Coresets for Wasserstein Distributionally Robust Optimization Problems

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Wasserstein distributionally robust optimization (WDRO) is a popular model to enhance the robustness of machine learning with ambiguous data. However, the complexity of WDRO can be prohibitive in practice since solving its "minimax" formulation requires a great amount of computation. Recently, several fast WDRO training algorithms for some specific machine learning tasks (e.g., logistic regression) have been developed. However, the research on designing efficient algorithms for general large-scale WDROs is still quite limited, to the best of our knowledge. *Coreset* is an important tool for compressing large dataset, and thus it has been widely applied to reduce the computational complexities for many optimization problems. In this paper, we introduce a unified framework to construct the ϵ-coreset for the general WDRO problems. Though it is challenging to obtain a conventional coreset for WDRO due to the uncertainty issue of ambiguous data, we show that we can compute a "dual coreset" by using the strong duality property of WDRO. Also, the error introduced by the dual coreset can be theoretically guaranteed for the original WDRO objective. To construct the dual coreset, we propose a novel grid sampling approach that is particularly suitable for the dual formulation of WDRO. Finally, we implement our coreset approach and illustrate its effectiveness for several WDRO problems in the experiments.

1 Introduction

In the past decades, a number of optimization techniques have been proposed for solving machine learning problems [43]. However, real-world optimization problems often suffer from the issue of data ambiguity that can be generated by natural data noise, potential adversarial attackers [2], or the constant changes of the underlying distribution (e.g., continual learning [37]). As a consequence, our obtained dataset usually cannot be fully trusted. Instead it is actually a perturbation of the true distribution. The recent studies have shown that even small perturbation can seriously destroy the final optimization result and could also yield unexpected error for the applications like classification and pattern recognition [18; 44].

The "**distributionally robust optimization** (DRO)" is an elegant model for solving the issue of ambiguous data. The idea follows from the intuition of game theory [36]. Roughly speaking, the DRO aims to find a solution that is robust against the worst-case perturbation within a range of possible distributions. Given an empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\xi_i}$ where δ_{ξ_i} is the Dirac point mass at the i-th data sample ξ_i, the **worst-case empirical risk** at the hypothesis β is defined as $R^{\mathbb{P}_n}(\beta) = \sup_{\mathbb{Q}\in\mathcal{U}(\mathbb{P}_n)} \mathbb{E}^{\mathbb{Q}}[\ell(\beta,\xi)]$. Here $\mathcal{U}(\mathbb{P}_n)$ is the ambiguity set consisting of all possible perturbated distributions, and $\ell(\cdot,\cdot)$ is the non-negative loss function. The DRO model has shown its

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

promising advantage for enhancing the robustness for many practical machine learning problems, such as logistic regression [40], support vector machine [24], convex regression [4], neural networks [39], etc.

In this paper, we consider one of the most representative DRO models that is defined by using optimal transportation [45]. **Wasserstein distance** is a popular measure for representing the difference between two distributions; it indicates the minimum cost for transporting one distribution to the other. For $p \geq 1$, the p-th order Wasserstein distance between two probability distributions $\mathbb{P}$ and $\mathbb{P}'$ supported on Ξ is

$$W_p(\mathbb{P}, \mathbb{P}') = \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{P}')} \int_{\Xi \times \Xi} \mathtt{d}^p(\xi, \xi') \pi \left(\mathrm{d}\xi, \mathrm{d}\xi' \right) \right)^{\frac{1}{p}}, \tag{1}$$

where $\mathtt{d}(\cdot, \cdot)$ is a metric on Ξ, and $\Pi(\mathbb{P}, \mathbb{P}')$ is the set of all joint probability distributions on $\Xi \times \Xi$ with the marginals $\mathbb{P}$ and $\mathbb{P}'$. By using the above Wasserstein distance, we can define the ambiguity set $\mathcal{U}(\mathbb{P}_n)$ to be the p-th order **Wasserstein ball** $\mathbb{B}_{\theta,p}(\mathbb{P}_n)$, which covers all the distributions that have the p-th order Wasserstein distance at most $\theta > 0$ to the given empirical distribution $\mathbb{P}_n$. The use of Wasserstein ball is a discrepancy-based approach for choosing the ambiguity set [36, section 5]. Also let

$$R_{\theta,p}^{\mathbb{P}_n}(\beta) = \sup_{\mathbb{Q} \in \mathbb{B}_{\theta,p}(\mathbb{P}_n)} \mathbb{E}^{\mathbb{Q}}[\ell(\beta, \xi)] \tag{2}$$

denote the corresponding worst-case empirical risk. The **Wasserstein distributionally robust optimization** (WDRO) problem [23] is to find the minimizer

$$\beta_* = \arg\min_{\beta \in \mathcal{H}} R_{\theta,p}^{\mathbb{P}_n}(\beta) = \arg\min_{\beta \in \mathcal{H}} \sup_{\mathbb{Q} \in \mathbb{B}_{\theta,p}(\mathbb{P}_n)} \mathbb{E}^{\mathbb{Q}}[\ell(\beta, \xi)], \tag{3}$$

where $\mathcal{H}$ is the feasible region in the hypothesis space. It is easy to see that the WDRO is a minimax optimization problem.

Compared with other robust optimization models, the WDRO model enjoys several significant benefits from the Wasserstein metric, especially for the applications in machine learning [47; 41; 5; 16]. The Wasserstein ball captures much richer information than the divergence-based discrepancies for the problems like pattern recognition and image retrieval [38; 28; 15]. It has also been proved that the WDRO model yields theoretical quality guarantees for the "out-of-sample" robustness [11].

