Provable Non-Convex Min-Max Optimization

Mingrui Liu[†], **Hassan Rafique**[‡], **Qihang Lin**[♮], **Tianbao Yang**[†]
[†]Department of Computer Science, The University of Iowa, Iowa City, IA, 52242
[‡]Department of Mathematics, The University of Iowa, Iowa City, IA, 52242
[♮]Department of Management Sciences, The University of Iowa, Iowa City, IA, 52242
mingrui-liu, hassan-rafique, qihang-lin, tianbao-yang@uiowa.edu

Abstract

In this paper, we propose an efficient stochastic subgradient method for solving a broad class of non-convex min-max problems and establish its iteration complexities for different convergence measures depending on whether the problem is concave in terms of the variable of maximization. When the objective is weakly convex in terms of min variable and concave in terms of the max variable, we prove that the proposed algorithm converges to a nearly ϵ-stationary solution of the equivalent minimization problem with a complexity of $O(1/\epsilon^6)$. When the objective is weakly convex in terms of the min variable and weakly concave in terms of the max variable, we prove the algorithm converges a nearly ϵ-stationary solution of the min-max problem with the same complexity of $O(1/\epsilon^6)$. To the best of our knowledge, these are the first non-asymptotic convergence results of stochastic optimization for solving non-convex min-max problems.

1 Introduction

The main goal of this paper is to design provably efficient algorithms for solving saddle-point (aka min-max) problems of the following form that exhibits non-convexity:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}), \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets, and $f(\mathbf{x}, \mathbf{y})$ is non-convex in terms of $\mathbf{x}$ and could be non-concave in terms of $\mathbf{y}$. This problem has broad applications in machine learning, e.g., generative adversarial networks [11, 1], distributionally robust optimization [16, 15], reinforcement learning [2], and adversarial learning [23]. For more details about these applications and their formulations, please refer to the long papers of this extended abstract [18, 14].

Although many previous studies have considered the min-max formulation and proposed efficient algorithms, most of them focus on the *convex-concave* family, in which $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ given $\mathbf{y}$ and is concave in $\mathbf{y}$ given $\mathbf{x}$. However, convex-concave formulations cannot cover some important new methods/technologies/paradigms arising in machine learning. Hence, it becomes an emergent task to design provably efficient algorithms for solving (1) that exhibits non-convexity structure.

Solving non-convex min-max problems is more challenging than solving non-convex minimization problems. Although there is increasing interest on non-convex optimization problems, most of the existing algorithms are designed for the minimization problem without considering the max-structure of the problem when it appears, and therefore are not directly applicable to (1). For example, stochastic gradient descent (SGD) for a minimization problem assumes that a stochastic gradient is available at each iteration for the objective of the minimization problem, which might be impossible for (1) if the maximization over $\mathbf{y}$ is non-trivial or if f contains expectations. When designing an optimization algorithm for (1), the important questions are whether the algorithm has a polynomial runtime and what quality it guarantees in the output. In the recent studies for non-convex minimization

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada.

problems [4, 5, 3, 6, 8, 7, 9, 10, 13, 17, 19, 20], polynomial-time algorithms have been developed for finding a nearly stationary point that is close to a point where the subdifferential of objective function almost contains zero. Following this stream of work, we would naturally ask a question *whether it is possible to design a polynomial time algorithm for (1) that finds a nearly first-order stationary point of the problem.* We provide affirmative answers in this extended abstract. In particular, we propose a stagewise primal-dual stochastic subgradient method that is motivated by the inexact proximal point method. At each stage the standard primal-dual stochastic subgradient method is employed to solve a constructed convex-concave min-max problem. We prove the iteration complexities of the proposed algorithm for two classes of problems for finding a nearly stationary solution.

- When $f(\mathbf{x}, \mathbf{y})$ is weakly convex in terms of $\mathbf{x}$ and concave in terms of $\mathbf{y}$, we prove the proposed algorithm can find a nearly ϵ-stationary solution for the equivalent minimization problem with an iteration complexity of $O(1/\epsilon^6)$.
- When $f(\mathbf{x}, \mathbf{y})$ is weakly convex in terms of $\mathbf{x}$ and weakly-concave in terms of $\mathbf{y}$, we prove the proposed algorithm can find a nearly ϵ-stationary solution for the min-max problem with an iteration complexity of $O(1/\epsilon^6)$.

