathcal{C}$ and this is indeed one of the main issues to investigate in this section.

3.1 Hoffman's Lemma

To quantify the difference between $\mathcal{C}_N$ and $\mathcal{C}$ we first need to study the error bound condition for the moment system in (11) which quantifies the deviation of any probability measure $Q \in \mathscr{P}(\Xi)$ from $\mathcal{C}$. Observe first that if we regard $\mathcal{C}$ as the set of solutions to system $\mathbb{E}_P[\Psi(\xi)] \in \mathcal{K}$, then this is essentially about Hoffman's lemma (Hoffman [9]) in the space of probability measures $\mathscr{P}(\Xi)$.

For the simplicity of notation, we write $\langle P, \Psi(\xi)\rangle$ for $\mathbb{E}_P[\Psi(\xi)]$ so that we can see P more clearly as a variable in the moment system and the expected value depends on P linearly. We need the following condition.

Assumption 1 (The Slater condition). The system (11) satisfies the *Slater condition*, that is, there exist $P_0 \in \mathscr{P}(\Xi)$ and a constant $\alpha > 0$ such that

$$\langle P_0, \Psi(\xi)\rangle + \alpha\mathcal{B} \subset \mathcal{K}, \tag{13}$$

where $\mathcal{B}$ is the unit ball in the space of $\mathcal{K}$.

At this point, it might be helpful to remind readers that the Slater condition (13) differs from the Slater type condition

$$\alpha\mathcal{B} \subset -\{\langle P, \Psi(\xi)\rangle : P \in \mathscr{P}(\Xi)\} + \mathcal{K}. \tag{14}$$

The latter has been widely used in the literature of distributionally robust optimization, see for example Shapiro [26], Xu et al. [35], Zhang et al. [36] and the references therein. It is well known that the Slater condition is stronger than the Slater type condition in that the former implies the latter but not conversely. In particular, the latter may hold in moment problems with equality constraints whereas the former does not.

For the given P_0 in Assumption 1, let

$$\Delta := \max_{P \in \mathscr{P}(\Xi)} \mathsf{dl}_{\mathscr{G}}(P, P_0). \tag{15}$$

Following our discussions in Section 2, it is easy to figure out a bound for Δ when the ζ-metric takes a specific form. For instance, Δ is bounded by 1 under the total variation metric dl_{TV} and the Bounded Lipschitz metric dl_L, by $1/2$ under the uniform (Kolmogorov) metric d_U. In the case when the support set Ξ is bounded, the constant is bounded by $\text{diam}(\Xi)$ under the Kantorovich/ Wasserstein metric and the Fortet–Mourier metric.

The theorem below states that the distance between Q and $\mathcal{C}$ under ζ-metric is linearly bounded by the residual of the moment system.

Theorem 2 (**Hoffman's Lemma**). *Let Δ be defined as in (15). Suppose that Assumption 1 holds. Then*

$$\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}) \le C \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi)\rangle\| \tag{16}$$

for any $Q \in \mathscr{P}(\Xi)$, where $C := \frac{\Delta}{\alpha}$ and α is the positive constant defined in the Slater condition (Assumption 1).

Proof. Let $\rho_Q := \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi)\rangle\|$. Since $\mathcal{K} \neq \emptyset$, $\rho_Q < \infty$. Define

$$\overline{Q} := \left(1 - \frac{\rho_Q}{\rho_Q + \alpha}\right) Q + \frac{\rho_Q}{\rho_Q + \alpha} P_0.$$

Obviously $\overline{Q} \in \mathscr{P}(\Xi)$. Moreover, by the definition of ρ_Q, for any $\epsilon > 0$, there exists $W \in \mathcal{K} - \langle Q, \Psi(\xi)\rangle$ such that $\|W\| \le \rho_Q + \epsilon$. Thus $(\rho_Q + \epsilon)^{-1} W \in \mathcal{B}$, the unit ball in the space where $\mathcal{K}$ is defined. The Slater condition (13) ensures

$$-\alpha(\rho_Q + \epsilon)^{-1} W \in \mathcal{K} - \langle P_0, \Psi(\xi)\rangle.$$

The inclusion, and the fact that $W \in \mathcal{K} - \langle Q, \Psi(\xi)\rangle$ and $\mathcal{K}$ is convex, give rise to

$$\mathcal{K} - \langle \overline{Q}, \Psi(\xi)\rangle \ni \left(1 - \frac{\rho_Q}{\rho_Q + \alpha}\right) W - \frac{\rho_Q}{\rho_Q + \alpha} \alpha (\rho_Q + \epsilon)^{-1} W = \frac{\alpha \epsilon}{(\rho_Q + \epsilon)(\rho_Q + \alpha)} W.$$

Driving ϵ to zero, we arrive at $0 \in \mathcal{K} - \langle \overline{Q}, \Psi(\xi)\rangle$, which means $\overline{Q} \in \mathcal{C}$. Subsequently,

$$\begin{aligned}
\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}) \le \mathsf{dl}_{\mathscr{G}}(Q, \overline{Q}) &= \sup_{g \in \mathscr{G}} \left\{ \langle Q, g\rangle - \left\langle \left(1 - \frac{\rho_Q}{\rho_Q + \alpha}\right) Q + \frac{\rho_Q}{\rho_Q + \alpha} P_0, g \right\rangle \right\} \\
&= \frac{\rho_Q}{\rho_Q + \alpha} \sup_{g \in \mathscr{G}} \{\langle Q, g\rangle - \langle P_0, g\rangle\} \\
&\le \frac{\Delta}{\alpha} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi)\rangle\|.
\end{aligned}$$

The last inequality is derived by replacing $\rho_Q + \alpha$ in the denominator with α, the definition of the ζ-metric in (3), and the definitions of Δ and ρ_Q. The proof is complete. □

It might be helpful to make a few comments about Theorem 2. Through Δ, the constant C in (16) depends on the ζ-metric whereas the residual error $\inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi)\rangle\|$ does not. For instance, C equals to $1/\alpha$ under total variation metric dl_{TV} and the Bounded Lipschitz metric dl_L, and $1/2\alpha$ under uniform (Kolmogorov) metric dl_U. If the support set Ξ is bounded, the constant is bounded by $\mathrm{diam}(\Xi)/\alpha$ under the Kantorovich/ Wasserstein metric and the Fortet–Mourier metric dl_{FM}.

Hoffman's lemma for the moment problem is first established by Sun and Xu [31] for classical moment problems with equality and inequality constraints and matrix moment constraints. The authors employ the total variation metric for characterizing the distance of probability measures and use the result for stability analysis of a general class of one stage distributionally robust optimization and equilibrium problems with moment constraints, see Sun and Xu [31] for details. Zhang et al. [36] extend the discussion to a general cone constrained moment system defined as in (32) below. In both works, the proof of Hoffman's lemma depends on the total variation metric in that it requires some delicate reformulation of the distance as a minimax linear programming problem through Lagrange duality. It is unclear whether or not similar results can be established under other metrics such as Kantorovich/ Wasserstein metric and Fortet–Mourier metric. Moreover, the total variation metric has its limitation particularly when it is used to measure the discrete approximation of a continuous probability measure because the distance is always equal to 2.

Theorem 2 is partly motivated to address the challenge. It turns out that by adopting Robinson [22], we are able to derive Hoffman's lemma for the moment system (see (11) directly without resorting to Lagrange duality as in Sun and Xu [31], Zhang et al. [36]. This is primarily because under the Slater condition, we can explicitly find a probability measure $\bar{Q}$ in the ambiguity set whose distance from P can be described through (3).

On one hand, this allows us to establish a new version of Hoffman's lemma under generic ζ-metric and this will serve us for the rest of the discussion on discrete approximation in this paper. On the other hand, as we can see from the proof, Theorem 2 requires stronger conditions, that is, the Slater condition (13) as opposed to the Slater type conditions in Sun and Xu [31], Zhang et al. [36], which means our new result cannot be applied to a moment problem with equality constraints. Moreover, we implicitly require boundedness of the support set Ξ of ξ to ensure boundedness of Δ in order for the result to be valid for the Kantorovich/ Wasserstein metric and Fortet-Mourier metric. Another important difference between this new Hoffman's lemma and the earlier ones in Sun and Xu [31], Zhang et al. [36] is that no explicit assumption on the weak compactness of the ambiguity set $\mathcal{C}$ is needed. This allows us to lift some integrability conditions on $\Phi(\xi)$ such as Assumption 2.1 in Zhang et al. [36].

In what follows, we explain how the constant C in the error bound (16) can be figured out in some concrete moment problems and show the limitations of Theorem 2.

