Wasserstein Distributionally Robust Inverse Multiobjective Optimization

Chaosheng Dong[1*], **Bo Zeng**[2]
[1]Amazon
[2]University of Pittsburgh
chaosd@amazon.com, bzeng@pitt.edu

Abstract

Inverse multiobjective optimization provides a general framework for the unsupervised learning task of inferring parameters of a multiobjective decision making problem (DMP), based on a set of observed decisions from the human expert. However, the performance of this framework relies critically on the availability of an accurate DMP, sufficient decisions of high quality, and a parameter space that contains enough information about the DMP. To hedge against the uncertainties in the hypothetical DMP, the data, and the parameter space, we investigate in this paper the distributionally robust approach for inverse multiobjective optimization. Specifically, we leverage the Wasserstein metric to construct a ball centered at the empirical distribution of these decisions. We then formulate a Wasserstein distributionally robust inverse multiobjective optimization problem (WRO-IMOP) that minimizes a worst-case expected loss function, where the worst case is taken over all distributions in the Wasserstein ball. We show that the excess risk of the WRO-IMOP estimator has a sub-linear convergence rate. Furthermore, we propose the semi-infinite reformulations of the WRO-IMOP and develop a cutting-plane algorithm that converges to an approximate solution in finite iterations. Finally, we demonstrate the effectiveness of our method on both a synthetic multiobjective quadratic program and a real world portfolio optimization problem.

Introduction

Inverse multiobjective optimization provides a compelling tool to learn humans' decision making scheme or emulate their behaviors (Dong and Zeng 2020). Its goal is to infer the parameters of the multiobjective decision making problem (DMP), based on a set of observed decisions from the human experts. More precisely, it seeks to learn θ given $\{\mathbf{y}_i\}_{i\in[N]}$ that are observations of the Pareto optimal solutions of the multiobjective optimization problem (MOP):

$$\begin{array}{ll}\min\limits_{\mathbf{x}} & \{f_1(\mathbf{x},\theta), f_2(\mathbf{x},\theta),\ldots,f_p(\mathbf{x},\theta)\}\\ s.t. & \mathbf{x}\in X(\theta),\end{array}$$

where θ indicates the true but unknown parameter for the expert's multiobjective DMP.

This tool is generally applicable in many scenerios. These underlying multiobjective decision making schemes, once obtained, would presumably play critical roles in various aspects, such as assisting agents in automating the process of providing professional services for clients. For example, the modern portfolio theory—risk and profit are two objectives to optimize—is often used by portfolio managers when buying or selling stocks on behalf of clients (Markowitz 1952). To automate the portfolio management, one could leverage the portfolio manager's investment records to learn the key parameters of this model, e.g., the risk aversion score or expected returns of the assets.

Despite its widely applications, inverse multiobjective optimization relies critically on the availability of an accurate decision making model, sufficient decisions of high quality, and a parameter space that contains as much information about the objective functions or constraints as possible. In practice, however, it is highly unlikely that all of these critical factors would be satisfied. For example, outliers in a limited amount of decisions would render the empirical distribution of decisions deviate from the true distribution, and thus significantly weaken the predictive power of the inverse optimization multiobjective estimator. We note that this issue is not unique for inverse multiobjective optimization and one can observe similar findings in inverse optimization models that has only one objective function (Keshavarz, Wang, and Boyd 2011; Bertsimas, Gupta, and Paschalidis 2015; Aswani, Shen, and Siddiq 2018; Esfahani et al. 2018; Dong, Chen, and Zeng 2018).

To hedge against these uncertainties contained in the hypothetical decision making model, the data and the selected parameter space, we investigate the distributionally robust approach for inverse multiobjective optimization. More specifically, motivated by Shafieezadeh-Abadeh, Esfahani, and Kuhn (2015); Gao and Kleywegt (2016); Esfahani and Kuhn (2018), etc., we use the Wasserstein metric (Villani 2008) to construct the uncertainty set centered at the empirical distribution of the observed decisions. Subsequently, we propose a distributionally robust inverse multiobjective optimization program that minimizes the worst-case risk of loss, where the worst case is taken over all distributions in the uncertainty set. By such a distributionally robust framework, we aim to bridge the discrepancy between the lack of certainties in the information and the expectation for the

[*]Work was done prior to joining Amazon

accurate prediction of human's or robot's future behavior.

Related Work

Our work is most related to Dong and Zeng (2020), which proposes the general framework of using inverse multiobjective optimization to infer the objective functions or constraints of the multiobjective DMP, based on observations of Pareto optimal solutions. Dong and Zeng (2020) takes the framework of empirical risk minimization for this unsupervised learning task, and generally works well when there are few uncertainties in the model, data or hypothetical parameter space. In contrast, we believe that those uncertainties inherently root in the applications of inverse multiobjective optimization, and we aim to hedge against their influences by adopting the distributionally robust optimization paradigm based on Wasserstein metric. We demonstrate both theoretically and experimentally that our method has out-of-sample performance guarantees under uncertainties.

Our work draws inspirations from Esfahani et al. (2018), who develop a distributionally robust approach for inverse optimization to infer the utility function from sequentially arrived observations. They aim to mitigate the poor performance of inverse optimization models (Ahuja and Orlin 2001; Keshavarz, Wang, and Boyd 2011; Bertsimas, Gupta, and Paschalidis 2015; Aswani, Shen, and Siddiq 2018; Esfahani et al. 2018; Bärmann, Pokutta, and Schneider 2017; Dong, Chen, and Zeng 2018) when the learner has imperfect information. They show that the associated distributionally robust inverse optimization approach offers out-of-sample performance guarantees under such a situation. However, their approach is specially designed for the simpler case where the DMP has only one objective function. Differently, our approach considers a more complex situation where the DMP has multiple objectives. Moreover, instead of using the suboptimality loss function, we consider another one that would better capture the learner's purpose to predict the decision maker's decisions. Due to the nonconvex nature of our loss function, extensive efforts are made to develop the algorithm for solving the resulting nonconvex minmax program.

