Non-convex Distributionally Robust Optimization: Non-asymptotic Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distributionally robust optimization (DRO) is a widely-used approach to learn models that are robust against distribution shift. Compared with the standard optimization setting, the objective function in DRO is more difficult to optimize, and most of the existing theoretical results make strong assumptions on the loss function. In this work we bridge the gap by studying DRO algorithms for general smooth non-convex losses. By carefully exploit the specific form of the DRO objective, we are able to provide non-asymptotic convergence guarantees even though the objective function is possibly non-convex, non-smooth and has unbounded gradient noise. In particular, we prove that a special algorithm called the mini-batch normalized gradient descent with momentum, can find an ϵ-first-order stationary point under $\mathcal{O}(\epsilon^{-4})$ gradient complexity. For the conditional value-at-risk (CVaR) objective, we propose a penalized DRO objective based on a smoothed version of the CVaR that allows us to obtain better complexity. We finally verify our theoretical results in a number of tasks and find that the proposed algorithm can consistently achieve prominent acceleration.

1 Introduction

For a classical machine learning problem, the goal is typically to train a model over a training set that achieves good performance on a test set, where both the training set and the test set are drawn from the *same* distribution P. While such assumption is reasonable and simple for theoretical analysis, it is often not the case in real applications. For example, this setting may be improper when there is a gap between training and test distribution (e.g. in domain adaptation tasks) [Zhang et al., 2021], when there is severe class imbalance in the training set [Sagawa et al., 2020], when fairness in minority groups is an important consideration [Hashimoto et al., 2018], or when the deployed model is exposed to adversarial attacks [Sinha et al., 2018].

Distributionally robust optimization (DRO), as a popular approach to deal with the above situations, has attracted great interest for the machine learning research communities in recent years. In contrast to classical machine learning problems, for DRO it is desired that the trained model still has good performance under distributional shift. Specifically, DRO proposes to minimize the worst-case loss over a set of probability distributions Q around P. This can be formulated as the following constrained optimization problem [Rahimian and Mehrotra, 2019, Shapiro, 2017]:

$$\text{minimize}_{x\in\mathcal{X}} \quad \Psi(x) := \sup_{Q\in\mathcal{U}(P)} \mathbb{E}_{\xi\sim Q}\left[\ell(x;\xi)\right] \tag{1}$$

where $x \in \mathcal{X}$ is the parameter to be optimized, ξ is a sample randomly drawn from distribution Q, and $\ell(x;\xi)$ is the loss function so that $\mathbb{E}_{\xi\sim Q}\left[\ell(x;\xi)\right]$ is the expectation loss over distribution Q. The DRO loss $\Psi(x)$ is therefore the worst-case loss when the distribution P is shifted to Q. The set $\mathcal{U}(P)$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

is called the uncertainty set and typically defined as

$$\mathcal{U}(P) := \{Q : d(Q, P) \leq \epsilon\} \tag{2}$$

where d measures the distance between two probability distributions, and the positive number ϵ corresponds to the magnitude of the uncertainty set.

Instead of imposing a hard constrained uncertainty set, sometimes it is more preferred to use a soft penalty term, resulting in the penalized DRO problem [Sinha et al., 2018]:

$$\text{minimize}_{x \in \mathcal{X}} \quad \Psi(x) := \sup_{Q} \left\{ \mathbb{E}_{\xi \sim Q} \left[\ell(x; \xi) \right] - \lambda d(Q, P) \right\} \tag{3}$$

where $\lambda > 0$ is the regularization coefficient.

There are many possible choices of d. A detailed discussion of different distance measures and their properties can be found in Rahimian and Mehrotra [2019]. In this paper we consider a general class of distances d called the ψ-divergence, which is a popular choice in DRO literature [Namkoong and Duchi, 2016, Shapiro, 2017]. Specifically, for a non-negative convex function ψ such that $\psi(1) = 0$ and two probability distributions P, Q such that Q is absolutely continuous w.r.t. P, the ψ-divergence between Q and P is defined as

$$d_\psi(Q, P) := \int \psi \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \right) \mathrm{d}P.$$

It follows that $d_\psi(Q, P) \geq 0$ and $d_\psi(Q, P) = 0$ iff $Q = P$ (almost surely under P).

The main focus of this paper is to study efficient first-order optimization algorithms for DRO problem (3) for *non-convex* losses $\ell(x, \xi)$. While non-convex models (especially deep neural networks) have been extensively used in DRO setting (e.g. Sagawa et al. [2020]), theoretical analysis about the convergence speed is still lacking. Most previous works (e.g. Levy et al. [2020]) assume the loss $\ell(\cdot, \xi)$ is convex, and in this case (3) is equivalent to a convex optimization problem (see Section 2 for details). Recently some works provide convergence rates of algorithms for non-convex losses in certain special cases, e.g. the divergence measure ψ is chosen as the conditional-value-at-risk (CVaR) and the loss function has some nice structural properties [Soma and Yoshida, 2020, Kalogerias, 2020]. Gürbüzbalaban et al. [2020] considered a more general setting but only proved an asymptotic convergence result for non-convex DRO.

