Large-Scale Non-convex Stochastic Constrained Distributionally Robust Optimization

Qi Zhang [1], Yi Zhou [2], Ashley Prater-Bennette [3], Lixin Shen [4], Shaofeng Zou [1]
[1]University at Buffalo
[2]The University of Utah
[3]Air Force Research Laboratory
[4]Syracuse University
qzhang48@buffalo.edu, yi.zhou@utah.edu, ashley.prater-bennette@us.af.mil, lshen03@syr.edu, szou3@buffalo.edu

Abstract

Distributionally robust optimization (DRO) is a powerful framework for training robust models against data distribution shifts. This paper focuses on constrained DRO, which has an explicit characterization of the robustness level. Existing studies on constrained DRO mostly focus on convex loss function, and exclude the practical and challenging case with non-convex loss function, e.g., neural network. This paper develops a stochastic algorithm and its performance analysis for non-convex constrained DRO. The computational complexity of our stochastic algorithm at each iteration is independent of the overall dataset size, and thus is suitable for large-scale applications. We focus on the general Cressie-Read family divergence defined uncertainty set which includes χ^2-divergences as a special case. We prove that our algorithm finds an ϵ-stationary point with an improved computational complexity than existing methods. Our method also applies to the smoothed conditional value at risk (CVaR) DRO.

1 Introduction

Machine learning algorithms typically employ the approach of Empirical Risk Minimization (ERM), which minimizes the expected loss under the empirical distribution P_0 of the training dataset and assumes that test samples are generated from the same distribution. However, in practice, there usually exists a mismatch between the training and testing distributions due to various reasons, for example, in domain adaptation tasks domains differ from training to testing (Blitzer, McDonald, and Pereira 2006; Daume III and Marcu 2006); test samples were selected from minority groups which are underrepresented in the training dataset (Grother et al. 2011; Hovy and Søgaard 2015) and there might exist potential adversarial attacks (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017). Such a mismatch may lead to a significant performance degradation.

This challenge spurred noteworthy efforts on developing a framework of Distributionally Robust Optimization (DRO) e.g., (Ben-Tal et al. 2013; Shapiro 2017; Rahimian and Mehrotra 2019). Rather than minimizing the expected loss under one fixed distribution, in DRO, one seeks to optimize the expected loss under the worst-case distribution in an uncertainty set of distributions.

Specifically, DRO aims to solve the following problem:

$$\inf_{x} \sup_{Q\sim\mathcal{U}(P_0)} \mathbb{E}_{S\sim Q}\, \ell(x;S), \tag{1}$$

where $\mathcal{U}(P_0)$ is an uncertainty set of distributions centered at P_0, P_0 is the empirical distribution of the training dataset, ℓ is the loss function, and x is the optimization variable. For example, the uncertainty set can be defined as

$$\mathcal{U}(P_0) := \{Q : D(Q\|P_0) \le \rho\}, \tag{2}$$

where D is some distance-like metric, e.g., Kullback-Leibler (KL) divergence and χ^2 divergence, and ρ is the uncertainty level. In practice, for ease of implementation and analysis, a relaxed formulation of eq. (1), which is referred to as the penalized DRO, is usually solved (Levy et al. 2020; Jin et al. 2021; Qi et al. 2021; Sinha et al. 2017):

$$\inf_{x} \sup_{Q} \mathbb{E}_{S\sim Q}\, \ell(x;S) - \lambda D(Q\|P_0), \tag{3}$$

where $\lambda > 0$ is a fixed hyperparameter that needs to be chosen manually. In contrast to constrained DRO in eq. (1), a regularization term is added to the objective function to keep the distribution Q and the distribution P_0 close, and the hyperparameter λ is manually chosen beforehand to control the tradeoff with minimizing the loss. Compared with the penalized DRO setting, the constrained DRO problem in eq. (1) requires that the distribution Q to be strictly in the uncertainty set, and searches for the optimal solution under the worst-case distribution in the uncertainty set. Therefore, the obtained solution from the constrained DRO is minimax optimal for distributions in the uncertainty set, whereas it is hard to get such a guarantee for the penalized DRO relaxation. In this paper, we focus on the challenging constrained DRO problem in eq. (1).

Existing studies on constrained DRO are limited to convex loss functions or require some additional assumptions (Soma and Yoshida 2020; Hashimoto et al. 2018; Levy et al. 2020; Duchi and Namkoong 2018; Duchi, Glynn, and Namkoong 2021; Qi et al. 2022; Wang, Gao, and Xie 2021). Little understanding on the practical non-convex loss functions, e.g., neural network, is known. In this paper, we focus on the constrained DRO problem with non-convex loss.

DRO problems under different uncertainty sets are fundamentally different. As will be discussed later in related

works, there is a rich literature on DRO with various uncertainty sets. In this paper, we focus on the general Cressie-Read family divergence defined uncertainty set (Duchi and Namkoong 2018; Jin et al. 2021), which includes, e.g., χ^2 divergence, as a special case (see Section 2 for more details). We also investigate the smoothed conditional value at risk (CVaR) DRO problem.

More importantly, we focus on the practical yet challenging large-scale scenario, where P_0 is the empirical distribution of N samples and N is very large. In classic stochastic optimization problems, e.g., ERM, it is easy to get an unbiased estimate of the gradient using only a few samples, and therefore the computational complexity at each iteration is independent of the training dataset size. However, in the DRO problems, due to taking the worst-case distributions in the objective, the problem becomes challenging. Many existing DRO algorithms incur a complexity that increases linearly (or even worse) in the training dataset size (Duchi and Namkoong 2018; Namkoong and Duchi 2016; Ghosh, Squillante, and Wollega 2018), which is not feasible for large-scale applications. In this paper, we will design a stochastic algorithm with computational complexity at each iteration being independent of the training dataset size (Qi et al. 2022; Wang, Gao, and Xie 2021; Levy et al. 2020).

1.1 Challenges and Contributions

The key challenges and main contributions in this paper are summarized as follows.

