On the complexity of the data-driven Wasserstein distributionally robust binary problem

Hyoseok Kim[1,2], Dimitri Watel[3,4], Alain Faye[2,3], Cédric Hervet[1]

[1] Kardinal, France
[2] CEDRIC-CNAM, France
[3] ENSIIE, France
[4] SAMOVAR, France
{hyoseok.kim, cedric.hervet}@kardinal.ai
{dimitri.watel, alain.faye}@ensiie.fr

1 Introduction

In this paper, we use a data-driven Wasserstein distributionally robust framework to consider uncertain parameters in optimization problems.

Distributionally robust optimization (DRO) is an approach to optimization under uncertainty that assumes only partial information on the probability distribution of the uncertain parameters. For example, in transportation optimization problems, the probability distribution of the uncertainties on road traffic is unknown (even if it can be estimated through historical records). DRO can be seen as the unification of stochastic and robust optimization. DRO offers more robustness than stochastic optimization because instead of considering that we know the real distribution, we optimize inside of a set of candidate distributions. In general, DRO is less conservative than optimizing the worst-case scenario in robust optimization because an historical data can contain outliers due to errors or bad measures. When the distribution set is reduced to one single distribution D, DRO is equivalent to stochastic optimization using D as the real distribution. Unlike stochastic optimization, we assume that the decision maker does not know the real distribution of the uncertainties. Instead, we consider that it belongs to an ambiguity set of distributions and we want to be robust on this set of distributions. In order to construct this set of distributions, we compute the empirical distribution with a training samples dataset, assuming that each sample can appear with the same probability. We then consider a ball of distributions around that distribution. The metric we use to define that ball is the Wasserstein metric that gives a distance value between two distributions. This particular case of DRO is called data-driven Wasserstein DRO (WDRO). As a generalization of the stochastic optimization and the robust optimization, WDRO is obviously NP-hard. Recently, data-driven Wasserstein DRO gained attention in operations research and machine learning literature [3, 4, 6].

In the case of a combinatorial optimization problem when only the cost function is affected by uncertainties, we show that WDRO counterpart of a polynomial problem remains polynomial. More precisely, we prove that, if the optimization problem can be written as a 0-1 integer linear program with n variables, the complexity of solving the distributionally robust counterpart is at most $n + 1$ times the complexity of solving the original problem. This means that every complexity results (related to polynomiality) of an optimization problem is kept for its WDRO counterpart. For example, the WDRO counterpart of any α-approximable NP-hard 0-1 discrete problem is also α-approximable. Our theoretical results show that under certain conditions, the WDRO counterpart of a combinatorial problem is not much harder than the original problem which means that WDRO is a powerful framework to consider uncertainties in combinatorial problems without being expensive.

2 Definition of the framework

This section is dedicated to defining the necessary notations and concepts. We first review models for optimizing under uncertainty, including the Distributionally Robust Optimization and the data-driven Wasserstein DRO. We finish by explaining how we transform a combinatorial problem into an instance of the WDRO problem.

2.1 Optimizing under uncertainty

Stochastic optimization and robust optimization are two classic frameworks to model uncertainty in optimization problems. Distributionally robust optimization is an alternative framework that unifies both approaches.

In this part, we consider an optimization where $\boldsymbol{x} \in \mathcal{X} \subset \{0;1\}^n$ is the decision vector and h is the cost function we want to optimize. This function is subject to uncertainty. We write $\boldsymbol{\xi} \in \Xi \subset \mathbb{R}^n$ the uncertain parameter and $h(\boldsymbol{x}, \boldsymbol{\xi})$ as the objective value of $\boldsymbol{x}$ given a fixed value $\boldsymbol{\xi}$ of uncertainty.

Stochastic optimization In stochastic optimization, we assume that the exact probability distribution P^* of $\boldsymbol{\xi}$ is known. The random variable associated to uncertainties is called $\widetilde{\boldsymbol{\xi}}$. We want to compute a feasible solution $\boldsymbol{x}$ minimizing the expected value of $h(\boldsymbol{x}, \widetilde{\boldsymbol{\xi}})$. In other words :

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{P^*}[h(\boldsymbol{x}, \widetilde{\boldsymbol{\xi}})]$$

Robust optimization In robust optimization, we consider that only the support of the uncertain parameters is known which means that we know all the different values that can be taken by the uncertain parameters. The objective is to compute a feasible solution $\boldsymbol{x}$ minimizing the maximum possible value of $h(\boldsymbol{x}, \boldsymbol{\xi})$.