Compared with these works, we provide the first *non-asymptotic* analysis of optimization algorithms for DRO with *general smooth non-convex* losses $\ell(x, \eta)$ and general ψ-divergence. In this setting, there are two major difficulties we must encounter: (i) the DRO objective $\Psi(x)$ is non-convex and can become arbitrarily *non-smooth*, causing standard techniques in smooth non-convex optimization fail to provide a good convergence guarantee; (ii) the noise of the stochastic gradient of $\Psi(x)$ can be arbitrarily large and unbounded even if we assume the gradient of the inner loss $\ell(x, \eta)$ has bounded variance. To tackle these challenges, we propose to optimize the DRO objective using *mini-batch normalized SGD with momentum*, and we are able to prove an $\mathcal{O}(\epsilon^{-4})$ complexity of this algorithm. The core technique here is to exploit the specific structure of $\Psi(x)$, which shows that (i) the DRO loss satisfies a generalized smoothness condition [Zhang et al., 2020a,b] and (ii) the variance of the stochastic gradient of $\Psi(x)$ can be bounded by the true gradient. This motivates us to adopt the special algorithm that combines gradient normalization and momentum techniques into SGD, by which both non-smoothness and the problem of unbounded noise can be tackled, finally resulting in an $\mathcal{O}(\epsilon^{-4})$ complexity similar to standard smooth non-convex optimization.

The above analysis applies to a broad class of divergence function ψ. In some special cases, we can continue to prove a better complexity bound in terms of problem-dependent parameters. As an example, we propose a new divergence ψ that can be seen as a smoothed variant of CVaR. We show that using smoothed CVaR we can obtain an improved complexity bound compared with the general case. While having similar behaviors as the original CVaR, it solves the non-differentiability issue and is more desirable from an optimization viewpoint.

Contributions. We summarize our main results and contributions below. Let ψ^* be the conjugate function of ψ (see Definition 2.3). For non-convex optimization problems, since obtaining the global minima is NP-hard in general, this paper adopts the commonly used (relaxed) criteria: to find an ϵ-approximate first-order stationary point such that $\|\nabla \Psi(x)\| \leq \epsilon$. We measure the complexity of optimization algorithms by the number of computations of the stochastic gradient $\nabla \ell(x, \xi)$ to reach an ϵ-stationary point.

- Assuming that ψ^* is smooth and the loss ℓ is Lipschitz and smooth (possibly non-convex or unbounded), we show in Section 3.2 that the mini-batch normalized momentum algorithm (cf. Algorithm 1) has a complexity of $\mathcal{O}(\epsilon^{-4})$.

- Assuming that ψ^* is further Lipschitz, in Section 3.4 we continue to prove a better complexity bound in terms of problem-dependent parameters. As a special case, we propose a new divergence which is a smoothed approximation of CVaR.

- We conduct experiments to verify our theoretical results. We discover that the normalized SGD with momentum algorithm can greatly accelerate the optimization process, and the DRO objective using smoothed CVaR is much easier to train compared with the vanilla CVaR.

1.1 Related work

Constrained DRO and Penalty-based DRO. There are two existing formulations of the DRO problem: the constrained DRO and the penalized DRO. The constrained DRO formulation (1) has been studied in a number of works [Namkoong and Duchi, 2016, Shapiro, 2017, Duchi and Namkoong, 2018], while other works consider the penalty-based formulation (3) [Sinha et al., 2018, Levy et al., 2020]. From a Lagrangian perspective, the two formulations are in fact equivalent; However, the dual objective of the constrained formulation is sometimes intractable as pointed out in [Namkoong and Duchi, 2016, Duchi and Namkoong, 2018]. In this paper we focus on the penalty-based version and provide the first non-asymptotic analysis in the non-convex setting. Moreover, we do not make the assumption that the loss is bounded, as assumed in Levy et al. [2020] in the convex setting.

DRO with ψ-divergence. ψ-divergence is one of the most standard choices in DRO literature to measure the distance between probability measures. It encompasses a variety of popular functions such as KL-divergence, χ^2-divergence, and the conditional-value-at-risk (CVaR). Table 1 gives detailed descriptions for these functions.

For CVaR, Namkoong and Duchi [2016] propose a mirror-descent method which achieves $\mathcal{O}(\sqrt{T})$ regret. Levy et al. [2020] proposes a stochastic gradient-based method with optimal convergence rate in the convex setting. They also discuss an alternative approach based on the dual formulation which they call *Dual SGM*. In the non-convex setting, Soma and Yoshida [2020] propose a smoothed approximation of CVaR and obtain an $\mathcal{O}(\epsilon^{-6})$ complexity. We contribute to this line of work by proposing a different divergence with similar behavior as CVaR, and we are able to prove an $\mathcal{O}(\epsilon^{-4})$ complexity.