Example 3 (Moment system due to Delage and Ye [5], So [30]). Consider the ambiguity set

$$\mathcal{C} := \left\{ P \in \mathscr{P}(\Xi) : \quad \begin{array}{l} \mathbb{E}_P[\xi - \mu_0]^T \Sigma_0^{-1} \mathbb{E}_P[\xi - \mu_0] \leq \gamma_1 \\ \mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}, \tag{17}$$

where γ_1 and γ_2 are nonnegative constants, μ_0 and Σ_0 are the sample mean and sample covariance. The ambiguity has been first considered by Delage and Ye [5] and further studied by So [30]. We may reformulate $\mathcal{C}$ in the form of (2) by employing Schur complement with

$$\Psi(\xi) := \left(\begin{array}{c} \left[\begin{array}{cc} -\Sigma_0 & \mu_0 - \xi \\ (\mu_0 - \xi)^T & -\gamma_1 \end{array} \right] \\ (\xi - \mu_0)(\xi - \mu_0)^T - \gamma_2 \Sigma_0 \end{array} \right),$$

$\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$ where $\mathcal{K}_1 = \mathcal{S}_-^{k+1}$ and $\mathcal{K}_2 = \mathcal{S}_-^k$, k is the dimension of ξ. Let $\gamma_1 > 0$, $\gamma_2 > 1$ and P_0 be the empirical probability measure. Then (17) satisfies the Slater constraint qualification with $\alpha = \min\{\gamma_1, (\gamma_2 - 1)\lambda_k, \lambda_k\}$, where λ_k denotes the smallest eigenvalue of Σ_0. Subsequently, the constant modulus of (16) is

$$C = \frac{\Delta}{\min\{\gamma_1, (\gamma_2 - 1)\lambda_k, \lambda_k\}}$$

and the corresponding norm of the residual is sum of spectral norm $\|\cdot\|_2$ on $\mathcal{S}^k$ and $\mathcal{S}^{k+1}$.

Example 4 (Variation of the moment system (17)). Consider the following ambiguity set

$$\mathcal{C} = \left\{ P \in \mathscr{P}(\Xi) : \quad \begin{array}{l} |\,\mathbb{E}_P[\xi - \mu_0]| \le \gamma_1 \\ \|\,\mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0\|_2 \le \gamma_2 \end{array} \right\},$$

where γ_1 and γ_2 are small positive numbers, μ_0 and Σ_0 are the sample mean and sample covariance, $|a|$ denotes the absolute value of a vector a with the absolute value taken componentwise. Using the property of the norm, we can reformulate Ψ in the form of (2) with

$$\Psi(\xi) = \left(\begin{array}{l} \xi - \mu_0 - \gamma_1 \\ \mu_0 - \xi - \gamma_1 \\ (\xi - \mu_0)(\xi - \mu_0)^T - \Sigma_0 - \gamma_2 I \\ -(\xi - \mu_0)(\xi - \mu_0)^T + \Sigma_0 - \gamma_2 I \end{array} \right)$$

and $\mathcal{K} = \mathbb{R}_-^k \times \mathbb{R}_-^k \times \mathcal{S}_-^k \times \mathcal{S}_-^k$, where k is the dimenstion of random variable ξ. If $\gamma_1 > 0$ and $\gamma_2 > 0$, Slater condition holds with $\alpha = \min\{\gamma_1, \gamma_2\}$ and then the constant modulus of (16) is

$$C = \frac{\Delta}{\min\{\gamma_1, \gamma_2\}}.$$

Similarly, the norm of the residual is the sum of L_1-norm on $\mathbb{R}^{2k}$ and two spectral norms on $\mathcal{S}^k$.

Example 5 (Moment system due to Liu et al. [13]). Let

$$\mathcal{C} := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{l} |\,\mathbb{E}_P[\xi - \mu_0]| \le \gamma_1, \\ \|\,\mathbb{E}_P[(\xi - \mu_0)(\xi - \mu_0)^T] - \Sigma_0\|_{\max} \le \gamma_2 \end{array} \right\},$$

where $\|A\|_{\max} = \max |a_{ij}|$. It is easy to verify that $\|\cdot\|_{\max}$ is a norm for the matrix but without the sub-multiplicative property. The ambiguity set has been considered in Liu et al.

[13]. Let k be the dimension of random vector ξ, $q = \frac{1}{2}(k^2 + 3k)$, $\psi_I(\xi) = \xi - \bar{\mu}$ and $\psi_J(\xi)$ denote the elements of the upper triangular of matrix $(\xi - \mu_0)(\xi - \mu_0)^T - \Sigma_0$. We may reformulate $\mathcal{C}$ in the form of (2) with

$$\Psi(\cdot) = \begin{pmatrix} \Psi_I(\cdot) - \gamma_1 \\ -\Psi_I(\cdot) - \gamma_1 \\ \Psi_J(\cdot) - \gamma_2 \\ -\Psi_J(\cdot) - \gamma_2 \end{pmatrix}$$

and $\mathcal{K} = \mathbb{R}_{-}^{\frac{k^2+3k}{2}}$. Analogously to Example 4, the Slater condition is satisfied when $\gamma_1 > 0$ and $\gamma_2 > 0$. The constant modulus of (16) is

$$C = \frac{\Delta}{\min\{\gamma_1, \gamma_2\}}$$

and the norm of the residual is the L_1-norm in the space of $\mathbb{R}^{\frac{k^2+3k}{2}}$.

Note that there is a limitation on the application of the established Hoffman's lemma. Consider for example an ambiguity set

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\mathbb{R}^{m_1} \times \mathbb{R}^{m_2})\colon\ \mathbb{E}_P[\mathbf{A}\xi + \mathbf{B}\tilde{\xi}] = \mathbf{b}, P\{(\xi, \tilde{\xi}) \in \mathbf{\Xi_i}\} \in [\underline{\mathbf{p}_i}, \overline{\mathbf{p}_i}], i \in \mathbf{I} \right\}, \tag{18}$$

where P represents a joint probability distribution of the random vector $\xi \in \mathbb{R}^{m_1}$ and some auxiliary random vectors $\tilde{\xi} \in \mathbb{R}^{m_2}$, $\mathbf{A} \in \mathbb{R}^{k \times m_1}$, $\mathbf{B} \in \mathbb{R}^{k \times m_2}$, $\mathbf{I} = \{1, \cdots, I\}$

$$\mathbf{\Xi_i} = \{(\xi, \tilde{\xi}) : \mathbf{C}_i\xi + \mathbf{D}_i\tilde{\xi} \preceq_{\mathcal{K}_i} \mathbf{c}_i\}$$

with $\mathbf{C}_i \in \mathbb{R}^{l_i \times m_1}$, $\mathbf{D}_i \in \mathbb{R}^{l_i \times m_1}$, $\mathbf{c}_i \in \mathbb{R}^{l_i}$, $\mathcal{K}_i$ being proper cone and $y'' \preceq_{\mathcal{K}_i} y'$ means $y' - y'' \in \mathcal{K}_i$. The ambiguity set was first considered by Wiesemann, Kuhn and Sim [33]. Unfortunately, Lemma 2 is not applicable here as the moment system contains equality constraints and it remains an open question if similar error bounds can be established for the moment system. On a positive footnote, the DRO problems with the ambiguity set defined as such can be reformulated as a numerical tractable optimization problem (see [33]) which does not require discretization approach as what Lemma 2 is aimed for in the later section.

A direct application of Hoffman's lemma is to the case when the ambiguity set is defined by a parametric moment system

$$\mathcal{C}_t := \{P \in \mathscr{P}(\Xi)\colon\ \mathbb{E}_P[\Psi_t(\xi)] \in \mathcal{K}\}, \tag{19}$$

where t is a parameter in Banach space and $\Psi_0(\xi) = \Psi(\xi)$. In practice, the parameter t may represent some statistical quantities of ξ such as mean value and standard deviation calculated through samples, for instance, in Example 4, the true mean value μ_0 and covariance matrix Σ_0 may be unknown but it is possible to obtain an estimate of them through empirical data. It is of both theoretical and computational interest to see how changing these estimates affects the ambiguity set; see Delage and Ye [5], Sun and Xu [31], Zhang et al. [36] for more comprehensive discussions.

Corollary 6 (**Application of the Hoffman's lemma to parametric moment systems**). *Suppose that*

(a) Assumption 1 holds for the system (19) with $t = 0$, that is, there exists a positive number α and $P_0 \in \mathscr{P}(\Xi)$ such that (13) holds,

(b) there exist positive constants ρ_0, L and a measurable function $k(\xi)$ such that

$$\|\Psi_{t_1}(\xi) - \Psi_{t_2}(\xi)\| \leq k(\xi)\|t_1 - t_2\|,$$

for all t_1, t_2 with $\|t_i\| \leq \rho_0$ for $i = 1, 2$ and $\langle Q, \kappa(\xi)\rangle \leq L$ for all $Q \in \mathcal{C}_t$ with $\|t\| \leq \rho_0$.

Then there exist positive numbers $\tilde{\alpha} < \alpha$ and $\rho^* \leq \rho_0$ such that

$$\mathbb{H}(\mathcal{C}_{t_1}, \mathcal{C}_{t_2}; \mathsf{dl}_{\mathscr{G}}) \leq \frac{\Delta}{\tilde{\alpha}} L\|t_1 - t_2\| \quad \textit{for all } t_1,\ t_2 \textit{ with } \|t_1\|,\ \|t_2\| \leq \rho^*,$$

where Δ is defined in (15).