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{\xi} \in \Xi} h(\boldsymbol{x}, \boldsymbol{\xi})$$

Distributionally robust optimization Distributionally robust optimization is an approach to optimization under uncertainty that assumes only partial distributional information. Unlike the classic approach of stochastic optimization, in DRO, the exact probability distribution P^* is unknown. Instead, we assume that it belongs to an ambiguity set $\mathcal{P}$ of distributions constructed from the partial information. We compute a feasible solution $\boldsymbol{x}$ minimizing the maximum expected value of $h(\boldsymbol{x}, \widetilde{\boldsymbol{\xi}})$ among all the possible distributions. In some way, DRO is a robust optimization model where Ξ is replaced by the set $\mathcal{P}$ of distributions.

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[h(\boldsymbol{x}, \widetilde{\boldsymbol{\xi}})]$$

2.2 Data-driven Wasserstein DRO

The exact probability distribution P^* can be estimated through a finite sample dataset. A natural method is the sample average approximation (SAA) where P^* is replaced by the empirical distribution $\widehat{P}_N$ obtained by averaging of the sample dataset.

Definition 1 *Given a list $\widehat{\Xi}$ of N values from Ξ. We define the empirical distribution $\widehat{P}_N$ with $\frac{1}{N} \sum_{\boldsymbol{\xi} \in \widehat{\Xi}} \delta_{\boldsymbol{\xi}}$ where $\delta_{\boldsymbol{\xi}}$ is the Dirac distribution at $\boldsymbol{\xi} \in \Xi$.*

From this empirical distribution, we build our set of distributions $\mathcal{P}$ with a ball of distributions centered at $\widehat{P}_N$. The metric used to define that ball is the p-Wasserstein distance. We define with $\mathcal{M}(\Xi)$ the set of all probability distributions on Ξ.

Definition 2 (p-Wasserstein distance) *Let $Q_1, Q_2 \in \mathcal{M}(\Xi)$ and $1 \leq p \leq +\infty$.*

$$d_{W_p}(Q_1, Q_2) = \inf_{\Pi \in \Gamma(Q_1, Q_2)} \int_{\Xi^2} \|\boldsymbol{\xi_1} - \boldsymbol{\xi_2}\|_p \Pi(d\xi_1, d\xi_2)$$

where $\Gamma(Q_1, Q_2)$ is the set of distributions on $\Xi \times \Xi$ with marginal Q_1 and Q_2.

The Wasserstein distance can be used to compare any two distributions Q_1 and Q_2 that can be discrete or continuous. It can be interpreted as the minimum transportation cost for moving from the probability density function of Q_1 to the one of Q_2. The p-norm is used to evaluate the cost of moving some probability from the vector $\boldsymbol{\xi_1}$ to $\boldsymbol{\xi_2}$.