Contributions

We summarize our contributions as follows:

- We present a novel Wasserstein distributionally robust framework for constructing inverse multiobjective optimization estimator. We use the prominent Wasserstein metric to construct the uncertainty set centered at the empirical distribution of observed decisions.
- We show that the proposed framework has statistical performance guarantees, and the excess risk of the distributionally robust inverse multiobjective optimization estimator would converge to zero with a sub-linear rate as the number of observed decisions approaches to infinity.
- We reformulate the resulting minmax problem as a semi-infinite program and develop a cutting-plane algorithm which converges to an approximate solution in finite iterations. We demonstrate the effectiveness of our method on both a multiobjective quadratic program and a portfolio optimization problem.

Problem Setting

Multiobjective Decision Making Problem

We consider a family of parametrized multiobjective decision making problems with p (≥ 2) objective functions,

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & \{f_1(\mathbf{x},\theta), f_2(\mathbf{x},\theta),\ldots,f_p(\mathbf{x},\theta)\} \\ s.t. & \mathbf{x}\in X(\theta), \end{array} \qquad \text{DMP}$$

where $\theta \in \Theta$ is the parameter for the multiobjective DMP. For easy exposition, we denote $\mathbf{f}(\mathbf{x},\theta)$ the vector valued function $(f_1(\mathbf{x},\theta), f_2(\mathbf{x},\theta),\ldots,f_p(\mathbf{x},\theta))^T$. Also, the feasible set $X(\theta)$ is characterized as $X(\theta) = \{\mathbf{x}\in\mathbb{R}^n : \mathbf{g}(\mathbf{x},\theta)\leq \mathbf{0}\}$, where $\mathbf{g}(\mathbf{x},\theta) = (g_1(\mathbf{x},\theta),\ldots,g_q(\mathbf{x},\theta))^T$ is another vector-valued function.

Following Dong and Zeng (2020), we consider a convex DMP where all objectives and constraints are convex in $\mathbf{x}$ for each $\theta\in\Theta$.

Definition 1 (Pareto optimality)**.** For a fixed θ, a decision $\mathbf{x}^*\in X(\theta)$ is said to be Pareto optimal if there exists no other decision $\mathbf{x}\in X(\theta)$ such that $f_i(\mathbf{x},\theta)\leq f_i(\mathbf{x}^*,\theta)$ for all $i\in[p]$, and $f_j(\mathbf{x},\theta) < f_j(\mathbf{x}^*,\theta)$ for at least one $j\in[p]$.

We denote $X_P(\theta)$ the Pareto optimal set that consists of all the Pareto optimal solutions. The weighted sum approach (Gass and Saaty 1955) is often taken to derive a Pareto optimal solution by solving

$$\begin{array}{ll} \min & w^T\mathbf{f}(\mathbf{x},\theta) \\ s.t. & \mathbf{x}\in X(\theta), \end{array} \qquad \text{WP}$$

where $w=(w^1,\ldots,w^p)^T$ is the nonnegative weight vector in the $(p-1)$-simplex $\mathscr{W}_p \equiv \{w\in\mathbb{R}^p_+ : \mathbf{1}^T w = 1\}$. When each $w\in\mathbb{R}^p_{++}$, such set is denoted by $\mathscr{W}_p^+$. We denote $S(w,\theta)$ the set of optimal solutions of WP, i.e.,

$$S(w,\theta) = \arg\min_{\mathbf{x}} \left\{ w^T\mathbf{f}(\mathbf{x},\theta) : \mathbf{x}\in X(\theta)\right\}.$$

We next make a few assumptions to simplify our understanding, which are actually mild and appear frequently in the inverse optimization literature.

Assumption 1. Set Θ is a convex compact set in $\mathbb{R}^{n_\theta}$. There exists $D>0$ such that $\sup_{\theta\in\Theta}\|\theta\|_2 \leq D$. In addition, $\mathbf{f}(\mathbf{x},\theta)$ and $\mathbf{g}(\mathbf{x},\theta)$ are convex in $\mathbf{x}$ for each $\theta\in\Theta$.

Inverse Multiobjective Optimization

Consider a learner who has access to decision makers' decisions, but does not know the underlying decision making model. In the inverse multiobjective optimization model, the learner aims to learn the parameter θ in DMP from observed noisy decisions only, and no information regarding decision makers' preferences over multiple objective functions is available. We denote $\mathbf{y}$ the observed noisy decision that might carry measurement error or is generated with bounded rationality of the decision maker, i.e., being suboptimal. Throughout the paper we assume that $\mathbf{y}$ is a random variable distributed according to an unknown distribution $\mathbb{P}_{\mathbf{y}}$ supported on $\mathcal{Y}$.

We next discuss the construction of the loss function for the unsupervised learning task in Dong and Zeng (2020).

Given a noisy decision $\mathbf{y}$ and a hypothesis θ, the loss function could ideally be defined as the minimum distance between $\mathbf{y}$ and the $X_P(\theta)$. For a general DMP, however, there typically exists no explicit way to characterize $X_P(\theta)$. Instead, a sampling approach is adopted to generate $w_k \in \mathscr{W}_p$ for each $k \in [K]$ and approximate $X_P(\theta)$ as $\bigcup_{k\in[K]} S(w_k, \theta)$. Then, the *loss function* is defined as

$$l_K(\mathbf{y}, \theta) = \min_{\mathbf{x} \in \bigcup_{k\in[K]} S(w_k,\theta)} \|\mathbf{y} - \mathbf{x}\|_2^2. \qquad \text{(loss function)}$$

By using binary variables, this loss function can be converted into the following problem.

$$\begin{aligned} l_K(\mathbf{y}, \theta) \quad = \quad & \min_{z_j \in \{0,1\}} \|\mathbf{y} - \sum_{k\in[K]} z_k \mathbf{x}_k\|_2^2 \\ & \text{s.t.} \quad \sum_{k\in[K]} z_k = 1, \ \mathbf{x}_k \in S(w_k, \theta) \end{aligned} \tag{1}$$

Constraint $\sum_{k\in[K]} z_k = 1$ ensures that exactly one of Pareto optimal solutions will be chosen to measure the distance from $\mathbf{y}$ to $X_P(\theta)$. Hence, solving this optimization problem identifies some w_k with $k \in [K]$ such that the corresponding Pareto optimal solution $S(w_k, \theta)$ is closest to $\mathbf{y}$.