However, due to the intractability of the the inner maximization problem (2), it is challenging to directly solve the minimax optimization problem (3). As shown in the work of Esfahani and Kuhn [11], the WDRO problem (3) usually has tractable reformulations [40; 30; 35; 24; 19; 3]. Although these reformulations are polynomial-time solvable, the off-the-shelf solvers can be costly for large-scale data (e.g., the interior-point method [6] takes $O(n^3)$ time each iteration). Another approach is to directly solve the minimization problem and the maximization problem alternatively [34] under a finite-support assumption. Gao and Kleywegt [15] proposed a routine to compute the finite structure of the worst-case distribution in theory. Nevertheless it still takes a high computational complexity if the Wasserstein ball has a large support size. Several fast WDRO training algorithms for some specific machine learning tasks, e.g., SVM and logistic regression by Li et al. [26, 27], have been developed recently; but it is unclear whether their methods can be generalized to solve other problems.

Therefore, it is urgent to develop efficient algorithmic techniques for reducing the computational complexity of the WDRO problems. **Coreset** is a popular tool for compressing large datasets, which was initially introduced by Agarwal et al. in computational geometry [1]. Intuitively, the coreset is an approximation of the original input data, but has a much smaller size. Thus any existing algorithm can run on the coreset instead and the computational complexity can be largely reduced. The coresets techniques have been widely applied for many optimization problems such as clustering and regression (we refer the reader to the recent surveys on coresets [31; 12]). Therefore a natural idea is to consider applying the coreset technique to deal with large-scale WDRO problems. Below we introduce the formal definition of the coreset for WDRO problems.

Definition 1 (ϵ-coreset) *Let ϵ be any given small number in $(0, 1)$. An ϵ-coreset for the WDRO problem (3) is a sparse nonnegative mass vector $W = [w_1, \ldots, w_n]$, such that the total mass*

$\sum_{i=1}^{n} w_i = 1$ *and the induced distribution* $\tilde{\mathbb{P}}_n = \sum_{i=1}^{n} w_i \delta_{\xi_i}$ *satisfies*

$$R_{\theta,p}^{\tilde{\mathbb{P}}_n}(\beta) \in (1 \pm \epsilon) R_{\theta,p}^{\mathbb{P}_n}(\beta), \forall \beta \in \mathcal{H}, \tag{4}$$

where $R_{\theta,p}^{\tilde{\mathbb{P}}_n}(\beta) := \sup_{\mathbb{Q} \in \mathbb{B}_{\theta,p}(\tilde{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{Q}}[\ell(\beta, \xi)]$ *is the worst-case empirical risk of the coreset.*

It is worth to emphasize that the above coreset for WDRO is fundamentally different from the conventional coresets [12]. The main challenge for constructing the coreset of WDRO is from the "**uncertainty**" issue, that is, we have to consider all the possible distributions in the Wasserstein ball $\mathbb{B}_{\theta,p}(\tilde{\mathbb{P}}_n)$; and more importantly, when the parameter vector β varies, the distribution that achieves the worst-case empirical risk also changes inside $\mathbb{B}_{\theta,p}(\tilde{\mathbb{P}}_n)$.

1.1 Our Contribution

In this paper, we propose a novel framework to construct the ϵ-coresets for general WDRO problems. To the best of our knowledge, this is the first coreset algorithm for Wasserstein distributionally robust optimization problems. Our main contributions are twofold.

-From coresets to dual coresets. As mentioned before, it is challenging to directly construct the coresets for the WDRO problems. Our key observation is inspired by the strong duality property of the WDRO model [11; 3; 15]. We introduce the "dual coreset" for the dual formulation of the WDRO problems. We can neatly circumvent the "uncertainty" issue in Definition 1 through the dual form. Also, we prove that the dual coreset can yield a theoretically quality-guaranteed coreset as Definition 1.

-How to compute the dual coresets. Further, we provide a unified framework to construct the dual coresets efficiently. The sensitive-sampling based coreset framework usually needs to compute the "pseudo-dimension" of the objective function and the "sensitivities" of the data items, which can be very difficult to obtain [12] (the pseudo-dimension measures how complicated the objective function is, and the sensitivity of each data item indicates its importance to the whole input data set). Therefore we consider to apply the spatial partition approach that was initiated by Chen [7]; roughly speaking, we partition the space into a logarithmic number of regions, and take a uniform sample from each region. This partition approach needs to compute the exact value of the Moreau-Yosida regularization [33], which is a key part in the dual formulation of WDRO (the formal definition is shown in Proposition 1). However, this value is often hard to obtain for general Ξ and general $\ell(\cdot, \cdot)$. For instance, suppose Ξ admits a conic representation and the learning model is SVM, then computing the Moreau-Yosida regularization is equivalent to solving a convex conic programming [41, corollary 3.12]. For some machine learning problems, it is usually relatively easier to estimate the bounds of the Moreau-Yosida regularization[41, Theorem 3.30]. Based on this observation, we generalize the spatial partition idea and propose a more practical "grid sampling" framework. By using this framework, we only need to estimate the upper and lower bounds of the Moreau-Yosida regularization instead of the exact value. We also prove that a broad range of objective functions can be handled under this framework.

1.2 Other Related Works

A number of coreset-based techniques have been studied before for solving robust optimization problems. For example, Mirzasoleiman et al. [29] designed an algorithm to generate coreset to approximate the Jacobian of a neural network so as to train against noisy labels. The outlier-resistant coresets were also studied for computing the robust center-based clustering problems [13; 14; 21; 10]. For the general continuous and bounded optimization problems [42], Wang et al. [46] proposed a dynamic framework to compute the coresets resisting against outliers. Several other techniques also have been proposed for dealing with large-scale DRO problems, such as the PCA based dimension reduction methods [8; 9] and the stochastic gradient optimization methods [25; 32].