- For large-scale applications, the number of training samples is large, and therefore directly computing the full gradient is not practical. Nevertheless, as discussed above, it is challenging to obtain an unbiased estimate of the gradient for DRO problems using only a few samples. For φ-divergence DRO problem, the distributions in the uncertainty set are continuous w.r.t. the training distribution. Thus the distributions in the uncertainty set can be parameterized by an N-dimensional vector (Namkoong and Duchi 2016). Then the DRO problem becomes a min-max problem and primal-dual algorithms (Rafique et al. 2022; Lin, Jin, and Jordan 2020; Xu et al. 2023) can be used directly. Subsampling methods in DRO were also studied in (Namkoong and Duchi 2016; Ghosh, Squillante, and Wollega 2018). However, all the above studies require a computational complexity linear or even worse in the training dataset size at each iteration and thus is prohibitive in large-scale applications. In (Levy et al. 2020), an efficient subsampling method was proposed, where the batch size is independent of the training dataset size. However, they only showed the sampling bias for χ^2 and CVaR DRO problems. In this paper, we generalize the analysis of the bias in (Levy et al. 2020) to the general Cressie-Read family. We further develop a Frank-Wolfe update on the dual variables in order to bound the gap between the objective and its optimal value given the optimization variable x and the biased estimate.
- The second challenge is due to the non-convex loss function. Existing studies for the Cressie-Read divergence family (Duchi and Namkoong 2018; Levy et al. 2020) are limited to the case with convex loss function, and their approach does not generalize to the non-convex case. The key difficulty lies in that the subgradient of the objective function can not be obtained via subdifferential for non-convex loss functions. Instead of explicitly calculating the worst-case distribution as in (Duchi and Namkoong 2018; Levy et al. 2020), we propose to design an algorithm for the dual problem which optimizes the objective under a known distribution. Thus the gradient of the objective can be efficiently obtained.
- The third challenge is that the dual form of constrained DRO is neither smooth nor Lipschitz, making the convergence analysis difficult. Existing studies, e.g., (Wang, Gao, and Xie 2021), assume that the optimal dual variable is bounded away from zero, i.e., $\lambda^* > \lambda_0$ for some $\lambda_0 > 0$, so that it is sufficient to consider $\lambda \geq \lambda_0$. However, this assumption may not necessarily be true as shown in (Wang, Gao, and Xie 2021; Hu and Hong 2013). In this paper, we generalize the idea in (Qi et al. 2022; Levy et al. 2020) to the general Cressie-Read divergence family. We design an approximation of the original problem, and show that it is smooth and Lipschitz. The approximation error can be made arbitrarily small so that the solution to the approximation is still a good solution to the original. We prove the strong duality of the approximated problem. We add a regularizer to the objective and at the same time we keep the hard constraint. In this way, we can guarantee that its dual variable λ has a positive lower bound. Moreover, our strong duality holds for any φ-divergence DRO problem.
- We design a novel algorithm to solve the approximated problem and prove it converges to a stationary point of the constrained DRO problem. The general Proximal Gradient Descent algorithm (Ghadimi, Lan, and Zhang 2016) can be used to solve this approximated problem directly. However, it assumes the objective is non-convex in all parameters and does not provide a tight bound on the bias due to subsampling. We take advantage of the fact that the objective function is convex in λ and thus the bias due to subsampling can be bounded in a tighter way. Our proposed algorithm converges to a stationary point faster than existing methods.

1.2 Related Work

Various Uncertainty Sets. φ-divergence DRO problems (Ali and Silvey 1966; Csiszár 1967) were widely studied, for example, CVaR in (Rockafellar, Uryasev et al. 2000; Soma and Yoshida 2020; Curi et al. 2020; Tamar, Glassner, and Mannor 2015), χ^2-divergence in (Hashimoto et al. 2018; Ghosh, Squillante, and Wollega 2018; Levy et al. 2020), KL-divergence in (Qi et al. 2021, 2022; Hu and Hong 2013) and Sinkhorn distance (Wang, Gao, and Xie 2021), a variant of Wasserstein distance based on entropic regularization. However, the above studies are for some specific divergence function and can not be extended directly to the general Cressie-Read divergence family.

Penalized DRO. The general φ-divergence DRO problem was studied in (Jin et al. 2021) where the proposed algorithm

works for any divergence function with a smooth conjugate. The authors also designed a smoothed version of the CVaR problem and showed their method converges to a stationary point. However, their method is for the penalized formulation and does not generalize to the constrained DRO. In this paper, we focus on the challenging constrained DRO, the solution of which is minimax optimal over the uncertainty set. Our proposed algorithm can also be applied to solve the smoothed CVaR problem in the constrained setting.

Constrained DRO With Convex Loss. The general φ-divergence constrained DRO problem was studied in (Namkoong and Duchi 2016). Instead of optimizing from the dual form, the authors treat the worst-case distribution as a N-dimentional vector and implement a stochastic primal-dual method to solve the min-max problem. However, the computational complexity at each iteration is linear in the number of the training samples and can not be used in large-scale applications. The same problem was further studied in (Duchi, Glynn, and Namkoong 2021). The authors pointed out that minimizing constrained DRO with φ-divergence is equivalent to adding variance regularization for the Empirical Risk Minimization (ERM) objective. The general Cressie-Read divergence family DRO problem was studied in (Duchi and Namkoong 2018), where the basic idea is to calculate the worst-case distribution for the constrained DRO first and then use the subdifferential to get the subgradient. Furthermore, the χ^2 and CVaR DRO problems were studied in (Levy et al. 2020). Compared with the method in (Duchi and Namkoong 2018), they calculate the worst-case distribution for the penalized DRO and then optimize both the Lagrange multiplier and the loss function. This approach converges to the optimal solution with a reduced complexity. Their method can be extended to the general Cressie-Read divergence family. However, all the above papers are limited to the case with convex loss function. To the best of our knowledge, our work is the first paper on large-scale non-convex constrained DRO with the general Cressie-Read divergence family. We note that the KL DRO was studied in (Qi et al. 2022), which however needs an exponential computational complexity. We achieve a polynomial computational complexity for the Cressie-Read divergence family.

