Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

COMPUTING WORST-CASE EXPECTATIONS GIVEN MARGINALS VIA SIMULATION

Jose Blanchet
Department of Management Science and Engineering
Stanford University
475 Via Ortega 3rd Floor
Stanford, CA 94305, USA

Fei He
Department of IEOR
Columbia University
500 W 120th St
New York, NY 10027, USA

Henry Lam
Department of IEOR
Columbia University
500 W 120th St
New York, NY 10027, USA

ABSTRACT

We study a direct Monte-Carlo-based approach for computing the worst-case expectation of two multidimensional random variables given a specification of their marginal distributions. This problem is motivated by several applications in risk quantification and statistics. We show that if one of the random variables takes finitely many values, a direct Monte Carlo approach allows to evaluate such worst case expectation with $O(n^{-1/2})$ convergence rate as the number of Monte Carlo samples, n, increases to infinity.

1 INTRODUCTION

We focus on the problem of computing lower and upper bounds among any dependence structure for a function of two random vectors whose marginal distributions are assumed to be known. This problem is motivated from several applications in risk quantification and statistics. Before discussing its applications, let us first describe it precisely.

Suppose that $X \in \mathbb{R}^d$ follows distribution μ and $Y \in \mathbb{R}^l$ follows distribution ν. We define $\Pi(\mu,\nu)$ to be the set of joint distributions π in $\mathbb{R}^{d\times l}$ such that the marginal of the first d entries coincides with μ and the marginal of the last l entries coincides with ν. In other words, for any probability measure π in $\mathbb{R}^{d\times l}$ (endowed with the Borel σ-field), if we let $\pi_X(A) = \pi\left(A\times\mathbb{R}^l\right)$ for any Borel measurable set $A \in \mathbb{R}^d$, and $\pi_Y(B) = \pi\left(\mathbb{R}^d\times B\right)$ for any Borel measurable set $B \in \mathbb{R}^l$, then $\pi \in \Pi(\mu,\nu)$ if and only if $\pi_X = \mu$ and $\pi_Y = \nu$. We are interested in the quantity (focusing on minimization)

$$V = \min\{\mathbb{E}_\pi[c(X,Y)] : \pi \in \Pi(\mu,\nu)\} \tag{1}$$

where $c(\cdot,\cdot) \in \mathbb{R}$ is some cost function. Formulation (1) is well-defined as the class $\Pi(\mu,\nu)$ is non-empty, because the product measure $\pi = \mu\times\nu$ belongs to $\Pi(\mu,\nu)$.

In operations research contexts, problem (1) arises as a means to obtain bounds for performance measures in situations where dependence information is ambiguous. Such situations occur because, in practice, accurately estimating the marginal distributions of random variables is often relatively easy, e.g., by goodness-of-fit against well-chosen parametric distributions. They also occur in scenarios where data from different stochastic sources are collected independently (i.e., rather than in pairs), in which case no dependence information between these sources can be inferred. Indeed, special (i.e., discrete) cases of (1) have been analyzed in the distributionally robust optimization literature (e.g., Doan et al. 2015). Variants of (1) to risk measures have also been studied, regarding both algorithmic approaches (e.g., Rüschendorf

978-1-5386-3428-8/17/$31.00 ©2017 IEEE

1983, Embrechts et al. 2013) and sharp bounds over specific geometric classes of marginals (e.g., Wang and Wang 2011, Puccetti 2013, Puccetti and Rüschendorf 2013).