2 Preliminaries

We assume the input-output space $\Xi = \mathbb{X} \times \mathbb{Y}$ with $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathbb{Y} \subseteq \mathbb{R}$, and let $\mathcal{P}(\Xi)$ denote the set of Borel probability distributions supported on Ξ. For $1 \leq i \leq n$, each data sample is a random vector $\xi_i = (x_i, y_i)$ drawn from some underlying distribution $\mathbb{P} \in \mathcal{P}(\Xi)$. The empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ is induced by the dataset $\{\xi_1, \ldots, \xi_n\}$, where δ_{ξ_i} is the Dirac point mass at ξ_i. We endow Ξ with the distance $\mathtt{d}(\xi_i, \xi_j) = \|x_i - x_j\| + \frac{\gamma}{2}|y_i - y_j|$, where $\|\cdot\|$ stands for an arbitrary norm of $\mathbb{R}^d$ and the positive parameter "γ" quantifies the transportation cost on the label. This distance function is used for defining the Wasserstein distance (1).

In the rest of this paper, we consider the WDRO problems satisfying the following two assumptions. The first assumption is on the smoothness and boundedness of β. Similar assumptions have been widely adopted in the machine learning field [48; 46].

Assumption 1 (Smoothness and Boundedness of β [42])

(i) (Boundedness) *The feasible region $\mathcal{H}$ of the parameter space for the WDRO problem (3) is within a closed Euclidean ball $\mathbb{B}(\beta_{\mathtt{anc}}, l_{\mathtt{p}})$ centered at some "anchor" point $\beta_{\mathtt{anc}} \in \mathbb{R}^d$ with radius $l_{\mathtt{p}} > 0$;*

(ii) (Lipschitz Smoothness[1]) *There exists a constant $L > 0$, such that for any $\xi \in \Xi$ and any $\beta_1, \beta_2 \in \mathbb{B}(\beta_{\mathtt{anc}}, l_{\mathtt{p}})$, we have $|\ell(\beta_1, \xi) - \ell(\beta_2, \xi)| \leq L \|\beta_1 - \beta_2\|_2$.*

The second assumption states that the loss function $\ell(\beta, \xi)$ is continuous and has a bounded growth rate on data ξ. The detailed growth rate functions are discussed in Section 5.

Assumption 2 (Continuity and Bounded Growth Rate of ξ)

(i) (Continuity) *The loss function $\ell(\beta, \cdot)$ is continuous for any $\beta \in \mathcal{H}$;*

(ii) (Bounded Growth Rate) *There exists some positive continuous growth rate function $\mathtt{C}(\beta)$ and $\zeta_0 \in \Xi$ such that*

$$\ell(\beta, \xi) \leq \mathtt{C}(\beta)\left(1 + \mathtt{d}^p(\xi, \zeta_0)\right)$$

for any $\beta \in \mathcal{H}$ and any $\xi \in \Xi$.

Now we state the strong duality for the WDRO, which is the cornerstone of our dual coreset method.

Proposition 1 (Strong duality [11; 3; 15]) *For any upper semi-continuous $\ell(\beta, \cdot)$, any β and any nominal distribution $\mathbb{P}$ with finite p-th moment, the worst-case risk satisfies*

$$R^{\mathbb{P}}_{\theta,p}(\beta) = \inf_{\lambda \geq 0}\{\lambda\theta^p + H^{\mathbb{P}}(\beta, \lambda)\}, \tag{5}$$

where $H^{\mathbb{P}}(\beta, \lambda) := \mathbb{E}^{\mathbb{P}}[h(\beta, \lambda, \xi)]$ and $h(\beta, \lambda, \xi) = \sup_{\zeta \in \Xi}\{\ell(\beta, \zeta) - \lambda \mathtt{d}^p(\zeta, \xi)\}$ is the Moreau-Yosida regularization [33]. We use $\lambda^{\mathbb{P}}_*(\beta)$ to denote the λ attaining the infimum in (5).

Remark 1 *By the definition of h, we have $h(\beta, \lambda, \xi) \geq \ell(\beta, \xi) - \lambda \mathtt{d}^p(\xi, \xi) = \ell(\beta, \xi)$. Therefore $h(\beta, \lambda, \xi)$ is always nonnegative and so is $H^{\mathbb{P}}(\beta, \lambda)$. As a consequence, $\lambda\theta^p + H^{\mathbb{P}}(\beta, \lambda)$ grows to infinity as $\lambda \to \infty$. Together with the fact that $\lambda\theta^p + H^{\mathbb{P}}(\beta, \lambda)$ is continuous on λ [15, lemma 3 (ii)], we can deduce that for each β, there always exists some $\lambda^{\mathbb{P}}_*(\beta) < \infty$ attaining the infimum of (5).*

For the sake of convenience, we abbreviate $R_{\theta,p}(\beta) = R^{\mathbb{P}_n}_{\theta,p}(\beta)$, $H(\beta, \lambda) = H^{\mathbb{P}}(\beta, \lambda)$, $\lambda_*(\beta) = \lambda^{\mathbb{P}}_*(\beta)$, and $h_i(\beta, \lambda) = h(\beta, \lambda, \xi_i)$. We define the asymptotic growth rate function

$$\kappa(\beta) := \limsup_{\mathtt{d}(\xi, \zeta_0) \to \infty} \frac{\ell(\beta, \xi) - \ell(\beta, \zeta_0)}{\mathtt{d}^p(\xi, \zeta_0)}$$

so as to conclude the continuity of $h_i(\cdot, \cdot)$ in the following two claims (the proofs are placed in our supplement). Here ζ_0 is the fixed point in Assumption 1 (ii). Set $\kappa \equiv 0$ if Ξ is bounded.

[1] The methods proposed in this paper can be easily extended to other types of smoothness, e.g., gradient Lipschitz continuity.

Claim 1 (Continuity of h_i on β) *For each $i \in \{1, \dots, n\}$ and any fixed $\lambda \geq 0$, we have*

$$|h_i(\beta, \lambda) - h_i(\beta, \lambda')| \leq L\|\beta - \beta'\|_2,$$

for any $\beta, \beta' \in \mathcal{H}$ with $\kappa(\beta), \kappa(\beta') \leq \lambda$.