2 Preliminaries and Problem Model

2.1 Notations

Let s be a sample in $\mathbb{S}$ and P_0 be the distribution on the points $\{s_i\}_{i=1}^N$, where N is the size of the support. Denote by $\Delta^n := \{\mathbf{p} \in \mathbb{R}^n | \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ the n-dimensional probability simplex. Denote by $x \in \mathbb{R}^d$ the optimization variable. We denote by $\mathbb{1}_{\mathbb{X}}(x)$ the indicator function, where $\mathbb{1}_{\mathbb{X}}(x) = 0$ if $x \in \mathbb{X}$, otherwise $\mathbb{1}_{\mathbb{X}}(x) = \infty$. Let $\ell : \mathbb{R}^d \times \mathbb{S} \to \mathbb{R}$ be a non-convex loss function. Let $\|\cdot\|$ be the Euclidean norm and $(t)_+ := \max\{t, 0\}$ be the positive part of $t \in \mathbb{R}$. Denote ∇_x by the gradient to x. For a function $f : \mathbb{R}^d \to \mathbb{R}$, a point $x \in \mathbb{R}^d$ is said to be an ϵ-optimal solution if $|f(x) - f(x^*)| \leq \epsilon$, where $f(x^*)$ is the optimal value of f. If the function f is differentiable, a point $x \in \mathbb{R}^d$ is said to be first-order ϵ-stationary if $\|\nabla f(x)\| \leq \epsilon$.

2.2 Assumptions

In this paper, we take the following standard assumptions that are commonly used in the DRO literature (Duchi and Namkoong 2018; Levy et al. 2020; Qi et al. 2021, 2022; Wang, Gao, and Xie 2021; Soma and Yoshida 2020):

- The non-convex loss function is bounded: $0 \leq \ell(x; s) \leq B$ for some $B > 0, \forall x \in \mathbb{R}^d, s \in \mathbb{S}$.
- The non-convex loss function is G-Lipschitz such that $|\ell(x_1; s) - \ell(x_2; s)| \leq G\|x_1 - x_2\|$ and L-smooth such that $\|\nabla_x \ell(x_1; s) - \nabla_x \ell(x_2; s)\| \leq L\|x_1 - x_2\|$ for any $x_1, x_2 \in \mathbb{R}^d$ and $s \in \mathbb{S}$.

2.3 DRO Objective and Its Dual Form

In empirical risk minimization (ERM), the goal is to solve

$$\inf_x \mathbb{E}_{S\sim P_0}\, \ell(x; S),$$

where the objective function is the expectation of the loss function with respect to the training distribution P_0. To solve the distributional mismatch between training data and testing data, the formulation of Distributionally Robust Optimization (DRO) (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017; Rahimian and Mehrotra 2019) was developed, where the goal is to minimize the expected loss with respect to the worst distribution in an uncertainty set $\mathcal{U}(P_0)$:

$$\inf_x \sup_{Q\sim\mathcal{U}(P_0)} \mathbb{E}_{S\sim Q}\, \ell(x; S). \tag{4}$$

DRO problems under different uncertainty sets are fundamentally different. Consider the uncertainty set defined by φ-divergence $D_\varphi(Q\|P_0)$, which is one of the most common choices in the literature and can be written as $D_\varphi(Q\|P_0) := \int \varphi\left(\frac{dQ}{dP_0}\right) dP_0$, where φ is a non-negative convex function such that $\varphi(1) = 0$ and $\varphi(t) = +\infty$ for ant $t < 0$. Then let the uncertainty set $\mathcal{U}(P_0) := \{Q : D_\varphi(Q\|P_0) \leq \rho\}$ where $\rho > 0$ is the radius of the uncertainty set.

In this paper, we study the general Cressie-Read family of φ-divergence (Cressie and Read 1984; Van Erven and Harremos 2014), where

$$\varphi_k(t) := \frac{t^k - kt + k - 1}{k(k-1)}, \tag{5}$$

$k \in (-\infty, +\infty) \setminus \{0, 1\}$. Let $k_* = \frac{k}{k-1}$. This family includes as special cases χ^2-divergence ($k = 2$) and KL divergence ($k \to 1$). When $k > 2$, the conjugate function of $\varphi_k(t)$ (which will be introduced later) is not smooth, thus the problem becomes hard to solve even in the penalized formulation (Jin et al. 2021). In this paper, we focus on $k \in (1, 2]$ ($k_* \in [2, \infty)$). The objective is

$$\inf_x \sup_{Q:D_{\varphi_k}(Q\|P_0)\leq\rho} \mathbb{E}_{S\sim Q}\, \ell(x; S). \tag{6}$$

Solving (6) directly is challenging due to the sup over Q. In (Namkoong and Duchi 2016), a finite-dimensional vector $\mathbf{q}$ was used to parameterize the distributions in the uncertainty set since $Q \ll P_0$ for φ-divergence. Then the DRO problem becomes a convex concave min-max problem. This

method can be extended to the case with non-convex loss function by applying the algorithms for non-convex concave min-max problems (Rafique et al. 2022; Lin, Jin, and Jordan 2020; Xu et al. 2023). However, the dimension of distribution in the uncertainty set is equal to the number of training samples. Thus, the computational complexity at each iteration is linear in the sample size and is prohibitive in large-scale applications.

To obtain a complexity independent of the sample size, one alternative is to use its dual. By duality, we can show that the DRO objective (6) can be equivalently written as (Levy et al. 2020; Shapiro 2017)

$$\inf_x \inf_{\lambda\geq 0,\tilde{\eta}\in\mathbb{R}} \mathbb{E}_{S\sim P_0}\left[\lambda\varphi_k^*\left(\frac{\ell(x;S)-\tilde{\eta}}{\lambda}\right)+\lambda\rho+\tilde{\eta}\right],$$

where $\varphi_k^*(t') = \sup_t\{t't - \varphi_k(t)\}$ is the conjugate function of $\varphi_k(t')$. In this way, the optimization problem under an unknown distribution is rewritten into one under a known distribution. The subsampling method can then be used, which leads to a complexity independent of the sample size (which will be introduced later). For the Cressie-Read family in (5), the corresponding conjugate function family is $\varphi_k^*(t) = \frac{1}{k}\left[((k-1)t+1)_+^{k_*} - 1\right]$. Therefore, the objective can be written as

$$\inf_{x,\lambda\geq 0,\tilde{\eta}\in\mathbb{R}} \mathbb{E}_{S\sim P_0}\left[\frac{\lambda}{k}\left((k-1)\frac{\ell(x;S)-\tilde{\eta}}{\lambda}+1\right)_+^{k_*}\right] + \lambda\left(\rho-\frac{1}{k}\right)+\tilde{\eta}.$$