In statistics and machine learning contexts, the value of (1) is the Wasserstein distance (of order 1) between X and Y when $c(\cdot,\cdot)$ is taken as a metric. The optimization can be viewed as the classical Kantorovich relaxation to Monge's problem in optimal transport (e.g., Rachev and Rüschendorf 1998, Villani 2008), where solutions based on differential properties have been extensively studied. Wasserstein distance is of central importance in probabilistic analysis (e.g., quantifying model discrepancies in Bayesian settings (Minsker et al. 2014) and convergence rates of ergodic processes (Boissard and Le Gouic 2014), among many others). The estimation of the distance itself is also suggested as a tool for statistical inference, including the use in goodness-of-fit tests (Del Barrio et al. 1999, Del Barrio et al. 2005) and in applications such as image recognition (Sommerfeld and Munk 2016). It has also been used to quantify model uncertainty in stochastic optimization problems (e.g., Esfahani and Kuhn 2015, Blanchet and Kang 2016, Blanchet and Murthy 2016, Gao and Kleywegt 2016) and in the application of distributionally robust optimization in machine learning settings (Blanchet et al. 2016). As such, there have been growing studies on the convergence behaviors of its empirical estimation. Central limit theorems (CLTs) on the empirical estimation of (1), based on representations using quantile functions, have been investigated in the one-dimensional case (e.g., Bobkov and Ledoux 2014, Del Barrio et al. 1999). More generally, concentration bounds have been studied in the line of work including Horowitz and Karandikar (1994), Bolley et al. (2007), Boissard (2011), Sriperumbudur et al. (2012), Trillos and Slepčev (2014) and Fournier and Guillin (2015), so do laws of large numbers in some special cases (e.g., Dobrić and Yukich 1995).

Since classical methods for solving (1), based for instance on Euler-Lagrange equations, may not yield straightforward computational schemes in general, we resort to Monte Carlo for an easy-to-implement approximation. Our contribution is precisely to quantify the rate of convergence of such Monte Carlo schemes. Our results also add to the literature of empirical Wasserstein estimation when these Monte Carlo samples are viewed as data. We focus on the setting where one of the marginals, say Y, is a finite-support distribution, and another, say X, is a multi-dimensional distribution that can be continuous. To approximate V, we consider the drawn samples from the continuous variable X, and replace the infinite-dimensional linear program (LP) in (1) by its sampled counterpart, which can be solved by standard LP solvers.

Our main result shows that the error of our procedure is $O(n^{-1/2})$ where n is the sample size, independent of the dimension d or l. We also identify the limiting distribution in the associated CLT. The closest work to our results, as far as we know, is the recent work of Sommerfeld and Munk (2016), who derive a CLT when both marginal distributions are finitely discrete. Our result here can be viewed as a generalization to theirs when one of the distributions is continuous. We remark that our obtained rate differs from the typical rate of $O(n^{-1/d})$ in high-dimensional empirical Wasserstein estimation where $d \geq 3$ is the dimension of the marginal distributions. As we will see, the finite-support property of one of the marginals plays a crucial role in applying classical results in sample average approximation (SAA) that maintain the standard Monte Carlo rate in our scheme.

In the rest of this paper, we will first describe our algorithm, followed by our main results on the convergence analysis.

2 ALGORITHMIC DESCRIPTION

Suppose that the distribution ν for Y has finite support $\{y_1,...y_m\} \subset \mathbb{R}^l$. Supposing that X can be simulated, we sample n i.i.d. observations $X_1,\ldots,X_n$ from μ, and approximate V by

$$V_n = \min\{\mathbb{E}_\pi [c(X,Y)] : \pi \in \Pi(\mu_n, \nu)\} \tag{2}$$

where μ_n is the empirical distribution of X constructed from the X_i's, i.e.,

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A)$$

for any Borel measurable A.

Note that (2) is a finite-dimensional LP, which can be written more explicitly as

$$\begin{array}{ll} \min & \sum_{i=1}^{n}\sum_{j=1}^{m} c(X_i,y_j)p_{ij} \\ \text{subject to} & \sum_{j=1}^{m} p_{ij} = \frac{1}{n} \quad \forall i=1,\ldots,n \\ & \sum_{i=1}^{n} p_{ij} = \nu\{y_j\} \quad \forall j=1,\ldots,m \\ & p_{ij} \geq 0 \quad \forall i=1,\ldots,n,\ j=1,\ldots,m \end{array} \tag{3}$$

where the decision variables p_{ij} represent the probability masses on (X_i,y_j), and $\nu\{y_j\}$ denotes the mass on y_j under ν. Problem (3) is an assignment problem, which is a special type of minimum cost problem and can be solved by, e.g., successive shortest path algorithms in polynomial time of order $O(n^2m+n(n+m)\log(n+m))$ (see, e.g., R.K.Ahuja et al. 2000 pp. 471, 500).