For χ^2 divergence, Hashimoto et al. [2018] consider a constrained formulation of DRO but do not provide theoretical guarantees. Levy et al. [2020] propose algorithms based on an multi-level Monte-Carlo stochastic gradient estimator, and provide convergence guarantees in the convex setting. In contrast, we consider general smooth non-convex loss function ℓ and provide convergence guarantee as a special case of Corollary 3.6.

Non-smooth non-convex optimization. Conventional non-convex optimization typically focuses on smooth objective functions. For general smooth non-convex stochastic optimization, it is already known that the best achievable gradient complexity is $\mathcal{O}(\epsilon^{-4})$ for finding an ϵ-approximate stationary point [Arjevani et al., 2019], and SGD based algorithm can reach the complexity [Ghadimi and Lan, 2013]. However, the optimization can be much harder for non-smooth non-convex objective functions, and there are limited results in this setting. Ruszczynski [2020] proposes a stochastic gradient-based method and shows that it converges to a stationary point with probability one, under the assumption that the feasible region is bounded. For unconstrained optimization, Zhang et al. [2020c] shows that if the objective function is Lipschitz continuous and Hadamard semi-differentiable, then it is intractable to find an ϵ-stationary point. When the function is weakly convex, Davis and Drusvyatskiy [2019] shows that the projected SGD converges to the stationary point of a Moreau envelope, and a recent work [Mai and Johansson, 2020] extends this result to SGD with momentum. In this paper, we show that for smooth non-convex loss ℓ, DRO can be formulated as a non-smooth non-convex optimization problem, but the special property of the DRO loss make it possible to find an ϵ-stationary point with $\mathcal{O}(\epsilon^{-4})$ complexity.

2 Preliminaries

2.1 Notations and Assumptions

Throughout this paper we use $\|\cdot\|$ to denote the ℓ_2-norm in an Euclidean space $\mathbb{R}^d$ and use $\langle\cdot,\cdot\rangle$ to denote the standard inner product. For a real number t, denote $(t)_+$ as $\max(t,0)$. For a set C, denote $\mathbb{I}_C(\cdot)$ as the indicator function such that $\mathbb{I}_C(x)=0$ if $x\in C$ and $\mathbb{I}_C(x)=+\infty$ otherwise. We first list some basic definitions in optimization literature, which will be frequently used in this paper.

Definition 2.1 *(Lipschitz continuity) A mapping $f:\mathcal{X}\to\mathbb{R}^n$ is G-Lipschitz continuous if for any $x,y\in\mathcal{X}$ we have $\|f(x)-f(y)\|\le G\|x-y\|$.*

Definition 2.2 *(Smoothness) A function $f:\mathcal{X}\to\mathbb{R}$ is L-smooth if it is differentiable on $\mathcal{X}$ and for any $x,y\in\mathcal{X}$ we have $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$.*

Definition 2.3 *(Conjugate function) For a function $\psi:\mathbb{R}\to\mathbb{R}$, the conjugate function ψ^* is defined as $\psi^*(t):=\max_{s\in\mathbb{R}}(st-\psi(s))$.*

Assumption 2.4 *We make the following assumptions throughout the paper:*

- *Given ξ, the loss function $\ell(x,\xi)$ is G-Lipschitz continuous and L-smooth in x.*
- *ψ is a valid divergence function, i.e. a non-negative convex function satisfying $\psi(1)=0$ and $\psi(t)=+\infty$ for all $t<0$. Furthermore the conjugate ψ^* is M-smooth.*

2.2 Equivalent formulation of the DRO objective

The original formulation (3) involves a max operation over distributions which brings optimization difficulties. A common way to handle the inner maximum is to use duality. Using the conjugate function, after some straightforward calculations the DRO objective (3) can be equivalently written as (see detailed derivations in [Levy et al., 2020, Section A.1.2])

$$\Phi(x)=\min_{\eta\in\mathbb{R}}\lambda\mathbb{E}_{\xi\sim P}\psi^*\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)+\eta. \tag{4}$$

Thus, to solve (3) it suffices to jointly minimize

$$\mathcal{L}(x,\eta):=\mathbb{E}_{\xi\sim P}\left[\lambda\psi^*\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)+\eta\right] \tag{5}$$

over $(x,\eta)\in\mathcal{X}\times\mathbb{R}\subset\mathbb{R}^{n+1}$. This is a standard stochastic optimization problem, and a lot of works use formula (5) to design efficient algorithms [Shapiro, 2017, Duchi and Namkoong, 2018, Levy et al., 2020].

The property of the objective function (5) heavily depends on ψ^*. We list some popular choices of ψ together with the corresponding ψ^* in Table 1. They serve as motivating examples of our subsequent results.

Table 1: Some commonly used divergences and the corresponding conjugates.