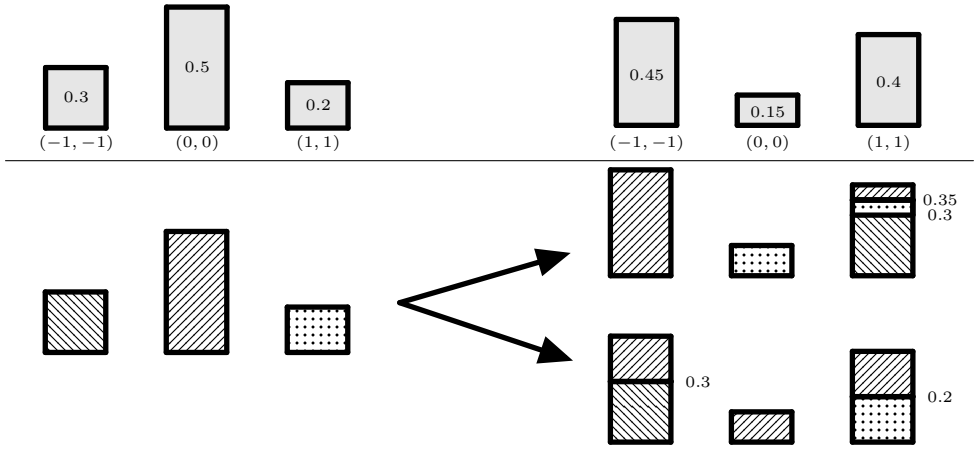

FIG. 1 – Transformations from histogram Q_1 (upper left) to histogram Q_2 (upper right) representing probabilities of apparition of three vectors $(-1,-1), (0,0)$ and $(1,1)$. Wasserstein distance can be visualized as the best transportation for moving from Q_1 to Q_2. In the three lower histograms, we can see two possible transformations of the histogram Q_1 to histogram Q_2. The second plot seems more efficient since it only moves a portion of the second bar of Q_1, whereas, in the first plot, all the three bars are moved, including the whole second bar. The cost from moving a portion from one bar to another depends on the norm we use. For instance, assuming we use the 2-norm, moving a fraction δ from $(1,1)$ to $(0,0)$ costs $\delta \cdot \|(1,1)-(0,0)\|_2 = \delta\sqrt{2}$. In this example, the cost of the upper transformation would be $0.3\sqrt{8}+0.15\sqrt{2}+0.5\sqrt{2} \simeq 1.77$. The cost of the lower transformation would be $0.35\sqrt{2} \simeq 0.5$. The p-Wasserstein distance is the minimum cost (using the p-norm) obtained by the best transformation among all the transformations possible from Q_1 to Q_2. In this case, the distance is at most 0.5.

Definition 3 (Wasserstein ball of radius $\varepsilon > 0$ centered at $\widehat{P}_N$)

$$B_{p,\varepsilon}(\widehat{P}_N) := \left\{ Q \in \mathcal{M}(\Xi) : d_{W_p}(\widehat{P}_N, Q) \leq \varepsilon \right\}$$

The Wasserstein ball of radius ε centered at $\widehat{P}_N$ contains all the probability distributions that are at most at a distance of ε using the Wasserstein metric. When $\varepsilon = 0$, it only contains the empirical distribution $\widehat{P}_N$.

2.3 Transform a problem into a WDRO problem

We consider a combinatorial optimization problem Pb written as a 0-1 ILP :

$$\text{Pb} : \inf_{\substack{\boldsymbol{x} \in [\![0;1]\!]^n \\ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}}} \boldsymbol{c}\boldsymbol{x}$$

The vector c is subject to uncertainty. As done previously, Ξ is the set of values that can be taken by this parameter and $\widetilde{\xi}$ is the random variable associated to this uncertainty. The exact distribution P^* is not known but a finite sample dataset $\widehat{\Xi}$ is given from which we deduce the empirical distribution $\widehat{P}_N$.

Finally, given an integer $1 \leq p \leq +\infty$ and a real $\varepsilon > 0$, we define the (p, ε)-WDRO-Pb problem.

$$(p,\varepsilon)\text{-WDRO-Pb} : \inf_{\substack{\boldsymbol{x}\in[\![0;1]\!]^n \\ \boldsymbol{A}\boldsymbol{x}\leq\boldsymbol{b}}} \sup_{Q\in B_{p,\varepsilon}(\widehat{P}_N)} \mathbb{E}_Q[\widetilde{\boldsymbol{\xi}}\boldsymbol{x}]$$

3 Reformulation

We reformulate the problem by simplifying the expression $\sup_{Q\in B_{p,\varepsilon}(\widehat{P}_N)} \mathbb{E}_Q[\widetilde{\boldsymbol{\xi}}\boldsymbol{x}]$ into a deterministic term. We first give some intermediate lemmas. Proofs of the lemmas are omitted for lack of space.