3 CONVERGENCE ANALYSIS

Our main result is a convergence analysis on V_n to V. We impose the assumptions:

Assumption 1 For each y_j, $c(.,y_j)$ is non-negative and lower semicontinuous.

Assumption 2 Suppose that ν has finite support $\{y_1,...,y_m\} \subset \mathbb{R}^l$. We have

$$\mathbb{E}_\mu[c(X,y_j)^2] < \infty, \forall j=1,\ldots,m.$$

Denote

$$V' = \max_{\beta_1,\ldots,\beta_m\in\mathbb{R}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^{m}\beta_j\nu\{y_j\}\Big] \tag{4}$$

which is the dual problem of (1) (see Lemma 1 for an explanation in the special case of finite-dimensional settings). Under Assumptions 1 and 2, strong duality (known as the Kantorovich duality) holds and $V'=V$; see, e.g., Theorem 5.10 in Villani (2008).

In order to state our main result, we need to introduce a Gaussian random field $G(\cdot):\mathbb{R}^m\to\mathbb{R}$ with covariance structure given by

$$\mathrm{Cov}(G(\beta),G(\beta')) = \mathrm{Cov}\Big(\min_{j=1,\ldots,m}\big\{c(X,y_j)-\beta_j\big\}, \min_{j=1,\ldots,m}\big\{c(X,y_j)-\beta_j'\big\}\Big)$$

for any $\beta=(\beta_j)_{j=1}^m$ and $\beta'=(\beta_j')_{j=1}^m$. Our main result is the following.

Theorem 1 Under Assumption 2, $V_n \xrightarrow{p} V'$ as $n\to\infty$. Moreover,

$$n^{1/2}\left(V_n - V'\right) \Rightarrow G^*$$

as $n\to\infty$, where

$$G^* = \max_{\beta=(\beta_1,\ldots,\beta_m)\in\mathbb{S}} G(\beta).$$

Here $\mathbb{S}$ is the set of all optimal solutions $\beta=(\beta_j)_{j=1}^m\in\mathbb{R}^m$ for the convex optimization problem

$$\max_{\substack{\beta_1,\ldots,\beta_m\in\mathbb{R}\\ \sum_{j=1}^m\beta_j=0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^{m}\beta_j\nu\{y_j\}\Big]. \tag{5}$$

Remark 1 The significance of this result is that one can approximate worst-case expectations by sampling with a rate of convergence (as measured by the sample size of the continuous distribution) of order $O(n^{-1/2})$. As we mentioned earlier, this might be somewhat surprising given that standard empirical estimators for Wasserstein distances exhibit a degradation which becomes quite drastic in high dimensions.

3.1 Proof of Theorem 1

We first note that adding a constant to β_j in the objective function of the dual does not change the objective value. To remove this ambiguity we inroduce the next result.

Lemma 1 Define

$$\widehat{V_n} := \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m. \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\}. \tag{6}$$

We have $V_n = \widehat{V_n}$.

Proof. The dual formulation of V_n, depicted as the LP (3), is given by

$$\begin{array}{ll} \max & \frac{1}{n}\sum_{i=1}^n \alpha_i + \sum_{j=1}^m \beta_j \nu\{y_j\} \\ \text{subject to} & \alpha_i + \beta_j \le c(X_i, y_j)\ \forall i = 1, \ldots, n,\ j = 1, \ldots, m \end{array} \tag{7}$$

where $(\alpha_i)_{i=1}^m, (\beta_j)_{j=1}^m$ are the dual variables. Note that the constraint in (7) can be written as $\alpha_i \le \min_{j=1,\ldots,m}\{c(X_i, y_j) - \beta_j\}\ \forall i = 1, \ldots, n$, which implies that (7) is equivalent to

$$\max_{\beta_j \in \mathbb{R}, j=1,\ldots,m} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\} \tag{8}$$

