Single-Loop Variance-Reduced Stochastic Algorithm for Nonconvex-Concave Minimax Optimization

Xia Jiang, Linglingzhi Zhu, Taoli Zheng, Anthony Man-Cho So
Department of Systems Engineering and Engineering Management, *The Chinese University of Hong Kong*
xiajiang@cuhk.edu.hk, {llzzhu, tlzheng, manchoso}@se.cuhk.edu.hk

***Abstract*—Nonconvex-concave (NC-C) finite-sum minimax problems have broad applications in decentralized optimization and various machine learning tasks. However, the nonsmooth nature of NC-C problems makes it challenging to design effective variance reduction techniques. Existing vanilla stochastic algorithms using uniform samples for gradient estimation often exhibit slow convergence rates and require bounded variance assumptions. In this paper, we develop a novel probabilistic variance reduction updating scheme and propose a single-loop algorithm called the probabilistic variance-reduced smoothed gradient descent-ascent (PVR-SGDA) algorithm. The proposed algorithm achieves an iteration complexity of $\mathcal{O}(\epsilon^{-4})$, surpassing the best-known rates of stochastic algorithms for NC-C minimax problems and matching the performance of the best deterministic algorithms in this context. Finally, we demonstrate the effectiveness of the proposed algorithm through numerical simulations.**

***Index Terms*—nonconvex-concave minimax optimization, variance reduction, single-loop, stochastic algorithm.**

I. Introduction

In recent years, various applications in decentralized optimization [1], (distributionally) robust optimization [2]–[4], and reinforcement learning [5]–[7] have underscored the need to tackle nonconvex-concave (NC-C) smooth minimax problems. While the ultimate objective is to train models that perform well to unseen data, in practice, we deal with a finite dataset during training. This leads to the finite-sum minimax problem addressed in this paper:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) := \frac{1}{n}\sum_{i=1}^{n} f_i(x,y), \tag{1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ can be nonconvex with respect to x but concave with respect to y, f_i refers to the cost function associated with the i-th sample of a finite training dataset, and $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^d$ are nonempty convex compact sets. In large-scale network systems such as smart grids, UAV swarms, and intelligent transportation systems, decentralized nonconvex optimization problems commonly encountered over multi-agent networks can be reformulated as NC-C minimax problems using Lagrangian duality theory [1]. Consequently, developing efficient algorithms for NC-C minimax problems is essential to address the optimization challenges inherent in decentralized scenarios.

This work is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) project CUHK 14204823.

Stochastic first-order algorithms have attracted significant research interest for solving minimax problems due to their scalability and efficiency. Among these, the stochastic gradient descent-ascent (StocGDA) algorithm is the most widely used, extending the stochastic gradient descent approach to the minimax setting. For NC-C cases, the work [8] introduced a stochastic variant of the two-timescale GDA algorithm [9], which employs unequal step sizes and provides a non-asymptotic convergence guarantee. On another front, multi-loop type algorithms with acceleration in the subproblems [10] have advantages over GDA variants in terms of iteration complexity for general NC-C problems. Both of these algorithms emphasize the importance of updating y more frequently than x for solving minimax problems, while the two-timescale StocGDA is relatively easier to implement and generally demonstrates superior empirical performance compared to the multi-step algorithms.

To further capitalize on the performance advantages of two-timescale StocGDA, the work [11] explored the favorable convergence properties of the alternating version of the two-timescale GDA by proposing a stochastic alternating proximal gradient algorithm. Additionally, the work [12] introduced a stochastic algorithm based on the inexact proximal point method with unequal step sizes to address NC-C minimax problems. Both stochastic algorithms presented in [11], [12] require an iteration complexity of $\mathcal{O}(\epsilon^{-6})$ to solve NC-C minimax problems.

While these stochastic approaches have established non-asymptotic convergence guarantees for NC-C minimax problems, they generally exhibit significantly slower convergence rates compared to deterministic methods and require the additional assumption of bounded gradient variance. One effective technique to enhance the convergence performance of stochastic optimization algorithms is variance reduction. For nonconvex minimization problems, extensive research has demonstrated the effectiveness of variance reduction techniques in improving computational complexity and achieving faster convergence rates, with notable examples including the stochastic path integrated differential estimator (SPIDER) [13] and the stochastic recursive momentum (STORM) algorithm [14]. In the realm of minimax optimization, the works [12], [15] introduced double-loop variance-reduced stochastic algorithms based on the SPIDER technique for nonconvex-strongly

concave (NC-SC) minimax optimization, which necessitate large batch sizes during each periodic iteration. To the best of our knowledge, there are currently no efficient variance-reduced stochastic algorithms with low iteration complexity available for addressing the NC-C minimax problems to enhance the convergence performance of StocGDA algorithms.

Compared to NC-SC minimax problems, NC-C minimax problems are more challenging to solve due to the nonsmoothness introduced by the nonunique dual solutions caused by the lack of strong concavity. Hence, it is not a trivial task to combine variance reduction technique into the algorithm design. Classic variance reduction methods, such as SVRG and SAGA, are not applicable to nonsmooth loss functions [16]. Additionally, existing variance reduction techniques like SPIDER and STORM suffer from unavoidable variance in the stochastic gradient estimators, which impedes the assurance of recursive gradient descent for NC-C minimax problems.