Lemma 1 *Let P and Q be two distributions on Ξ. Then $\|\mathbb{E}_P[\widetilde{\boldsymbol{\xi}}] - \mathbb{E}_Q[\widetilde{\boldsymbol{\xi}}]\|_p \leq d_{W_p}(P, Q)$*

Lemma 1 can be proven using triangle inequality for integrals and the definition of the p-Wasserstein distance.

Lemma 2 $$\sup_{Q\in B_{p,\varepsilon}(\widehat{P}_N)} \mathbb{E}_Q[\boldsymbol{x}\widetilde{\boldsymbol{\xi}}] = \sup_{\|\boldsymbol{\Delta}\|_p\leq\varepsilon} \mathbb{E}_{\widehat{P}_N}[\boldsymbol{x}(\widetilde{\boldsymbol{\xi}} + \boldsymbol{\Delta})]$$

Lemma 2 transforms the constraint on the Wasserstein ball into a constraint on a norm of a vector $\boldsymbol{\Delta}$.

We now prove our main theorem that reformulates the probabilistic objective function of (p, ε)-WDRO-Pb into a deterministic objective function. Recall that the dual norm of $\|.\|_p$ is $\|.\|_q$ with q such as $\frac{1}{p} + \frac{1}{q} = 1$ (p (resp q) can be infinite if $q = 1$ (resp. $p = 1$)) which means that for any vector $\boldsymbol{z} \in \mathbb{R}^n$, $\|\boldsymbol{z}\|_q = \sup\{\boldsymbol{z}\boldsymbol{x} \mid \|\boldsymbol{x}\|_p \leq 1\}$.

Theorem 1 $$\inf_{\substack{\boldsymbol{x}\in[\![0;1]\!]^n \\ \boldsymbol{A}\boldsymbol{x}\leq\boldsymbol{b}}} \sup_{Q\in B_{p,\varepsilon}(\widehat{P}_N)} \mathbb{E}_Q[\boldsymbol{x}\widetilde{\boldsymbol{\xi}}] = \inf_{\substack{\boldsymbol{x}\in[\![0;1]\!]^n \\ \boldsymbol{A}\boldsymbol{x}\leq\boldsymbol{b}}} \|\boldsymbol{x}\|_q \cdot \varepsilon + \frac{1}{N}\sum_{\boldsymbol{\xi}\in\widehat{\Xi}} \boldsymbol{x}\boldsymbol{\xi}$$

Proof : Let $\boldsymbol{x} \in [\![0; 1]\!]^n$ such that $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$.

$$\begin{aligned}\sup_{Q\in B_{p,\varepsilon}(\widehat{P}_N)} \mathbb{E}_Q[\boldsymbol{x}\widetilde{\boldsymbol{\xi}}] &= \sup_{\|\boldsymbol{\Delta}\|_p\leq\varepsilon} \mathbb{E}_{\widehat{P}_N}[\boldsymbol{x}(\widetilde{\boldsymbol{\xi}} + \boldsymbol{\Delta})] = \mathbb{E}_{\widehat{P}_N}[\boldsymbol{x}\widetilde{\boldsymbol{\xi}}] + \sup_{\|\boldsymbol{\Delta}\|_p\leq\varepsilon} \boldsymbol{x}\boldsymbol{\Delta} = \frac{1}{N}\sum_{\xi\in\widehat{\Xi}} \boldsymbol{x}\boldsymbol{\xi} + \sup_{\|\boldsymbol{\Theta}\|_p\leq 1} \boldsymbol{x}\boldsymbol{\Theta}\cdot\varepsilon \\ &= \frac{1}{N}\sum_{\xi\in\widehat{\Xi}} \boldsymbol{x}\boldsymbol{\xi} + \|\boldsymbol{x}\|_q \cdot \varepsilon\end{aligned}$$

□

Remark 1 *The new term $\|\boldsymbol{x}\|_q$ added from the reformulation can be seen as a penalty from the number of uncertain elements we choose in our solution $\boldsymbol{x}$. The more uncertain elements we use to construct our solution, the more penalty we get from this term $\|\boldsymbol{x}\|_q$. What this reformulation means is that we want a good trade-off between the evaluation of a solution using the past data and its uncertainty.*

Remark 2 *Note that this result can also be shown using the duality theory from the original problem. It can be seen as a special case of the reformulation using duality theory done in [3].*