Let $\eta = \tilde{\eta} - \frac{\lambda}{k-1}$ and the corresponding objective is

$$\inf_{x,\lambda\geq 0,\eta\in\mathbb{R}} \mathbb{E}_{S\sim P_0}\left[\frac{(k-1)^{k_*}}{k}(\ell(x;S)-\eta)_+^{k_*}\lambda^{1-k_*}\right] + \lambda\left(\rho+\frac{1}{k(k-1)}\right)+\eta.$$

Define

$$f(x;\lambda;\eta;s) = \frac{(k-1)^{k_*}}{k}(\ell(x;S)-\eta)_+^{k_*}\lambda^{1-k_*} + \lambda\left(\rho+\frac{1}{k(k-1)}\right)+\eta. \quad (7)$$

Thus the goal is to solve

$$\inf_x \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F(x;\lambda;\eta), \quad (8)$$

where $F(x;\lambda;\eta)$ is defined as $F(x;\lambda;\eta) = \mathbb{E}_{S\sim P_0}\Big[f(x;\lambda;\eta;S)\Big]$. Therefore, we reformulate the DRO problem as one to minimize an objective function under a known distribution, where subsampling method could be used to reduce the complexity.

3 Analysis of Constrained DRO

In this section, we analyze the constrained DRO problems under Cressie-Read family divergence uncertainty sets with general smooth non-convex loss function. We first discuss the challenges appearing in constrained formulations, then we present how to construct the corresponding approximated problem in order to overcome these challenges.

3.1 Smooth and Lipschitz Approximation

For $\lambda \in [0,+\infty), \eta \in \mathbb{R}$, the objective function $F(x;\lambda;\eta)$ is neither smooth nor Lipschitz. Thus it is difficult to implement gradient-based algorithms. In the following, we will construct an approximation of the original problem so that the objective function $F(x;\lambda;\eta)$ becomes smooth and Lipschitz by constraining both λ and η in some bounded intervals.

Denote by $\omega = (k(k-1)\rho+1)^{\frac{1}{k_*}}$. Since the loss function is bounded such that $0 \leq \ell \leq B$, we can show that there exists an upper bound $\bar{\lambda} = (k-1)\omega^{-1}\left(1+\frac{(\frac{1}{\omega})^{\frac{1}{k_*-1}}}{1-(\frac{1}{\omega})^{\frac{1}{k_*-1}}}\right)B$ which only depends on k, ρ and B such that the optimal value $\lambda^* \leq \bar{\lambda}$. In this paper, we do not assume that $\lambda^* \geq \lambda_0 > 0$ as in (Wang, Gao, and Xie 2021). Instead, we consider an approximation with $\lambda \in [\lambda_0, \bar{\lambda}]$, and show that the difference between the orignial and the approximation can be bounded. We can show corresponding optimal $\eta^* \in [-\bar{\eta}, B]$, where $\bar{\eta} = \bar{\lambda}\left(\frac{k}{(k-1)^{k_*}k_*}\right)^{\frac{1}{k_*-1}}$. The proof can be found in Appendix A. The challenge lies in that the value of η can be negative. Thus given this η, the optimal value of λ can be quite large then it is hard to upper bound λ. In our proof, we change the objective to the function that only depends on η and find the lower bound on η. Based on this lower bound, we solve this challenge and further get the bound on λ.

We show that the difference between the original and the approximation can be bounded in the following lemma.

Lemma 1. *$\forall x \in \mathbb{R}^d, 0 \leq \lambda_0 \leq \bar{\lambda}$,*

$$\left|\inf_{\lambda\in[\lambda_0,\bar{\lambda}],\eta\in[-\bar{\eta},B]} F(x;\lambda;\eta) - \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F(x;\lambda;\eta)\right| \leq 2\lambda_0\rho.$$

The proof can be found in Appendix B. Note in the proof of this lemma, we provide the strong duality of

$$\sup_{D_{\varphi_k}(Q\|P_0)\leq\rho} \mathbb{E}_{S\sim Q}\,\ell(x;S) - \lambda_0 D_{\varphi_k}(Q\|P_0),$$

where both the hard constraint and regulator are kept. This is different from the approach in Section 3.2 of (Shapiro 2017). Note this strong duality holds for any φ-divergence DRO problem.

Lemma 1 demonstrates that the non-smooth objective function can be approximated by a smooth objective function. A smaller λ_0 makes the gap smaller but the function "less smooth".

3.2 Convexity and Smoothness on Parameters

The advantage of our approximated problem is that the function is smooth in all x, λ, and η. Moreover, We find that the objective function is convex in λ and η though the loss function is non-convex in x.

Lemma 2. *Define $z = (\lambda,\eta) \in \mathcal{M}$, where $\mathcal{M} = \{(\lambda,\eta) : \lambda \in [\lambda_0,\bar{\lambda}], \eta \in [-\bar{\eta},B]\}$. Then $\forall x \in \mathbb{R}^d, z \in \mathcal{M}$, the objective function $F(x;z)$ is convex and L_z-smooth in z, where $L_z = \frac{1}{\lambda_0} + \frac{2(B+\bar{\eta})}{\lambda_0^2} + \frac{(B+\bar{\eta})^2}{2\lambda_0^3}$ if $k_* = 2$ and $L_z = \frac{(k-1)^{k_*}}{k}k_*(k_*-1)\left(\frac{(B+\bar{\eta})^{k_*}}{\lambda_0^{k_*+1}} + \frac{(B+\bar{\eta})^{k_*-2}}{\lambda_0^{k_*-1}}\right)$ if $k_* > 2$.*

Algorithm 1: SFK-DRO

Input: Iteration number K, initial point (x_1, z_1), sample numbers n_x, n_z, step size α, and one constant C