Proof. Observe first that under condition (b),

$$\langle P_0, \Psi_t(\xi)\rangle \leq \langle P_0, \Psi_0(\xi)\rangle + \|t\|\langle P_0, \kappa(\xi)\rangle$$

and $\langle P_0, \kappa(\xi)\rangle \leq L$. Thus we can set ρ^* sufficiently small such that for any $t \in B(0, \rho^*)$, the ball centered at 0 with radius ρ^*, the moment system (19) satisfies the Slater condition

$$\langle P_0, \Psi_t(\xi)\rangle + \tilde{\alpha}\mathcal{B} \subset \mathcal{K}.$$

With the Slater condition above, we can apply Theorem 2 to the parametric moment system, that is, for any $t \in B(0, \rho^*)$ and $P \in \mathscr{P}(\Xi)$,

$$\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}_t) \leq \frac{\Delta}{\tilde{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi_t(\xi)\rangle\|.$$

Let $t_1, t_2 \in B(0, \rho^*)$. For any $Q \in \mathcal{C}_{t_1}$,

$$\begin{aligned}\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}_{t_2}) &\leq \frac{\Delta}{\tilde{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi_{t_2}(\xi)\rangle\| \\ &\leq \frac{\Delta}{\tilde{\alpha}} \|\langle Q, \Psi_{t_1}(\xi)\rangle - \langle Q, \Psi_{t_2}(\xi)\rangle\| \leq \frac{\Delta}{\tilde{\alpha}} L\|t_1 - t_2\|,\end{aligned}$$

where the second inequality follows from $\langle Q, \Psi_{t_1}(\xi)\rangle \in \mathcal{K}$ and the last inequality follows from condition (b) of the corollary. Exchanging the position between t_1 and t_2 we deduce

$$\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}_{t_1}) \le \frac{\Delta}{\tilde{\alpha}} L\|t_1 - t_2\| \quad \text{for all } Q \in \mathcal{C}_{t_2}.$$

The rest follows from the definition of $\mathbb{H}(\cdot, \cdot, \mathsf{dl}_{\mathscr{G}})$. □

Corollary 6 gives an error bound on change of of ambiguity set $\mathcal{C}_t$ against perturbation of parameter value t. Compared to a similar result, i.e., Zhang et al. [36, Theorem 2.1], our new error bound is established for all metrics with ζ-structure. Of course, our results are derived under the Slater condition rather than Slater type conditions, which means that they are not necessarily applicable to moment problems with equality constraints.

3.2 Discrete approximation of the ambiguity set under Kantorovich metric

With the new Hoffman's lemma, we are ready to discuss quantification of the difference between the ambiguity sets $\mathcal{C}$ and $\mathcal{C}_N$. Here we give a sketch of ideas. Since $\mathcal{C}_N \subset \mathcal{C}$, it suffices to estimate the deviation of $\mathcal{C}$ from $\mathcal{C}_N$, that is, to estimate the distance (metric) from any $P \in \mathcal{C}$ to $\mathcal{C}_N$. We proceed the discussion in two steps:

(a) estimate the distance from $P \in \mathscr{P}(\Xi)$ to $\mathscr{P}\left(\Xi^N\right)$ and

(b) estimate the distance from a point in $\mathscr{P}\left(\Xi^N\right)$ to $\mathcal{C}_N$.

The distance from P to $\mathcal{C}_N$ is then bounded by the sum of the two distances described above through triangle inequality.

The quantification is not possible under generic ζ-metric because the total variation metric between a continuous probability measure and a discrete probability measure over Ξ is always equal to 1 (see, e.g., Gibbs and Su [7]). Thus, we restrict our discussion to the Kantorovich/ Wasserstein metric.

Recall that in the literature of probability theory, the Kantorovich/ Wasserstein metric is also defined as

$$\mathsf{dl}_K(P,Q)^r = \inf_{\pi \in \Pi} \left\{ \iint_{\Xi\times\Xi} d(\xi,\xi')^r \pi(d\xi, d\xi') \,\middle|\, \begin{array}{l} P(A) = \pi(A \times \Xi'),\ \forall A \in \mathscr{B}, \\ Q(B) = \pi(\Xi \times B),\ \forall B \in \mathscr{B}' \end{array} \right\} \tag{20}$$

where Π denotes the set of all probability measures in the space $(\Xi, \mathscr{B}) \times (\Xi', \mathscr{B}')$, $r \ge 1$ and d is a metric on Ξ, which is usually assumed to be induced by the Euclidean norm.

By the Kantorovich–Rubinstein theorem, the collection of all Lipschitz-1 functions (cf. (5)) induces the Kantorovich/ Wasserstein metric dl_K whenever $r = 1$; see Rachev [21] for this duality relation. A nice interpretation of formulation (20) is Kantorovich's representation of Monge's transportation problem where $\mathsf{dl}_K(P,Q)$ represents the minimal cost of transference of goods spread over Ξ under distribution P to that of Q.

The formulation (20) highlights an important property of the Wasserstein distance. Indeed, the problem formulation (20) is linear in π, and thus the complete machinery from linear programming is available to compute or investigate the distance. The Kantorovich–Rubinstein theorem provides the dual formulation, which is linear as well. This setting is perfectly adapted to stochastic optimization problems, as the bounds obtained are tight, which is particularly important for numerical approximations. This is in significant contrast to other metrics (as the Prokhorov metric), which metrize weak convergence as well.

Discrete approximations

Let $\Xi^N := \{\xi_1, \dots, \xi_N\}$ be a subset of Ξ and set

$$\beta_N := \max_{\xi \in \Xi} \min_{1 \le i \le N} d(\xi, \xi_i). \tag{21}$$

Since $\Xi^N \subset \Xi$, it is easy to see that β_N is indeed the Hausdorff distance between Ξ and Ξ^N.

For a given $\Xi^N = \{\xi_1, \dots, \xi_N\}$, let $\{\Xi_1, \dots, \Xi_N\}$ be a Voronoi tessellation of Ξ, that is,

$$\Xi_i \subseteq \left\{ y \in \Xi : \|y - \xi^i\| = \min_k \|y - \xi^k\| \right\} \qquad \text{for} \quad i = 1, \dots, N$$

are pairwise disjoint subsets forming a partition of Ξ. For a fixed $P \in \mathscr{P}(\Xi)$, let $p_i = P(\Xi_i)$ for $i = 1, \dots, N$ and define

$$P_N(\cdot) := \sum_{i=1}^N p_i \, \delta_{\xi_i}(\cdot). \tag{22}$$

The following result provides an upper bound for the discrete approximation.

Proposition 7 (cf. Pflug and Pichler [18, Lemma 4.9]). *Let $P \in \mathscr{P}(\Xi)$ be fixed and P_N be defined as in (22). Then*

$$\mathsf{dl}_K(P, P_N) = \int \min_{1 \le i \le N} d(\xi, \xi_i) P(d\xi) = \sum_{i=1}^N \int_{\Xi_i} d(\xi, \xi_i) P(d\xi) \le \beta_N. \tag{23}$$

In what follows, we call P_N defined by (22) the Voronoi projection of the probability measure P on space $\mathscr{P}(\Xi^N)$. P_N converges to P under Kantorovich metric when β_N tends to zero.

Remark 8 (Quantizers). The centers $\xi^1, \dots, \xi^N$ may be samples or selected in a deterministic manner. In either case, we are able to estimate the rate of convergence for β_N.

Note that finding the best locations of $\xi^1, \dots, \xi^N$ is a *facility location* problem which is non-convex, non-linear and NP hard. However, Dudley [6, p. 42] establishes a tight bound for $\mathsf{dl}_K(P, P_N)$ as follows:

$$\mathsf{dl}_K(P, P_N) \sim N^{-1/n},$$

where n is the dimension of the state space, i.e., $\xi^i \in \mathbb{R}^n$. This bound cannot be improved asymptotically as N goes to infinity. Graf and Luschgy [8, Section 7.2] gave a description on the density of the optimal quantization points $\xi^1, \dots, \xi^N$.

There are other schemes which provide sub-optimal locations (i.e., Graf and Luschgy [8, Example 4.17] establish the error bound for equi-partition case). Convergence is guaranteed as long as the diameter of the largest ball tends to 0. Specific tessellations such as power diagrams and multiplicatively weighted Voronoi diagrams are therefore reasonable choices with lower computational costs.

We shall investigate the case of randomly chosen quantization points further. For example, if $\xi^1, \dots, \xi^N$ is an independent and identically distributed (iid) sample, then we are able to employ a large deviation theorem to establish an exponential rate of convergence as stated in the proposition below.

Proposition 9. *Let $\xi_1, \dots, \xi_N$ be iid copies of ξ. Assume:*

(a) Ξ is bounded and

(b) the true probability distribution of ξ is continuous and there exist positive constants C, ν and δ_0 such that
$$P(\|\xi - \xi_0\| \le \delta) > C\delta^\nu$$
for any fixed point $\xi_0 \in \Xi$ and $\delta \in (0, \delta_0)$.

Then for any small number $\epsilon > 0$, there exist positive constants $\beta(\epsilon)$ and $C(\epsilon)$ depending on ϵ such that
$$\text{Prob}(\beta_N \ge \epsilon) \le C(\epsilon)e^{-\beta(\epsilon)N} \tag{24}$$
when N is sufficiently large. Here the probability measure "Prob" is understood as the product of the true (unknown) probability measure of P over the measurable space $\Xi \times \Xi \times \dots$ with product Borel sigma-algebra $\mathscr{B} \times \mathscr{B} \times \dots$.