4 Main algorithm

In the following section, we note $\mathcal{J} = (\boldsymbol{A}, \boldsymbol{b}, P)$ an instance of (p, ε)-WDRO-Pb where $\boldsymbol{A}, \boldsymbol{b}$ are the constraint coefficients matrix and vector which defines the set of feasible solutions and P is the empirical distribution, which is the center of the distribution ambiguity set. We also note $\mathcal{I} = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$ an instance of Pb where $\boldsymbol{A}, \boldsymbol{b}$ are the constraint coefficients matrix and vector which defines the set of feasible solutions and $\boldsymbol{c}$ is the cost coeffients vector.

Theorem 1 shows that (p, ε)-WDRO-Pb can be reformulated into a problem with a deterministic objective function, without any notion of probability distributions.

As $\boldsymbol{x}$ is a binary vector, we can rewrite $\|\boldsymbol{x}\|_q = \sqrt[q]{\sum_i x_i^q} = \sqrt[q]{\sum_i x_i} = \sqrt[q]{|\boldsymbol{x}|}$ where $|\boldsymbol{x}| = \sum_i x_i$.

Suppose that an algorithm $\mathcal{A}$ can return a solution of (Pb). We provide a method to solve its WDRO counterpart using the same algorithm $\mathcal{A}$. Notice that unlike (Pb), the objective function in (p, ε)-WDRO-Pb is not linear. Solving such a problem can be hard in general cases. The idea of our method is to linearize the non linear part of the objective function. To do so, we want to replace $f(|\boldsymbol{x}|) = \sqrt[q]{|\boldsymbol{x}|}$ by tangents $g_i : k \mapsto a_i k + b_i$ where g_i is the tangent of f at value $|\boldsymbol{x}| = i$. The idea of using tangents is the fact that it gives a linear function which dominates f at all the evaluated points. Since we only evaluate f at integer values, instead of using tangents, we can take any $a_i \in [\sqrt[q]{i+1} - \sqrt[q]{i}; \sqrt[q]{i} - \sqrt[q]{i-1}]$.

When the objective function is linearized, we are able to solve the linearized problem using the algorithm $\mathcal{A}$. This procedure is described with Algorithm 1.

Remark 3 *In this paper, we will use $a_i = \sqrt[q]{i+1} - \sqrt[q]{i}$ and consider that it can be computed in polynomial time to avoid technical details. However, in order to have a complete proof, we must show that we can choose a rational $a_i \in [\sqrt[q]{i+1} - \sqrt[q]{i}; \sqrt[q]{i} - \sqrt[q]{i-1}]$ in polynomial time that keeps the same properties.*

Algorithm 1 Algorithm to solve (p, ε)-WDRO-Pb

Require: An algorithm $\mathcal{A}$ for Pb returning a feasible solution and an instance $\mathcal{J} = (\boldsymbol{A}, \boldsymbol{b}, \widehat{P}_N)$ of (p, ε)-WDRO-Pb

Ensure: A feasible solution $\boldsymbol{x}$ of $\mathcal{J}$

if $p = 1$ **then**

$\boldsymbol{x} \leftarrow \mathcal{A}(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c} = \frac{1}{N} \sum_{\xi \in \widehat{\Xi}} \boldsymbol{\xi})$

return $\boldsymbol{x}$ if ($\boldsymbol{c} \cdot \boldsymbol{x} + \varepsilon < 0$ or $\mathbf{0}$ is unfeasible) and the vector $\mathbf{0}$ otherwise

else

$q \leftarrow \frac{p}{p-1}$

for i from 0 to n **do**

$a_i \leftarrow \sqrt[q]{i+1} - \sqrt[q]{i}$

$\boldsymbol{x}^{(\boldsymbol{i})} \leftarrow \mathcal{A}(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c} = \frac{1}{N} \sum_{\xi \in \widehat{\Xi}} \boldsymbol{\xi} + \varepsilon a_i \cdot \mathbf{1})$

return the $\boldsymbol{x}^{(\boldsymbol{i})}$ that minimizes the cost

Lemma 3 *Assuming $\mathcal{A}$ is a polynomial algorithm, the complexity of Algorithm 1 is also polynomial.*