We make the following assumptions as those in (Dong and Zeng 2020).

Assumption 2. **(a)** For each $\theta \in \Theta$, $X(\theta)$ is compact, and has a nonempty relative interior. Namely, there exists $B > 0$ such that $\|\mathbf{x}\|_2 \leq B$ for all $\mathbf{x} \in X(\theta)$. The support $\mathcal{Y}$ of the noisy decisions $\mathbf{y}$ is contained within a ball of radius R, where $R < \infty$.

(b) Each function in $\mathbf{f}$ is strongly convex on $\mathbb{R}^n$, that is for each $l \in [p]$, $\exists \lambda_l > 0$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\Big(\nabla f_l(\mathbf{y}, \theta_l) - \nabla f_l(\mathbf{x}, \theta_l)\Big)^T (\mathbf{y} - \mathbf{x}) \geq \lambda_l \|\mathbf{x} - \mathbf{y}\|_2^2.$$

Regarding Assumption 2 (a), we note that assuming the compactness of the feasible region is very common in inverse optimization. The finite support of the observations is needed since we do not hope outliers have too strong impacts in our learning process. Let $\lambda = \min_{l\in[p]}\{\lambda_l\}$. It follows that $w^T\mathbf{f}(\mathbf{x}, \theta)$ is strongly convex with parameter λ for each $w \in \mathscr{W}_p$. Therefore, Assumptions 2 (a) - (b) together ensure that $S(w, \theta)$ is a single-valued set for each w and θ.

Given observations $\{\mathbf{y}_i\}_{i\in[N]}$ drawn i.i.d. according to the distribution $\mathbb{P}_\mathbf{y}$, the inverse multiobjective optimization program is given in the following.

$$\min_{\theta\in\Theta} \quad \tfrac{1}{N} \sum_{i\in[N]} l_K(\mathbf{y}_i, \theta). \qquad \text{(IMOP)}$$

Statistical properties, algorithm developments, and connections to other unsupervised learning tasks have been extensively investigated in Dong and Zeng (2020).

Wasserstein Ambiguity Set

Let $\mathcal{Y} \subseteq \mathbb{R}^n$ be the observation space where the observed noisy decisions take values. Denote $\mathscr{P}(\mathcal{Y})$ be the set of all probability distributions on $\mathcal{Y}$. From now on, we let the Wasserstein ambiguity set $\mathcal{P}$ be the 1-Wasserstein ball of radius ϵ centered at P_0:

$$\mathcal{P} = \mathbb{B}_\epsilon(P_0) := \{Q \in \mathscr{P}(\mathcal{Y}) : \mathcal{W}(Q, P_0) \leq \epsilon\}, \tag{2}$$

where P_0 is the nominal distribution on $\mathcal{Y}$, $\epsilon > 0$ is the radius of the set, and $\mathcal{W}(Q, P_0)$ is the wasserstein distance metric of order 1 defined as (Villani 2008; Esfahani and Kuhn 2018; Gao and Kleywegt 2016)

$$\mathcal{W}(Q, P_0) = \inf_{\pi\in\Pi(Q,P_0)} \int_{\mathcal{Y}\times\mathcal{Y}} \|z_1 - z_2\|_2 \pi(dz_1, dz_2),$$

where $\Pi(Q, P_0)$ is the set of probability distributions on $\mathcal{Y}\times\mathcal{Y}$ with marginals Q and P_0.

Wasserstein Distributionally Robust IMOP

In this section, we propose the Wasserstein distributionally robust IMOP, and show its equivalence to a semi-infinite program. Subsequently, we present an algorithm to handle the resulting reformulations, and show its convergence in finite steps. Finally, we establish the statistical performance guarantees for the distributionally robust IMOP.

Given observations $\{\mathbf{y}_i\}_{i\in[N]}$ drawn i.i.d. according to the distribution $\mathbb{P}_\mathbf{y}$, the corresponding distributionally robust program of (IMOP) equipped with the Wasserstein ambiguity set is constructed as follows

$$\min_{\theta\in\Theta} \sup_{Q\in\mathbb{B}_\epsilon(\widehat{P}_N)} \mathbb{E}_{\mathbf{y}\sim Q}\left[l_K(\mathbf{y}, \theta)\right], \qquad \text{(WRO-IMOP)}$$

which minimizes the worst case expected loss over all the distributions in the Wasserstein ambiguity set. Here $\mathbb{B}_\epsilon(\widehat{P}_N)$ is defined in (2), and $\widehat{P}_N$ is the empirical distribution satisfying: $\widehat{P}_N(\mathbf{y}_i) = 1/N, \forall i \in [N]$.

Semi-infinite Reformulations

WRO-IMOP involves minimizing a supremum over infinitely many distributions, making it difficult to solve. In this section, we establish the reformulation of WRO-IMOP into a semi-infinite program.

The performance of WRO-IMOP depends on how the change of θ affects the objective values. For $\forall w \in \mathscr{W}_p, \theta_1, \theta_2 \in \Theta$, we consider the following function

$$h(\mathbf{x}, w, \theta_1, \theta_2) = w^T\mathbf{f}(\mathbf{x}, \theta_1) - w^T\mathbf{f}(\mathbf{x}, \theta_2).$$

Assumption 3. $\exists \kappa > 0$, $\forall w \in \mathscr{W}_p$, $\forall \theta_1 \neq \theta_2 \in \Theta$, $h(\cdot, w, \theta_1, \theta_2)$ is Lipschitz continuous on $\mathcal{Y}$:$\forall \mathbf{x}, \mathbf{y} \in \mathcal{Y}$,

$$|h(\mathbf{x}, w, \theta_1, \theta_2) - h(\mathbf{y}, w, \theta_1, \theta_2)| \leq \kappa \|\theta_1 - \theta_2\|_2 \|\mathbf{x} - \mathbf{y}\|_2.$$

Basically, this assumption requires that the objective functions will not change much when either the parameter θ or the variable $\mathbf{x}$ is perturbed. It actually holds in many common situations, including the multiobjective linear program (MLP) and multiobjective quadratic program (MQP). As a motivating example, we give the κ for an MQP.