Proof. We use Xu et al. [35, Lemma 3.1] to prove the result with $G(x, \xi) = -\|x - \xi\|$. It suffices to verify the conditions there. Condition (a) is satisfied when Ξ is bounded. The so-called tail behaviour condition is guaranteed by our condition (b) through Anderson et al. [2, Proposition 1]. The rest follows from Xu et al. [35, Lemma 3.1]. □

Remark 10. It might be interesting to note that using the samples generated by the true probability distribution P as we described above is not the most efficient way to reduce β_N; Graf and Luschgy [8, Section 7.2] describe the most efficient distribution.

Further, if $\xi^1, \dots, \xi^N$ are generated by a uniform distribution over the support Ξ, denoted by P', then β_N follows an extreme value distribution (a Gumbel distribution) with
$$\lim_{N\to\infty} \frac{N(2\beta_N)^d - \log N}{\log\log N} = d - 1 \quad \text{with probability 1;}$$

see Zhigljavsky and Žilinskas [38, Section 2.2] and the details on the theory of maximum spacing. This means that β_N depends on the dimension of ξ. In the case when ξ has several components, we might need a large N to get a moderate bound for β_N.

We now move on to estimate the distance between $P \in \mathscr{P}(\Xi^N)$ to $\mathcal{C}_N$. This amounts to Hoffman's lemma for the discretized moment system (12).

Corollary 11 (**Hoffman's lemma for the discrete moment problem** (12))**.** *Let β_N be defined as in (21). Suppose:*

(a) Assumption 1 holds,

(b) Ξ is bounded and β_N tends to zero as $N \to \infty$, and

(c) $\Psi(\cdot)$ is continuous and bounded on Ξ.

Then for any positive number $\hat{\alpha}$, there exists a positive integer N_0 such that when $N \geq N_0$

$$\mathsf{dl}_K(Q, \mathcal{C}_N) \leq \frac{\delta_\Xi}{\hat{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi) \rangle\| \quad \textit{for all } Q \in \mathscr{P}(\Xi^N)$$

where δ_Ξ denotes the diameter of Ξ.

Proof. Let P_N be the Voronoi projection of P_0 where P_0 is the probability measure satisfying the Slater condition (13). Since β_N converges to zero as N tends to infinity, it follows by (23), $\mathsf{dl}_K(P_N, P_0) \to 0$. The latter implies that P_N converges to P_0 weakly because the Kantorovich metric metrizes weak convergence. Moreover, under the Slater condition (13), there exists a sufficiently large N^* such that for $N \geq N^*$,

$$\langle P_N, \Psi(\xi) \rangle + \hat{\alpha}\mathcal{B} \subset \mathcal{K}, \tag{25}$$

which means that the system (12) satisfies the Slater condition when N is sufficiently large. By Theorem 2, for any $Q \in \mathscr{P}(\Xi^N)$

$$\mathsf{dl}_K(Q, \mathcal{C}_N) \leq \frac{\Delta}{\hat{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi) \rangle\| \leq \frac{\delta_\Xi}{\hat{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle Q, \Psi(\xi) \rangle\|,$$

where Δ is defined in (15) (with $\mathscr{G}$ being defined by (5)) and δ_Ξ is the diameter of Ξ. This completes the proof. □

With Proposition 7 and Corollary 11 we are ready to present our main result in this section which quantifies the approximation of $\mathcal{C}_N$ to $\mathcal{C}$ under Kantorovich metric.

Theorem 12 (Quantification of discrete approximation of the ambiguity set)**.** *Suppose:*

(a) Assumption 1 holds,

(b) Ξ is bounded and β_N tends to zero as $N \to \infty$ and

(c) each component of $\Psi(\cdot)$ is Lipschitz continuous on Ξ with Lipschitz modulus L.

Then for N sufficiently large

$$\mathbb{H}_K(\mathcal{C}_N, \mathcal{C}; \mathsf{dl}_K) \leq \left(1 + \frac{L\rho\delta_\Xi}{\hat{\alpha}}\right)\beta_N, \tag{26}$$

where β_N is defined in (21), δ_Ξ denotes the diameter of Ξ, $\rho = \|\mathbf{E}\|$ and $\mathbf{E}$ denotes a matrix of size $\Psi(\xi)$ with each component being 1.

Proof. By the definition of $\mathcal{C}$ and $\mathcal{C}_N$, $\mathcal{C}_N \subset \mathcal{C}$ in that $\Xi^N \subset \Xi$. It is sufficient to show (26) for $\mathbb{D}(\mathcal{C}, \mathcal{C}_N; \mathsf{dl}_K)$. For fixed $P \in \mathcal{C}$, let P_N be the Voronoi projection of P. If $P_N \in \mathcal{C}_N$, then

$$\mathsf{dl}_K(P, \mathcal{C}_N) \leq \mathsf{dl}_K(P, P_N) \leq \beta_N, \tag{27}$$

where the second inequality follows from (23). Thus, we are left to the case with that $P_N \notin \mathcal{C}_N$. Let $Q_N \in \arg\min \mathsf{dl}_K(P_N, \mathcal{C}_N)$. Existence of Q_N is due to the fact that $\mathcal{C}_N$ is a compact set. By the definition of Q_N, P_N and the triangle inequality of the Kantorovich metric,

$$\mathsf{dl}_K(P, \mathcal{C}_N) \leq \mathsf{dl}_K(P, Qd_N) \leq \mathsf{dl}_K(P, P_N) + \mathsf{dl}_K(P_N, Q_N) = \mathsf{dl}_K(P, P_N) + \mathsf{dl}_K(P_N, \mathcal{C}_N). \tag{28}$$

From (27), $\mathsf{dl}_K(P, P_N)$ is bounded by β_N. On the other hand, by Corollary 11,

$$\mathsf{dl}_K(P_N, \mathcal{C}_N) \leq \frac{\delta_\Xi}{\hat{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle P_N, \Psi(\xi)\rangle\| \tag{29}$$

for N sufficiently large (such that (25) holds). Combining (28) and (29) and observing that $\langle P, \Psi(\xi)\rangle \in \mathcal{K}$, we have

$$\begin{aligned}\mathsf{dl}_K(P, \mathcal{C}_N) &\leq \beta_N + \frac{\delta_\Xi}{\hat{\alpha}} \inf_{K \in \mathcal{K}} \|K - \langle P_N, \Psi(\xi)\rangle\| \\ &\leq \beta_N + \frac{\delta_\Xi}{\hat{\alpha}} \|\langle P, \Psi(\xi)\rangle - \langle P_N, \Psi(\xi)\rangle\|. \end{aligned}\tag{30}$$

Moreover, under condition (c), every component $\Psi_{i,j}$ of Ψ is Lipschitz continuous on Ξ with modulus being bounded by L, which means $\Psi_{i,j}/L$ is Lipschitz continuous with modulus bounded by 1. By the definition of Kantorovich metric (cf. (3) and (5)), we obtain

$$\langle Q, \Psi_{i,j}(\xi)/L\rangle - \langle P, \Psi_{i,j}(\xi)/L\rangle \leq \mathsf{dl}_K(Q, P),$$

hence

$$\|\langle P_N, \Psi(\xi)\rangle - \langle P, \Psi(\xi)\rangle\| \leq L\rho\, \mathsf{dl}_K(P_N, P) \leq L\rho\beta_N, \tag{31}$$

where $\|\cdot\|$ denotes the Frobenius norm. Combining (30) and (31) we arrive at

$$\mathsf{dl}_K(P, \mathcal{C}_N) \leq \beta_N + \frac{L\rho\delta_\Xi}{\hat{\alpha}}\beta_N = \left(1 + \frac{L\rho\delta_\Xi}{\hat{\alpha}}\right)\beta_N.$$

This completes the proof. □

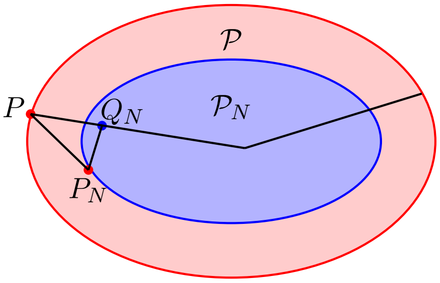

Figure 1: $\mathsf{dl}_K(P, \mathcal{C}_N) \leq \mathsf{dl}_K(P, P_N) + \mathsf{dl}_K(P_N, Q_N)$

Figure 1 gives a geometric interpretation of the relationship between $\mathcal{C}_N$ and $\mathcal{C}$. Clearly the distance between P and $\mathcal{C}_N$ is bounded by $\mathsf{dl}_K(P, Q_N)$. On the other hand, $\mathsf{dl}_K(P, Q_N)$ is bounded by the sum of $\mathsf{dl}_K(P, P_N)$ and $\mathsf{dl}_K(P_N, Q_N)$ which are respectively bounded by β_N and $\frac{\delta_\Xi}{\alpha} \inf_{K \in \mathcal{K}} \|K - \langle P_N, \Psi(\xi)\rangle\|$ through Hoffman's Lemma (Theorem 2).

