# Sharp Analysis of Simple Restarted Stochastic Gradient Methods for Min-Max Optimization

**Yan Yan**[1], **Yi Xu**[2], **Qihang Lin**[3], **Wei Liu**[4], **Tianbao Yang**[1]
[1]Department of Computer Science, University of Iowa
[2]Machine Intelligence Technology, Alibaba Group
[3]Department of Management Sciences, University of Iowa
[4]Tecent AI Lab

## Abstract

Recently, min-max saddle-point problems have garnered increasing attention due to their applications in machine learning (e.g., GAN, distributionally robust optimization, AUC maximinzation). However, the research of min-max optimization is still far behind that of the minimization problems. It is still unclear (i) how to achieve a fast rate of $O(1/T)$ for the duality gap of strongly-convex strongly-concave (SCSC) min-max problems given that the fast rate for strongly convex minimization is well studied; (ii) how to design a practical algorithm that can achieve the same complexity of $O(1/\epsilon^4)$ for finding an $\epsilon$-stationary solution to a weakly-convex strongly-concave (WCSC) min-max problem similar to that for weakly-convex minimization problem. In this paper, we aim to fill these gaps by providing sharp analysis of *simple restarted stochastic gradient* methods. For SCSC problems, to the best of our knowledge, we are the first time to establish an iteration complexity of $O(1/\epsilon)$ for reaching *$\epsilon$-duality gap*, instead of primal gap in existing studies. For WCSC problems, we prove that the proposed algorithm achieves an iteration complexity of $O(1/\epsilon^4)$ for finding an *$\epsilon$-nearly stationary point* unlike existing studies that require special structure of the objective function and large mini-batch size.

## 1 Introduction

In this paper, we consider **stochastic algorithms** for solving the following min-max saddle-point problem with a general objective function $f$ without smoothness:

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

where $X$ and $Y$ are closed convex sets and $f : X \times Y \to \mathbb{R}$ is continuous.

Problem (1) covers a number of applications in machine learning, including distributionally robust optimization (DRO), learning with non-decomposable loss functions, etc. For example, in [11, 10, 18], variance regularized problems are formulated as DRO by setting $f(x,y) = \sum_{i=1}^n y_i \ell_i(x)$ and $Y = \{y \in \mathbb{R}^n | D_\phi(y, \frac{\mathbf{1}}{n}) \leq \frac{\rho}{n}, \sum_{i=1}^n y_i = 1, y_i > 0\}$ where $D_\phi(p\|q) = \int \phi(\frac{dp}{dq})dq$ and $\phi : \mathbb{R}_+ \to \mathbb{R}$ is convex with $\phi(1) = 0$. Another example is AUC maximization, which is a typical non-decomposable loss. It is shown that AUC maximization with a square loss is equivalent to a min-max problem $\min_{x,a,b} \max_\alpha \frac{1}{n} \sum_{i=1}^n F(x, a, b, \alpha; (\mathbf{c}_i, d_i))$ with the objective function $F(x, a, b, \alpha; (\mathbf{c}, d)) = (1-p)(x^\top \mathbf{c} - a)^2 I_{[d=1]} + p(x^\top \mathbf{c} - d)^2 I_{[d=-1]} - p(1-p)\alpha^2 + 2(1+\alpha)(px^\top \mathbf{c} I_{[d=-1]} - (1-p)x^\top \mathbf{c} I_{[d=1]})$ where $(\mathbf{c}_i, d_i)$ denotes a feature-label pair, $p$ is the percentage of positive example and $I_{[\cdot]}$ is the indicator function [18, 9, 4, 19].

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Table 1: Summary of complexity results of our algorithms and existing algorithms for finding an $\epsilon$-optimal solution for SCSC or an $\epsilon$-stationary solution for WCSC min-max problems.

| Setting | Works | Restrictions | Complexity |
|---|---|---|---|
| SCSC | [18] | Special Structure, Primal Gap | $O\left(1/\epsilon + \text{Compl}(\mathcal{A})\log(1/\epsilon)\right)$ |
| | This paper | No | $O\left(1/\epsilon\right)$ (Duality gap) |
| WCSC | [14] | Special Structure | $\widetilde{O}\left(1/\epsilon + n/\epsilon^2\right)$ |
| | [8] | Large mini-batch & Smoothness | $O\left(1/\epsilon^4\right)$ |
| | This paper | No | $\widetilde{O}\left(1/\epsilon^4\right)$ |