Lemma 4 *Let $b_i = \sqrt[q]{i} - a_i \cdot i$. Let $\mathcal{J} = (\boldsymbol{A}, \boldsymbol{b}, \widehat{P}_N)$ be an instance of (p, ε)-WDRO-Pb. We assume $p \neq 1$, let $\boldsymbol{x}^*$ be an optimal solution of $\mathcal{J}$ with value ω^* and, for $i \in [\![0; n]\!]$, let $\boldsymbol{x}_i^*$ be an optimal solution of $\mathcal{I}_i$, the instance of (Pb) defined as $\mathcal{I}_i = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c} = \frac{1}{N} \sum_{\xi \in \widehat{\Xi}} \boldsymbol{\xi} + \varepsilon \cdot a_i \cdot \mathbf{1})$ and let ω_i^* be its optimal value. Then $\min_{i \in [\![0;n]\!]} \omega_i^* + \varepsilon \cdot b_i = \omega^*$.*

The idea of the proof of Lemma 4 is the fact that if x_i an optimal solution of $\mathcal{I}_i$ then either $|x_i| = i$ or $|x_i| \neq i$ but the objective value of $|x_{|x_i|}|$ is lower than the value of x_i. This means that any solutions x_i such that $|x_i| \neq i$ can not be an optimal solution of $\mathcal{J}$.

Lemma 3 describes the time complexity of Algorithm 1 and Lemma 4 proves the correctness of Algorithm 1. Using both Lemmas 3 and 4, we can show that (p, ε)-WDRO-Pb can be solved in polynomial time if $\mathcal{A}$ is a polynomial algorithm solving (Pb).

Theorem 2 *If Pb is α-approximable in polynomial time, then, for any $p \in \mathbb{R}_+^* \cup \{+\infty\}$ and $\varepsilon > 0$, (p, ε)-WDRO-Pb is α-approximable in polynomial time.*

Remark 4 *When our framework is applied to the shortest path problem, we obtain a polynomial distributionally robust shortest path problem whereas most of the robust versions of the shortest path problem are known to be NP-hard. This result is similar to the one in [1] with a different framework.*

5 Conclusion and perspectives

In this paper, we describe distributionally robust optimization paradigm to take account of uncertainties, in particular, the data-driven Wasserstein distributionally robust optimization framework. Knowing a black box algorithm to solve a 0-1 discrete optimization problem, we propose an algorithm to solve its distributionally robust counterpart. Complexity results related to polynomiality of the original problem are still available for the new problem. We are aware that the purpose of our algorithm is only to give a theoretical complexity result on the WDRO problems. From a practical point of view, this algorithm should obviously be adapted to each combinatorial optimization problem on a case-by-case basis. For example, adapting the Dijkstra algorithm should provide a better time complexity than our Algorithm 1 for the WDRO shortest path problem.

An interesting perspective is to assume that the constraint coefficients are subject to uncertainty. In this case, we have to describe another framework that can take account of the modification of the feasible solution set by the uncertain coefficients. For example, one paradigm that can be used is the distributionally robust chance constrained programs studied in [2, 5, 7].

Références

[1] D. Bertsimas and Melvyn Sim. Robust discrete optimization and network flows. *Mathematical Programming*, 98 :49–71, 2003.

[2] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls, 2018. http ://arxiv.org/abs/1809.00210 arXiv :1809.00210.

[3] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric : Performance guarantees and tractable reformulations, 2017. http ://arxiv.org/abs/1505.05116 arXiv :1505.05116.

[4] Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance, 2016. http ://arxiv.org/abs/1604.02199 arXiv :1604.02199.

[5] Ran Ji and Miguel A Lejeune. Data-driven distributionally robust chance-constrained optimization with wasserstein metric. *Journal of Global Optimization*, 79(4) :779–811, 2021.

[6] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization : Theory and applications in machine learning, 2019. http ://arxiv.org/abs/1908.08729 arXiv :1908.08729.

[7] Weijun Xie. On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming*, 186(1) :115–155, 2021.