Remark 13. The significance of Theorem 12 is that it gives a quantitative description for difference between two ambiguity sets $\mathcal{C}_N$ and $\mathcal{C}$ in terms of Kantorovich/ Wasserstein metric and an explicit bound for the discrepancy in terms of Hausdorff distance between Ξ^N and Ξ.

If Ξ^N is constructed in a deterministic way, then we can easily figure out β_N. On the other hand, if Ξ^N is composed of iid samples, then, through Proposition 9, we can establish an exponential rate of convergence for $\mathcal{C}_N \to \mathcal{C}$ under the Kantorovich/ Wasserstein metric.

Theorem 12 may be applied to parametric moment system (19). Under the conditions of Corollary 6, we may establish

$$\mathbb{H}_K(\mathcal{C}_N^{t_N}, \mathcal{C}^{t_0}; \mathsf{dl}_K) \leq \mathbb{H}_K(\mathcal{C}_N^{t_N}, \mathcal{C}^{t_N}; \mathsf{dl}_K) + \mathbb{H}_K(\mathcal{C}^{t_N}, \mathcal{C}^{t_0}; \mathsf{dl}_K) \leq \left(1 + \frac{L\rho\delta_\Xi}{\hat{\alpha}}\right)\beta_N + \frac{\Delta}{\tilde{\alpha}} L\|t_N - t_0\|.$$

Here we write the parameter t_N explicitly to indicate that the parameter may also depend on $\{\xi_1, \cdots, \xi_N\}$. We leave the details to the interested reader.

4 One-stage distributionally robust optimization problem

With quantification of the difference between $\mathcal{C}_N$ and $\mathcal{C}$ in the preceding section, we now move on to investigate how the discrepancy propagate in the resulting one-stage distributionally robust optimization problems in terms of the optimal value.

Let us start by rewriting (1) as

$$\textbf{(DRO)} \quad \min_{x \in X} \max_{P \in \mathcal{C}} \mathbb{E}_P[f(x, \xi)] \tag{32}$$

so that we can focus on probability measures over Ξ. In the literature of distributionally robust optimization, a lot of research has been focused on the case that f, Ψ and $\mathcal{K}$ take a specific structure, and problem (32) is reformulated as a tractable semidefinite programming (SDP) problem, see for instance Delage and Ye [5], Popescu [20], Wiesemann et al. [33], Zymler et al. [40] and references therein. Here we consider general cases without assuming any specific structure of f, Ψ or Ξ. Let $\mathcal{C}_N$ be defined as in (12). We consider a discrete distributionally robust optimization problem

$$\textbf{(DDRO)} \quad \min_{x\in X} \max_{P\in\mathcal{C}_N} \mathbb{E}_P[f(x,\xi)] \tag{33}$$

and regard DDRO (33) as an approximation of DRO (32).

The program DDRO (33) can be written as

$$\textbf{(DDRO')} \quad \begin{array}{ccl} \min\limits_{x\in X} & \max\limits_{(p_1,\ldots,p_N)\in\mathbb{R}^N_+} & \sum_{j=1}^N p_j f(x,\xi^j) \\ & \text{s.t.} & \sum\limits_{j=1}^N p_j \Psi(\xi^j) \in \mathcal{K}, \\ & & \sum_{i=1}^N p_i = 1. \end{array}$$

The latter is a deterministic saddle point problem for which various existing numerical methods can be applied, see Xu et al. [35] and references therein. Our focus here is not on numerical solutions of DDRO (33), instead we are interested in deriving a quantitative description on the difference between DDRO (33) and DRO (32) in terms of the optimal value.

For the purpose of the stability analysis, we need to introduce another metric which is closely related to the objective function $f(x,\xi)$. Let

$$\overline{\mathscr{G}} := \{g(\cdot) := f(x,\cdot) : x \in X\}. \tag{34}$$

For any two probability measures $P, Q \in \mathscr{P}(\Xi)$, let

$$\mathscr{D}(P,Q) := \sup_{g\in\overline{\mathscr{G}}} |\mathbb{E}_P[g] - \mathbb{E}_Q[g]|. \tag{35}$$

Here we implicitly assume that $\mathscr{D}(P,Q) < \infty$. From (35), we can see immediately that $\mathscr{D}(P,Q) = 0$ if and only if

$$\mathbb{E}_P[g] = \mathbb{E}_Q[g] \quad \text{for all } g \in \overline{\mathscr{G}},$$

which means that weak convergence of a sequence of probability measures $\{P_N\}$ to P entails uniform convergence of $\mathbb{E}_{P_N}[f(x,\xi)]$ to $\mathbb{E}_P[f(x,\xi)]$. This kind of distance has been widely used for stability analysis in stochastic programming and it is known as *pseudometric* in that it satisfies all properties of a metric except that $\mathscr{D}(Q,P) = 0$ does not necessarily imply $P = Q$ unless the set of functions $\overline{\mathscr{G}}$ is sufficiently large (in this case, the distance $\mathscr{D}$ is said to be strict). For a comprehensive discussion of the concept and related issues, see Römisch [23, Sections 2.1–2.2]. Obviously the pseudo-metric is a special case of a metric with ζ structure (3).

Employing the pseudometric instead of a metric in (8) one may define a pseudo-distance from a single probability measure $Q \in \mathscr{P}(\Xi)$ to a set of probability measures $\mathcal{A}_1 \subset \mathscr{P}(\Xi)$. This setting generalizes the deviation (excess, cf. (9)) and the Hausdorff (pseudo)distance between two sets of probability measures $\mathcal{A}_1$ and $\mathcal{A}_2$, cf. (10).

We are now ready to state our main stability result.

Theorem 14. *Let ϑ and ϑ_N denote the optimal values of (32) and (33) respectively, and they are attained by the corresponding optimal solutions x^*, P^* and x^N, P^N. Assume the setting and conditions of Theorem 12. Assume further:*

(a) for each fixed x, there exists a positive constant κ independent of x such that for all $x \in X$ it holds

$$|f(x,\xi') - f(x,\xi'')| \leq \kappa \|\xi' - \xi''\| \quad \textit{for all } \xi', \xi'' \in \Xi;$$

(b) X is bounded and for each $x \in X$, $\mathbb{E}_P[f(x,\xi)] < \infty$ for all $P \in \mathcal{C}$.

Then

(i) there exists a positive constant C_1 such that

$$|\vartheta - \vartheta_N| \leq C_1 \beta_N \tag{36}$$

where β_N is defined as in (21).

(ii) If, in addition, $\mathbb{E}_{P^*}[f(\cdot,\xi)]$ satisfied the second order growth condition at point x^*, that is,

$$\mathbb{E}_{P^*}[f(x,\xi)] - \mathbb{E}_{P^*}[f(x^*,\xi)] \geq r\|x - x^*\|^2 \qquad \textit{for all } x \in X, \tag{37}$$

for a positive constant r, then there exists a positive constant C_2 such that

$$\|x^* - x^N\| \leq C_2 \beta_N^{\frac{1}{2}}. \tag{38}$$

Proof. Part (i). Under condition (a), $f(x,\cdot)/\kappa$ is uniformly Lipschitz continuous over Ξ with modulus bounded by 1. By the definition of pseudometric

$$\mathscr{H}(\mathcal{C}_N, \mathcal{C}) \leq \kappa \, \mathbb{H}(\mathcal{C}_N, \mathcal{C}; \mathsf{dl}_K).$$

On the other hand, under the condition of Theorem 12, we have from the theorem

$$\mathscr{H}(\mathcal{C}_N, \mathcal{C}) \leq \kappa \left(1 + \frac{L\rho\delta_\Xi}{\hat{\alpha}}\right) \beta_N. \tag{39}$$

Consequently,

$$\begin{aligned}\vartheta - \vartheta^N &= \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(x^*,\xi)] - \sup_{P\in\mathcal{C}_N} \mathbb{E}_P[f(x^N,\xi)]\\ &\leq \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(x^N,\xi)] - \sup_{P\in\mathcal{C}_N} \mathbb{E}_P[f(x^N,\xi)]\\ &\leq \mathscr{H}(\mathcal{C}_N,\mathcal{C})\\ &\leq \kappa\left(1+\tfrac{L\rho\delta_\Xi}{\hat{\alpha}}\right)\beta_N.\end{aligned}$$

Note that since $\mathcal{C}_N \subset \mathcal{C}$ it holds that $\vartheta - \vartheta_N \geq 0$. This shows (36) with $C_1 = \kappa\left(1+\frac{L\rho\delta_\Xi}{\hat{\alpha}}\right)$.

