Non-convex Distributionally Robust Optimization: Non-asymptotic Analysis

Bohang Zhang

March 8, 2022

Introduction

- Classical machine learning setting: both the training set and the test set are drawn from the *same* distribution *P*.
- This setting may be problematic in many situations:
 - in domain adaptation tasks;
 - when there is severe class imbalance in the training set;
 - when fairness in minority groups is an important consideration;
 - when the model is exposed to adversarial attacks.



Introduction

Distributionally robust optimization (DRO) setting: the trained model still has good performance under distribution shift.

 DRO minimizes the worst-case loss over a set of probability distributions Q around P:

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

- $x \in \mathcal{X}$: the parameter to be optimized;
- ▶ $\mathbb{E}_{\xi \sim Q}[\ell(x;\xi)]$: the expected loss over distribution Q;
- ▶ $\mathcal{U}(P)$: the uncertainty set typically defined as $\mathcal{U}(P) := \{Q : d(Q, P) \leq \epsilon\}$ and d is a distance measure between Q and P.



Introduction

• The soft-penalized DRO problem with regularization $\lambda > 0$:

$$\operatorname{minimize}_{x \in \mathcal{X}} \quad \Psi(x) := \sup_{Q} \left\{ \mathbb{E}_{\xi \sim Q} \left[\ell(x; \xi) \right] - \lambda d(Q, P) \right\} \tag{2}$$

▶ A commonly used distance function is the ϕ -divergence:

$$d_{\psi}(Q, P) := \int \psi\left(\frac{\mathrm{d}\,Q}{\mathrm{d}\,P}\right) \mathrm{d}P. \tag{3}$$

This paper studies efficient first-order optimization algorithms for DRO problem (2) and provides the first *non-asymptotic* analysis for *non-convex* losses $\ell(x,\xi)$ and *general* ϕ -divergence.



Contributions

Bohang Zhang

- We tackle two main challenges in non-convex DRO:
 - The DRO objective (2) is non-convex and can become arbitrarily non-smooth, causing standard techniques in smooth optimization fail to provide a convergence guarantee;
 - 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 $\ell(x,\xi)$ has bounded variance.
- We propose a novel algorithm called *mini-batch normalized SGD with momentum* and prove an $\mathcal{O}(\epsilon^{-4})$ gradient complexity.
 - Our analysis clearly demonstrates the effectiveness of gradient normalization and momentum techniques in optimizing ill-conditioned objective functions.
 - More importantly, the algorithm and analysis are not limited to DRO setting, and are described in the context of a general class of optimization problems.
 - Our result can shed light on why some popular optimizers, in particular Adam, often exhibit superior performance in real applications.

Non-convex DRO

5/20

March 8, 2022

Definitions

- (Lipschitz continuity) A mapping $f: \mathcal{X} \to \mathbb{R}^m$ is G-Lipschitz continuous if for any $x, y \in \mathcal{X}$, $||f(x) f(y)|| \le G||x y||$.
- (Smoothness) A function $f\colon \mathcal{X} \to \mathbb{R}$ is L-smooth if it is differentiable on \mathcal{X} and the gradient ∇f is L-Lipschitz continuous, i.e. $\|\nabla f(x) \nabla f(y)\| \leq L \, \|x-y\|$ for all $x,y \in \mathcal{X}$. We say f is non-smooth if such L does not exist.
- (Conjugate function) For a function $\psi: \mathbb{R} \to \mathbb{R}$, the conjugate function ψ^* is defined as $\psi^*(t) := \sup_{s \in \mathbb{R}} (st \psi(s))$.
- (ϵ -stationary point) For a differentiable function $f \colon \mathcal{X} \to \mathbb{R}$, a point $x \in \mathcal{X}$ is said to be first-order ϵ -stationary if $\|\nabla f(x)\| \le \epsilon$.



Assumptions

- Given ξ , the loss function $\ell(x,\xi)$ is G-Lipschitz and L-smooth w.r.t. 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 ψ^* is M-smooth.
- The stochastic loss with distribution P has bounded variance, i.e. $\mathbb{E}_{\xi \sim P} \left(\ell(x,\xi) \ell(x) \right)^2 \leq \sigma^2$ where $\ell(x) = \mathbb{E}_{\xi \sim P} \ell(x,\xi)$.

The aim of this paper is to find an ϵ -stationary point of problem (2).



7/20

Equivalent formulation of the DRO objective

- The original formulation (2) involves a sup over distribution Q which is hard to solve.
- By duality arguments the DRO objective (2) can be equivalently written as

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

• This corresponds to jointly minimizing

$$\mathcal{L}(x,\eta) := \mathbb{E}_{\xi \sim P} \left[\lambda \psi^* \left(\frac{\ell(x,\xi) - \eta}{\lambda} \right) + \eta \right] \text{ over } (x,\eta) \in \mathcal{X} \times \mathbb{R} \subset \mathbb{R}^{n+1}.$$

• This can be seen as a standard stochastic optimization problem.