2 Preliminaries

We use $\|\cdot\|$ to denote the Euclidean norm. Given a function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we define the *(Fréchet) subdifferential* of h as $\partial h(\mathbf{x}) = \{\boldsymbol{\zeta} \in \mathbb{R}^d | h(\mathbf{x}') \geq h(\mathbf{x}) + \boldsymbol{\zeta}^\top(\mathbf{x}' - \mathbf{x}) + o(\|\mathbf{x}' - \mathbf{x}\|), \quad \mathbf{x}' \to \mathbf{x}\}$, where each element in $\partial h(\mathbf{x})$ is called a (Fréchet) subgradient of h at $\mathbf{x}$. We define $\partial_x f(\mathbf{x}, \mathbf{y})$ as the subgradient of $f(\mathbf{x}, \mathbf{y})$ with respective to $\mathbf{x}$ for a fixed $\mathbf{y}$ and $\partial_y[-f(\mathbf{x}, \mathbf{y})]$ as the subgradient of $-f(\mathbf{x}, \mathbf{y})$ with respective to $\mathbf{y}$ for a fixed $\mathbf{x}$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $1_{\mathcal{Z}}(\mathbf{z})$ denote the indicator function, $\pi_{\mathcal{X}}[\mathbf{x}]$ denote the projection onto $\mathcal{X}$.

A function $h : \mathcal{X} \to \mathbb{R}$ is *ρ-weakly convex* if $h(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x}\|^2$ is convex. Similarly, a function $h : \mathcal{Y} \to \mathbb{R}$ is *ρ-weakly concave* if $h(\mathbf{y}) - \frac{\rho}{2}\|\mathbf{y}\|^2$ is concave. Define $\psi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. $\text{Dist}(\mathbf{x}, S)$ denotes the Euclidian distance from a point $\mathbf{x}$ to a set S. The assumptions made for the optimization problem (1) are the following:

Assumption 1. *(1) $\mathcal{X}$ and $\mathcal{Y}$ are compact sets, and $f(\mathbf{x}, \mathbf{y})$ is finite on $\mathcal{X} \times \mathcal{Y}$; (2) $f(\mathbf{x}, \mathbf{y})$ is ρ-weakly convex in $\mathbf{x}$ for any $\mathbf{y} \in \mathcal{Y}$; (3) $f(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y}$ or is ρ-weakly concave in $\mathbf{y}$ for any $\mathbf{x} \in \mathcal{X}$; (7) For any $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and any realization of $\boldsymbol{\xi}$, we can compute $(\mathbf{g}_x, -\mathbf{g}_y) \in \partial_x F(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi}) \times \partial_y[-F(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi})]$ such that $(\mathbb{E}\mathbf{g}_x, -\mathbb{E}\mathbf{g}_y) \in \partial_x f(\mathbf{x}, \mathbf{y}) \times \partial_y[-f(\mathbf{x}, \mathbf{y})]$; (8) $\mathbb{E}\|\mathbf{g}_x^{(j)}\|_2^2 \leq M_x^2$ and $\mathbb{E}\|\mathbf{g}_y^{(j)}\|_{y,*}^2 \leq M_y^2$ for any $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ for some $M_x > 0$ and $M_y > 0$.*

It should be noted that we do not assume the smoothness of $f(\mathbf{x}, \mathbf{y})$ in terms of $\mathbf{x}$ and $\mathbf{y}$. However, a smooth function $f(\mathbf{x}, \mathbf{y})$ is a weakly convex function in terms of $\mathbf{x}$ and weakly concave in terms of $\mathbf{y}$. We differentiate the case that $f(\mathbf{x}, \mathbf{y})$ is concave in terms of $\mathbf{y}$ from the case that $f(\mathbf{x}, \mathbf{y})$ is weakly-concave in terms of $\mathbf{y}$. This is because that in the former case, we can prove a stronger convergence. In particular, when $f(\mathbf{x}, \mathbf{y})$ is concave in terms of $\mathbf{y}$, we establish convergence to a nearly stationary point for the equivalent minimization problem $\min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x})$; and when $f(\mathbf{x}, \mathbf{y})$ is weakly-concave in terms of $\mathbf{y}$, we establish convergence to a nearly stationary point for the min-max saddle-point problem $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. To this end, we first introduce nearly stationarity for the minimization problem and the min-max problem.