Part (ii). By definition

$$\begin{aligned}\vartheta - \vartheta_N &= \mathbb{E}_{P^*}[f(x^*,\xi)] - \mathbb{E}_{P^N}[f(x^N,\xi)]\\ &= \mathbb{E}_{P^*}[f(x^*,\xi)] - \mathbb{E}_{P^*}[f(x^N,\xi)] + \mathbb{E}_{P^*}[f(x^N,\xi)] - \mathbb{E}_{P^N}[f(x^N,\xi)]\\ &\leq -r\|x^*-x^N\|^2 + \mathbb{E}_{P^*}[f(x^N,\xi)] - \mathbb{E}_{P^N}[f(x^N,\xi)]\\ &\leq -r\|x^*-x^N\|^2 + \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(x^N,\xi)] - \sup_{P\in\mathcal{C}_N} \mathbb{E}_P[f(x^N,\xi)]\\ &\leq -r\|x^*-x^N\|^2 + \kappa\left(1+\tfrac{L\rho\delta_\Xi}{\hat{\alpha}}\right)\beta_N,\end{aligned}$$

where the first inequality follows from the growth condition (37), the third inequality follows from (39). The last inequality is due to the fact that

$$\begin{aligned}\sup_{P\in\mathcal{C}} \mathbb{E}_P[f(x^N,\xi)] - \sup_{P\in\mathcal{C}_N} \mathbb{E}_P[f(x^N,\xi)] &= \sup_{P\in\mathcal{C}} \inf_{Q\in\mathcal{C}_N} \mathbb{E}_P[f(x^N,\xi)] - \mathbb{E}_Q[f(x^N,\xi)]\\ &\leq \sup_{P\in\mathcal{C}} \inf_{Q\in\mathcal{C}_N} \sup_{x\in X} \mathbb{E}_P[f(x,\xi)] - \mathbb{E}_Q[f(x,\xi)]\\ &\leq \sup_{P\in\mathcal{C}} \inf_{Q\in\mathcal{C}_N} \kappa \mathsf{dl}_K(P,Q)\\ &\leq \mathbb{D}(\mathcal{C},\mathcal{C}_N;\mathsf{dl}_K).\end{aligned}$$

Recall that $\vartheta - \vartheta_N \geq 0$, we have

$$\|x^* - x^N\| \leq \sqrt{\kappa/r\left(1+\frac{L\rho\delta_\Xi}{\hat{\alpha}}\right)\beta_N},$$

which means (38) holds with $C_2 = \sqrt{\kappa/r\left(1+\frac{L\rho\delta_\Xi}{\hat{\alpha}}\right)}$. □

Theorem 14 is a step forward from Xu et al. [35, Theorem 4.2], where the former presents an explicit bound for $|\vartheta - \vartheta_N|$ in terms of β_N, whereas the latter only states that $|\vartheta - \vartheta_N|$ tends to zero as N increases. It also complements the stability results in Sun and Xu [31, Section 4] where the bound established for $|\vartheta - \vartheta_N|$ does not apply to the case when $\mathcal{P}_N$ is a discretization of $\mathcal{P}$. Note that since β_N depends on the dimension of ξ, the error bounds established in this theorem may be coarse when ξ has several components unless N is sufficiently large. In other words, the discretized DRO might suffer from curse of dimensionality [16]. This differs from the well-known sample average approximation method in stochastic programming where the error bounds are independent of the dimension of the underlying random vector, see [29].

Note also that in the statement of Theorem 14 we explicitly assume existence of optimal solutions x^* and P^*. The existence is guaranteed by compactness of X, weak compactness of $\mathcal{C}$ and continuity of the function f in (x, ξ). The assumption can be avoided by using an ϵ-optimal solution argument but the latter will perhaps blur up the key arguments in the proof. We leave interested readers to verify.

5 On the distributionally robust multistage problem

This section extends the discussion to multistage distributionally robust optimization. Multistage stochastic programming has wide applications such as long term financial planning, pension fund management, energy production and trading, supply chain management and inventory control. Compared to one stage and two stage stochastic programming, the mathematical structure of multistage stochastic programming is far more complicated due to its nested structure in the decision making process and hence requires new numerical methods and underlying theory including discrete approximation, scenario reduction and stability analysis (see the recent monograph by Pflug and Pichler [18] for a comprehensive treatment of the topic).

Mathematical approach. The following treatment of the multistage situation follows a similar strategy as for the two stage problem in the previous section. We discuss the general multistage problem formulation first, which we then extend to the ambiguous situation including moment conditions.

Our main result is an upper bound for multistage problem, which is formulated in terms of the Hausdorff distance with respect to the nested distance.

Increasing information, which gets known to the decision maker gradually, is intrinsic to *multistage* stochastic optimization. The evolution of information is modelled by a filtration

$$\mathscr{F} := (\mathcal{F}_0, \dots \mathcal{F}_T)$$

of increasing sigma-algebras, $\mathcal{F}_t \subset \mathcal{F}_{t+1}$; $\mathcal{F}_0 := \{\emptyset, \Omega\}$ is the trivial sigma algebra. The multistage stochastic optimization problem is thus formulated on the filtered probability space $(\Omega, \mathscr{F}, P)$ as

$$\inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \mathbb{E}\, f\big(x(\omega), \xi(\omega)\big), \tag{40}$$

where $\mathbb{X}$ collects all feasible control processes. In particular, the controlling process $\mathbf{x} = (\mathbf{x}_0, \dots \mathbf{x}_T)$ needs to be adapted, i.e., each component $\mathbf{x}_t$ satisfies the *nonanticipativity constraint*

$$\mathbf{x}_t \text{ is measurable with respect to } \mathcal{F}_t.$$

Alternatively, one may also consider a stochastic process $\xi : \Omega \to \Xi$, where

$$\xi = (\xi_0, \dots \xi_T)$$

represents the vector of sequential observations. In this situation the natural filtration is generated by the process ξ itself, that is,

$$\mathcal{F}_t = \sigma(\xi_0, \dots \xi_t). \tag{41}$$

It follows from the Doob–Dynkin Lemma (cf. Kallenberg [11, Lemma 1.13]) that $\mathbf{x}(\cdot)$, which is measurable with respect to the sigma algebra (41), can be expressed as a function of $(\xi_0, \dots, \xi_t)$,

$$\mathbf{x}_t(\omega) = \mathbf{x}_t(\xi_0(\omega), \dots \xi_t(\omega)).$$

The multistage stochastic optimization problem thus reads

$$\inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \mathbb{E}\, f\big[(\mathbf{x}(\xi), \xi)\big] \tag{42}$$

on the state space Ξ.

To formulate a *robust* version of the multistage stochastic optimization problem it is essential to separate the process ξ and the filtration $\mathscr{F}$. To this end we consider the projection

$$\begin{aligned} \pi_t : \Xi &\to \Xi_0 \times \cdots \times \Xi_t, \\ (\xi_0, \dots \xi_T) &\mapsto (\xi_0, \dots \xi_t) \end{aligned} \tag{43}$$

so that $\mathbf{x}_t(\cdot)$ is a function of the coordinates only,

$$\mathbf{x}_t = \mathbf{x}_t(\xi_0, \dots, \xi_t)$$

and the natural filtration is generated by the projections,

$$\mathcal{F}_t := \sigma(\pi_1, \dots, \pi_t) = \mathscr{B}(\Xi_0 \times \cdots \times \Xi_t) \times \Xi_{t+1} \times \cdots \times \Xi_T, \tag{44}$$

where $\mathscr{B}$ is the Borel sigma-algebra.

In the literature of stochastic programming, the true probability measure P is either known or can be approximated through samples. In many practical applications, the true probability distribution of the stochastic process may be unknown at each stage but it is possible to use partial information such as samples and prior moment conditions to construct a set of distributions which either contains the true probability distribution or approximates it with some confidence.

Specifically, we may formulate the robust multistage stochastic optimization problem as

$$\inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{P\in\mathcal{P}} \mathbb{E}_P f\big((\mathbf{x}(\xi), \xi)\big), \tag{45}$$

where $\mathcal{P}$ is a collection of probability measures.

Remark 15 (**Formulation of the robustified, multistage problem**). To formulate a robustified version of the multistage stochastic optimization problem it is notably essential to fix the filtration. Indeed, the robust control policy $\mathbf{x}(\cdot)$ has to remain measurable for every inner model, cf. (45). The particular choice (44) and the formulation (45) ensure, that $\mathbf{x}(\cdot)$ stays measurable in this sense.

Remark 16 (**Discrete measures and tree structures**). Discrete measures have a finite support and are thus often considered together with a filtration consisting of finite sigma algebras. This is indeed the standard numerical implementation of trees (i.e., finite-space and finite-time processes).

However, discrete measures are well-defined even on the power set. Here we consider all measures on the Borel filtration (44) induced by the projection (43). Note that this setting includes tree structures. Indeed, individual elements of the support set are vectors of length T. If any of these vectors have the first t, say, components in common, then they share a common history up to time t. These vectors cannot be distinguished under the filtration $\mathcal{F}_t$ induced by the projection π_t. The tree structure thus is naturally encoded by considering the filtration (44). Even more, the support of a discrete measure and the filtration (44) naturally constitute a tree process.