Since shifting any $(\beta_j)_{j=1}^m$ to $(\beta_j + \lambda)_{j=1}^m$ by an arbitrary constant λ does not affect the objective value of (8), we can always set $\lambda = -\frac{1}{m}\sum_{j=1}^m \beta_j$ to enforce the constraint $\sum_{j=1}^m \beta_j = 0$, so that (8) is equal to (6). Finally, since (3) is feasible by choosing an independent distribution, strong duality holds. We therefore conclude the lemma. □

Next we show that $\widehat{V_n}$ can be further reduced to a problem with compact feasible region, which will subsequently facilitate the invocation of classical results in SAA:

Proposition 2 Define

$$\widehat{V_n^b} := \max_{\substack{\beta_j \in \mathbb{R}, |\beta_j| \le b, j=1,\ldots,m \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\}. \tag{9}$$

There exists some large enough constant $b > 0$ such that

$$V_n = \widehat{V_n^b} \tag{10}$$

eventually, i.e., holds for any $n > N$ for some $N < \infty$ almost surely.

Proof. By Lemma 1, we have

$$\begin{aligned} V_n &= \widehat{V_n} \\ &= \max \Bigg\{ \max_{\substack{\beta_j \in \mathbb{R}, |\beta_j| \le b, j=1,\ldots,m \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\}, \\ &\qquad \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j| > b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\} \Bigg\}. \end{aligned} \tag{11}$$

Note that the first term inside the outer max is $\widehat{V_n^b}$ by our definition (9). We will show that there exists a deterministic $b > 0$ such that the first term dominates the second term eventually, which will then conclude the proposition.

To this end, consider the second term in (11)

$$\begin{aligned}
&\max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n}\sum_{i=1}^n \min_{j=1,\ldots,m} \{c(X_i,y_j) - \beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\} \\
\leq\ & \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \min_{j=1,\ldots,m} \{-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\} + \frac{1}{n}\sum_{i=1}^n \max_{j=1,\ldots,m} c(X_i,y_j). \qquad (12)
\end{aligned}$$

We analyze

$$\max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \min_{j=1,\ldots,m} \{-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\}. \qquad (13)$$

Denote $M = \max_{j=1,\ldots,m} |\beta_j|$, so that $M > b$ for any β inside the feasible region. There must exist either a $\beta_{j^*} = M$ or $\beta_{j^*} = -M$. In the first case, we have

$$\begin{aligned}
&\max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \min_{j=1,\ldots,m} \{-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\} \\
\leq\ & -M + \left\{ \begin{array}{ll} \max & \sum_{j=1}^m \beta_j \nu\{y_j\} \\ \text{subject to} & \beta_j \leq M \ \forall j = 1,\ldots,m \\ & \sum_{j=1}^m \beta_j = 0 \end{array} \right\} \\
=\ & -M + M \times \left\{ \begin{array}{ll} \max & \sum_{j=1}^m \beta_j \nu\{y_j\} \\ \text{subject to} & \beta_j \leq 1 \ \forall j = 1,\ldots,m \\ & \sum_{j=1}^m \beta_j = 0 \end{array} \right\} \qquad (14)
\end{aligned}$$

where the last equality follows by a change of variable from β_j to β_j/M in the optimization. Note that the optimal value of

$$\begin{array}{ll} \max & \sum_{j=1}^m \beta_j \nu\{y_j\} \\ \text{subject to} & \beta_j \leq 1 \ \forall j = 1,\ldots,m \\ & \sum_{j=1}^m \beta_j = 0 \end{array}$$

is strictly less than 1. To see this, observe that the optimal value is at most 1 by using the first constraint. The value of exactly 1 is attained under the first constraint by the unique solution $\beta_j = 1, j = 1,\ldots,m$, which is ruled out because it would violate the second constraint. With this claim, we conclude that (14) is equal to θM for some $\theta < 0$, which is bounded from above by θb.

In the second case, we have $\beta_{j^*} = -M$. Let $\tilde{j}^* = \operatorname{argmax}_{j=1,\ldots,m}\{\beta_j\}$. By the constraint $\sum_{j=1}^m \beta_j = 0$ in (13), we must have $\beta_{\tilde{j}^*} \geq M/(m-1)$. Therefore, applying our argument for the first case gives that (13) is bounded from above by $\theta M/(m-1) \leq \theta b/(m-1)$ for the same $\theta < 0$ chosen before.