Under Assumption 1 with $f(\mathbf{x}, \mathbf{y})$ being concave in terms of $\mathbf{y}$, $\psi(\cdot)$ is ρ-weakly convex (and thus non-convex) so that finding the global optimal solution in general is difficult. An alternative goal is to find a *stationary* point of (1) which is defined as a point $\mathbf{x}_* \in \mathcal{X}$ such that $\mathbf{0} \in \partial\psi(\mathbf{x}_*)$. Because of the iterative nature of optimization algorithms, such a stationary point generally can only be approached in the limit as the number of iterations increases to infinity. With finitely many iterations, a more realistic goal is to find an *ϵ-stationary* point, i.e., a point $\hat{\mathbf{x}} \in \mathcal{X}$ satisfying $\text{Dist}(\mathbf{0}, \partial\psi(\hat{\mathbf{x}})) := \min_{\boldsymbol{\zeta} \in \partial\psi(\hat{\mathbf{x}})} \|\boldsymbol{\zeta}\|_2 \leq \epsilon$. However, when the objective function is non-smooth, computing an ϵ-stationary point is still not an easy task even for convex optimization problem. A simple example is $\min_{x \in \mathbb{R}} |x|$ where the only stationary point is 0 but any $x \neq 0$ is not an ϵ-stationary point ($\epsilon < 1$) no matter how close it is to 0. This situation is likely to occur in problem (1) because of the potential non-smoothness of f and the presence of domain constraints. Therefore, following [5, 6, 3, 24], we consider the *Moreau envelope* of ψ in (1), which is

$$\psi_\gamma(\mathbf{x}) := \min_{\mathbf{z} \in \mathbb{R}^p} \left\{ \psi(\mathbf{z}) + \frac{1}{2\gamma}\|\mathbf{z} - \mathbf{x}\|_2^2 \right\} \tag{2}$$

Algorithm 1 A Stagewise Primal-Dual Stochastic Subgradient Method

Input: step size η_k, integers T_k and non-decreasing weights θ_k, $\mathbf{z}_0 \in \mathcal{Z}$, $0 < \gamma < \rho^{-1}$
for $k = 0, \ldots, K-1$ **do**
  Let $F_k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}_k\|^2$ if $f(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y}$, otherwise let
  $F_k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{y}_k\|^2$
  $\mathbf{z}_{k+1} := (\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \text{PDSG}(F_k, \mathbf{z}_k, \eta_k, T_k)$
end for
Sample τ randomly from $\{0, 1, \ldots, K-1\}$ with $\text{Prob}(\tau = k) = \frac{\theta_k}{\sum_{k=0}^{K-1} \theta_k}$.
Output: $\mathbf{z}_\tau$.

Algorithm 2 PDSG: Primal-Dual Stochastic Subgradient Method: PDSG($F, \mathbf{z}_0, \eta, T$)

for $t = 0, ..., T-1$ **do**
  Sample $\boldsymbol{\xi}^t$, and compute stochastic subgradients $(\mathbf{g}_x^t, -\mathbf{g}_y^t)$ of $F(\mathbf{x}, \mathbf{y})$ at $\mathbf{z}_t$
  Compute

$$\mathbf{x}_{t+1} = \pi_{\mathcal{X}}[\mathbf{x}_t - \eta \mathbf{g}_x^t], \quad \mathbf{y}_{t+1} = \pi_{\mathcal{Y}}[\mathbf{y}_t + \eta \mathbf{g}_y^t]$$