Example 1. Suppose that $\mathbf{f}(\mathbf{x}, \theta) = \begin{pmatrix} \frac{1}{2}\mathbf{x}^T Q_1 \mathbf{x} + \mathbf{c}_1^T \mathbf{x} \\ \frac{1}{2}\mathbf{x}^T Q_2 \mathbf{x} + \mathbf{c}_2^T \mathbf{x} \end{pmatrix}$, where $\theta = (Q_1, Q_2, \mathbf{c}_1, \mathbf{c}_2)$. Under Assumption 2, we know that $\|\mathbf{y}\|_2 \leq R$. Then, $h(\cdot, w, \theta_1, \theta_2)$ is $2R\|\theta_1 - \theta_2\|_2$-Lipschitz continuous on $\mathcal{Y}$. That is, we can set $\kappa = 2R$.

Under the previous assumptions, we will establish several properties of the loss function $l_K(\mathbf{y}, \theta)$, which are ensential for our reformulation for WRO-IMOP.

Lemma 1. Under Assumptions 1 - 3, the loss function $l_K(\mathbf{y}, \theta)$ has the following properties:

(a) $\forall \mathbf{y} \in \mathcal{Y}, \theta \in \Theta, 0 \leq l_K(\mathbf{y}, \theta) \leq (B+R)^2$.

(b) $l_K(\mathbf{y}, \theta)$ is uniformly $2(B+R)$-Lipschitz continuous in $\mathbf{y}$. That is, $\forall \theta \in \Theta, \forall \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, we have

$$|l_K(\mathbf{y}_1, \theta) - l_K(\mathbf{y}_2, \theta)| \leq 2(B+R)\|\mathbf{y}_1 - \mathbf{y}_2\|_2.$$

(c) $l_K(\mathbf{y}, \theta)$ is uniformly $\frac{4(B+R)\kappa}{\lambda}$-Lipschitz continuous in θ. That is, $\forall \mathbf{y} \in \mathcal{Y}, \forall \theta_1, \theta_2 \in \Theta$, we have

$$|l_K(\mathbf{y}, \theta_1) - l_K(\mathbf{y}, \theta_2)| \leq \frac{4(B+R)\kappa}{\lambda}\|\theta_1 - \theta_2\|_2.$$

(a) and (b) of Lemma 1 are built upon direct analysis of the loss function $l_K(\mathbf{y}, \theta)$. Proof of (c) is much more involved and needs the key observation that the perturbation of $S(w, \theta)$ due to θ is bounded by the perturbation of θ by applying Proposition 6.1 in (Bonnans and Shapiro 1998). Details of the proof are provided in the supplementary material.

Let

$$\mathcal{V} := \Big\{ \mathbf{v} \in \mathbb{R}^{N+1} : V_1 \leq v_i \leq (m+1)V_2 - mV_1, \forall i \in [N], \\ 0 \leq v_{N+1} \leq (V_2 - V_1)/\epsilon \Big\}.$$

where V_1 and V_2 are the lower and upper bounds for the loss function $l_K(\mathbf{y}, \theta)$, respectively. By part (a) of Lemma 1, we will set $V_1 = 0$, and $V_2 = (B+R)^2$ throughout the remainder of the paper.

The following theorem presents a tractable reformulation of the distributionally robust optimization problem WRO-IMOP and thus constitutes the first main result of this paper.

Theorem 1 (Semi-infinite Reformulation)**.** Under Assumptions 1 - 3, WRO-IMOP is equivalent to the following semi-infinite program:

$$\begin{aligned} \min_{\theta, \mathbf{v}} \quad & \epsilon \cdot v_{N+1} + \tfrac{1}{N} \sum_{i \in [N]} v_i \\ s.t. \quad & \sup_{\widetilde{\mathbf{y}} \in \mathcal{Y}} (l_K(\widetilde{\mathbf{y}}, \theta) - v_{N+1} \cdot \|\widetilde{\mathbf{y}} - \mathbf{y}_i\|_2) \leq v_i, \quad \forall i \in [N], \\ & \theta \in \Theta, \mathbf{v} \in \mathcal{V} \end{aligned} \tag{3}$$

Proof. Under Assumption 1, we know that Θ is compact. Similarly, $\mathcal{Y}$ is also compact under Assumption 2 (a). By lemma 1 (a), $\forall \mathbf{y} \in \mathcal{Y}, \theta \in \Theta, 0 \leq l_K(\mathbf{y}, \theta) \leq (B+R)^2$, and thus $l_K(\mathbf{y}, \theta)$ is bounded. In addition, by lemma 1 (b), $l_K(\mathbf{y}, \theta)$ is continuous in $\mathbf{y}$ for any $\mathbf{y} \in \Theta$. Finally, by Lemma 1 (c), $l_K(\mathbf{y}, \theta)$ is uniformly $\frac{4(B+R)\kappa}{\lambda}$-Lipschitz continuous in θ. Hence, applying Corollary 3.8 of (Luo and Mehrotra 2017) yields the result. □

Remark 1. The establishment of Theorem 1 relies on those properties of $l_K(\mathbf{y}, \theta)$ stated in Lemma 1. Although $l_K(\mathbf{y}, \theta)$ might not be convex in θ or $\mathbf{y}$, these properties ensure that strong (Kantorovich) duality holds for the inner problem of WRO-IMOP.

Next, we will discuss how to incorporate the explicit form of $l_K(\mathbf{y}, \theta)$ into the constraints of (3). For each $i \in [N]$, constraints in (3) is equivalent to: $\forall \widetilde{\mathbf{y}} \in \mathcal{Y}$,

$$\begin{aligned} & \|\widetilde{\mathbf{y}} - \mathbf{x}_k\|_2^2 - v_{N+1} \cdot \|\widetilde{\mathbf{y}} - \mathbf{y}_i\|_2 - v_i \leq M z_{ik}, \\ & \sum_{k \in [K]} z_{ik} = K - 1, \end{aligned} \tag{4}$$

where the additional constraint $\sum_{k \in [K]} z_{ik} = K-1$, is imposed to ensure that $\|\widetilde{\mathbf{y}} - \mathbf{x}_k\|_2^2 - v_i - v_{N+1} \cdot \|\widetilde{\mathbf{y}} - \mathbf{y}_i\|_2 \leq 0$ for at least one $k \in [K]$. M is an uniform upper bound for the left-hand side of the first constraint in (4). An appropriate M could be $(B+R)^2$, since $\forall i \in [N], k \in [K]$,

$$\begin{aligned} & \|\widetilde{\mathbf{y}} - \mathbf{x}_k\|_2^2 - v_{N+1} \cdot \|\widetilde{\mathbf{y}} - \mathbf{y}_i\|_2 - v_i \leq \|\widetilde{\mathbf{y}} - \mathbf{x}_k\|_2^2 \\ & \leq (B+R)^2. \end{aligned}$$

One can verify that (4) is indeed equivalent to the first set of constraints in (3) without much effort.