Although stochastic algorithms for min-max problems have been studied extensively in the literature, their research is still far behind that for stochastic minimization problems. Below, we highlight some of these gaps to motivate the present work. When $f$ is convex in terms of $x$ and concave in terms of $y$, many studies have designed and analyzed stochastic primal-dual algorithms for solving the min-max problems [12, 21, 15, 5, 7, 16, 20, 13, 18]. The standard stochastic primal-dual gradient method suffers from a convergence rate of $O(1/\sqrt{T})$ for convex-concave min-max problems [12], which is similar to that for stochastic convex minimization. However, a fast rate of $O(1/T)$ for the duality gap of a stochastic algorithm is still unknown even for a strongly-convex and strongly-concave problem without imposing special structure and smoothness for the objective function. In contrast, the fast rate of $O(1/T)$ has been established for stochastic strongly convex minimization problems [6]. Recently, Yan et al. [18] has considered stochastic algorithms for SCSC min-max problems. They considered a special family of problems where $f(x,y) = y^\top \ell(x) - \phi^*(y) + g(x)$ is strongly convex and strongly concave, and proposed a restarted algorithm that runs standard stochastic gradient updates for each stage and computes a restarted dual solution by $\mathcal{A}(\bar{x}) = \nabla\phi(\ell(\bar{x}))$ where $\bar{x}$ is the averaged solution for restarting the primal update. They established an complexity of $O(1/\epsilon + \text{Compl}(\mathcal{A})\log(1/\epsilon))$ for finding a solution with $\epsilon$-primal objective gap, where $\text{Compl}(\mathcal{A})$ denotes the complexity for computing $\mathcal{A}(\bar{x})$. **In this paper**, we advance the research for solving SCSC min-max problems significantly. We do not impose any special structure except for strong-convexity and strong concavity, analyze a simple restarted stochastic gradient method that restarts the primal and dual updates with averaged solution from the previous stage, and establish $O(1/\epsilon)$ iteration complexity for finding a solution with $\epsilon$-duality gap in high probability. In a word, our algorithm is simpler, our assumption is weaker, and our theoretical result is stronger. The key to achieving this stronger result lies at the sharp analysis of the simple restarted stochastic gradient method. To the best of our knowledge, this is the first work that establishes $O(1/T)$ convergence rate of a stochastic algorithm for solving SCSC problems without imposing special structure and smoothness condition of the objective function.

When $f$ is non-convex in terms of $x$, there are some recent studies trying to find first-order stationary point [14, 8]. Rafique et al. [14] is the first work that considers weakly-convex concave min-max problems and proposed stochastic primal-dual algorithms with theoretical guarantee. Without strong concavity, their algorithm enjoys an iteration complexity of $O(1/\epsilon^6)$. When the objective function is strongly concave in terms of $y$ and has a special structure as $f(x,y) = \frac{1}{n}\sum_i y^\top c_i(x) - r(y) + g(x)$, their algorithm suffers from a complexity of $O(1/\epsilon^4 + n/\epsilon^2)$ for reaching a stationary point. Lin et al. [8] analyzed stochastic gradient ascent algorithm for smooth non-convex and concave problems. Their complexity is $O(1/\epsilon^4)$ when the dual part becomes strongly concave. However, their algorithm requires a large number of mini-batch size in the order of $O(1/\epsilon^2)$, which is not practical. In contrast, stochastic algorithms for weakly convex minimization problems do not require any special structure of the problem and do not necessarily require a large mini-batch size [2, 1, 3]. **To fill this gap**, we present a simple restarted gradient method and provide a sharp analysis for finding a nearly $\epsilon$-stationary solution. Our algorithm does not require a large mini-batch size and achieve the same iteration complexity of $O(1/\epsilon^4)$ for weakly-convex and strongly-concave problems without smoothness assumption. Finally, we summarize our results and the comparison with existing results in Table 1.