Claim 2 (Continuity of h_i on λ) *For each $i \in \{1, \dots, n\}$ and any fixed $\beta \in \mathbb{R}^d$, we have*

$$|h_i(\beta, \lambda) - h_i(\beta, \lambda')| \leq \max\{r_i^p(\beta, \lambda), r_i^p(\beta, \lambda')\}|\lambda - \lambda'|, \ \forall \lambda, \lambda' \geq \kappa(\beta),$$

where $r_i(\beta, \lambda) := \min_{\zeta \in \Xi}\{\mathtt{d}(\zeta, \xi_i) \colon \ell(\beta, \zeta) - \lambda \mathtt{d}^p(\zeta, \xi_i) = h_i(\beta, \lambda)\}$ is the closest distance between ξ_i and all the ζs that attain the supremum of $\ell(\beta, \zeta) - \lambda \mathtt{d}^p(\zeta, \xi_i)$ in Ξ.

Remark 2 *The reason that we let $\lambda \geq \kappa(\beta)$ in the above claims is that each $h_i(\beta, \lambda)$ goes to infinity if $\lambda < \kappa(\beta)$. Without loss of generality[2], we suppose $h_i(\beta, \kappa(\beta)) < \infty$ in this paper.*

3 From Coresets to Dual Coresets

In this section, we provide the concept of "dual coreset" and prove that it is sufficient to guarantee the correctness with respect to the WDRO coreset. First, we show the dual coreset definition via directly combining Proposition 1 and Definition 1. Suppose I is an interval within $[\kappa(\beta), \infty)$ that will be discussed in detail later.

Definition 2 (Dual ϵ-Coreset) *A dual ϵ-coreset for the WDRO problem (3) is a sparse non-negative mass vector $W = [w_1, \dots, w_n]$ such that the total mass $\sum_{i=1}^n w_i = 1$ and*

$$\tilde{H}(\beta, \lambda) := \sum_{i=1}^n w_i h_i(\beta, \lambda) \in (1 \pm \epsilon) H(\beta, \lambda) \tag{6}$$

for any $\beta \in \mathcal{H}$ and $\lambda \in I$.

Remark 3 *Note that we require the approximation guarantee holds not only for any $\beta \in \mathcal{H}$, but also for any $\lambda \in I$ in the above definition. This is also a key difference to the traditional coresets.*

If we let $I = [\kappa(\beta), \infty)$, the dual coreset of Definition 2 requires to approximate the queries from all $\lambda \geq \kappa(\beta)$, which is too **strong** and can be even troublesome for the coreset construction. Below we show that such requirement on the mass vector W is not necessary for obtaining a qualified WDRO coreset.

Given a non-negative mass vector $W = [w_1, \cdots, w_n]$, the corresponding weighted empirical distribution is $\tilde{\mathbb{P}}_n = \sum_{i=1}^n w_i \delta_{\xi_i}$. Recall that we define a parameter $\lambda_*^{\mathbb{P}}(\beta)$ for duality in Proposition 1. Together with Assumption 2, we show the boundedness of the $\lambda_*^{\tilde{\mathbb{P}}_n}(\beta)$ (abbreviated as $\tilde{\lambda}_*(\beta)$ for convenience) for $\tilde{\mathbb{P}}_n$. Let $[n] = \{1, 2, \cdots, n\}$. The following result is a key to relax the requirement for the dual coreset in Definition 2.

Lemma 1 (Boundedness of $\tilde{\lambda}_*$) *Given the empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$, we define the value $\rho = \max_{i \in [n]}\{\mathtt{d}(\xi_i, \zeta_0)\}$ that is the largest distance from the data samples to ζ_0. Here ζ_0 is defined in Assumption 2 (ii). For any $\beta \in \mathcal{H}$ and any mass vector W, the $\tilde{\lambda}_*(\beta)$ of the corresponding weighted empirical distribution $\tilde{\mathbb{P}}_n$ is **no larger than***

$$\mathtt{C}(\beta) \cdot \left(2^{p-1} + \frac{1 + 2^{p-1}\rho^p}{\theta^p}\right), \tag{7}$$

where $\mathtt{C}(\beta)$ is defined in Assumption 2 (ii). We use $\tau(\beta)$ to denote this upper bound $\mathtt{C}(\beta)\left(2^{p-1} + \frac{1+2^{p-1}\rho^p}{\theta^p}\right)$.

[2] It is possible that $h_i(\beta, \kappa(\beta)) = \infty$, e.g., $\ell(\beta, \xi)$ is the loss function of ordinary linear regression. In this case, the argument in this paper still holds with slight modification.

Proof. First, by Assumption 2 we have

$$\ell(\beta,\zeta) - \lambda \mathtt{d}^p(\zeta,\xi) \le \mathtt{C}(\beta)\left(1 + \mathtt{d}^p\left(\zeta,\zeta_0\right)\right) - \lambda \mathtt{d}^p(\zeta,\xi), \forall \zeta \in \Xi. \tag{8}$$

If letting $\lambda = 2^{p-1}\mathtt{C}(\beta)$, the above inequality yields

$$\ell(\beta,\zeta) - 2\mathtt{C}(\beta)\mathtt{d}^p(\zeta,\xi) \le \mathtt{C}(\beta)(1 + \mathtt{d}^p(\zeta,\zeta_0)) - 2^{p-1}\mathtt{C}(\beta)\mathtt{d}^p(\zeta,\xi), \forall \zeta \in \Xi. \tag{9}$$

Further, combining the triangle inequality and Jensen's inequality [22], we know that

$$\mathtt{d}^p(\zeta,\zeta_0) \le 2^{p-1}\mathtt{d}^p(\zeta,\xi) + 2^{p-1}\mathtt{d}^p(\xi,\zeta_0). \tag{10}$$

We then substitute (10) into (9), and based on the definition of $h(\beta,\lambda,\xi)$ we have

$$h(\beta, 2^{p-1}\mathtt{C}(\beta),\xi) \le \mathtt{C}(\beta)(1 + 2^{p-1}\mathtt{d}^p(\xi,\zeta_0)). \tag{11}$$