Divergence	$\psi(t)$	$\psi^*(t)$
χ^2	$\frac{1}{2}(t-1)^2$	$-1+\frac{1}{4}(t+2)_+^2$
K-L	$t\log t-t+1$	e^t-1
CVaR	$\mathbb{I}_{[0,\alpha^{-1}]}$	$\alpha^{-1}(t)_+$
Cressie-Read	$\frac{t^k-kt+k-1}{k(k-1)}, k\in\mathbb{R}$	$\frac{1}{k}\left(((k-1)t+1)_+^{\frac{k}{k-1}}-1\right)$

Finally, we write the stochastic gradients of $\mathcal{L}$ as follows

$$\nabla_x\mathcal{L}(x,\eta;\xi)=(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\cdot\nabla\ell(x,\xi),\quad \nabla_\eta\mathcal{L}(x,\eta;\xi)=-(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)+1 \tag{6}$$

which will be used in subsequent analysis.

3 Analysis of general non-convex DRO

In this section we will analyze the DRO problem with general smooth non-convex loss functions ℓ. We first discuss the challenges appearing in our analysis, then show how to leverage the specific structure of the objective function in order to bypass these challenges. Finally, we are able to provide a non-asymptotic complexity bound using a special algorithm called the normalized SGD with momentum.

3.1 Challenges in non-convex DRO

The optimization theory has pointed out that if the objective function is smooth and the stochastic gradient is unbiased and has bounded variance (i.e. $\mathbb{E}_{\xi\sim P}\|\nabla_x \ell(x,\xi) - \nabla_x \ell(x)\|^2 \le \sigma^2$ for some σ and all $x \in \mathcal{X}$ where $\ell(x) = \mathbb{E}_{\xi\sim P}\ell(x,\xi)$), then standard stochastic gradient descent (SGD) algorithm can provably find a ϵ-first-order stationary point under $\mathcal{O}(\epsilon^{-4})$ gradient complexity [Ghadimi and Lan, 2013]. Here the smoothness and bounded variance property is crucial for the convergence of SGD [Zhang et al., 2019]. However, we find that both assumptions are violated in non-convex DRO, even if the inner loss $\ell(x,\xi)$ is smooth and the stochastic noise is bounded for both $\ell(x,\cdot)$ and $\nabla_x \ell(x,\cdot)$. We present a counter example to illustrate this point, in which we can gain some insight about the structure of the DRO objective.

Example 3.1 Consider the loss $\ell(x;\xi) = x^2\left(1 + \frac{\xi}{x^2+1}\right)^2$ which is a quadratic-like function with noise ξ, where ξ is a Rademacher variable drawn from $\{-1,+1\}$ with equal probabilities. Then a straightforward calculation shows that the loss ℓ has the following properties:

- (Smoothness) For any $\xi \in \{-1,+1\}$, $\ell(x,\xi)$ is L-smooth with respect to x for $L = 8$;
- (Bounded variance) For any $x \in \mathbb{R}$, $\mathbb{E}_\xi\left[\left(\ell(x,\xi) - x^2\right)^2\right] = \frac{4x^4}{(x^2+1)^2} + \frac{x^4}{(x^2+1)^4} \le 4$. It then follows that $\mathrm{Var}[\ell(x,\xi)] \le 4$;
- (Bounded variance for gradient) Similarly we can check that the gradient of ℓ also has bounded variance. Moreover, the variance tends to zero when $x \to \infty$.

Now consider the DRO where the ψ is chosen as the commonly used χ^2-divergence. Fix $\lambda = 1$ and $\eta = 0$. Based on the expression of $\psi^*(t)$ in Table 1, The DRO objective function (3) thus takes the form $\mathcal{L}(x,0;\xi) = \frac{1}{4}\left[x^2\left(1 + \frac{\xi}{x^2+1}\right)^2 + 2\right]^2 - 1$, which is a quartic-like function. It then follows that

- $\mathcal{L}(x,0;\xi) = \Theta(x^4)$ for large x and therefore $\mathcal{L}(x,0;\xi)$ is not globally smooth;
- $\nabla_x \mathcal{L}(x,0;\xi) = x^3 + 2x\xi + 2x + \mathcal{O}(1)$ for large x and the stochastic gradient variance $\mathrm{Var}[\mathcal{L}(x,0;\xi)] = \Theta(x^2)$ which is unbounded globally.

As we can see from the above example, both the local smoothness and the gradient variance of $\mathcal{L}$ strongly rely on x. Indeed, in general settings both the two quantities have a positive correlation with the magnitude of ℓ; as shown in Appendix, if we make the additional assumption that ℓ is bounded by a small constant, then the smoothness and gradient noise can be controlled in a straightforward way, and we show that a projected stochastic gradient method can be applied in this setting. However, such assumption is quite restrictive and not satisfactory.

3.2 Main results

In this section, we present the main theoretical result of this paper. All proofs can be founded in the Appendix. We make the following assumption on the noise of the stochastic loss:

Assumption 3.2 *We assume that for all $x \in \mathcal{X}$, the stochastic loss has bounded variance, i.e. $\mathbb{E}_{\xi\sim P}\left(\ell(x,\xi) - \ell(x)\right)^2 \le \sigma^2$ where $\ell(x) = \mathbb{E}_{\xi\sim P}\ell(x,\xi)$.*

We now provide formal statements of the important properties mentioned above, which show that both the gradient variance and the local smoothness strongly correlate to the gradient scale.