end for
Return $(\sum_{t=0}^{T-1} \mathbf{x}_t, \sum_{t=0}^{T-1} \mathbf{y}_t)/T$ if $f(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y}$ or $(\mathbf{x}_\tau, \mathbf{y}_\tau)$ if $f(\mathbf{x}, \mathbf{y})$ is weakly concave in $\mathbf{y}$ where τ is randomly sampled from $\{0, \ldots, T-1\}$

where $\gamma > 0$. The above problem is well-defined when $\psi(\cdot)$ is ρ-weakly convex and $\frac{1}{\gamma} > \rho$, whose optimal solution denoted by $\text{prox}_{\gamma\psi}(\mathbf{x})$ is unique. Moreover, $\psi_\gamma(\cdot)$ is a smooth function whose gradient is

$$\nabla\psi(\mathbf{x}) = \gamma^{-1}(\mathbf{x} - \text{prox}_{\gamma\psi}(\mathbf{x})). \tag{3}$$

The definition of the Moreau envelope directly implies that for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}_+ - \mathbf{x}\| = \gamma\|\nabla\psi_\gamma(\mathbf{x})\|$, $\psi(\mathbf{x}_+) \leq \psi(\mathbf{x})$, $\text{Dist}(\mathbf{0}, \partial\psi(\mathbf{x}_+)) = \|\nabla\psi_\gamma(\mathbf{x})\|$, where $\mathbf{x}_+ = \text{prox}_{\gamma\psi}(\mathbf{x})$ [5, 6, 3, 24], Hence, the norm of the gradient $\|\nabla\psi_\gamma(\mathbf{x})\|$ can be used as measure of the quality of a solution $\bar{\mathbf{x}}$. This lead us to the definition of nearly ϵ-stationary solution to the problem $\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x})$, i.e., $\mathbf{x}$ is nearly ϵ-stationary solution if $\|\mathbf{x} - \text{prox}_{\gamma\psi}(\mathbf{x})\| \leq O(\epsilon)$ such that $\psi_\gamma(\mathbf{x})\| \leq \epsilon$.

When $f(\mathbf{x}, \mathbf{y})$ is weakly concave in terms of $\mathbf{y}$, we consider nearly stationarity for the min-max saddle-point problem (1). A point $\mathbf{z} \in \mathcal{Z}$ is called first-order stationary point of the min-max saddle-point problem if

$$\mathbf{z} \in \mathcal{F}_* := \{(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} : 0 \in \partial_x(f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Z}}(\mathbf{x}, \mathbf{y})), 0 \in \partial_y(f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Z}}(\mathbf{x}, \mathbf{y}))\}.$$

An iterative algorithm can be expected to find an ϵ-stationary solution such that $\text{Dist}^2(0, \partial(f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Z}}(\mathbf{x}, \mathbf{y})) \leq \epsilon^2$. Similar to the previous argument, the non-smoothness nature of the problem makes it challenging to find an ϵ-stationary solution. To address this, we introduce the notion of nearly stationary point for a min-max saddle-point problem. A point $\mathbf{w} = (\mathbf{u}, \mathbf{v})^\top \in \mathcal{Z}$ is called a nearly ϵ-stationary solution to (1) if

$$\|\mathbf{w} - \mathbf{w}^+\| \leq O(\epsilon), \quad \text{Dist}^2(0, \partial(f(\mathbf{u}^+, \mathbf{v}^+) + 1_{\mathcal{Z}}(\mathbf{u}^+, \mathbf{v}^+))] \leq \epsilon^2.$$

where $\mathbf{w}^+ = (\mathbf{u}^+, \mathbf{v}^+)^\top$ is the optimal solution to the convex-concave saddle-point problem $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{u}\|^2 - \frac{\rho}{2}\|\mathbf{y} - \mathbf{v}\|^2$.