Note that $\tilde{H}(\beta,\lambda) = \mathbb{E}^{\tilde{\mathbb{P}}_n}[h(\beta,\lambda,\xi)]$ and $\mathtt{d}(\xi_i,\zeta_0) \le \rho$ for all $i \in [n]$. So we have

$$\tilde{H}(\beta, 2^{p-1}\mathtt{C}(\beta)) \le \mathtt{C}(\beta)(1 + 2^{p-1}\rho^p) \tag{12}$$

for any mass vector W. Therefore we obtain

$$\inf_{\lambda \ge 0}\{\lambda\theta^p + \tilde{H}(\beta,\lambda)\} \le 2^{p-1}\mathtt{C}(\beta)\theta^p + \mathtt{C}(\beta)(1 + 2^{p-1}\rho^p), \tag{13}$$

which implies that

$$\tilde{\lambda}^*(\beta) \le \mathtt{C}(\beta)\left(2^{p-1} + \frac{1 + 2^{p-1}\rho^p}{\theta^p}\right) \tag{14}$$

for any $\beta \in \mathcal{H}$ and any mass vector W. □

Remark 4 ***(i)*** *In practice, we usually normalize the dataset before training a machine learning model, which implies that ρ is not large.* ***(ii)*** *It is worth noting that the above lemma can help us to compute an upper bound for $\lambda_*(\beta)$. For example, if letting $W = [\frac{1}{n},\ldots,\frac{1}{n}]$, (7) directly yields an upper bound.*

The following theorem shows that the query region $I = [\kappa(\beta),\tau(\beta)]$ is sufficient for obtaining a coreset of the WDRO problem (3). Due to the space limit, we leave the proof of Theorem 1 to our supplement.

Theorem 1 (Sufficiency of the bounded query region) *If we let query region $I = [\kappa(\beta),\tau(\beta)]$ in Definition 2, the dual ϵ-coreset defined in such way also satisfies the coreset of Definition 1.*

Therefore in the rest of this paper, we let $I = [\kappa(\beta),\tau(\beta)]$ in Definition 2. Theorem 1 also implies the following corollary. So we can only focus on solving the dual WDRO problem (3) on the obtained dual ϵ-coreset.

Corollary 1 *Let $\alpha \ge 1$ and suppose the parameter vector β_0 yields an α-approximation obtained on the dual ϵ-coreset. Then β_0 is also an $(\alpha \cdot \frac{1+\epsilon}{1-\epsilon})$-approximation of the original WDRO (3).*

To end this section, similar to $\mathbb{B}(\beta_{\mathtt{anc}}, l_{\mathtt{p}})$ in Assumption 2 (i), we define an interval $[\lambda_{\mathtt{anc}} - l_{\mathtt{d}}, \lambda_{\mathtt{anc}} + l_{\mathtt{d}}]$ centered at some "anchor" point $\lambda_{\mathtt{anc}} > 0$ with radius $l_{\mathtt{d}} > 0$. To ensure $[\kappa(\beta),\tau(\beta)]$ is always within the interval $[\lambda_{\mathtt{anc}} - l_{\mathtt{d}}, \lambda_{\mathtt{anc}} + l_{\mathtt{d}}]$ for all $\beta \in \mathcal{H}$, we let $\lambda_{\mathtt{anc}} := \max_{\beta\in\mathcal{H}}\{\kappa(\beta_{\mathtt{anc}}), \frac{\tau(\beta)}{2}\}$ and $l_{\mathtt{d}} := \lambda_{\mathtt{anc}}$.

4 The Construction of Dual Coresets

Following the results of Section 3, we show how to compute a qualified dual coreset in this section. Suppose we can evaluate the lower and upper bounds for each $h_i(\cdot,\cdot)$ with respect to a given couple of $\lambda_{\mathtt{anc}}$ and $\beta_{\mathtt{anc}}$, namely, we have

$$a_i(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}}) \le h_i(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}}) \le b_i(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}})$$

for $1 \le i \le n$. We defer the details for obtaining such upper and lower bounds for each application to Section 5.

Algorithm 1 Dual ϵ-Coreset Construction

Input: The empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$, the Lipschitz constant L, the "anchors" $\beta_{\mathtt{anc}}$ and $\lambda_{\mathtt{anc}}$, and corresponding radii $l_{\mathtt{p}}$ and $l_{\mathtt{d}}$; the parameter $\epsilon \in (0,1)$; lower bound oracle $a_i(\cdot,\cdot)$ and upper bound oracle $b_i(\cdot,\cdot)$ for $i \in [n]$.

1. Compute $A = \frac{1}{n}\sum_{i=1}^n a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$ and $B = \frac{1}{n}\sum_{i=1}^n b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$.
2. Let $N = \lceil \log n \rceil$; initialize $W = [0, 0, \cdots, 0] \in \mathbb{R}^n$.
3. The dataset $\{\xi_1, \ldots, \xi_n\}$ is partitioned into $(N+1)^2$ cells $\{C_{ij} | 0 \le i, j \le N\}$ as (19).
4. For each $C_{ij} \neq \emptyset$, $0 \le i, j \le N$:
   (a) take a sample Q_{ij} from C_{ij} uniformly at random, where the size $|Q_{ij}|$ depends on the parameters ϵ, $l_{\mathtt{p}}$, $l_{\mathtt{d}}$ and L (the exact value will be discussed in our following analysis in Section 4.2);
   (b) for each sample $\xi_k \in Q_{ij}$, assign the mass of quantity $w_k = \frac{|C_{ij}|}{n|Q_{ij}|}$;

Output: the mass vector $W = [w_1, w_2, \cdots, w_n]$ as the dual ϵ-coreset.

4.1 The Construction Algorithm

We show the dual ϵ-coreset construction procedure in Algorithm 1, where the high-level idea is based on the following grid sampling.