Let $t = 1$
while $t \leq K$ **do**
  randomly select n_x samples and compute $\nabla_x f_x(x_t, z_t) = \sum_{i=1}^{n_x} \frac{\nabla_x f(x_t; z_t; s_i)}{n_x}$.
  $x_{t+1} = x_t - \alpha \nabla_x f_x(x_t, z_t)$
  randomly select n_z samples and compute $\nabla_z f_z(x_{t+1}, z_t) = \sum_{j=1}^{n_z} \frac{\nabla_z f(x_{t+1}; z_t; s_j)}{n_z}$
  $e_t = \arg\min_{e \in \mathcal{M}} \langle e, \nabla_z f_z(x_{t+1}; z_t) \rangle$
  $d_t = e_t - z_t$
  $g_t = \langle d_t, -\nabla_z f_z(x_{t+1}; z_t) \rangle$
  $\gamma_t = \min\left\{\frac{g_k}{C}, 1\right\}$
  $z_{t+1} = z_t + \gamma_t d_t$
  $t = t + 1$
end while
$t' = \arg\min_t \|\nabla_x f_x(x_t; z_t)\|^2 + g_t^2$

Output: $(x_{t'+1}, z_{t'})$

Moreover, the objective function $F(x; z)$ is L_x-smooth in x, where $L_x = \frac{(k-1)^{k_*}}{k} k_* \lambda_0^{1-k_*} (B + \bar{\eta})^{k_*-2} ((k_* - 1)G^2 + (B + \bar{\eta})L)$.

The proof can be found in Appendix C. Note the first-order gradient of the objective function is non-differential at some point when $k_* = 2$. Therefore, we discuss in two cases: $k_* > 2$ and $k_* = 2$. In the first case, we can get the Hessian matrix of the objective. In the second case, we show the smoothness and convexity.

4 Mini-Batch Algorithm

Existing constrained stochastic algorithm for general non-convex functions (Ghadimi, Lan, and Zhang 2016) can be used to solve the approximated problem directly. However, their method optimizes $y = (x; \lambda; \eta)$ as a whole. It can be seen that the objective function is non-convex in y and the computation complexity to get the ϵ-stationary point is $\mathcal{O}(\epsilon^{-3k_*-5})$.

In the previous section, we show that $F(x; z)$ is L_z-smooth in z and L_x-smooth in x. Moreover, $L_z \sim \mathcal{O}(\lambda_0^{-k_*-1})$, which is much larger then L_x when λ_0 is small, since $L_x \sim \mathcal{O}(\lambda_0^{-k_*+1})$. If we optimize all the parameters together, we need to implement non-convex algorithms to optimize a smooth function with a large smooth constant, which is not computationally efficient. However, if we optimize x and z separately, though $L_z > L_x$ which requires more resources to optimize z, the convexity in z makes it faster to converge to the optimal value of z.

This motivates us to consider a stronger convergence criterion. Instead of finding the ϵ- stationary point for $F(y)$, we can find (x, λ, η) such that

$$|\nabla_x F(x; \lambda; \eta)| \leq \epsilon,$$
$$|F(x; \lambda; \eta) - \inf_{\lambda' \geq 0; \eta'} F(x; \lambda'; \eta')| \leq \epsilon.$$

We then provide our Stochastic gradient and Frank-Wolfe DRO algorithm (SFK-DRO), which optimizes x and z separately (see Algorithm 1). Define $D = \max_{z_1, z_2 \in \mathcal{M}} \|z_1 - z_2\|$, $\sigma = \frac{(k-1)^{k_*}}{k} k_* (B + \bar{\eta})^{k_*-1} G \lambda_0^{1-k_*}$, $\Delta = F(x_1; z_1) - \inf_{x,z \in \mathcal{M}} F(x; z)$ and C is a constant such that $C \geq DL_z$. The convergence rate is then provided in the following theorem.

Theorem 1. *With a mini-batch size $n_x = \frac{8L_x\sigma^2}{C\epsilon^2} \sim \mathcal{O}(\lambda_0^{-2k_*+4}\epsilon^{-2}), n_z \sim \mathcal{O}(\epsilon^{-k_*})$ such that $3B\sqrt{1 + k(k-1)\rho}\sqrt{\frac{4+\log(n_z)}{4n_z}} < \frac{\epsilon}{4}$ if $k_* = 2$ or $3B(1 + k(k-1)\rho)^{\frac{1}{k}} \left(\frac{1}{n_z} + \frac{1}{2^{k_*-1}(k_*-2)n_z}\right)^{\frac{1}{k_*}} < \frac{\epsilon}{4}$ if $k_* > 2$ and $\alpha = \frac{1}{C}, \lambda_0 = \frac{\epsilon}{8\rho}$, for any small $\epsilon > 0$ such that $\frac{DL_z}{L_x} \sim \mathcal{O}(\epsilon^{-2}) \geq 2$ and $\frac{g}{C} \sim \mathcal{O}(\epsilon) \leq 1$, at most $T = 16C\Delta\epsilon^{-2} \sim \mathcal{O}(\lambda_0^{-k_*-1}\epsilon^{-2})$ iterations are needed to guarantee a stationary point $(x_{t'+1}; z_{t'})$ in expectation:*

$$\mathbb{E}\|\nabla_x F(x_{t'+1}; z_{t'})\| \leq \epsilon,$$
$$\mathbb{E}\left[\left|F(x_{t'+1}; z_{t'}) - \inf_{\lambda \geq 0; \eta \in \mathbb{R}} F(x_{t'+1}; \lambda; \eta)\right|\right] \leq \epsilon.$$

The detailed proof can be found in Appendix D and a proof sketch will be provided later. Before that, we introduce a lemma for our subsampling method. Via this lemma, we can show the complexity is independent of the sample size and thus is suitable for our large-scale setting. When we optimize z, an estimator $f_z(x, z) = \sum_{j=1}^{n_z} \frac{f(x; z; s_j)}{n_z}$ is build to estimate $F(x; z) = \mathbb{E}_{S \sim P_0}\left[f(x; z; S)\right]$. Though the estimator is unbiased, in our Frank-Wolfe update process (Jaggi 2013; Frank, Wolfe et al. 1956; Lacoste-Julien 2016) we need to estimate $\min F(x; z)$ via $\mathbb{E} \min f_z(x; z)$. Obviously, the expectation of minimum is not equal to the minimum of expectation, thus it is a biased estimator. In the following lemma, we show that this gap can be bounded by a decreasing function of the sample batch n_z.