Bohang Zhang

Equivalent formulation of the DRO objective

We present a theorem that gives connection of the gradient of $\Psi(x)$ to the gradient of $\mathcal{L}(x,\eta)$.

Lemma 1.

 $\Psi(x)$ is differentiable, and $\nabla \Psi(x) = \nabla_x \mathcal{L}(x, \eta)$ for any $\eta \in \arg \min_{n'} \mathcal{L}(x, \eta')$.

Theorem 2.

Define a rescaled function

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

then $\|\nabla \widehat{\mathcal{L}}(x,\eta)\| \le \epsilon/\sqrt{2}$ implies that x is an ϵ -stationary point of $\Psi(x)$.

It suffices to find an ϵ -stationary point of $\widehat{\mathcal{L}}(x,\eta)$.



Challenges in Non-convex DRO

- A standard result in optimization: for smooth objective function with bounded variance of the stochastic gradient, SGD can provably find an ϵ -stationary point under $\mathcal{O}(\epsilon^{-4})$ gradient complexity.
- 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 $\nabla_x \ell(x,\cdot)$.



A Motivating Example

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 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 $\operatorname{Var}_{\xi}[\ell(x,\xi)] \leq 4$;

• (Bounded variance for gradient) Similarly we can check that the gradient of ℓ also has bounded variance. Moreover, the variance tends to zero when $x \to \infty$.



Bohang Zhang Non-convex DRO

A Motivating Example

Now consider the DRO where ψ is chosen as the commonly used χ^2 -divergence. Fix $\lambda=1$ and $\eta=0$. Based on the expression of

$$\psi^*(t) = -1 + \frac{1}{4}(t+2)_+^2$$

the DRO objective function (5) thus takes the form

$$\widehat{\mathcal{L}}(x,0;\xi) = \frac{1}{4} \left[x^2 \left(1 + \frac{\xi}{x^2 + 1} \right)^2 + 2 \right]^2 - 1$$

, which is a quartic-like function. It follows that

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



Properties of Non-convex DRO

We prove that both the gradient variance and the local smoothness can be controlled in terms of the gradient norm.

Lemma 3.

The gradient estimators of (5) satisfies the following property:

$$\mathbb{E}_{\xi} \|\nabla \widehat{\mathcal{L}}(x, \eta, \xi) - \nabla \widehat{\mathcal{L}}(x, \eta)\|^{2} \le 11 G^{2} M^{2} \lambda^{-2} \sigma^{2} + 8(G^{2} + \|\nabla \widehat{\mathcal{L}}(x, \eta)\|^{2})$$
 (6)

Lemma 4.

For any pair of parameters (x,η) and (x',η') , we have

$$\|\nabla \widehat{\mathcal{L}}(x,\eta) - \nabla \widehat{\mathcal{L}}(x',\eta')\| \le \left(K + \frac{L}{G} \|\nabla \widehat{\mathcal{L}}(x,\eta)\|\right) \|(x - x',\eta - \eta')\| \tag{7}$$

13 / 20

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

Note that (7) reduces to the standard notion of smoothness if the term $\frac{L}{G}\|\nabla\widehat{\mathcal{L}}(x,\eta)\| \text{ is absent. Thus the inequality (7) can be seen as a generalized smoothness condition.}$

Algorithm

Algorithm 1: Mini-batch Normalized SGD with Momentum

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

1 for $t \leftarrow 1$ to T do

$$\begin{array}{c|c} \mathbf{2} & & \hat{\nabla} F(w_{t-1}) \leftarrow \frac{1}{S} \sum_{i=1}^{S} \nabla F(w_{t-1}, \xi_{t-1}^{(i)}) \text{ where } \{\xi_{t-1}^{(i)}\}_{i=1}^{S} \text{ are i.i.d. samples} \\ & & \text{drawn from } P \\ \mathbf{3} & & & m_{t} \leftarrow \beta m_{t-1} + (1-\beta) \hat{\nabla} F(w_{t-1}) \\ \mathbf{4} & & & w_{t} \leftarrow w_{t-1} - \gamma \frac{m_{t}}{\|m_{t}\|} \end{array}$$

←□▶←□▶←□▶←□▶ □ ♥Q@

Results

Theorem 5. (Main Theorem: A general Convergence Guarantee)

Suppose that F satisfies the following conditions:

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

Let $\{w_t\}$ be the sequence produced by Algorithm 1. Then with a mini-batch size $S=\Theta(\Gamma^2)$ and a suitable choice of parameters γ and β , for any small $\epsilon=\mathcal{O}(\min(K_0/K_1,\Lambda/\Gamma))$, we need at most $\mathcal{O}\left(\Delta K_0\Lambda^2\epsilon^{-4}\right)$ gradient complexity to guarantee that we find an ϵ -stationary point in expectation, i.e.