## 2 Preliminaries

This section gives notations and assumptions used in the paper. We let $\|\cdot\|$ denote the Euclidean norm of a vector. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote the Fréchet subgradients and limiting Fréchet gradients by $\hat{\partial} f$ and $\partial f$ respectively, i.e., at $x$, $\hat{\partial} f(x) = \{y \in \mathbb{R}^d : \lim_{x \to x'} \inf \frac{f(x) - f(x') - y^\top}{\|x - x'\|} \geq$

**Algorithm 1** Restarted Stochastic Gradient Method (RSG-MM-1) for SCSC

```
1: Init.: $x_0^1 = x_0 \in X$, $y_0^1 = y_0 \in Y$.
2: for $s = 1, 2, ..., S$ do
3:     for $t = 0, 1, 2, ..., T_s - 1$ do
4:         Compute stochastic gradients $\mathcal{G}_{x,t}^s = \partial_x f(x_t^s, y_t^s; \xi_t^s)$ and $\mathcal{G}_{y,t}^s = \partial_y f(x_t^s, y_t^s; \xi_t^s)$.
5:         $x_{t+1}^s = \Pi_{X \cap \mathcal{B}(x_0^s, R_s)}(x_t^s - \eta_{x,s}\mathcal{G}_{x,t}^s)$
6:         $y_{t+1}^s = \Pi_{Y \cap \mathcal{B}(y_0^s, R_s)}(y_t^s + \eta_{y,s}\mathcal{G}_{y,t}^s)$
7:     end for
8:     $x_0^{s+1} = \bar{x}_s = \frac{1}{T}\sum_{t=0}^{T-1} x_t^s$ , $y_0^{s+1} = \bar{y}_s = \frac{1}{T}\sum_{t=0}^{T-1} y_t^s$
9:     $\eta_{x,s+1} = \frac{\eta_{x,s}}{2}$, $\eta_{y,s+1} = \frac{\eta_{y,s}}{2}$, $R_{s+1} = R_s/\sqrt{2}$, $T_{s+1} = 2T_s$.
10: end for
11: Return $\bar{x}_S$.
```

$0\}$, and $\partial f(x) = \{y \in \mathbb{R}^d : \exists x_k \xrightarrow{f} x, v_k \in \hat{\partial} f(x_k), v_k \to v\}$. Here $x_k \xrightarrow{f} x$ represents $x_k \to x$ and $g(x_k) \to g(x)$. A function $f(x)$ is $\mu$-strongly convex on $X$ if for any $x, x' \in X$, $\partial f(x')^\top (x - x') + \frac{\mu}{2}\|x - x'\|^2 \leq f(x) - f(x')$. A function $f(x)$ is $\rho$-weakly convex on $X$ for any $x, x' \in X$ $\partial f(x')^\top (x - x') - \frac{\mu}{2}\|x - x'\|^2 \leq f(x) - f(x')$. Let $\mathcal{G}_x = \partial_x f(x, y; \xi)$ denote a stochastic subgradient of $f$ at $x$ given $y$, where $\xi$ is used to denote the random variable. Similarly, let $\mathcal{G}_y = \partial_y f(x, y; \xi)$ denote a stochastic sugradient of $f$ at $y$ given $x$. Let $\Pi_\Omega[\cdot]$ denote the projection onto the set $\Omega$, and let $\mathcal{B}(x, R)$ denotes an Euclidean ball centered at $\mathbf{x}$ with a radius $R$.

For saddle-point problems with SCSC functions, we use *duality gap* to measure the convergence. Let us define $\text{Gap}(x, y) = \max_{y' \in Y} f(x, y') - \min_{x' \in X} f(x', y)$, which is the duality gap at $(x, y)$. For WCSC functions, we use *nearly $\epsilon$-stationarity* as the measure of convergence, which is defined as follows.

**Definition 1.** *A solution $x$ is a nearly $\epsilon$-stationary point of* $\min_x \psi(x)$ *if there exists $z$ and a constant $c > 0$ such that $\|z - x\| \leq c\epsilon$ and dist$(0, \partial\psi(z)) \leq \epsilon$.*

The following assumptions will be imposed in our analysis and we suppose that Assumption 1 holds throughout the paper.