Lemma 3. *For any bounded loss function ℓ, if $k_* = 2$,*

$$\left|\inf_{z \in \mathcal{M}} [F(x_{t+1}; z)] - \mathbb{E}\left[\inf_{z \in \mathcal{M}} f_z(x_{t+1}; z)\right]\right| \leq 3B\sqrt{1 + k(k-1)\rho}\sqrt{\frac{4 + \log(n_z)}{4n_z}};$$

and if $k_* > 2$,

$$\left|\inf_{z \in \mathcal{M}} [F(x_{t+1}; z)] - \mathbb{E}\left[\inf_{z \in \mathcal{M}} f_z(x_{t+1}; z)\right]\right| \leq 3B(1 + k(k-1)\rho)^{\frac{1}{k}} \left(\frac{1}{n_z} + \frac{1}{2^{k_*-1}(k_*-2)n_z}\right)^{\frac{1}{k_*}}.$$

The detailed proof can be found in Appendix E. Note that (Levy et al. 2020) only shows this lemma when $k_* = 2$, and we extend the results to $k_* > 2$. This lemma shows that the gap is in the order of $\mathcal{O}({n_z}^{-\frac{1}{k_*}})$ and is independent of the total number of samples.

4.1 Proof Sketch of Theorem 1

We use a stochastic gradient descent method (Moulines and Bach 2011; Gower et al. 2019; Robbins and Monro 1951) to update x. Since the objective function is L_x-smooth in x, if $\alpha \leq \frac{1}{2L_x}$ we have that:

$$\frac{\alpha}{2}\mathbb{E}\left[\|\nabla_x F_x(x_t;z_t)\|^2\right] \leq \mathbb{E}[F(x_t;z_t)] - \mathbb{E}[F(x_{t+1};z_t)] + \alpha^2 L_x \mathbb{E}[\|\nabla_x f_x(x_t;z_t) - \nabla_x F_x(x_t;z_t)\|^2], \quad (9)$$

where $f_x(x,z) = \sum_{j=1}^{n_x} \frac{f(x;z;s_j)}{n_x}$. Define $\sigma = \frac{(k-1)^{k_*}}{k} k_* (B+\bar{\eta})^{k_*-1} G \lambda_0^{1-k_*}$ and we can show that

$$\mathbb{E}[\|\nabla_x f_x(x_t;z_t) - \nabla_x F_x(x_t;z_t)\|^2] \leq \frac{\sigma^2}{n_x}. \quad (10)$$

Since $z \in \mathcal{M}$, instead of the stochastic gradient descent method, we employ the Frank-Wolfe method (Frank, Wolfe et al. 1956) to update z. Define $e_t = \arg\min_{e\in\mathcal{M}} \langle e, \nabla_z f_z(x_{t+1};z_t)\rangle$ and $g_t = \langle e_t - z_t, -\nabla_z f_z(x_{t+1};z_t)\rangle$. In addition, we have

$$g_t \geq f_z(x_{t+1};z_t) - \min_{z\in\mathcal{M}} f_z(x_{t+1};z)$$

since $f(x;z)$ is convex in z (Jaggi 2013). We can show that $\frac{g_t}{C} \sim \mathcal{O}(\lambda_0)$ thus for small λ_0 we have $\frac{g_t}{C} \leq 1$. Then due to the fact that the objective is L_z-smooth in z ((9) of (Lacoste-Julien 2016)), we have that

$$\mathbb{E}\left[\frac{g_t^2}{2C}\right] \leq \mathbb{E}[F(x_{t+1};z_t)] - \mathbb{E}[F(x_{t+1};z_{t+1})]. \quad (11)$$

By recursively adding (9) and (11), we have that

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\alpha}{2}\mathbb{E}\left[\|\nabla_x F_x(x_t;z_t)\|^2\right] + \mathbb{E}\left[\frac{g_t^2}{2C}\right] \leq \frac{F(x_1;z_1) - \mathbb{E}[F(x_{T+1};z_{T+1})]}{T} + \alpha^2 L_x \frac{\sigma^2}{n_x}. \quad (12)$$

Since $L_z \sim \mathcal{O}(\lambda_0^{-k_*-1})$ and $L_z \sim \mathcal{O}(\lambda_0^{-k_*-1})$, for small λ_0 we have $C \geq DL_z \geq 2L_x$. Then we set $\alpha = \frac{1}{C} \leq \frac{1}{2L_x}$, $T = 16C\Delta\epsilon^{-2} \sim \mathcal{O}(\lambda_0^{-k_*-1}\epsilon^{-2})$, $n_x = \frac{8L_x\sigma^2}{C\epsilon^2}$, and denote $\Delta = F(x_1;z_1) - \min_{x,z\in\mathcal{M}} F(x;z)$, for some $t \in [1,T]$ we have

$$\mathbb{E}\left[\|\nabla_x F_x(x_t;z_t)\|\right] \leq \frac{\epsilon}{2}, \quad (13)$$

$$\mathbb{E}\left[F(x_{t+1};z_t) - \inf_{z\in\mathcal{M}} f_z(x_{t+1};z)\right] \leq \mathbb{E}[g_t] \leq \frac{\epsilon}{2}. \quad (14)$$

We choose (x_{t+1},z_t) as our output and we need to bound $\mathbb{E}\left[\|\nabla_x F_x(x_{t+1};z_t)\|\right]$ and $\mathbb{E}\left[F(x_{t+1};z_t) - \inf_{z\in\mathcal{M}} F_z(x_{t+1};z)\right]$. Since $F(x;z)$ is L_x-smooth in x, we have that $\mathbb{E}[\|\nabla_x F_x(x_{t+1};z_t)\|] \leq \epsilon$.

By Lemma 3, we pick $n_z \sim \mathcal{O}(\epsilon^{-k_*})$ such that

$$\left|\inf_{z\in\mathcal{M}}[F(x_{t+1};z)] - \mathbb{E}\left[\inf_{z\in\mathcal{M}} f_z(x_{t+1};z)\right]\right| \leq \frac{\epsilon}{4}.$$

By Lemma 1, when $\lambda_0 = \frac{\epsilon}{8\rho}$, we have

$$\left|\inf_{\lambda\in[\lambda_0,\bar{\lambda}],\eta\in[-\bar{\eta},B]} F(x;\lambda;\eta) - \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F(x;\lambda;\eta)\right| \leq \frac{\epsilon}{4}.$$

Thus we have

$$F(x_{t+1};z_t) - \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F(x;\lambda;\eta) \leq \epsilon, \quad (15)$$

which completes the proof.