Therefore, in either case (13) is bounded from above by $\theta b/(m-1)$. Note that the first term inside the outer max in (11), namely $\widehat{V_n^b}$, satisfies $\widehat{V_n^b} \geq (1/n)\sum_{i=1}^n \min_{j=1,\ldots,m} c(X_i,y_j)$ by plugging in the feasible solution given by $\beta_j = 0, j = 1,\ldots,m$. Thus, with the law of large numbers, by choosing $b > 0$ large enough such that

$$\frac{\theta b}{m-1} + \mathbb{E}_\mu \left[\max_{j=1,\ldots,m} c(X,y_j) \right] < \mathbb{E}_\mu \left[\min_{j=1,\ldots,m} c(X,y_j) \right] \qquad (15)$$

the first term dominates the second term inside the outer max in (11) as $n \to \infty$ almost surely.

$\square$

We are now ready to prove Theorem 1:

Proof of Theorem 1. Note that the function

$$F(X,\beta) := \min_{j=1,\dots,m} \left\{ c(X,y_j) - \beta_j \right\} + \sum_{j=1}^{m} \beta_j \nu\{y_j\} \tag{16}$$

on $\beta = (\beta_j)_{j=1}^m \in \mathbb{R}^m$ is Lipschitz continuous in the sense that

$$|F(X,\beta) - F(X,\beta')| \leq (1 + \|\nu\|)\|\beta - \beta'\|$$

where $\|\cdot\|$ denotes the L_2-norm, and ν is interpreted as a vector $(\nu\{y_j\})_{j=1}^m$. This follows since

$$\left| \min_{j=1,\dots,m} \left\{ c(X,y_j) - \beta_j \right\} - \min_{j=1,\dots,m} \left\{ c(X,y_j) - \beta'_j \right\} \right| \leq \|\beta - \beta'\|_\infty$$

and

$$\left| \sum_{j=1}^{m} \beta_j \nu\{y_j\} - \sum_{j=1}^{m} \beta'_j \nu\{y_j\} \right| \leq \|\nu\| \|\beta - \beta'\|$$

by the Cauchy-Schwarz inequality. Since the set $\mathbb{B} := \{\beta \in \mathbb{R}^m : \sum_{j=1}^m \beta_j = 0, |\beta_j| \leq b, \forall j = 1,\dots,m\}$ is compact and $\mathbb{E}_\mu[F(X,\beta)^2] < \infty$ by Assumption 2, by using Theorem 5.7 in Shapiro et al. (2009), we have

$$\widehat{V_n^b} \xrightarrow{p} V^b \tag{17}$$

and

$$\sqrt{n}(\widehat{V_n^b} - V^b) \Rightarrow G^{*,b} \tag{18}$$

where

$$V^b = \max_{\substack{\beta_j \in \mathbb{R}, |\beta_j| \leq b, j=1,\dots,m \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu \left[\min_{j=1,\dots,m} \{ c(X,y_j) - \beta_j \} + \sum_{j=1}^{m} \beta_j \nu\{y_j\} \right] \tag{19}$$

and

$$G^{*,b} = \max_{\beta = (\beta_1,\dots,\beta_m) \in \mathbb{S}^b} G(\beta)$$

with $\mathbb{S}^b$ denoting the set of optimal solutions for (19) and $G(\cdot)$ is defined as in Theorem 1 but restricted to the domain $\mathbb{B}$.

By Proposition 2, we have $\sqrt{n}(\widehat{V_n^b} - V_n) \xrightarrow{p} 0$ as $n \to \infty$. Thus, together with (17), we have

$$V_n \xrightarrow{p} V^b$$

and together with (18), we have

$$\sqrt{n}(V_n - V^b) \Rightarrow G^{*,b}$$

by Slutsky's Theorem.