Lemma 3.3 *Under the Assumptions 2.4 and 3.2, the gradient estimators in* (6) *satisfies the following property:*

$$\mathbb{E}\left\|\nabla\mathcal{L}(x,\eta,\xi)-\nabla\mathcal{L}(x,\eta)\right\|^2 \leq (10G^2+1)M^2\lambda^{-2}\sigma^2 + 8G^2\left(1+\left\|\nabla\mathcal{L}(x,\eta)\right\|^2\right) \tag{7}$$

Lemma 3.4 *Under the Assumptions 2.4 and 3.2, for any pair of parameters* (x,η) *and* (x_2,η_2)*, we have the following gradient property:*

$$\|\nabla\mathcal{L}(x,\eta)-\nabla\mathcal{L}(x_2,\eta_2)\| \leq (K+L\|\nabla\mathcal{L}(x,\eta)\|)\|(x-x_2,\eta-\eta_2)\| \tag{8}$$

where $K=\frac{(G^2+1)M}{\lambda}+L$.

Note that if the term $L\|\nabla\mathcal{L}(x,\eta)\|$ in (8) is absent, Lemma 3.4 becomes exactly the K-smoothness property. Thus the inequality (8) can be seen as a generalized smoothness assumption. Zhang et al. [2020b] for the first time proposed such generalized smoothness for twice-differentiable functions under a different form, and Zhang et al. [2020a] further gave a comprehensive analysis of algorithms for generalized smooth functions. However, all these works make strong assumptions on the gradient noise and can not be applied in our setting.

Instead, we propose to use the *mini-batch normalized SGD with momentum* algorithm for non-convex DRO, shown in Algorithm 1. The algorithm has been theoretically analysed in Cutkosky and Mehta [2020] under a different setting. Compared with Cutkosky and Mehta [2020], we use mini-batches in each iteration in order to ensure convergence in our setting.

Algorithm 1: Mini-batch Normalized SGD with Momentum

Input : The objective function $F(w)$, distribution P, initial point w_0, initial momentum m_0, learning rate γ, momentum factor β, batch size S and the total number of iterations T

for $t \leftarrow 1$ ***to*** T **do**

  $\hat{\nabla}F(w_{t-1}) \leftarrow \frac{1}{S}\sum_{i=1}^{S}\nabla F(w_{t-1},\xi_{t-1}^{(i)})$ where $\{\xi_{t-1}^{(i)}\}_{i=1}^{S}$ are i.i.d. samples drawn from P

  $m_t \leftarrow \beta m_{t-1}+(1-\beta)\hat{\nabla}F(w_{t-1})$

  $w_t \leftarrow w_{t-1}-\gamma\dfrac{m_t}{\|m_t\|}$

The following main theorem establishes convergence guarantee of Algorithm 1. We further provide a sketch of proof in Section 3.3, in which we can gain insight on how gradient normalization and momentum techniques help tackle the difficulties in Lemma 3.3 and Lemma 3.4.

Theorem 3.5 *Suppose that* F *satisfies the following conditions:*

- *(Generalized smoothness)* $\|\nabla F(w_1)-\nabla F(w_2)\| \leq (K_0+K_1\|\nabla F(w_1)\|)\|w_1-w_2\|$ *holds for any* w_1,w_2*;*
- *(Gradient variance) The stochastic gradient* $\nabla F(w,\xi)$ *is unbiased (*$\nabla F(w)=\mathbb{E}[\nabla F(w,\xi)]$*) and satisfies* $\mathbb{E}\left\|\nabla F(w,\xi)-\nabla F(w)\right\|^2 \leq \Gamma^2\left\|\nabla F(w)\right\|^2+\Lambda^2$ *for some* Γ *and* Λ.

Let $\{w_t\}$ *be the sequence produced by Algorithm 1. Then with mini-batch size* $S=\mathcal{O}(\Gamma^2)$ *and suitable choice of parameters* γ *and* β*, for any small* $\epsilon=\mathcal{O}(K/L)$*, we need at most* $\mathcal{O}\left(\max\left\{1,\Gamma^2\right\}\Delta K\Lambda^2\epsilon^{-4}\right)$ *gradient complexity to guarantee that we find an* ϵ*-first-order stationary point in expectation, i.e.* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w_t)\| \leq \epsilon$ *where* $\Delta=F(w_0)-\inf_{w\in\mathbb{R}^d}F(w)$.