5 Smoothed CVaR

Our algorithm can also solve other DRO problems efficiently, for example, the Smoothed CVaR proposed in (Jin et al. 2021). The CVaR DRO is an important φ-divergence DRO problem, where $\varphi(t) = \mathbb{1}_{[0,\frac{1}{\mu})}$ if $0 \leq t < \frac{1}{\mu}$, and $0 < \mu < 1$ is some constant. The dual expression of CVaR can be written as

$$\mathcal{L}_{CVaR}(x;P_0) = \inf_{\eta\in\mathbb{R}} \frac{1}{\mu}\mathbb{E}_{S\sim P_0}\left[(\ell(x;S) - \eta)_+ + \eta\right].$$

The dual of CVaR is non-differentiable, which is undesirable from an optimization viewpoint. To solve this problem, (Jin et al. 2021) proposed a new divergence function, which can be seen as a smoothed version of the CVaR. Their experiment results show the optimization of smoothed CVaR is much easier. However, (Jin et al. 2021)'s method only works for the penalized formulation of DRO. We will show that our method can solve the constrained smoothed CVaR.

Here, the divergence function is

$$\varphi_s(t) = \begin{cases} t\log(t) + \frac{1-\mu t}{\mu}\log(\frac{1-\mu t}{1-\mu}), & t \in [0,\frac{1}{\mu}); \\ +\infty, & \text{otherwise.} \end{cases} \quad (16)$$

The corresponding conjugate function is

$$\varphi_s^*(t) = \frac{1}{\mu}\log(1-\mu+\mu\exp(t)). \quad (17)$$

The objective function is then written as

$$\inf_x \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F_s(x;\lambda;\eta) = \mathbb{E}_{S\sim P_0}\left[\lambda\varphi_s^*(\frac{\ell(x;S)-\eta}{\lambda}) + \lambda\rho + \eta\right]. \quad (18)$$

We can show that there exist upper bounds for the optimal values λ^* and η^*. There exists a $\bar{\lambda} > 0$ only depends on μ, B and ρ such that $\lambda^* \in [0,\bar{\lambda}]$ and $\eta^* \in [0,B]$. The proof can be found in Appendix F.

This objective function is non-smooth when $\lambda \to 0$. Therefore, we take a similar approach as the one in Section 3.1 to approximate the original problem with $\lambda \in [\lambda_0,\bar{\lambda}]$. We bound the difference in the following lemma.

Lemma 4. *$\forall x \in \mathbb{R}^d, \lambda_0 \geq 0$,*

$$\left|\inf_{\lambda\in[\lambda_0,\bar{\lambda}],\eta\in[0,B]} F_s(x;\lambda;\eta) - \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F_s(x;\lambda;\eta)\right| \leq 2\lambda_0\rho.$$

The proof is similar to Lemma 1 thus is omitted here. In addition, we can show that $F_s(x;z)$ is L_z'-smooth and convex in z, where $L_z' \sim \mathcal{O}(\lambda_0^{-3})$ if $\lambda \in [\lambda_0,\bar{\lambda}]$. Also it is easy to get $F_s(x;z)$ is L_x'-smooth in x, where $L_x' \sim \mathcal{O}(\lambda_0^{-2})$.

Similar to eq. (42) and Remark 1 in (Levy et al. 2020), we can prove that $\left|\min_{z\in\mathcal{M}}[F_s(x_{t+1};z)] - \mathbb{E}\left[\min_{z\in\mathcal{M}} f_s(x_{t+1};z)\right]\right| \sim \mathcal{O}(n_s^{-0.5})$. We then use Algorithm 1 directly and the complexity to get the ϵ-stationary point is $\mathcal{O}(\epsilon^{-7})$. The detailed proof can be found in Appendix F.

Class	0	1	2	3	4	5	6	7	8	9
EMR	77.64	86.19	69.33	54.03	51.53	47.05	87.66	85.35	87.12	83.15
SFK-DRO	76.11	84.71	66.18	**54.95**	**58.65**	49.36	89.06	84.03	88.41	83.09
PAN-DRO	74.92	85.62	65.72	52.69	55.83	**49.50**	88.85	84.06	88.68	81.29

Table 1: Test Accuracy of each class for imbalanced CIFAR 10.

6 Numerical Results

In this section, we verify our theoretical results in solving an imbalanced classification problem. In the experiment, we consider a non-convex loss function and k is set to be 2 for the Cressie-Read family. We will show that 1) to optimize the same dual objective function, our proposed algorithm converges faster than the general Proximal Gradient Descent(PGD) algorithm (Ghadimi, Lan, and Zhang 2016); 2) The performance proposed algorithm for the constrained DRO problem outperforms or is close to the performance of the penalized DRO with respect to the worst classes. Both of them outperform the baseline.

Tasks. We conduct experiments on the imbalanced CIFAR-10 dataset, following the experimental setting in (Jin et al. 2021; Chou et al. 2020). The original CIFAR-10 test dataset consists of 10 classes, where each of the classes has 5000 images. We randomly select training samples from the original set for each class with the following sampling ratio: $\{0.804, 0.543, 0.997, 0.593, 0.390, 0.285, 0.959, 0.806, 0.967, 0.660\}$. We keep the test dataset unchanged.

Models. We learn the standard Alexnet model in (Krizhevsky, Sutskever, and Hinton 2012) with the standard cross-entropy (CE) loss. For the comparison of convergence rate, we optimize the same dual objective with the PGD algorithm in (Ghadimi, Lan, and Zhang 2016). To compare robustness, we optimize the ERM via vanilla SGD. In addition, we propose an algorithm PAN-DRO, which fixes λ and only optimizes η and the neural network. Thus it gets the solution for the penalized DRO problem.

Training Details. We set $\lambda_1 = 1, \eta_1 = 0$, $\lambda_0 = 0.1, -\bar{\eta} = -10$, and the upper bounds $\bar{\lambda} = 10, B = 10$. To achieve a faster optimization rate, we set the learning rate $\alpha = 0.01$ before the first 40 epochs and $\alpha = 0.001$ after. The mini-batch size is chosen to be 128. All of the results are moving averaged by a window with size 5. The simulations are repeated by 4 times.