To conclude the theorem, we show that $V^b = V'$, and $\mathbb{S}^b = \mathbb{S}$ so that $G^{*,b} = G^*$. By using essentially the same argument as for Proposition 2 (with the empirical expectation replaced by $\mathbb{E}_\mu[\cdot]$) and choosing the same b as in (15), we have

$$\begin{aligned}
V' &= \max_{\beta_j \in \mathbb{R}, j=1,\ldots,m} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Big] \\
&= \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Big] \\
&\quad \text{by shifting any } (\beta_j)_{j=1}^m \text{ to } (\beta_j - (1/m)\sum_{k=1}^m \beta_k)_{j=1}^m \text{ which does not affect the objective value and} \\
&\quad \text{enforces the constraint } \sum_{j=1}^m \beta_j = 0 \\
&= \max\Bigg\{ \max_{\substack{\beta_j \in \mathbb{R}, |\beta_j| \le b, j=1,\ldots,m \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Big], \\
&\qquad \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j| > b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Big] \Bigg\}
\end{aligned}$$

where

$$\begin{aligned}
& V^b \\
= & \max_{\substack{\beta_j \in \mathbb{R}, |\beta_j| \le b, j=1,\ldots,m \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Big] \\
> & \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m, |\beta_j| > b \text{ for some } j \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Big]
\end{aligned}$$

so that $V' = V^b$ and $\mathbb{S}^b = \mathbb{S}$. □

4 ADDITIONAL DISCUSSION AND EXTENSIONS

Finally, we briefly discuss the challenge in generalizing our procedure to the case when both X and Y are continuous. Here, one may attempt to sample both variables (assuming both can be simulated) and formulate a sampled program like (2) or (3). However, the analog of its reformulation in (6) and (9) will have a growing number of variables β_j and an analogous limit in (5) that involves an infinite-dimensional variable, which challenges the use of standard SAA machinery. In fact, consider a special example where $X, Y \sim U[0,1]^d$ and $c(x,y) = \|x-y\|$. In this case, (1) corresponds to the Wasserstein distance (of order 1) between X and Y, which is of course 0. It is known that sampling X and keeping Y continuous will give, for $d \ge 3$, an expected optimal value of (2) that is of order $n^{-1/d}$, i.e., $C_1 n^{-1/d} \le EV_n \le C_2 n^{-1/d}$ for all n for some $C_1, C_2 > 0$ (e.g., Problem 5.11 in van Handel 2014). Thus, the convergence rate deteriorates with the dimension and the standard Monte Carlo rate $O(n^{-1/2})$ cannot be maintained without assuming additional structure or infomation available to the modeler on the primal problem. It is of interest to investigate reasonable assumptions which are useful in applications and which would mitigate such rate-of-convergence deterioration.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation (NSF) under grants CMMI-1542020, CMMI-1523453 and CAREER CMMI-1653339. Additional support from NSF grants CMMI-1436700, DMS-1320550, and CMMI-1538217 is gratefully acknowledged.

REFERENCES

Blanchet, J., and Y. Kang. 2016. "Sample Out-Of-Sample Inference Based on Wasserstein Distance". *arXiv preprint arXiv:1605.01340*.

Blanchet, J., Y. Kang, and K. Murthy. 2016. "Robust Wasserstein Profile Inference and Applications to Machine Learning". *arXiv preprint arXiv:1610.05627v2*.

Blanchet, J., and K. Murthy. 2016. "Quantifying distributional model risk via optimal transport".

Bobkov, S., and M. Ledoux. 2014. "One-dimensional empirical measures, order statistics and Kantorovich transport distances". *preprint*.

Boissard, E. 2011. "Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance". *Electronic Journal of Probability* 16:2296–2333.

Boissard, E., and T. Le Gouic. 2014. "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance". In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Volume 50, 539–563. Institut Henri Poincaré.

Bolley, F., A. Guillin, and C. Villani. 2007. "Quantitative concentration inequalities for empirical measures on non-compact spaces". *Probability Theory and Related Fields* 137 (3): 541–593.

Del Barrio, E., J. A. Cuesta-Albertos, and C. Matrán. 1999. "Tests of goodness of fit based on the L_2-Wasserstein distance". *The Annals of Statistics* 27 (4): 1230–1239.

Del Barrio, E., E. Giné, and F. Utzet. 2005. "Asymptotics for L2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances". *Bernoulli* 11 (1): 131–189.