Grid sampling. Let $N = \lceil \log n \rceil$. Given the anchor $(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$, we can conduct the partitions over the dataset based on the lower bounds $a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$ and upper bounds $b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$ separately. Let $A = \frac{1}{n}\sum_{i=1}^n a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$ and $B = \frac{1}{n}\sum_{i=1}^n b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$. Then we have the following partitions.

$$A_0 = \big\{\xi_i \mid a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) \le A\big\}, \tag{15}$$

$$A_j = \big\{\xi_i \mid 2^{j-1}A < a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) \le 2^j A\big\}, 1 \le j \le N. \tag{16}$$

$$B_0 = \big\{\xi_i \mid b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) \le B\big\}, \tag{17}$$

$$B_j = \big\{\xi_i \mid 2^{j-1}B < b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) \le 2^j B\big\}, 1 \le j \le N. \tag{18}$$

We denote the lower bound and upper bound partitions as $\mathcal{A} = \{A_0, \cdots, A_N\}$ and $\mathcal{B} = \{B_0, \cdots, B_N\}$ respectively. Then, we compute the intersections over $\mathcal{A}$ and $\mathcal{B}$ to have the "grid":

$$\mathcal{C} = \{C_{ij} | C_{ij} = A_i \cap B_j, 0 \le i, j \le N\}. \tag{19}$$

It is easy to see that $\mathcal{C}$ is a collection of disjoint "cells" and $\bigcup_{i,j} C_{ij} = P$. For each $\xi_k \in C_{ij}$, we have

$$\mu_i \cdot 2^{i-1}A \le h_k(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) \le 2^j B \tag{20}$$

where $\mu_i = 0$ if $i = 0$ and $\mu_i = 1$ otherwise. Through the grid partition $\mathcal{C}$, we can take a set of samples Q_{ij} from C_{ij} uniformly at random, and assign the weight $\frac{|C_{ij}|}{n|Q_{ij}|}$ to each sample.

Remark 5 ***(i)*** *The grid sampling is a variance reduction technique in the Monte-Carlo methods [17], since the grid partition is also a stratification for $h_k(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$ as is shown in (20). If we only use upper bounds $b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$ or lower bounds $a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$, the resulted partition is not a valid stratification for $h_k(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$.* ***(ii)*** *If we can obtain the exact value of $h_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$, i.e., $a_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) = b_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}}) = h_i(\beta_{\mathtt{anc}}, \lambda_{\mathtt{anc}})$, then the grid partition is exactly the spatial partition that was studied before [7; 46].*

4.2 Theoretical Analysis

In this section we analyze the complexity of Algorithm 1 in theory. Recall that we define $r_i(\beta, \lambda) = \min_{\zeta \in \Xi}\{\mathtt{d}(\zeta, \xi_i) \colon \ell(\beta, \zeta) - \lambda \mathtt{d}^p(\zeta, \xi_i) = h_i(\beta, \lambda)\}$ in Claim 2. We define two constants $R \coloneqq \max_{\substack{i \in [n] \\ \beta \in \mathcal{H}}}\{r_i^p(\kappa(\beta), \beta)\}$ and $H \coloneqq \min_{\substack{\beta \in \mathcal{H} \\ \lambda \in [\lambda_{\mathtt{anc}} - l_{\mathtt{d}}, \lambda_{\mathtt{anc}} + l_{\mathtt{d}}]}} H(\beta, \lambda)$.

Theorem 2 *With probability $1-\frac{1}{n}$, the Algorithm 1 returns a qualified dual ϵ-coreset with size $\tilde{O}\left(\left(B\cdot\frac{B+Ll_{\mathtt{p}}+Rl_{\mathtt{d}}}{AH}\right)^2\cdot\frac{d}{\epsilon^2}\right)$[3]. The construction time is $O(n\cdot\mathtt{time}_{ab})$ where $\mathtt{time}_{ab}$ is the time complexity for computing the lower bound $a_i(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}})$ and the upper bound $b_i(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}})$ for each $h_i(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}})$.*

Remark 6 *Note that the value $H\geq\min_{\beta\in\mathcal{H}}\mathbb{E}^{\mathbb{P}_n}\ell(\beta,\xi)$, which is not too small since the loss function $\ell(\cdot,\cdot)$ usually contains positive penalty terms. The value of R will be discussed in Section 5.*

We show the sketched proof of Theorem 2 below, and leave the full proof to our supplement. Based on the continuity of $h_i(\cdot,\cdot)$ and the Hoeffding's inequality [20], for a fixed couple (β,λ), we provide an upper bound on the sample complexity first. The bound ensures that the estimation for each cell C_{ij} has a bounded deviation with high probability

Lemma 2 *Let σ be a given positive number. We fix a couple $(\beta,\lambda)\in\mathbb{B}(\beta_{\mathtt{anc}},l_{\mathtt{p}})\times[\kappa(\beta),\tau(\beta)]$ and take a uniform sample Q_{ij} from C_{ij} with the sample size*

$$|Q_{ij}|=O\left((2^jB-\mu_i\cdot 2^{i-1}A+2Ll_{\mathtt{p}}+2Rl_{\mathtt{d}})^2\sigma^{-2}\log\frac{1}{\eta}\right). \tag{21}$$

Then, we have the probability

$$\mathtt{Prob}\left[\left|\frac{1}{|Q_{ij}|}\sum_{\xi_k\in Q_{ij}}h_k(\beta,\lambda)-\frac{1}{|C_{ij}|}\sum_{\xi_k\in C_{ij}}h_k(\beta,\lambda)\right|\geq\sigma\right]\leq\eta. \tag{22}$$

We aggregate the deviations from all the cells to obtain a total error on the coreset. To guarantee the approximation in (6), we need to design a sufficiently small value of the deviation σ for each cell C_{ij} based on the grid partition structure.