Remark 2. We admit that the semi-infinite reformulation in Theorem 1 might still be valid if some assumption is not satisfied. Consider, for example, one of the objective functions is known to be strongly convex and the decision makers always has a positive preference for it.

Algorithm and Analysis of Convergence

Theorem 1 shows that the Wasserstein distributionally inverse multiobjective program WRO-IMOP is equivalent to the semi-infinite program (3). Now, any existing method for solving the general semi-infinite program can be employed to solve (3). In particular, we are interested in exchange methods (Hettich and Kortanek 1993; Joachims, Finley, and Yu 2009), since our algorithms inherits the spirit of these methods when applied to solve the minmax problem. The basic idea is to approximate the infinite set of constraints in (3) with a sequence of finite sets of constraints. Iteratively, new constraints are added to the previous set of constraints by solving a maximum constraint violation problem. This is repeated until certain stopping criterion is satisfied.

Next, we discuss how to construct the finite problem.

Let $\widetilde{\mathcal{Y}}_i = \{\widetilde{\mathbf{y}}_{i1}, \cdots, \widetilde{\mathbf{y}}_{iJ_i}\} \subseteq \mathcal{Y}, \forall i \in [N]$ be a collection of finite subsets of $\mathcal{Y}$, where each subset has J_i samples. Then, the associated finite problem of (3) is

$$\begin{aligned} & \min_{\theta, \mathbf{v}} \ \epsilon \cdot v_{N+1} + \tfrac{1}{N} \sum_{i \in [N]} v_i, \\ & s.t. \ l_K(\widetilde{\mathbf{y}}_{ij}, \theta) - v_{N+1} \cdot \|\widetilde{\mathbf{y}}_{ij} - \mathbf{y}_i\|_2 \leq v_i, \forall j \in [J_i], i \in [N], \\ & \quad \theta \in \Theta, v \in \mathcal{V}. \end{aligned} \tag{5}$$

By the same arguments for the transformation from constraints in (3) to those in (4), constraints in (5) are equivalent to

$$\begin{aligned} & \|\widetilde{\mathbf{y}}_{ij} - \mathbf{x}_k\|_2^2 - v_{N+1} \cdot \|\widetilde{\mathbf{y}}_{ij} - \mathbf{y}_i\|_2 - v_i \leq M z_{ijk}, \\ & \sum_{k \in [K]} z_{ijk} = K - 1, \ \ \forall i \in [N], j \in [J_i]. \end{aligned}$$

Algorithm 1 Wasserstein Distributionally Robust IMOP

1: **Input:** noisy decisions $\{\mathbf{y}_i\}_{i\in[N]}$, weights $\{w_k\}_{k\in K}$, radius ϵ of Wasserstein ball, and stopping tolerance δ
2: **Initialize** $\widetilde{\mathcal{Y}}_i \leftarrow \emptyset, \forall i \in [N]$
3: **repeat**
4: solve (6) with $\widetilde{\mathcal{Y}}_i, \forall i \in [N]$, and return an optimal solution $(\widehat{\theta}, \widehat{\mathbf{v}})$
5: **for** $i = 1, \ldots, N$ **do**
6: solve maximum constraint violation problem (7)
7: **if** $CV_i > 0$ **then** let $\widetilde{\mathcal{Y}}_i \leftarrow \widetilde{\mathcal{Y}}_i \cup \{\widetilde{\mathbf{y}}_i\}$ **end if**
8: **end for**
9: **until** $\max_{i\in[N]} CV_i \le \delta$
10: **Output:** a δ-optimal solution $\widehat{\theta}_N$ of (3)

Using the above transformation, (5) can be further cast into the following finite problem with finitely many constraints:

$$\begin{aligned} \min_{\theta,\mathbf{v},\mathbf{x}_k,z_{ijk}} \quad & \epsilon \cdot v_{N+1} + \tfrac{1}{N}\sum_{i\in[N]} v_i, \\ s.t. \quad & \|\widetilde{\mathbf{y}}_{ij} - \mathbf{x}_k\|_2^2 - v_{N+1}\cdot\|\widetilde{\mathbf{y}}_{ij} - \mathbf{y}_i\|_2 - v_i \le M z_{ijk}, \\ & \mathbf{x}_k \in S(w_k, \theta), \\ & \sum_{k\in[K]} z_{ijk} = K-1, \\ & \theta \in \Theta, v \in \mathcal{V}, z_{ijk} \in \{0,1\}, \forall i \in [N], j \in [J_i], k \in [K]. \end{aligned} \tag{6}$$

At each iteration, new constraints are determined to add to the previous set of constraints in (6) by solving the following **Maximum constraint violation problem:** $\forall i \in [N]$,

$$CV_i = \max_{\widetilde{\mathbf{y}}\in\mathcal{Y}} l_K(\widetilde{\mathbf{y}}, \widehat{\theta}) - \widehat{v}_{N+1}\cdot\|\widetilde{\mathbf{y}} - \mathbf{y}_i\|_2 - \widehat{v}_i. \tag{7}$$

Denote $\widetilde{\mathbf{y}}_i$ the optimal solution of (7) for each $i \in [N]$. Whenever we find that $CV_i > 0$, we append $\widetilde{\mathbf{y}}_i$ to $\widetilde{\mathcal{Y}}_i$. As a result, we tighten our approximation for the infinite set of constraints in (3) by imposing the additional constraint $l_K(\widetilde{\mathbf{y}}_i, \widehat{\theta}) - \widehat{v}_{N+1}\cdot\|\widetilde{\mathbf{y}}_i - \mathbf{y}_i\|_2 - \widehat{v}_i \le 0$ in the next iteration.