Substituting Lemmas 3.4 and 3.3 into Theorem 3.5 immediately yields the final result:

Corollary 3.6 *Suppose the DRO problem* (5) *satisfies Assumptions 2.4 and 3.2. Then the gradient complexity of Algorithm 1 for finding an* ϵ*-first-order stationary point is*

$$\mathcal{O}\left(G^2\left(G^2+1\right)\left(M^2\sigma^2\lambda^{-2}+1\right)\left(\lambda^{-1}M(1+G^2)+L\right)\Delta\epsilon^{-4}\right).$$

The above result shows that Algorithm 1 finds and ϵ-stationary point with complexity $\mathcal{O}(\epsilon^{-4})$, which is the same as standard smooth non-convex optimization. While the above bound is satisfactory in terms of ϵ, it may also be interesting to improve the complexity on other problem-dependent parameters such as G. In Section 3.4 we will obtain an improved complexity in terms of G for a special class of divergences ψ. We leave the improvement in the general case as an open problem.

3.3 Proof sketch of Theorem 3.5

Below we present our proof sketch. Similar to standard analysis in non-convex optimization, we first derive a descent inequality for functions satisfying the generalized smoothness:

Lemma 3.7 *(Descent Inequality) Let $F(x)$ be a function satisfying the generalized smoothness condition (i.e. for any x_1, x_2 we have $\|\nabla F(x_1) - \nabla F(x_2)\| \leq (K_0 + K_1\|\nabla F(x_1)\|)\|x_1 - x_2\|$). Then for any point x and direction z the following holds:*

$$F(x - z) \leq F(x) - \langle \nabla F(x), z \rangle + \frac{1}{2}(K_0 + K_1 \|\nabla F(x)\|) \|z\|^2 . \tag{9}$$

The above lemma suggests that the algorithm should take a small step size when $\|\nabla F(x)\|$ is large in order to decrease F. This is the main motivation of considering a normalized update. Indeed we can prove the following result:

Lemma 3.8 *Consider the algorithm that starts at w_0 and makes updates $w_{t+1} = w_t - \gamma \frac{m_{t+1}}{\|m_{t+1}\|}$ where $\{m_t\}$ is an arbitrary sequence of points. Define $\delta_t := m_{t+1} - \nabla F(w_t)$ be the estimation error. If $\gamma = O(1/K_1)$, then*

$$F(w_t) - F(w_{t+1}) \geq \left(\gamma - \frac{1}{2}K_1\gamma^2\right) \|\nabla F(w_t)\| - \frac{1}{2}K_0\gamma^2 - 2\gamma\|\delta_t\|$$

which is $\gamma\|\nabla F(w_t)\| - 2\gamma\|\delta_t\| + o(\gamma)$ for small γ. Therefore to decrease the objective function F it suffices to control the term $\|\delta_t\|$. However, δ_t is related to the stochastic gradient noise which is very large due to Lemma 3.3. This motivates us to the use the momentum technique for the choice of $\{m_t\}$ to reduce the noise. Formally, let β be the momentum factor and define $\delta'_t = \hat{\nabla} F(w_t) - \nabla F(w_t)$, then using the recursive equation of momentum m_t we can show that

$$\delta_t = \beta \sum_{\tau=0}^{t-1} \beta^\tau (\nabla F(w_{t-\tau-1}) - \nabla F(w_{t-\tau})) + (1-\beta) \sum_{\tau=0}^{t-1} \beta^\tau \delta'_{t-\tau} + (1-\beta)\beta^t \delta'_0.$$

The first term can be bounded using Lemma 3.4, and the second term can be bounded using the independence of noises $\{\delta'_t\}$ and Cauchy-Schwartz inequality. Finally, the use of mini-batches of size $\mathcal{O}(\Gamma^2)$ ensures that $\sum_{t=0}^{T-1} \|\delta_t\| < \frac{1}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w_t)\|$. By combining all the inequalities we obtain the desired result.

3.4 Deal with the CVaR case

Previous analysis applies to any divergence function ψ as long as ψ^* is smooth. It encompasses some popular choices such as the χ^2-divergence, but not the CVaR. In the case of CVaR, ψ^* is not differentiable as shown in Table 1, which is undesirable from an optimization viewpoint. In this section we introduce a smoothed version of CVaR. The conjugate function of the smoothed CVaR is also smooth, so that the results in Section 3.2 can be directly applied in this setting.

For standard CVaR at level α, $\psi_\alpha(t)$ takes zero when $t \in [0, 1/\alpha)$ and takes infinity otherwise. Instead, we consider the following smoothed version of CVaR:

$$\psi_\alpha^{\text{smo}}(t) = \begin{cases} t \log t + \frac{1-\alpha t}{\alpha} \log \frac{1-\alpha t}{1-\alpha} & t \in [0, 1/\alpha) \\ +\infty & \text{otherwise} \end{cases} \tag{10}$$

It is easy to see that ψ_α^{smo} is a valid divergence. The corresponding conjugate function is

$$\psi_\alpha^{\text{smo},*}(t) = \frac{1}{\alpha} \log(1 - \alpha + \alpha \exp(t)). \tag{11}$$

The following propositions demonstrate that ψ_α^{smo} is indeed a smoothed approximation of CVaR.