Lemma 3 *In Lemma 2, we set the deviation $\sigma=\epsilon_1(2^{j-1}+2^{i-1})A$ for $0\leq i,j\leq N$. Then we have*

$$\mathtt{Prob}\left[|\tilde{H}(\beta,\lambda)-H(\beta,\lambda)|\leq 3\epsilon_1H(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}})\right]\geq 1-(N+1)^2\eta. \tag{23}$$

To generalize the result of Lemma 3 to the whole feasible region $\mathbb{B}(\beta_{\mathtt{anc}},l_{\mathtt{p}})\times[0,2l_{\mathtt{d}}]$, we apply the discretization idea. Imagine to generate the axis-parallel grid with side length $\frac{\epsilon_3 l_{\mathtt{p}}}{\sqrt{d}}\times\epsilon_2 l_{\mathtt{d}}$ inside $\mathbb{B}(\beta_{\mathtt{anc}},l_{\mathtt{p}})\times[0,2l_{\mathtt{d}}]$; the parameters ϵ_2 and ϵ_3 are two small numbers that will be determined in our following analysis. For each grid cell we arbitrarily take a (β,λ) as its representative point. Let G be the set of the selected representative points; it is easy to see the cardinality $|G|=\frac{1}{\epsilon_2}\cdot O(\frac{1}{\epsilon_3^d})$. Through taking the union bound over all $(\beta,\lambda)\in G$ for (23), we obtain the following Lemma 4.

Lemma 4 *With probability at least $1-(N+1)^2|G|\eta$, we have*

$$|\tilde{H}(\beta,\lambda)-H(\beta,\lambda)|\leq 3\epsilon_1H(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}})\ \textit{for all}\ (\beta,\lambda)\in G. \tag{24}$$

By using the above lemmas, we are ready to prove Theorem 2.

Proof.(**of Theorem 2**) For any $(\beta,\lambda)\in\mathbb{B}(\lambda_{\mathtt{anc}},l_{\mathtt{d}})\times[\kappa(\beta),\tau(\beta)]$, we let $(\beta',\lambda')\in G$ be the representative point of the cell containing (β,λ). Then we have $\|\beta-\beta'\|_2\leq\epsilon_3 l_{\mathtt{p}}$ and $|\lambda'-\lambda|\leq\epsilon_2 l_{\mathtt{d}}$. Without loss of generality, we assume $\lambda'\geq\lambda$. By using the triangle inequality, we have

$$\begin{aligned}&|h_k(\beta,\lambda)-h_k(\beta',\lambda')|\\ \leq&|h_k(\beta',\lambda')-h_k(\beta,\lambda')|+|h_k(\beta,\lambda)-h_k(\beta,\lambda')|\\ \leq&L\epsilon_3 l_{\mathtt{p}}+R\epsilon_2 l_{\mathtt{d}}.\qquad\text{(By Claim 2, Claim 1 and }\lambda'\geq\lambda\geq\kappa(\beta))\end{aligned} \tag{25}$$

The above inequality implies

$$|H(\beta,\lambda)-H(\beta',\lambda')|\leq R\epsilon_2 l_{\mathtt{d}}+L\epsilon_3 l_{\mathtt{p}} \tag{26}$$

[3] $\tilde{O}(g):=O(g\cdot\mathtt{polylog}(\frac{nLl_{\mathtt{p}}\bar{R}l_{\mathtt{d}}}{\epsilon h}))$

and

$$|\tilde{H}(\beta,\lambda)-\tilde{H}(\beta',\lambda')|\leq R\epsilon_2 l_{\mathtt{d}}+L\epsilon_3 l_{\mathtt{p}}. \tag{27}$$

Overall we have $|\tilde{H}(\beta,\lambda)-H(\beta,\lambda)|$

$$\begin{aligned}&\leq\left|\tilde{H}(\beta,\lambda)-\tilde{H}\left(\beta',\lambda'\right)\right|+\left|\tilde{H}\left(\beta',\lambda'\right)-H\left(\beta',\lambda'\right)\right|+\left|H\left(\beta',\lambda'\right)-H\left(\beta,\lambda\right)\right|\\&\leq 3\epsilon_1 H\left(\beta_{\mathtt{anc}},\lambda_{\mathtt{anc}}\right)+2\times\left(R\epsilon_2 l_{\mathtt{d}}+L\epsilon_3 l_{\mathtt{p}}\right)\quad\text{(By Lemma 4, (26) and (27))}\end{aligned} \tag{28}$$

By setting $\epsilon_1=\frac{H\epsilon}{9B}$, $\epsilon_2=\frac{H\epsilon}{6l_{\mathtt{d}}R}$, $\epsilon_3=\frac{H\epsilon}{6Ll_{\mathtt{p}}}$ and $\eta=\frac{1}{n(N+1)^2|G|}$, we obtain the total sample complexity as stated in Theorem 2. □

5 Applications

In this section, we show several WDRO problems that their complexities can be reduced by using our dual coreset method.

5.1 Binary Classification

For the binary classification, $\mathbb{Y}=\{-1,1\}$ and the loss function $\ell(\beta,\xi)=L(y\cdot\beta^\top x)$ where $L(\cdot)$ is a non-negative and non-increasing function. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$ on $\mathbb{R}^d$. We consider the **Support Vector Machine (SVM)** and **logistic regression** problems. The SVM takes the hinge loss $L(z)=\max\{0,1-z\}$ and the logistic regression takes the logloss $L(z)=\log(1+\exp(-z))$. If $\mathbb{X}=\mathbb{R}^d$ and $p=1$, for both of these two problems by [41] we have :

- $R\leq\gamma$, $\kappa(\beta)=\mathtt{C}(\beta)=\|\beta\|_*$;
- $a_i(\beta,\lambda)=b_i(\beta,\lambda)=h_i(\beta,\lambda)=\max\{L(y_i\cdot\beta^\top x_i),L(-y_i\cdot\beta^\top x_i)-\lambda\gamma\}$ for any $\lambda\geq\kappa(\beta)$.