With the above assumptions and analysis, we now present our method to solve WRO-IMOP in Algorithm 1. We also illustrate the general scheme of Algorithm 1 in Figure 1.

Remark 3. In Step 6, the maximum constraint violation problem can be solved exactly and efficiently by invoking solver such as Baron (Sahinidis 1996). Nevertheless, it can also be solved approximately by decomposing into K subproblems, each of which is a possibly nonconvex program when $\widehat{v}_{N+1} < 1$. Nevertheless, we note that this nonconvex problem is a quadratically constrained quadratic program (QCQP) with a single constraint, and thus can be solved exactly and efficiently through the so-called S-procedure (Boyd and Vandenberghe 2004; Pólik and Terlaky 2007). Additionally, K different subproblems can be solved independently and in parallel, allowing a linear speedup of Step 6.

For completeness, we give the convergence proof of Algorithm 1 in the following theorem.

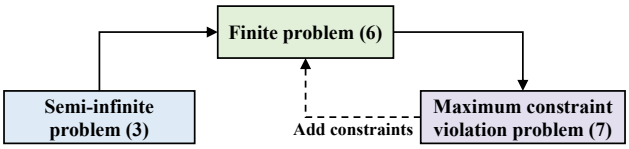


Figure 1: General scheme of Algorithm 1.

Theorem 2. Under Assumptions 1 - 3, Algorithm 1 converges within $(\frac{GR_0}{\delta}+1)^{n_\theta+N+1}$ iterations. Here,

$$G = (1 + 2R + \frac{4(B+R)\kappa}{\lambda}),$$

$$R_0 = \sqrt{D^2 + N\big((m+1)V_2 - mV_1\big)^2 + \left(\frac{V_2 - V_1}{\epsilon}\right)^2}.$$

Remark 4. The proof of convergence is in spirit similar to that of the cutting plane methods for robust optimization and distributionally robust optimization (Mutapcic and Boyd 2009; Luo and Mehrotra 2017). In practice, we mention that the actual number of iterations typically required is much smaller than $(\frac{GR_0}{\delta}+1)^{n_\theta+N+1}$ as can be seen in the experiments.

Variants of Algorithm 1 Note that we add $\widetilde{\mathbf{y}}_i$ to $\widetilde{\mathcal{Y}}_i$ whenever $V_i > 0$ for each $i \in [N]$. Nevertheless, from the convergence proof, it suffices to add only one $\widetilde{\mathbf{y}}_i$ corresponding to the biggest V_i that are positive. Consequently, we dramatically ease the computational burden in each iteration.

Performance Guarantees

One of the main goals of statistical analysis of learning algorithms is to understand how the excess risk of a data dependent decision rule output by the empirical risk minimization depends on the sample size of the observations and on the "complexity" of the class Θ. Next, we provide a performance guarantee for WRO-IMOP by showing below that the excess risk of the estimator obtained by solving WRO-IMOP would converge sub-linearly to zero.

Theorem 3 (Excess risk bound)**.** Define the minimax risk estimator

$$\theta^* \in \arg\min_{\theta\in\Theta}\left\{\sup_{Q\in\mathbb{B}_\epsilon(P)} \mathbb{E}_{\mathbf{y}\sim Q}\left[l_K(\mathbf{y},\theta)\right]\right\},$$

where P is the distribution from which the observations $\{\mathbf{y}_i\}_{i\in[N]}$ are drawn, and the minimax empirical risk estimator

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta}\left\{\sup_{Q\in\mathbb{B}_\epsilon(\widehat{P}_N)} \mathbb{E}_{\mathbf{y}\sim Q}\left[l_K(\mathbf{y},\theta)\right]\right\}.$$

and $\widehat{P}_N$ is the empirical distribution of the observations $\{\mathbf{y}_i\}_{i\in[N]}$.

Under Assumptions 1 - 3, $\forall 0 < \delta < 1$, the following holds with probability at least $1-\delta$:

$$\begin{aligned} \sup_{Q\in\mathbb{B}_\epsilon(P)} \mathbb{E}_{\mathbf{y}\sim Q}\left[l_K(\mathbf{y},\widehat{\theta}_N)\right] - \sup_{Q\in\mathbb{B}_\epsilon(P)} \mathbb{E}_{\mathbf{y}\sim Q}\left[l_K(\mathbf{y},\theta^*)\right] \\ \le \tfrac{H}{\sqrt{N}} + \tfrac{3(B+R)^2\sqrt{\log(2/\delta)}}{\sqrt{2N}}, \end{aligned}$$

where H is a constant depending only on $D, B, R, n_\theta, \kappa$:

$$H = 96\left(\frac{3D\sqrt{n_\theta}}{\kappa} + 2R\right)(B+R).$$

Remark 5 (Performance Guarantees)**.**

- The bounded support assumption for $\mathcal{Y}$ of the noisy decisions is restrictive but seems to be unavoidable for any a priori guarantees of the type described in Theorem 3. In future work, we will investigate whether we could obtain other types of performace guarantees while relaxing $\mathbb{P}_{\mathbf{y}}$ to be light-tailed.
- Analogous to the convergence rate of empirical risk minimization when $\epsilon = 0$, we get an $\mathcal{O}(1/\sqrt{N})$ excess risk bound. However, the obtained excess risk bound does not depend on the radius ϵ of the Wasserstein ambiguity set. Similar to Lee and Raginsky (2018), this phenomenon is due to the fact that we are using the Lipschitz continuity of the loss function $l_K(\mathbf{y}, \theta)$.
- The right terms in the excess risk bound inequality increase as either D, B, R, n_θ grow or κ shrinks, indicating that the learnability of the decision making model decreases. This is consistent with our observation that uncertainties in the model, data, and parameter space will enhance the difficulty of learning the parameters through inverse multiobjective optimization in general.