**Assumption 1.** *$X$ and $Y$ are closed convex sets. There exists initial solution $x_0 \in X, y_0 \in Y$ and $\epsilon_0 > 0$ such that Gap$(x_0, y_0) \leq \epsilon_0$.*

**Assumption 2.** *(1) $f(x, y)$ is $\mu$-strongly convex in $x$ for any $y \in Y$ and $\lambda$-strongly concave in $y$ for any $x \in X$. (2) $\|\mathcal{G}_x\| \leq B_1$ and $\|\mathcal{G}_y\| \leq B_2$.*

**Assumption 3.** *(1) $f(x, y)$ is $\rho$-weakly convex in $x$ for any $y \in Y$ and is $\lambda$-strongly concave in $y$ for any $x \in X$. (2) $\mathrm{E}[\|\mathcal{G}_x\|^2] \leq M_1$ and $\mathrm{E}[\|\mathcal{G}_y\|^2] \leq M_2$.*

## 3 Restarted Stochastic Gradient Methods for Min-Max Problems

**SCSC Problems.** The proposed algorithm for SCSC min-max problems is shown in Algorithm 1, which is simply a restarted version of stochastic primal-dual gradient method. It is worth mentioning that this algorithm can be considered as a primal-dual variant of the stochastic algorithm proposed in [17]. We first give convergence analysis of the inner loop (Line 4 to 6) in Lemma 1 and Lemma 2 (we omit the $s$ index of outer loop for simplicity). Then we analyze the convergence of the outer loop in Corollary 1 and Corollary 1.

**Lemma 1.** *Let Line 4 to 6 of Algorithm 1 run for $T$ iterations by fixed step size $\eta_x$ and $\eta_y$. Then with the probability at least $1 - \tilde{\delta}$, for any $x \in X \cap \mathcal{B}(x_0, R)$ and $y \in Y \cap \mathcal{B}(y_0, R)$, we have*

$$f(\bar{x}, y) - f(x, \bar{y}) \leq \frac{||x - x_0||^2}{\eta_x T} + \frac{||y - y_0||^2}{\eta_y T} + \frac{5\eta_x B_1^2}{2} + \frac{5\eta_y B_2^2}{2} + \frac{8(B_1 + B_2)R\sqrt{2\log\frac{1}{\tilde{\delta}}}}{\sqrt{T}}, \tag{2}$$

*where $\bar{x} = \sum_{t=0}^{T-1} x_t/T$, $\bar{y} = \sum_{t=0}^{T-1} y_t/T$.*

**Lemma 2.** *Suppose Assumption 2 holds. Denote $(x^*, y^*)$ the unique optimal solution of $f(x, y)$, $\hat{x}_R(y) = \arg\min_{x \in X \cap \mathcal{B}(x_0, R)} f(x, y)$ and $\hat{y}_R(x) = \arg\max_{y \in Y \cap \mathcal{B}(y_0, R)} f(x, y)$. Assume the initial duality gap Gap$(x_0, y_0) \leq \epsilon_0$. Let Line 4 to 6 of Algorithm 1 run*

**Algorithm 2** Restarted Stochastic Gradient Method (RSG-MM-2) for WCSC

```
Init.: $x_0^1 = x_0 \in X$, $y_0^1 = y_0 \in Y$, $\gamma = 2\rho$.
for $k = 1, 2, ..., K$ do
    Set $T_k = \frac{106(k+1)}{3}$, $\eta_x^k = \frac{2}{\rho k}$, $\eta_y^k = \frac{2}{\lambda k}$, $\mathcal{G}_{x,t}^k = \partial_x f(x_t^k, y_t^k; \xi_t^k)$, $\mathcal{G}_{y,t}^k = \partial_y f(x_t^k, y_t^k; \xi_t^k)$.
    for $t = 1, 2, ..., T_k$ do
        $x_{t+1}^k = \arg\min_{x \in X} x^\top \mathcal{G}_{x,t}^k + \frac{1}{2\eta_x^k}\|x - x_t^k\|^2 + \frac{\gamma}{2}\|x - x_0^k\|^2$
        $y_{t+1}^k = \arg\min_{y \in Y} -y^\top \mathcal{G}_{y,t}^k + \frac{1}{2\eta_y^k}\|y - y_t^k\|^2$
    end for
    $x_0^{k+1} = \bar{x}_k = \frac{1}{T}\sum_{t=0}^{T-1} x_t^k$, $y_0^{k+1} = \bar{y}_k = \frac{1}{T}\sum_{n=0}^{T-1} y_t^k$
end for
Return $x_0^\tau$ by $\tau$ randomly sampled from $\{1, ..., K\}$.
```