3 Stagewise Primal-Dual Stochastic Subgradient Method

The proposed stochastic algorithm is presented in Algorithm 1 and Algorithm 2. The main Algorithm 1 is motivated by the proximal point method, and is running with multiple stages. At each stage, depending on whether $f(\mathbf{x}, \mathbf{y})$ is concave in terms of $\mathbf{y}$ or not, a strongly convex term in terms of $\mathbf{x}$ and a strongly concave term in terms of $\mathbf{y}$ is added to the objective function. The proximal center $\mathbf{x}_k, \mathbf{y}_k$ in the added strongly convex terms are updated using the returned solution from the last stage. Then the newly formed objective function $F_k(\mathbf{x}, \mathbf{y})$ is convex in terms of $\mathbf{x}$ and concave in terms of $\mathbf{y}$. Thus, the standard primal-dual stochastic subgradient method can be employed to solve the constructed convex-concave problem. The step size η_k and the number of iterations T_k for each stage is dynamically changed. We present the convergence of Algorithm 1 for solving the min-max problem in the following theorems.

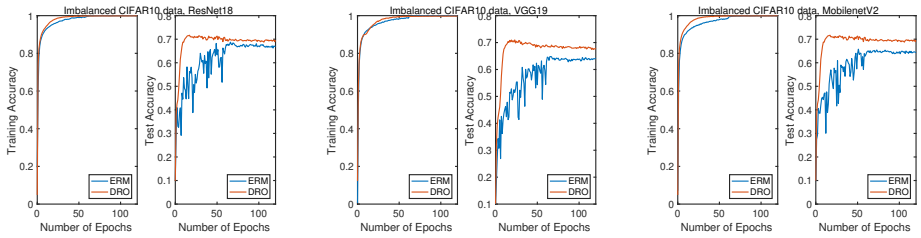

Figure 1: Comparison of ERM and DRO for ResNet18, VGG19, and MobileNetV2.

Theorem 1. *Suppose Assumption 1 holds with $f(\mathbf{x}, \mathbf{y})$ being concave in terms of $\mathbf{y}$. Let $\mathbf{x}_k^+ = prox_{\gamma\psi}(\mathbf{x}_k)$, where $\mathbf{x}_k$ is generated in Algorithm 1. . By running Algorithm 1 with $\gamma = 1/(2\rho)$, $\theta_k = (k+1)^\alpha$ with $\alpha > 1$, $\eta_k = c/(k+1), T_k = (k+1)^2$ with $c > 0$, and a total of stages $K = O(1/\epsilon^2)$, then we have*

$$\mathbb{E}[\|\mathbf{x}_\tau - \mathbf{x}_\tau{}^+\|^2] \leq O(\epsilon^2), \quad \mathbb{E}[Dist(0, \partial\psi(\mathbf{x}_\tau^+))^2] \leq \epsilon^2.$$

The total iteration complexity is $O(1/\epsilon^6)$.

Theorem 2. *Suppose Assumption 1 holds with $f(\mathbf{x}, \mathbf{y})$ being ρ-weakly-concave in terms of $\mathbf{y}$. Let $\mathbf{z}_k^+ = (\mathbf{x}_k^+, \mathbf{y}_k^+)^\top$ is the solution to $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{\rho}{2}\|\mathbf{y} - \mathbf{y}_k\|^2$, where $\mathbf{z}_k = (\mathbf{x}_k, \mathbf{y}_k)^\top$ is generated in Algorithm 1. By running Algorithm 1 with $\gamma = 1/(2\rho)$, $\theta_k = (k+1)^\alpha$ with $\alpha > 1$, $\eta_k = c/(k+1), T_k = (k+1)^2$ with $c > 0$, and a total of stages $K = O(1/\epsilon^2)$ we have*

$$\mathbb{E}[\|\mathbf{z}_\tau - \mathbf{z}_\tau^+\|^2] \leq O(\epsilon^2), \quad \mathbb{E}[Dist^2(0, \partial(f(\mathbf{x}_\tau^+, \mathbf{y}_\tau^+) + 1_{\mathcal{Z}}(\mathbf{x}_\tau^+, \mathbf{y}_\tau^+)))] \leq \epsilon^2.$$

The total iteration complexity is $O(1/\epsilon^6)$.