Proposition 3.9 *Fix $0 < \alpha < 1$. When $\lambda \to 0$, the solution of the DRO problem* (5) *for smoothed CVaR tends to the solution for the standard CVaR.*

Proposition 3.10 *$\psi_\alpha^{smo,*}(t)$ is $\frac{1}{\alpha}$-Lipschitz and $\frac{1}{4\alpha}$-smooth.*

Note that $\psi_\alpha^{\text{smo},*}(t)$ is not only smooth but also Lipschitz. In the special case when ψ^* is Lipschitz, we can in fact obtain a better complexity in terms of the key problem-dependent parameters than the general results provided in Corollary 3.6. In particular, the gradient noise and smoothness of the objective function $\mathcal{L}(x, \eta, \xi)$ can be bounded, as shown in the following lemma:

Lemma 3.11 *For smoothed CVaR, the DRO objective* (5) *satisfies*

$$\mathbb{E}\left\|\nabla \mathcal{L}(x, \eta, \xi)\right\|^2 \leq \alpha^{-2}\left(1 + G^2\right). \tag{12}$$

Moreover, $\mathcal{L}(x, \eta)$ is K-smooth with $K = \frac{1}{\alpha}L + \frac{1}{2\lambda\alpha}\left(G^2 + 1\right)$.

Equipped with the above lemma, we can obtain the following convergence guarantee for smoothed CVaR:

Theorem 3.12 *Suppose that $\psi = \psi_\alpha$. If we run SGD with properly selected parameters, then the complexity of finding ϵ-stationary point of $\mathcal{L}$ is $\mathcal{O}\left(\alpha^{-3}\lambda^{-1}\Delta G^2(G^2 + \lambda L)\epsilon^{-4}\right)$, where $\Delta = \mathcal{L}(x_0, \eta_0) - \mathcal{L}(x^*, \eta^*)$.*

4 Experiments

We perform two sets of experiments to verify our theoretical results. In the first set of experiments, we consider the setting in Section 3.2, where the loss $\ell(x; \xi)$ is highly non-convex and unbounded, and ψ is chosen to be the commonly used χ^2-divergence such that its conjugate is smooth. We will show that 1) the vanilla SGD algorithm cannot optimize this loss efficiently due to the non-smoothness of the DRO objective; 2) by simply adopting the normalized momentum algorithm, the optimization process can be greatly accelerated. In the second set of experiments, we deal with the CVaR setting in Section 3.4. We will show that by employing the smooth approximation of CVaR defined in (10) and (11), the optimization can be greatly accelerated.

4.1 Experimental settings

Tasks. We consider two tasks: the classification task and the regression task. While classification is more common in machine learning, here we may also highlight the regression task, since recent studies show that DRO may be more suitable for non-classification problems in which the metric of interest is continuous as opposed to the 0-1 loss [Hu et al., 2018, Levy et al., 2020].

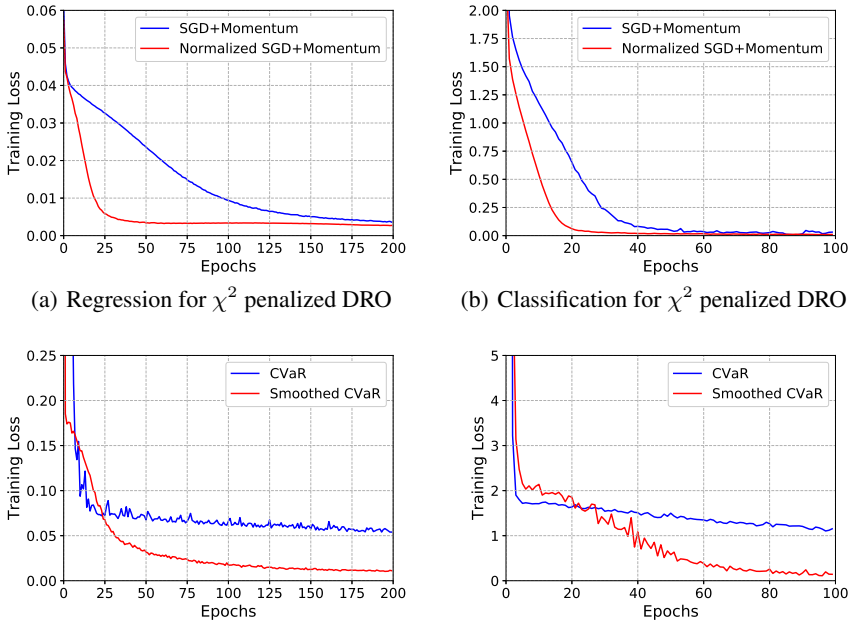
Datasets. We choose the AFAD-Full dataset for regression and CIFAR-10 dataset for classification. AFAD-Full [Niu et al., 2016] is a regression task to predict the age of human from the facial information, which contains more than 160K facial images and the corresponding age labels ranging from 15 to 75. Note that AFAD-Full is an imbalanced dataset where the ages of two thirds of the whole dataset are between 18 and 30. Following the experimental setting in [Chang et al., 2011, Chen et al., 2013, Niu et al., 2016], we randomly split the whole dataset into a training set comprised of 80% data and a test set comprised of the remaining 20% data. CIFAR-10 dataset is a classification task consisting of 10 classes with 5000 images for each class. To demonstrate the effectiveness of our method in DRO setting, we adopt the setting in Chou et al. [2020] to construct an imbalanced CIFAR-10 by randomly sampling each category at different ratio. See Appendix for more details.