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w_t)\| \leq \epsilon \text{ where } \Delta = F(w_0) - \inf_{w \in \mathbb{R}^d} F(w).$$



Bohang Zhang

Non-convex DRO

Results

Corollary 6.

For sufficiently small ϵ , the gradient complexity for finding an ϵ -stationary point of $\Psi(x)$ is

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

- Algorithm 1 finds an ϵ -stationary point with complexity $\mathcal{O}(\epsilon^{-4})$.
- The bound in Theorem 5 does not depend on K_1 and Γ as long as ϵ is sufficiently small. In other words, Algorithm 1 is well-adapted to the non-smoothness and unbounded noise setting.
- Although the batch size is chosen propositional to Γ^2 , the required number of iterations T is inversely propositional to Γ^2 , therefore the total number of stochastic gradient computations remains the same.
- The general result (Theorem 5) is not limited to DRO, and can shed light on optimizers such as Adam often show superior performance in real applications.

 Bohang Zhang
 Non-convex DRO
 March 8, 2022
 16 / 20

4 日 ト 4 周 ト 4 ヨ ト 4

Proof Sketch

Lemma 7. (Descent inequality)

Let F(x) be a function satisfying the generalized smoothness condition in Theorem 5. Then for any point x and direction z the following holds:

$$F(x-z) \le F(x) - \langle \nabla F(x), z \rangle + \frac{1}{2} (K_0 + K_1 \|\nabla F(x)\|) \|z\|^2.$$
 (8)

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.



17/20

Proof Sketch

Lemma 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

$$F(w_t) - F(w_{t+1}) \ge \left(\gamma - \frac{1}{2}K_1\gamma^2\right) \|\nabla F(w_t)\| - \frac{1}{2}K_0\gamma^2 - 2\gamma\|\delta_t\|$$
 (9)

- This leads to $\gamma \|\nabla F(w_t)\| 2\gamma \|\delta_t\| \mathcal{O}(\gamma^2)$ for small γ .
- Therefore the objective function F(w) decreases if $\|\delta_t\| < 1/2 \cdot \|\nabla F(w_t)\|$, i.e. a small estimation error.
- However, δ_t is related to the stochastic gradient noise which can be very large due to unbounded variance property. This motivates us to the use the momentum technique for the choice of $\{m_t\}$ to reduce the noise.



Bohang Zhang Non-convex DRO

Proof Sketch

Formally, let β be the momentum factor and define $\hat{\delta}_t = \hat{\nabla} F(w_t) - \nabla F(w_t)$, then using the recursive equation of momentum m_t we can show that

$$\delta_t = \beta \sum_{\tau=0}^{t-1} \beta^{\tau} (\nabla F(w_{t-\tau-1}) - \nabla F(w_{t-\tau})) + (1-\beta) \sum_{\tau=0}^{t-1} \beta^{\tau} \hat{\delta}_{t-\tau} + (1-\beta) \beta^{t} \hat{\delta}_{0}. \tag{10}$$
 The first term of the right hand side in (10) can be bounded using the generalized

The first term of the right hand side in (10) can be bounded using the generalized smoothness condition, and the core procedure is to bound the second term using a careful analysis of conditional expectation and the independence of noises $\{\hat{\delta}_t\}$. Finally, the use of mini-batches of size $\Theta(\Gamma^2)$, a carefully chosen β and a small enough γ ensure that T=1

$$\sum_{t=0}^{T-1} \|\delta_t\| < c \sum_{t=0}^{T-1} (\mathbb{E} \|\nabla F(w_t)\| + \mathcal{O}(\epsilon)) \text{ where } c < 1/2. \text{ This guarantees that the right }$$

hand side of (9) is overall positive, and by taking summation over $\it t$ in (9) we have that

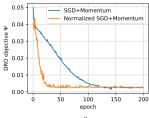
$$F(w_0) - F(w_T) \ge (1 - 2c)\gamma \sum_{t=0}^{T-1} \|\nabla F(w_t)\| - \mathcal{O}(\gamma^2 T - \gamma T\epsilon).$$

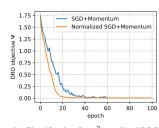
namely,
$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(w_t)\| \leq \mathcal{O}\left(\frac{\Delta}{\gamma T} + \gamma + \epsilon\right).$$

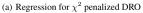
Finally, for a suitable choice of γ we can obtain the minimum gradient complexity bound on T.

Bohang Zhang Non-convex DRO March 8, 2022 19 / 20

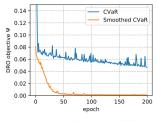
Experiments

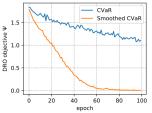












(c) Regression for smoothed CVaR

(d) Classification for smoothed CVaR



Thank You!