Experiments

In this section, we provide an MQP and a portfolio optimization problem to illustrate the performance of Algorithm 1. The MISOCPs for IMOP are solved by Gurobi. All the algorithms are programmed with Julia (Bezanson et al. 2017).

Sythetic Data: Learning the Objective Functions of an MQP

Consider the following multiobjective quadratic optimization problem.

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2_+} \quad & \begin{pmatrix} f_1(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q_1 \mathbf{x} + \mathbf{c}_1^T \mathbf{x} \\ f_2(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q_2 \mathbf{x} + \mathbf{c}_2^T \mathbf{x} \end{pmatrix} \\ s.t. \quad & A\mathbf{x} \leq \mathbf{b}, \end{aligned}$$

where the parameters of the two objective functions are

$$Q_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \mathbf{c}_1 = \begin{bmatrix} -0.5 \\ -1 \end{bmatrix}, Q_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{c}_2 = \begin{bmatrix} -5 \\ -2.5 \end{bmatrix},$$

and the parameters for the feasible region are

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}.$$

We seek to learn $\mathbf{c}_1$ and $\mathbf{c}_2$ in this experiment. The data is generated as follows. We first compute Pareto optimal solutions $\{\mathbf{x}_i\}_{i\in[N]}$ by solving WP with weight samples $\{w_i\}_{i\in[N]}$ that are uniformly chosen from $\mathscr{W}_2$. Next, the noisy decision $\mathbf{y}_i$ is obtained by adding noise to $\mathbf{x}_i$ for each $i \in [N]$. More precisely, $\mathbf{y}_i = \mathbf{x}_i + \epsilon_i$, where each element of ϵ_i has a uniform distribution supporting on $[-0.25, 0.25]$ with mean 0 for all $i \in [N]$. We assume that $\mathbf{c}_1$ and $\mathbf{c}_2$ are within $[-6, 0]^2$, and the first elements for them are given. $K = 6$ weights from $\mathscr{W}_2$ are evenly sampled. The radius ϵ of the Wasserstein ambiguity set is selected from the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We report below the results with lowest prediction error across all candidate radii. The stopping criteria δ is set to be 0.1. Then, we implement Algorithm 1 with different N.

To illustrate the performance of the algorithm in a statistical way, we run 10 repetitions of the experiments. Figure 2a shows the maximum constraint violation $\max_{i\in[N]} V_i$ versus iteration for one repetition when $N = 10$. As can be seen in the figure, the algorithm converges very fast. In Figure 2b, we report the prediction errors averaged over 10 repetitions with both the robust and non-robust approaches for different N. Here, we use an independent validation set that consists of 10^5 noisy decisions generated in the same way as the training data to compute the prediction error. The experiments suggest that the Wasserstein distributionally robust approach can significantly reduce the prediction error, especially when N is small, i.e., we have a very limited number of observations.

To further illustrate the performance of Algorithm 1, we randomly pick one repetition and plot the estimated Pareto optimal sets using both approaches in Figure 2c. We can see clearly that the estimated Pareto optimal set by the distributionally robust approach is closer to the real Pareto optimal set than that of the non-robust approach. Also, one could expect that the estimated Pareto optimal sets will get closer and closer to the true one as K and N increase.

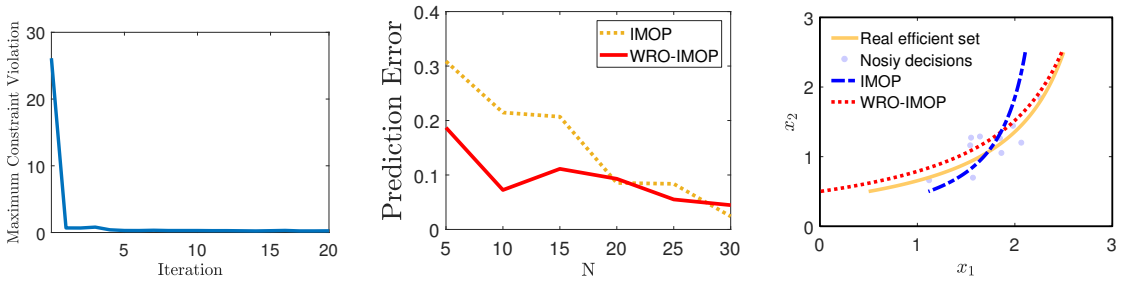
Real World Case Study: Learning the Expected Returns

We consider a portfolio selection problem, where investors need to determine the fraction of their wealth to invest in each security in order to maximize the total return and minimize the total risk. The classical Markovitz mean-variance portfolio selection (Markowitz 1952) in the following is frequently employed by analysts.

$$\begin{aligned} \min \quad & \begin{pmatrix} f_1(\mathbf{x}) & = -\mathbf{r}^T\mathbf{x} \\ f_2(\mathbf{x}) & = \mathbf{x}^T Q \mathbf{x} \end{pmatrix} \\ s.t. \quad & 0 \leq x_i \leq b_i, \qquad \forall i \in [n], \\ & \sum_{i=1}^{n} x_i = 1, \end{aligned}$$

where $\mathbf{r} \in \mathbb{R}^n_+$ is a vector of individual security expected returns, $Q \in \mathbb{R}^{n\times n}$ is the covariance matrix of securities returns, $\mathbf{x}$ is a portfolio specifying the proportions of capital to be invested in the different securities, and b_i is an upper bound on the proportion of security i, $\forall i \in [n]$.

Dataset: The dataset is derived from monthly total returns of 30 stocks from a blue-chip index which tracks the performance of top 30 stocks in the market when the total investment universe consists of thousands of assets. The true expected returns and true return covariance matrix for the first 8 securities are given in the supplementary material. Suppose a learner seeks to learn the expected return for the first



(a) (b) (c)

Figure 2: Learning the objective functions of an MQP. (a) Maximum constraint violation versus iteration for $N = 15$. (b) Prediction errors for two methods with different N. Results are averaged over 10 repetitions. (c) The Pareto optimal set and estimated Pareto optimal sets by using IMOP and WRO-IMOP with $N = 10$.