Model. For all experiments in this paper, we use the standard ResNet-18 model in [He et al., 2016]. The output has 10 logits for CIFAR-10 classification task, and has a single logit for regression.

Training details. We choose the penalty coefficient $\lambda = 0.1$ and the CVaR coefficient $\alpha = 0.02$ in all experiments. We tune the learning rate hyper-parameter from a grid search and pick the one that achieves the fastest optimization speed. The momentum factor in all experiments is taken to 0.9. We train the model for 100 epochs on CIFAR-10 dataset and 200 epochs on AFAD-Full dataset. Other training details can be found in Appendix.

4.2 Experimental results

Experimental result for χ^2 penalized DRO. Figure 1(a) and 1(b) plot the training curve of the distributionally robust loss using different algorithms. It can be clearly seen that in both regression and classification, the vanilla SGD converges slowly, and using normalized momentum algorithm



(a) Regression for χ^2 penalized DRO (b) Classification for χ^2 penalized DRO

(c) Regression for smoothed CVaR (d) Classification for smoothed CVaR

Figure 1: Training curve of χ^2 penalized DRO and smoothed CVaR in regression and classification task.

significantly improves the convergence speed. For example, in regression task SGD does not converge after 200 epochs while normalized momentum algorithm converges just after 40 epochs. These results highly consist with our theoretical findings, which shows that due to the non-smoothness of the DRO loss, vanilla SGD may not be able to optimize the loss well; In contrast, normalized momentum utilizes the relationship between local smoothness and gradient magnitude, and achieves better performance.

Experimental result for smoothed CVaR. Figure 1(c) and 1(d) plot the training curve for different training losses: CVaR and smoothed CVaR. Note that the evaluation metrics (y-axis) in these figures are both chosen to be CVaR so that we can make a fair comparison of optimization speed between these training curves. Firstly, it can be seen that the optimization of CVaR is very hard due to the non-smoothness, and the loss can easily be stuck after a few epochs (for the first few epochs the decrease of loss may mainly be due to the optimization of η). In contrast, the optimization of smoothed CVaR is much easier for both tasks, and the final loss is significantly lower. Such experimental results show the benefit of our proposed smoothed CVaR for optimization.

5 Discussions

Conclusion. In this paper we provide non-asymptotic analysis of first-order algorithms for the DRO problem with unbounded and non-convex loss. Specifically, we write the original DRO problem as a non-smooth non-convex optimization problem, and we propose an efficient normalization-based algorithm to solve it. The general result of Theorem 3.5 might be of independent value and is not limited to DRO setting. We hope that this work can also bring inspiration to the study of other non-smooth non-convex optimization problems.

Limitations. Despite the theoretical grounds and the promising experimental justifications, there are some interesting questions that remain unexplored. Firstly, it may be possible to obtain complexities that is better on problem-dependent parameters. Secondly, while this paper mainly considers the smooth ψ^*, in some cases ψ^* may be non-smooth (e.g. for KL-divergence) or even not continuous. In future we hope to discover approaches that can deal with more general classes of ψ-divergence. Finally, we are also looking forward to seeing more applications of DRO in real-world problems.

Potential negative societal impacts. Our paper presents a theoretical work without any foreseeable impact in the society.

References

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2011.

Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *European Conference on Computer Vision*, 2020.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International Conference on Machine Learning*, 2020.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 2019.

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

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Mert Gürbüzbalaban, Andrzej Ruszczyński, and Landi Zhu. A stochastic subgradient method for distributionally robust non-convex learning. *arXiv preprint arXiv:2006.04873*, 2020.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, 2018.

Dionysios S Kalogerias. Noisy linear convergence of stochastic gradient descent for cv@r statistical learning under polyak-ojasiewicz conditions. *arXiv preprint arXiv:2012.07785*, 2020.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Conference on Neural Information Processing Systems*, 2020.

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*, 2020.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Conference on Neural Information Processing Systems*, 2016.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Andrzej Ruszczynski. A stochastic subgradient method for nonsmooth nonconvex multi-level composition optimization. *arXiv preprint arXiv:2001.10669*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

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

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Conference on Neural Information Processing Systems*, 2020a.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why adam beats sgd for attention models. In *International Conference on Learning Representations*, 2019.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020b.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, 2020c.

Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021.

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? [Yes]
   (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]
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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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]
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]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

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

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