Results. In Figure 1, we plot the value of the CE loss using different algorithms through the training process. It can be seen that to optimize the same dual objective function with the same learning rate, the PGD algorithm converges slower than our proposed DRO algorithms, which matches our theoretical results. Moreover, compared with ERM, the DRO algorithms have higher training losses but lower test losses, which demonstrates they are robust.

We also provide the test accuracy of trained models in Table 1. It can be shown that for class $3, 4, 5$, the accuracies are the lowest due to the limited samples. For these classes, the performance of our SFK-DRO algorithm for the constrained DRO is better or close to the performance of PAN-DRO for the penalized DRO. Both DRO algorithms outperform the vanilla ERM algorithm.

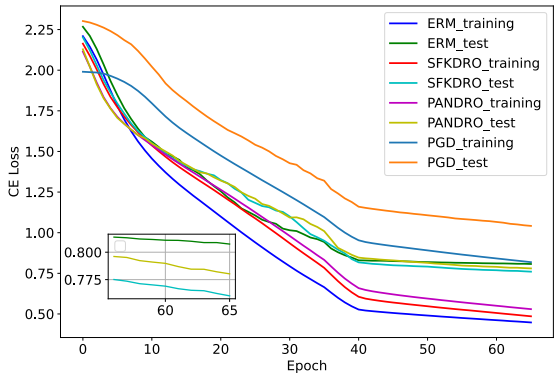

Figure 1: Training curve of classification task.

7 Conclusion

In this paper, we developed the first stochastic algorithm for large-scale non-convex stochastic constrained DRO problems in the literature with theoretical convergence and complexity guarantee. We developed a smooth and Lipschitz approximation with bounded approximation error to the original problem. Compared with existing algorithms, the proposed algorithm has an improved convergence rate. The computational complexity at each iteration is independent of the size of the training dataset, and thus our algorithm is applicable to large scale applications. Our results hold for a general family of Cressie-Read divergences.

Acknowledgments

This work of Q. Zhang and S. Zou is supported by the National Science Foundation under Grants CCF-2106560. Department of Education. Y. Zhou's work is supported by the National Science Foundation under Grants CCF-2106216, DMS-2134223 and ECCS-2237830 (CAREER). L. Shen's work is supported by the NSF under Grant DMS-2208385.

This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award No. 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S.

References

Ali, S. M.; and Silvey, S. D. 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1): 131–142.

Ben-Tal, A.; Den Hertog, D.; De Waegenaere, A.; Melenberg, B.; and Rennen, G. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 120–128.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: rebalanced mixup. In *Proceedings of Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 95–110. Springer.

Cressie, N.; and Read, T. R. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3): 440–464.

Csiszár, I. 1967. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2: 299–318.

Curi, S.; Levy, K. Y.; Jegelka, S.; and Krause, A. 2020. Adaptive sampling for stochastic risk-averse learning. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, 1036–1047.

Daume III, H.; and Marcu, D. 2006. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26: 101–126.

Duchi, J.; and Namkoong, H. 2018. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*.

Duchi, J. C.; Glynn, P. W.; and Namkoong, H. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969.

Frank, M.; Wolfe, P.; et al. 1956. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110.

Ghadimi, S.; Lan, G.; and Zhang, H. 2016. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2): 267–305.

Ghosh, S.; Squillante, M.; and Wollega, E. 2018. Efficient stochastic gradient descent for distributionally robust learning. *arXiv preprint arXiv:1805.08728*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gower, R. M.; Loizou, N.; Qian, X.; Sailanbayev, A.; Shulgin, E.; and Richtárik, P. 2019. SGD: General analysis and improved rates. In *Proceedings of International conference on machine learning*, 5200–5209. PMLR.

Grother, P. J.; Grother, P. J.; Phillips, P. J.; and Quinn, G. W. 2011. *Report on the evaluation of 2D still-image face recognition algorithms*. US Department of Commerce, National Institute of Standards and Technology.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of International Conference on Machine Learning*, 1929–1938. PMLR.

Hovy, D.; and Søgaard, A. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, 483–488.

Hu, Z.; and Hong, L. J. 2013. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2): 9.

Jaggi, M. 2013. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of International conference on machine learning*, 427–435. PMLR.

Jin, J.; Zhang, B.; Wang, H.; and Wang, L. 2021. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Proceedings of Advances in Neural Information Processing Systems*, volume 34, 2771–2782.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in neural information processing systems*, volume 25.

Lacoste-Julien, S. 2016. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*.

Levy, D.; Carmon, Y.; Duchi, J. C.; and Sidford, A. 2020. Large-scale methods for distributionally robust optimization. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, 8847–8860.

Lin, T.; Jin, C.; and Jordan, M. 2020. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of International Conference on Machine Learning*, 6083–6093. PMLR.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Moulines, E.; and Bach, F. 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Namkoong, H.; and Duchi, J. C. 2016. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Proceedings of Advances in neural information processing systems*, volume 29.

Nesterov, Y. 2003. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Qi, Q.; Guo, Z.; Xu, Y.; Jin, R.; and Yang, T. 2021. An online method for a class of distributionally robust optimization with non-convex objectives. In *Proceedings of Advances in Neural Information Processing Systems*, volume 34, 10067–10080.

Qi, Q.; Lyu, J.; Bai, E. W.; Yang, T.; et al. 2022. Stochastic constrained dro with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740*.

Rafique, H.; Liu, M.; Lin, Q.; and Yang, T. 2022. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3): 1087–1121.

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

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Rockafellar, R. T.; Uryasev, S.; et al. 2000. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42.

Rockafellar, R. T.; and Wets, R. J. 1998. *Variational Analysis*. Springer.

Shapiro, A. 2017. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4): 2258–2275.

Sinha, A.; Namkoong, H.; Volpi, R.; and Duchi, J. 2017. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Soma, T.; and Yoshida, Y. 2020. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*.

Tamar, A.; Glassner, Y.; and Mannor, S. 2015. Optimizing the CVaR via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Van Erven, T.; and Harremos, P. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7): 3797–3820.

Wang, J.; Gao, R.; and Xie, Y. 2021. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*.

Xu, Z.; Zhang, H.; Xu, Y.; and Lan, G. 2023. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 1–72.