Doan, X. V., X. Li, and K. Natarajan. 2015. "Robustness to dependency in portfolio optimization using overlapping marginals". *Operations Research* 63 (6): 1468–1488.

Dobrić, V., and J. E. Yukich. 1995. "Asymptotics for transportation cost in high dimensions". *Journal of Theoretical Probability* 8 (1): 97–118.

Embrechts, P., G. Puccetti, and L. Rüschendorf. 2013. "Model uncertainty and VaR aggregation". *Journal of Banking & Finance* 37 (8): 2750–2764.

Esfahani, P. M., and D. Kuhn. 2015. "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations". *arXiv preprint arXiv:1505.05116*.

Fournier, N., and A. Guillin. 2015. "On the rate of convergence in Wasserstein distance of the empirical measure". *Probability Theory and Related Fields* 162 (3-4): 707–738.

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

Horowitz, J., and R. L. Karandikar. 1994. "Mean rates of convergence of empirical measures in the Wasserstein metric". *Journal of Computational and Applied Mathematics* 55 (3): 261–273.

Minsker, S., S. Srivastava, L. Lin, and D. Dunson. 2014. "Scalable and robust Bayesian inference via the median posterior". In *International Conference on Machine Learning*, 1656–1664.

Puccetti, G. 2013. "Sharp bounds on the expected shortfall for a sum of dependent random variables". *Statistics & Probability Letters* 83 (4): 1227–1232.

Puccetti, G., and L. Rüschendorf. 2013. "Sharp bounds for sums of dependent risks". *Journal of Applied Probability* 50 (01): 42–53.

Rachev, S. T., and L. Rüschendorf. 1998. *Mass Transportation Problems: Volume I: Theory*, Volume 1. Springer Science & Business Media.

R.K.Ahuja, T.L.Magnanti, and J.B.Orlin. 2000. "Network Flows". *Prentice Hall, Inc.*.

Rüschendorf, L. 1983. "Solution of a statistical optimization problem by rearrangement methods". *Metrika* 30 (1): 55–61.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2009. *Lectures on stochastic programming: modeling and theory*. SIAM.

Sommerfeld, M., and A. Munk. 2016. "Inference for Empirical Wasserstein Distances on Finite Spaces". *arXiv preprint arXiv:1610.03287*.

Sriperumbudur, B. K., K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. 2012. "On the empirical estimation of integral probability metrics". *Electronic Journal of Statistics* 6:1550–1599.

Trillos, N. G., and D. Slepčev. 2014. "On the rate of convergence of empirical measures in ∞-transportation distance". *arXiv preprint arXiv:1407.1157*.

van Handel, R. 2014. "Probability in high dimension". Technical report, DTIC Document.

Villani, C. 2008. *Optimal transport: old and new*, Volume 338. Springer Science & Business Media.

Wang, B., and R. Wang. 2011. "The complete mixability and convex minimization problems with monotone marginal densities". *Journal of Multivariate Analysis* 102 (10): 1344–1360.

AUTHOR BIOGRAPHIES

JOSE BLANCHET is a faculty member in the Department of Management Science and Engineering at Stanford University and of IEOR and Statistics at Columbia University. He holds a Ph.D. in Management Science and Engineering from Stanford University. From 2004 to 2008 he was a faculty member in the Department of Statistics at Harvard University. Jose is a recipient of the 2009 Best Publication Award given by the INFORMS Applied Probability Society and of the 2010 Erlang Prize. He also received a PECASE award given by NSF in 2010. He worked as an analyst in Protego Financial Advisors, a leading investment bank in Mexico. He has research interests in applied probability and Monte Carlo methods. His email is jose.blanchet@stanford.edu.

FEI HE is a PhD candidate in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on stochastic simulation, mathematical finance and risk management. His email address is fh2293@columbia.edu.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Ph.D. degree in statistics at Harvard University, and was a faculty member in the Department of Mathematics and Statistics at Boston University and the Department of Industrial and Operations Engineering at the University of Michigan before joining Columbia. His research focuses on stochastic simulation, risk analysis, and simulation optimization. His email address is khl2114@columbia.edu.