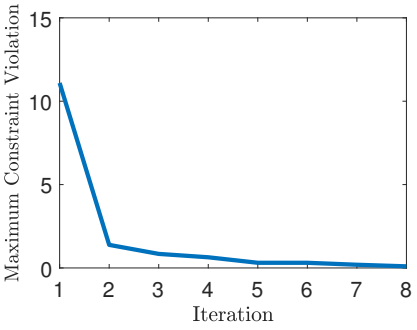


Figure 3: Maximum constraint violation versus iteration.

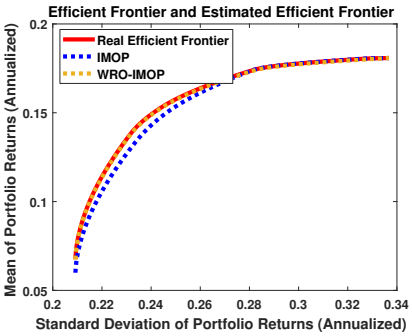


Figure 4: The red line indicates the real efficient frontier. The yellow dots indicates the estimated efficient frontier by solving WRO-IMOP. The blue dots indicates the estimated efficient frontier using the non-robust approach.

four securities that an analyst uses based on 20 noisy decisions from investors that the analyst serves.

The noisy decision for each investor $i \in [20]$ is generated as follows. We set each upper bound for the proportion of the 8 securities to $b_i = 1.0, \forall i \in [8]$. Then, we uniformly sample 20 weights and use them to generate optimal portfolios on the efficient frontier that is plot in Figure 4. Subsequently, each component of these portfolios is rounded to the nearest thousandth, which can be seen as measurement error. The radius ϵ of the Wasserstein ambiguity set is selected from the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The stopping criteria δ is set to be 0.1.

Figure 3 shows that our algorithm converges in 8 iterations. We also plot the estimated efficient frontiers using both the robust and non-robust approaches with $K = 6$ in Figure 4. We can see that the estimated efficient frontier of the Wasserstein distributionally robust approach is closer to the real one than the non-robust approach, showing that our method in this paper allows for a lower prediction error when a limited number of decisions observed are accessible. Note that the first function is not strongly convex. The experiment results suggest that our reformulation is generalizable to a broader class of problems.

Conclusions and Future Work

In this paper, we present a novel Wasserstein distributionally robust framework for constructing inverse multiobjective optimization estimator. We show that the proposed framework has statistical performance guarantees, and the excess risk of the distributionally robust inverse multiobjective optimization estimator would converge to zero with a sub-linear rate as the number of observed decisions approaches to infinity. To solve the resulting minmax problem, we reformulate it as a semi-infinite program and develop a cutting-plane algorithm which converges to an approximate solution in finite iterations. We demonstrate the effectiveness of our method on both a multiobjective quadratic program and a portfolio optimization problem.

We note that the technical assumptions of strong convexity of the objective functions in DMP and bounded support of the observations might not be fully satisfied in many real world scenarios. It remains to be seen what will happen if we relax these assumptions. Without them, these formulations and algorithm are still valid but performances are unlikely to be completely guaranteed. For example, we can only guarantee that the optimal objective of the Semi-infinite reformulation in Theorem 1 provides an upper bound for WRO-IMOP. In future, we will work on extending the current framework and analysis to scenarios such as when DMP has at least one strongly convex objective function or data that follows light-tailed distribution.

References

Ahuja, R. K.; and Orlin, J. B. 2001. Inverse optimization. *Operations Research* 49(5): 771–783.

Aswani, A.; Shen, Z.-J.; and Siddiq, A. 2018. Inverse optimization with noisy data. *Operations Research* .

Bärmann, A.; Pokutta, S.; and Schneider, O. 2017. Emulating the expert: Inverse optimization through online learning. In *International Conference on Machine Learning*, 400–410.

Bertsimas, D.; Gupta, V.; and Paschalidis, I. C. 2015. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming* 153(2): 595–633.

Bezanson, J.; Edelman, A.; Karpinski, S.; and Shah, V. B. 2017. Julia: A fresh approach to numerical computing. *SIAM Review* 59(1): 65–98.

Bonnans, J. F.; and Shapiro, A. 1998. Optimization problems with perturbations: A guided tour. *SIAM Review* 40(2): 228–264.

Boyd, S.; and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge university press.

Dong, C.; Chen, Y.; and Zeng, B. 2018. Generalized Inverse Optimization through Online Learning. In *NeurIPS*.

Dong, C.; and Zeng, B. 2020. Expert Learning through Generalized Inverse Multiobjective Optimization: Models, Insights, and Algorithms. In *ICML*.

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

Esfahani, P. M.; Shafieezadeh-Abadeh, S.; Hanasusanto, G. A.; and Kuhn, D. 2018. Data-driven inverse optimization with imperfect information. *Mathematical Programming* 167(1): 191–234.

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

Gass, S.; and Saaty, T. 1955. The computational algorithm for the parametric objective function. *Naval Research Logistics* 2(1-2): 39–45.

Hettich, R.; and Kortanek, K. O. 1993. Semi-infinite programming: theory, methods, and applications. *SIAM Review* 35(3): 380–429.

Joachims, T.; Finley, T.; and Yu, C.-N. J. 2009. Cutting-plane training of structural SVMs. *Machine learning* 77(1): 27–59.

Keshavarz, A.; Wang, Y.; and Boyd, S. 2011. Imputing a convex objective function. In *Intelligent Control (ISIC), 2011 IEEE International Symposium on*, 613–619. IEEE.

Lee, J.; and Raginsky, M. 2018. Minimax Statistical Learning with Wasserstein distances. In *NeurIPS*.

Luo, F.; and Mehrotra, S. 2017. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920* .

Markowitz, H. 1952. Portfolio selection. *The Journal of Finance* 7(1): 77–91.

Mutapcic, A.; and Boyd, S. 2009. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software* 24(3): 381–406.

Pólik, I.; and Terlaky, T. 2007. A survey of the S-lemma. *SIAM Review* 49(3): 371–418.

Sahinidis, N. V. 1996. BARON: A general purpose global optimization software package. *Journal of global optimization* 8(2): 201–205.

Shafieezadeh-Abadeh, S.; Esfahani, P. M.; and Kuhn, D. 2015. Distributionally robust logistic regression. In *NIPS*.

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