*$T$ iterations with $\eta_x = \frac{\min\{\mu,\lambda\}R^2}{200B_1^2}, \eta_y = \frac{\min\{\mu,\lambda\}R^2}{200B_2^2}, R \geq 2\sqrt{\frac{2\epsilon_0}{\min\{\mu,\lambda\}}}$ and $T \geq \frac{\max\left\{640^2(B_1+B_2)^2 2\log(\frac{1}{\tilde{\delta}}), 16000\max\{B_1^2, B_2^2\}\right\}}{\mu^2 R^2}$. We have the following results: $(i)\|x_0 - x^*\| \leq \frac{R}{2}$, $(ii)\|y_0 - y^*\| \leq \frac{R}{2}$, $(iii)\|\hat{x}_R(\bar{y}) - x^*\| \leq \frac{R}{2\sqrt{2}}$ and $\|\hat{y}_R(\bar{x}) - y^*\| \leq \frac{R}{2\sqrt{2}}$ with probability $1 - \tilde{\delta}$.*

*Remark.* Lemma 2 shows that with probability $1 - \tilde{\delta}$, we have $\|x_0 - \hat{x}_R(\bar{y})\| \leq \frac{R}{2} + \frac{R}{2\sqrt{2}} < R$ and $\|y_0 - \hat{y}_R(\bar{x})\| \leq \frac{R}{2} + \frac{R}{2\sqrt{2}} < R$. Therefore, supposing $\hat{x}(\bar{y}) = \arg\min_{x \in X} f(x, \bar{y})$ and $\hat{y}(\bar{x}) = \arg\max_{y \in Y} f(\bar{x}, y)$, we have $\hat{x}(\bar{y}) = \hat{x}_R(\bar{y})$ and $\hat{y}(\bar{x}) = \hat{y}_R(\bar{x})$. Then it holds that $\hat{x}(\bar{y}) \in \mathcal{B}(x_0, R)$ and $\hat{y}(\bar{x}) \in \mathcal{B}(y_0, R)$. That is, (2) holds for $x = \hat{x}(\bar{y})$ and $y = \hat{y}(\bar{x})$.

Lemma 2 is the key to our analysis. At the $s$-th outer loop, to derive Gap$(\bar{x}_s, \bar{y}_s)$ by (2), we have to plug in $x = \hat{x}(\bar{y}_s)$ and $y = \hat{y}(\bar{x}_s)$. However, it is not necessary that $\hat{x}(\bar{y}_s) \in \mathcal{B}(x_0^s, R_s)$ and $\hat{y}(\bar{x}_s) \in \mathcal{B}(y_0^s, R_s)$. Lemma 2 critically uses $x^*$ and $y^*$ as the connected points and shows the above two conditions are satisfied when properly setting $\eta_{x,s}, \eta_{y,s}, T_s$ and $R_s$. Then the following Theorem gives the relation between duality gaps of two consecutive outer loops by using the proved conditions. This approach is also employed when analyzing our algorithm for WCSC problems in Theorem 2.

**Theorem 1.** *Consider the $s$-th outer loop of Algorithm 1 with an initial solution $(x_0^s, y_0^s)$ and the ending averaged solution $(\bar{x}_s, \bar{y}_s)$. Suppose Assumption 2 and parameters setting in Lemma 2 hold and in particular Gap$(x_0^s, y_0^s) \leq \epsilon_0 \leq \frac{\min\{\mu,\lambda\}R^2}{8}$. We have with probability $1 - \tilde{\delta}$, Gap$(\bar{x}_s, \bar{y}_s) \leq \frac{\min\{\mu,\lambda\}R^2}{16}$.*

**Corollary 1.** *Suppose Assumption 2 and parameter setting in Lemma 2 hold and $R_1 \geq 2\sqrt{\frac{2\epsilon_0}{\min\{\mu,\lambda\}}}$. The total number of iterations of Algorithm 1 to achieve $\epsilon$-duality gap with probability $1 - \delta$ is*

$$T_{tot} = \frac{\max\left\{640^2(B_1+B_2)^2 2\log(\frac{1}{\tilde{\delta}}), 16000\max\{B_1^2, B_2^2\}\right\}}{8\min\{\mu,\lambda\}\epsilon},$$

*where $S = \lceil\log(\frac{\epsilon_0}{\epsilon})\rceil$ and $\delta = S\tilde{\delta}$.*