If $\mathbb{X}=[0,l]^d$ is a d-dimensional hypercube with side length $l>0$, $h_i(\beta,\lambda)$ is the optimal objective value of a convex constrained programming. Therefore we **cannot** obtain the exact value of $h_i(\beta,\lambda)$ easily. To remedy this issue, we can invoke the upper and lower bounds of $h_i(\beta,\lambda)$ and conduct the grid sampling to efficiently construct the coreset. Please see the details in our supplement.

5.2 Regression

For the regression problem, $\mathbb{Y}=\mathbb{R}$ and the loss function $\ell(\beta,\xi)=L(\beta^\top x-y)$, where $L(\cdot)$ is a non-negative function. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$ on $\mathbb{R}^{d+1}$. We consider the **robust regression** problem that takes the Huber loss $L(z)=\frac{1}{2}\delta^2$ if $|z|\leq\delta$ and $L(z)=\delta\left(|z|-\frac{1}{2}\delta\right)$ otherwise for some $\delta\geq 0$. If $p=1$ and $\mathbb{X}=\mathbb{R}^d$, by [41] we have

- $R=0,\kappa(\beta)=\mathtt{C}(\beta)=\delta\|(\beta,-1)\|_*$;
- $a_i(\beta,\lambda)=b_i(\beta,\lambda)=h_i(\beta,\lambda)=L(\beta^\top x_i-y_i)$ for any $\lambda\geq\kappa(\beta)$.

6 Conclusion

In this paper, we consider reducing the high computational complexity of WDRO via coreset method. We relate the coreset to its dual coreset by using the strong duality property of WDRO, and propose a novel grid sampling approach for the construction. To the best of our knowledge, our work is the first systematically study on the coreset of WDRO problems in theory. We also implement our proposed coreset algorithm and conduct the experiments to evaluate its performance for several WDRO problems (including the applications mentioned in Section 5). **Due to the space limit, we leave the full experimental results to our supplement**. Following our work, there also exist several important problems deserving to study in future. For example, it is interesting to consider the coresets construction for other robust optimization models, especially for adversarial training [18].

References

[1] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004.

[2] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit.*, 84:317–331, 2018.

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

[4] J. Blanchet, P. W. Glynn, J. Yan, and Z. Zhou. Multivariate distributionally robust convex regression under absolute error loss. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[6] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] K. Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[8] J. Cheng, R. L. Chen, H. N. Najm, A. Pinar, C. Safta, and J. Watson. Distributionally robust optimization with principal component analysis. *SIAM J. Optim.*, 28(2):1817–1841, 2018.

[9] M. Cheramin, J. Cheng, R. Jiang, and K. Pan. Computationally efficient approximations for distributionally robust optimization. *Optimization online*, 2020.

[10] H. Ding, H. Yu, and Z. Wang. Greedy strategy works for k-center clustering with outliers and coreset construction. In M. A. Bender, O. Svensson, and G. Herman, editors, *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Germany*, volume 144 of *LIPIcs*, pages 40:1–40:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[11] P. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171, 05 2015. doi: 10.1007/s10107-017-1172-1.

[12] D. Feldman. Core-sets: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(1), 2020.

[13] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In L. Fortnow and S. P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578. ACM, 2011.

[14] D. Feldman and L. J. Schulman. Data reduction for weighted and outlier-resistant clustering. In Y. Rabani, editor, *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 1343–1354. SIAM, 2012.

[15] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[16] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *CoRR*, abs/1712.06050, 2017.

[17] E. Gobet. *Monte-Carlo methods and stochastic processes: from linear to non-linear*. Chapman and Hall/CRC, 2016.

[18] I. Goodfellow, P. McDaniel, and N. Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[19] G. A. Hanasusanto and D. Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

[20] W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[21] L. Huang, S. H. Jiang, J. Li, and X. Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In M. Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 814–825. IEEE Computer Society, 2018.

[22] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[23] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[24] C. Lee and S. Mehrotra. A distributionally-robust approach for finding support vector machines. *Available from Optimization Online*, 2015.

[25] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[26] J. Li, S. Huang, and A. M.-C. So. A first-order algorithmic framework for distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] J. Li, C. Chen, and A. M.-C. So. Fast epigraphical projection-based incremental algorithms for wasserstein distributionally robust support vector machine. *Advances in Neural Information Processing Systems*, 33:4029–4039, 2020.

[28] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5):840–853, 2007.

[29] B. Mirzasoleiman, K. Cao, and J. Leskovec. Coresets for robust training of deep neural networks against noisy labels. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

[31] A. Munteanu and C. Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intell.*, 32(1):37–53, 2018.

[32] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2208–2216, 2016.

[33] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

[34] G. C. Pflug and A. Pichler. The problem of ambiguity in stochastic optimization. In *Multistage Stochastic Optimization*, pages 229–255. Springer, 2014.

[35] K. Postek, D. den Hertog, and B. Melenberg. Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650, 2016.

[36] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[37] M. B. Ring. Child: A first step towards continual learning. In *Learning to learn*, pages 261–292. Springer, 1998.

[38] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[39] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

[40] S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28, 2015.

[41] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[42] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

[43] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. Mit Press, 2012.

[44] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[45] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[46] Z. Wang, Y. Guo, and H. Ding. Robust and fully-dynamic coreset for continuous-and-bounded learning (with outliers) problems. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14319–14331, 2021.

[47] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

[48] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In T. Fawcett and N. Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 928–936. AAAI Press, 2003.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Some proofs are left to our supplement due to the space limit.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The full settings of experiments are placed in our supplement.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The full settings of experiments are placed in our supplement.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The full experiments results are left to our supplement.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The full settings of experiments are placed in our supplement.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] The full settings of experiments are placed in our supplement.

(b) Did you mention the license of the assets? [Yes] The full settings of experiments are placed in our supplement.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]