Moment conditions. To provide an example of a *robust* multistage problem formulation we consider the robust formulation (45) with moment conditions imposed. To this end one specifies the set $\mathcal{P}$ further by

$$\mathcal{P} := \{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\phi_t(\xi)|\mathcal{F}_t] \leq \mu_t \text{ for } t = 1, \dots, T\}, \tag{46}$$

where $\mu_t : \Xi^t \to \mathbb{R}^d$ is a continuous function on Ξ^t, $t = 1, \dots, T$, $\phi : \Xi \to \mathbb{R}^d$ is a continuous function on Ξ and

$$\int_{A_t} \mathbb{E}_P[\phi_t(\xi)|\mathcal{F}_t]dP = \int_{A_t} \phi_t(\xi)dP \leq \int_{A_t} \mu_t(\xi^t)dP \text{ for all } A_t \in \mathcal{F}_t,\ t = 1, \dots, T. \tag{47}$$

For a discrete probability measure P, in particular, the latter moment conditions (47) are equivalent to

$$\int_{\{\omega:\xi^t(\omega)=\xi^t\}} \mathbb{E}_P[\phi_t(\xi)|\mathcal{F}_t]dP = \int_{\{\omega:\xi^t(\omega)=\xi^t\}} \phi_t(\xi^t, \xi_{t+1}(\omega), \dots, \xi_T(\omega))P(d\omega)$$

and consequently

$$\int_{\{\xi^t(\cdot)=\xi^t\}} \phi\left(\xi^t, \xi_{t+1}(\omega), \dots, \xi_T(\omega)\right) P(d\omega) \leq \mu_t(\xi^t) \cdot P\left(\xi^t(\cdot) = \xi^t\right) \text{ for } t = 1, \dots, T.$$

The moment condition means that the conditional expected value of $\phi_t(\xi)$ at stage t does not exceed $\mu(\xi^t)$, a quantity which depends only on the realization of ξ^t up to stage t.

If we regard ξ_t as a loss and set $\phi_t(\xi) := \frac{1}{t+1}\sum_{\tau=1}^{t+1}\xi_\tau$ and $\mu(\xi^t) := \frac{C}{t}\sum_{\tau=1}^{t}\xi_\tau$, then the condition means that the average expected loss up to stage $t+1$ does not exceed C times the average expected loss up to stage t. In other words, we are looking at events (corresponding to probability distributions) where at each stage the expected loss of reference index in future falls within the range of the observed average up to the stage.

Remark 17 (Example). An example of a moment condition introduced in (46) is an extension of the classical Markowitz model for portfolio optimization; this model minimizes the risk, given that a certain return is to be achieved. The problem formulation (46) includes formulations which require a minimal return at each intermediate stage.

In what follows we investigate the multistage, distributionally robust optimization problem (45) (MDRO) further.

Definition 18 (Nested distance)**.** Let $\mathbb{P} := (\Xi, \mathscr{F}, P)$ and $\tilde{\mathbb{P}} := (\tilde{\Xi}, \tilde{\mathscr{F}}, \tilde{P})$ be filtered probability spaces. The nested distance of order $r \geq 1$, denoted by $\mathsf{dl}_r\left(\mathbb{P}, \tilde{\mathbb{P}}\right)$, is defined as

$$\mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}}) = \left(\inf_\pi \iint_{\Xi\times\tilde{\Xi}} d\left(\xi, \tilde{\xi}\right)^r \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi})\right)^{1/r},$$

where π is a probability measure defined over space $\Xi \times \tilde{\Xi}$ with *conditional* marginals P and $\tilde{P}$, i.e.,

$$\begin{aligned} P(A|\,\mathcal{F}_t) &= \pi(A \times \tilde{\Xi}|\,\mathcal{F}_t \times \tilde{\mathcal{F}}_t) && \text{for all } A \in \mathcal{F}_T \text{ and} \\ \tilde{P}(B|\,\tilde{\mathcal{F}}_t) &= \pi(\Xi \times B|\,\mathcal{F}_t \times \tilde{\mathcal{F}}_t) && \text{for all } B \in \tilde{\mathcal{F}}_T. \end{aligned}$$

From the definition, we see that a positive nested distance is influenced by different filtrations $\mathscr{F}$ and $\tilde{\mathscr{F}}$.

Note that for fixed $\mathscr{F}$, each pair of $P \in \mathcal{C}$ and $\tilde{P} \in \tilde{\mathcal{C}}$ from two sets of reference probability measures ($\mathcal{C}$ and $\tilde{\mathcal{C}}$) gives rise to a value of the nested distance $\mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})$. We can imagine there are two sets of processes induced by $\mathcal{C}$ and $\tilde{\mathcal{C}}$.

What we are interested in is

$$\mathbb{D}(\tilde{\mathcal{C}}, \mathcal{C}; \mathsf{dl}_r) := \sup_{\tilde{P}\in\tilde{\mathcal{C}}} \inf_{P\in\mathcal{C}} \mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}}), \tag{48}$$

which is a kind of excess nested distance of the set of processes induced by $\tilde{\mathcal{C}}$ over the set of processes induced by $\mathcal{C}$.

Theorem 19. *Let f be uniformly continuous in ξ, that is, there is a positive constant L such that*

$$f(x, \tilde{\xi}) - f(x, \xi) \leq L\,\mathsf{d}(\xi, \tilde{\xi}) \qquad \textit{for all } \xi,\ \tilde{\xi} \textit{ and } x$$

and uniformly convex in x,

$$f\big((1-\lambda)x + \lambda\tilde{x}, \xi\big) \leq (1-\lambda)f(x, \xi) + \lambda f(\tilde{x}, \xi) \qquad \textit{for all } \lambda \in [0,1] \textit{ and } \xi \in \Xi. \tag{49}$$

If $\mathbb{X}$ is a convex set, then

$$\inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(\xi,x(\xi))] - \inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{\tilde{P}\in\tilde{\mathcal{C}}} \mathbb{E}_{\tilde{P}}[f(\xi,x(\xi))] \leq \mathbb{D}(\tilde{\mathcal{C}},\mathcal{C};\, \mathsf{dl}_r),$$

where the Hausdorff deviation $\mathbb{D}$ is with respect to the nested distance, dl_r.

Proof. Let $\mathbf{x}^*(\cdot)\in\mathbb{X}$ be a feasible solution and $P^*\in\mathcal{P}$ be chosen so that

$$\inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(\xi,\mathbf{x}(\xi))] > \mathbb{E}_{P^*}[f(\xi,\mathbf{x}^*(\xi))] - \varepsilon.$$

Given $\tilde{P}\in\tilde{\mathcal{C}}$, let π have all conditional marginals from P^* and $\tilde{P}$. Then

$$\mathbb{E}_{P^*}[f(\xi,\mathbf{x}^*(\xi))] = \iint f(\xi,\mathbf{x}^*(\xi))\pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}) = \mathbb{E}_\pi[f(\xi,\mathbf{x}^*(\xi))]. \tag{50}$$

Define now

$$\tilde{\mathbf{x}}_t(\tilde{\xi}) := \int \mathbf{x}_t^*(\xi)\,\pi(\mathrm{d}\xi|\,\tilde{\xi}) \tag{51}$$

and note that $\tilde{\mathbf{x}}_t$ is measurable with respect to the filtration $\tilde{\mathcal{F}}_t$ by $\tilde{\mathbf{x}}$'s definition. Then we derive from uniform convexity (49) and Jensen's inequality that

$$f\left(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi})\right) = f\left(\tilde{\xi},\int \mathbf{x}^*(\xi)\,\pi(\mathrm{d}\xi|\,\tilde{\xi})\right) \leq \int f\left(\tilde{\xi},\mathbf{x}^*(\xi)\right)\,\pi(\mathrm{d}\xi|\,\tilde{\xi}),$$

where Jensen's inequality is applied on each fiber separately. Integrating the latter inequality with respect to $\tilde{P}$ one obtains that

$$\begin{aligned}\mathbb{E}_{\tilde{P}}[f(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi}))] &= \int f(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi}))\tilde{P}(\mathrm{d}\tilde{\xi}) \leq \int\int f\left(\tilde{\xi},\mathbf{x}^*(\xi)\right)\,\pi(\mathrm{d}\xi|\,\tilde{\xi})\,\tilde{P}(\mathrm{d}\tilde{\xi})\\ &= \iint f\left(\tilde{\xi},\mathbf{x}^*(\xi)\right)\,\pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}) = \mathbb{E}_\pi[f(\tilde{\xi},\mathbf{x}^*(\xi))],\end{aligned}$$

and together with (50) it follows that

$$\begin{aligned}\mathbb{E}_{\tilde{P}}[f(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi}))] - \mathbb{E}_P[f(\xi,\mathbf{x}^*(\xi))] &\leq \iint f\left(\tilde{\xi},\mathbf{x}^*(\xi)\right) - f(\xi,\mathbf{x}^*(\xi))\pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi})\\ &\leq L\cdot\iint \mathsf{d}\left(\xi,\tilde{\xi}\right)\,\pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}),\end{aligned}$$

and by taking the infimum with respect to π over all probability measures with adapted conditional marginals finally

$$\mathbb{E}_{\tilde{P}}[f(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi}))] - \mathbb{E}_P[f(\xi,\mathbf{x}^*(\xi))] \leq L\cdot\mathsf{dl}(\mathbb{P},\tilde{\mathbb{P}}).$$