**WCSC Problems.** The restarted stochastic gradient method for solving WCSC problems is summarized in Algorithm 2. It is similar to the stochastic algorithm PG-SMD proposed in [14] that consists of solving a sequence of proximally guided SCSC problems but with several differences: (i) the step size is different from PG-SMD for weakly convex and concave problems; (ii) the restarted solution for the dual variable is simple average that is different from that in PG-SMD, which requires solving a maximization problems with high time complexity. We first present the convergence result for each stage regarding the duality gap for the regularized function $\hat{f}_{x_0}(x, y) = f(x, y) + \frac{\gamma}{2}\|x - x_0\|^2$, and then use it to prove the convergence of the proposed algorithm for finding a nearly stationary solution.

**Lemma 3.** *Suppose Assumption 3 holds. Let* $\hat{x}_{x_0^k}(\bar{y}_k) = \arg\min_{x\in X} \hat{f}_{x_0^k}(x, \bar{y}_k)$ *and* $y^*(\bar{x}_k) = \arg\max_{y\in Y} f(\bar{x}_k, y)$. *For* $k \geq 1$, *Line 3 to 8 of Algorithm 2 guarantee*

$$\begin{aligned} &\mathrm{E}[\max_{y\in Y} \hat{f}_{x_0^k}(\bar{x}_k, y) - \min_{x\in X} \hat{f}_{x_0^k}(x, \bar{y}_k)] \\ &\leq \frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{1}{T}\Big\{(\frac{1}{\eta_x} + \frac{\rho}{2})\mathrm{E}[\|\hat{x}_{x_0^k}(\bar{y}_k) - x_0^k\|^2] + \frac{1}{\eta_y}\mathrm{E}[\|y^*(\bar{x}_k) - y_0^k\|^2]\Big\}. \end{aligned} \tag{3}$$

**Theorem 2.** *Define* $\hat{x}_{x_0}^* = \arg\min_{x\in X}[\psi(x) := \max_{y\in Y} \hat{f}_{x_0}(x, y)]$. *Algorithm 2 guarantees*

$$\mathrm{E}[\textit{Dist}(0, \partial\psi(\hat{x}_{x_0^\tau}^*))] \leq \gamma^2\mathrm{E}[\|\hat{x}_{x_0^\tau}^* - x_0^\tau\|^2] \leq \epsilon$$

*after* $K = \max\{\frac{1696\gamma(\frac{M_1^2}{\rho} + \frac{M_2^2}{\lambda})}{\epsilon^2}\ln(\frac{1696\gamma(\frac{M_1^2}{\rho} + \frac{M_2^2}{\lambda})}{\epsilon^2}), \frac{1376\gamma\epsilon_0}{5\epsilon^2}\}$. *The total number of iteration is* $\sum_{k=1}^K T_k = \tilde{O}(\frac{1}{\epsilon^4})$.

## References

[1] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *CoRR*, abs/1803.06523, 2018.

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

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

[4] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems 30*, pages 497–505. 2017.

[5] Gauthier Gidel, Hugo Berard, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *CoRR*, abs/1802.10551, 2018.

[6] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 421–436, 2011.

[7] Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3059–3067, 2014.

[8] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *CoRR*, abs/1906.00331, 2019.

[9] Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with o(1/n)-convergence rate. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[10] 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.

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

[12] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

[13] Balamurugan Palaniappan and Francis R. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1408–1416, 2016.

[14] 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.

[15] Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research (JMLR)*, 2013.

[16] Conghui Tan, Tong Zhang, Shiqian Ma, and Ji Liu. Stochastic primal-dual method for empirical risk minimization with o (1) per-iteration complexity. In *Advances in Neural Information Processing Systems*, pages 8366–8375, 2018.

[17] Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3821 – 3830, 2017.

[18] Yan Yan, Yi Xu, Qihang Lin, Lijun Zhang, and Tianbao Yang. Stochastic primal-dual algorithms with faster convergence than $o(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

[19] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in neural information processing systems*, pages 451–459, 2016.

[20] Adams Wei Yu, Qihang Lin, and Tianbao Yang. Doubly stochastic primal-dual coordinate method for regularized empirical risk minimization with factorized data. *CoRR*, abs/1508.03390, 2015.

[21] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *CoRR*, abs/1409.3257, 2014.