4 Numerical Experiments

For experiments, we consider the following distributionally robust optimization proposed by [16]:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \sum_{i=1}^n y_i f_i(\mathbf{x}) - r(\mathbf{y}), \tag{4}$$

where $f_i(\mathbf{x})$ denotes the loss of the model denoted by $\mathbf{x}$ on the i-th data point, $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set, $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^n \mid \sum_{i=1}^n y_i = 1,\ y_i \geq 0\ i = 1, ..., n\}$ is a simplex, and $r : \mathcal{Y} \to \mathbb{R}$ is a closed convex function. They showed that when $r(\mathbf{y})$ is the indicator function of the constraint set $\{\mathbf{y} : \sum_{i=1}^n (y_i - 1/n)^2 \leq \rho\}$ for some $\rho > 0$, the above min-max formulation achieves an effect that minimizes not only the bias but also the variance of the prediction, which could yield better generalization in some cases. In practice, one may also consider a regularized variant where $r(\mathbf{y}) = \lambda V(\mathbf{y}, \mathbf{1}/n)$ for some $\lambda > 0$, where $V(\cdot, \cdot)$ denotes some distance measure of two vectors (e.g., KL divergence, Euclidean distance). While optimization algorithms for solving convex-concave fromulation (4) were developed [15], it is still under-explored for problems with non-convex losses. When $f_i(\mathbf{x})$ is a non-convex loss function (e.g., the loss function associated with a deep neural network), (4) is non-convex in terms of $\mathbf{x}$ but is concave in $\mathbf{y}$.

We conduct a classification experiment to compare the standard empirical risk minimization (ERM) with distributionally robust optimization (DRO) (4). We use SGD to solve ERM and use Algorithm 1 to solve (4). We use imbalanced CIFAR10 data and three popular deep neural networks (ResNet18 [12], VGG19 [22], MobileNetV2 [21]) for the experiments. The original training data of CIFAR10 has 10 classes, each of which has 5000 images. We remove 4900 images for 5 classes to make the training data imbalanced. The test data remains intact. For ERM, we use SGD with stepsize 0.1 for epochs $1 \sim 60$, and 0.01 for epochs $61 \sim 120$. For our robust optimization, we use Algorithm 2 with $\eta_x = 0.1$ for epochs $10 \sim 60$ and $\eta_x = 0.001$ for epochs $61 \sim 120$, $\gamma = 0.2$, $r(\mathbf{y}) = \lambda \sum_{i=1}^n y_i \log y_i$ with $\lambda = 5$, $\eta_y = 10^{-5}$. We use 128 training examples as a minibatch for both methods. The training and testing curves are plotted in Figure 1, which show that robust optimization scheme is considerably better than ERM when dealing with imbalanced data.

5 Conclusion

In this paper, we have proposed a novel stagewise stochastic algorithm for solving a class of non-convex min-max optimization problems with polynomial time complexity for finding nearly stationary points. Experimental results verify its effectiveness for distributionally robust optimization in deep learning.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, International Convention Centre, Sydney, Australia, 2017.

[2] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1133–1142, 2018.

[3] Damek Davis and Dmitriy Drusvyatskiy. Complexity of finding near-stationary points of convex functions stochastically. *arXiv preprint arXiv:1802.08556*, 2018.

[4] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *arXiv preprint arXiv:1803.06523*, 2018.

[5] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[6] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *arXiv preprint arXiv:1707.03505*, 2017.

[7] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, Jul 2018.

[8] Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

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

[10] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[12] 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*, pages 770–778, 2016.

[13] Guanghui Lan and Yu Yang. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *CoRR*, abs/1805.05411, 2018.

[14] Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *CoRR*, abs/1810.10207, 2018.

[15] Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2208–2216, 2016.

[16] Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2975–2984, 2017.

[17] Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst for gradient-based nonconvex optimization. pages 1–10, 2018.

[18] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *CoRR*, abs/1810.02060, 2018.

[19] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczós, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pages 314–323. JMLR.org, 2016.

[20] Sashank J. Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Fast incremental method for smooth nonconvex optimization. In *55th IEEE Conference on Decision and Control (CDC)*, pages 1971–1977, 2016.

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifiable distributional robustness with principled adversarial training. *CoRR*, abs/1710.10571, 2017.

[24] Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.