Now recall that $\tilde{\mathbf{x}}$ depends on $\mathbf{x}^*$, P^* and $\tilde{P}$ by means of (51). We can hence take the infimum over P and then the supremum over $\tilde{P}$

and we get

$$\sup_{\tilde{P}\in\tilde{\mathcal{C}}} \mathbb{E}_{\tilde{P}}[f(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi}))] - \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(\xi,\mathbf{x}^*(\xi))] \le L\cdot \sup_{\tilde{P}\in\tilde{\mathcal{C}}} \inf_{P\in\mathcal{C}} \mathsf{dl}_r(\tilde{\mathcal{C}},\mathcal{C}) = \mathbb{D}(\tilde{\mathcal{C}},\mathcal{C};\,\mathsf{dl}_r),$$

where $\mathbb{D}(\tilde{\mathcal{C}},\mathcal{C};\,\mathsf{dl}_r)$ is defined in (9). Finally note that by taking the supremum among $\mathbf{x}\in\mathbb{X}$ to get

$$\sup_{\tilde{P}\in\tilde{\mathcal{C}}} \mathbb{E}_{\tilde{P}}\, f[(\tilde{\xi},\tilde{\mathbf{x}}(\tilde{\xi}))] - \inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(\xi,\mathbf{x}(\xi))] \le L\cdot\mathbb{D}(\tilde{\mathcal{C}},\mathcal{C};\mathsf{dl}_r)+\varepsilon,$$

and consequently

$$\begin{aligned}
&\inf_{\hat{\mathbf{x}}(\cdot)\in\mathbb{X}} \sup_{\tilde{P}\in\mathcal{C}_N} \mathbb{E}_{\tilde{P}}[f(\tilde{\xi},\hat{\mathbf{x}}(\tilde{\xi}))] - \inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(\xi,\mathbf{x}(\xi))]\\
&\le \inf_{\tilde{\mathbf{x}}(\cdot)\in\mathbb{X}} \sup_{\tilde{P}\in\tilde{\mathcal{C}}} \mathbb{E}_{\tilde{P}}[f(\tilde{\xi},\hat{\mathbf{x}}(\tilde{\xi}))] - \inf_{\mathbf{x}(\cdot)\in\mathbb{X}} \sup_{P\in\mathcal{C}} \mathbb{E}_P[f(\xi,\mathbf{x}(\xi))]\\
&\le L\cdot\mathbb{D}(\tilde{\mathcal{C}},\mathcal{C};\mathsf{dl}_r)+\varepsilon,
\end{aligned}$$

from which we conclude the assertion. □

Theorem 19 gives a quantitative description on the discrepancy between the optimal values of problems (45) in terms of the nested distance $\mathbb{D}(\tilde{\mathcal{C}},\mathcal{C};\,\mathsf{dl}_r)$. The latter is entirely determined by $\tilde{\Xi}$ and the ambiguity set.

6 Summary and future work

This paper explores discretization of ambiguity set defined by prior moment conditions in distributionally robust optimization problems. Discretization is important because it concerns numerical solvability of the DRO problems. It is also relevant to data-driven optimization problems where the number of samples of the underlying uncertainty is often limited.

A key issue to be addressed is to quantify the difference between discretized ambiguity set and its original under some appropriate metric. We have managed to do so in this paper by deriving a new form of Hoffman's lemma under ζ-metric and then use it to quantify discrepancy of the ambiguity sets under Kantorovich/ Wasserstein metric. The quantification allows one to assess the number of samples needed for prescribed accuracy.

The second part of the paper investigates propagation of the discrepancy (approximation error) in a one stage decision making problem under DRO structure. We have demonstrated how the optimal value and the optimal solutions are affected in a quantitative manner against variation on the ambiguity set and hence the change of sample size in practice. This effectively paves the way for numerical implementation of the discretization methods for solving DRO problems.

We finally extend the robust setting to the multistage environment. As for the unconstrained situation, the results can be formulated in terms of the nested distance for the robust multistage stochastic optimization problem, which involves moment constraints. Explicit bounds similar to Hoffman's Lemma, however, are not available in this situation.

References

[1] B. Analui and G. Ch. Pflug. On distributionally robust multiperiod stochastic optimization. *Computational Management Science*, 11:197–220, 2014. 4

[2] E. Anderson, H. Xu, and D. Zhang. Varying confidence levels for CVaR risk measures and minimax limits. unpublished, 2016. 16

[3] K. B. Athreya and S. N. Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006. 5

[4] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15:780–804, 2005. 3

[5] E. Delage and Y. Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58:595–612, 2010. 3, 10, 13, 20

[6] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969. 15

[7] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70:419–435, 2002. 5, 6, 7, 14

[8] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. doi:10.1007/BFb0103945. 16

[9] A. J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49:263–265, 1952. 8

[10] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30:257–280, 2005. 4

[11] O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, 2002. doi:10.1007/b98838. 24

[12] J. E. Kelley. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8:703–712, 1960. 3

[13] Y. Liu, R. Meskarian, and H. Xu. A semi-infinite programming approach for distributionally robust reward-risk ratio optimization with matrix moments constraints. *SIAM Journal on Optimization*, 27:957–985, 2017. 11, 12

[14] S. Mehrotra and D. Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization*, 24:1670–1697, 2014. 3

[15] A. Nilim and L. E. Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53:780–798, 2005. 4

[16] T. S. P. M. Esfahani and J. Lygeros. Performance bounds for the scenario approach and an extension to a class of non-convex program. *IEEE Transactions on Automatic Control*, 60:46–58, 2015. 22

[17] G. Ch. Pflug and A. Pichler. Approximations for probability distributions and stochastic optimization problems. In M. Bertocchi, G. Consigli, and M. A. H. Dempster, editors, *Stochastic Optimization Methods in Finance and Energy*, volume 163 of *International Series in Operations Research & Management Science*, chapter 15, pages 343–387. Springer, New York, 2011. ISBN 978-1-4419-9586-5. doi:10.1007/978-1-4419-9586-5. 2

[18] G. Ch. Pflug and A. Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014. doi:10.1007/978-3-319-08843-3. 2, 4, 15, 23

[19] G. Ch. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007. doi:10.1080/14697680701455410. 3

[20] I. Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30:632–657, 2005. 20

[21] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. John Wiley and Sons, West Sussex, England, 1991. URL `http://books.google.com/books?id=5grvAAAAMAAJ`. 3, 5, 14

[22] S. M. Robinson. An application of error bounds for convex programming in a linear space. *SIAM Journal on Control*, 13:271–273, 1975. 10

[23] W. Römisch. Stability of stochastic programming problems. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming, Handbooks in Operations Research and Management Science*, volume 10, chapter 8. Elsevier, Amsterdam, 2003. 20

[24] S. L. Savage. *The Flaw of Averages*. John Wiley & Sons, 2009. 2

[25] H. Scarf. A min-max solution of an inventory problem. In K. J. Arrow and S. Karlin, editors, *Studies in the mathematical theory of inventory and production*, chapter 10, pages 201–209. Stanford University Press, 1958. 3

[26] A. Shapiro. On duality theory of conic linear problems. In M. Á. Goberna and M. A. López, editors, *Semi-Infinite Programming: Recent Advances*, pages 135–165. Springer US, 2001. ISBN 978-1-4757-3403-4. doi:10.1007/978-1-4757-3403-4. 8

[27] A. Shapiro. Monte Carlo sampling methods. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*, Handbooks in Operations Research and Management Science. Elsevier, 2003. doi:10.1016/S0927-0507(03)10006-0. 2

[28] A. Shapiro. Minimax and risk averse multistage stochastic programming. *European Journal of Operational Research*, 219(3):719–726, 2012. doi:10.1016/j.ejor.2011.11.005. 4

[29] A. Shapiro and H. Xu. Stochastic mathematical programs with equilibrium constraints, modeling and sample average approximation. *Optimization*, 57:395–418, 2008. 23

[30] A. M.-C. So. Moment inequalities for sums of random matrices and their applications in optimization. *Mathematical programming*, 130:125–151, 2011. 10

[31] H. Sun and H. Xu. Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, 41:377–401, 2016. 3, 9, 10, 13, 22

[32] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38:153–183, 2013. 4

[33] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62:1358–1376, 2014. 3, 12, 20

[34] L. Xin, D. Goldberg, and A. Shapiro. Distributionally robust multistage inventory models with moment constraint. Optimization online, 2013. 4

[35] H. Xu, Y. Liu, and H. Sun. Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane method. *Mathematical Programming*, 2017. doi:10.1007/s10107-017-1143-6. 3, 7, 8, 16, 20, 22

[36] J. Zhang, H. Xu, and L. W. Zhang. Quantitative stability analysis for distributionally robust optimization with moment constraints. *SIAM Journal on Optimization*, 26:1855–1882, 2016. 3, 7, 8, 10, 13, 14

[37] C. Zhao and Y. Guan. Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Optimization Online, 2015. 6

[38] A. Zhigljavsky and A. Žilinskas. *Stochastic Global Optimization*. Springer US, 2008. doi:10.1007/978-0-387-74740-8. 17

[39] V. M. Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28: 264–287, 1983. 5

[40] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137:167–198, 2013. 3, 20

output