To address the challenge of nonsmoothness in gradient estimation, it is crucial to carefully select and analyze an appropriate variance reduction technique. A probabilistic variance reduction method is proposed, based on the structure of the NC-C minimax problem. To be specific, we develop a novel probability-based gradient updating scheme that eliminates variance terms in the stochastic gradient estimators and ensures the recursive gradient descent property. Building on this approach, we develop a novel probabilistic variance-reduced smoothed gradient descent-ascent (PVR-SGDA) algorithm for solving NC-C minimax problems. By integrating the probabilistic variance reduction technique with Moreau-Yosida smoothing acceleration, our proposed stochastic algorithm achieves the iteration complexity of $\mathcal{O}(\epsilon^{-4})$, where ϵ denotes the desired optimization accuracy. This result matches the best-known iteration complexity of deterministic counterparts and represents a significant improvement over the $\mathcal{O}(\epsilon^{-6})$ complexity of stochastic algorithms in [12] for NC-C minimax problems. In contrast to existing stochastic algorithms that rely on bounded gradient variance [8], [11], as well as multi-loop variance-reduced stochastic algorithms [10], [15], the proposed algorithm adopts a single-loop structure, which simplifies implementation and remains robust to gradient variance.

The notation we use in this paper is standard. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ for any positive integer n. We use I_d to denote a $d \times d$ identity matrix and $\otimes$ to denote the Kronecker product. Let the Euclidean space of all real vectors be equipped with the inner product $\langle x, y\rangle := x^\top y$ for any vectors x, y and $\|\cdot\|$ denote the induced norm. For a differentiable function f, the gradient of f is denoted as ∇f.

II. Motivating Application — Consensus in Decentralized Learning

The popular optimal consensus problem and resource allocation problem in networked systems can be formulated as NC-C minimax problems [1]. Consider a connected network graph $\mathcal{G}$ of N agents, where each agent $i \in [N]$ only has access to a local nonconvex objective function f_i and can communicate with its neighbors. The optimal consensus problem over multi-agent networks with set constraints is described by

$$\min_{x\in\Omega} \sum_{i=1}^{N} f_i(x_i), \quad \text{s.t. } (\mathcal{L}\otimes I_n)\, x = 0, \tag{2}$$

where $\mathcal{L} \in \mathbb{R}^{N\times N}$ is the Laplacian matrix of graph $\mathcal{G}$, $x_i \in \Omega_i \subseteq \mathbb{R}^n$ with $\Omega = \prod_{i=1}^N \Omega_i$ being the Cartesian product of the local convex compact constraint sets Ω_i for $i \in [N]$, and $x = \operatorname{col}(x_i)_{i=1}^N \in \mathbb{R}^{nN}$. Each agent i owns a local variable estimate x_i. Since the graph $\mathcal{G}$ is connected, $(\mathcal{L}\otimes I_n)\, x = 0$ implies that the consensus condition $x_i = x_j$ holds for all $i, j \in [N]$.

The augmented Lagrangian function of the consensus problem (2) is $\mathscr{L}(x, v) := \sum_{i=1}^N f_i(x_i) + v^T(\mathcal{L}\otimes I_n)\, x + \frac{1}{2}x^T(\mathcal{L}\otimes I_n)\, x$, where $v := \operatorname{col}(v_i)_{i=1}^N \in \mathbb{R}^{nN}$ is the dual variable. Since $\mathscr{L}$ is an NC-C function, problem (2) can be transformed into the target constrained minimax problem $\min_{x\in\Omega}\max_{v\in\mathcal{V}} \mathscr{L}(x, v)$, where the convex compact set $\mathcal{V} \subseteq \mathbb{R}^{nN}$ is chosen to be sufficiently large.

III. Probabilistic Variance-reduced smoothed gradient descent-ascent

For solving the general smooth NC-C problem (1), a straightforward approach is to use the two-timescale GDA algorithm, and the work [17] further utilizes the Moreau-Yosida smoothing technique to accelerate the algorithms. To be specific, the smoothing technique introduces an auxiliary variable z and defines a regularized function

$$K(x, z; y) := f(x, y) + \frac{r}{2}\|x - z\|^2. \tag{3}$$

The additional quadratic term smooths the primal update and facilitates a better trade-off between the primal and dual updates when running GDA on this regularized function.

Utilizing the regularized function (3), we propose a stochastic gradient descent-ascent algorithm with a probabilistic variance reduction technique in the following Algorithm 1. Here, the stochastic gradients of the function K are given by

$$\begin{aligned} \nabla_x \tilde{K}(x, z; y) &:= G_x(x, y, \xi_1) + r(x - z), \\ \nabla_y \tilde{K}(x, z; y) &:= G_y(x, y, \xi_2), \end{aligned} \tag{4}$$

where $G_x(x, y, \xi_1)$ and $G_y(x, y, \xi_2)$ are stochastic estimators of $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ using random samples ξ_1 and ξ_2, respectively. For simplicity, we denote $\nabla K_t := \nabla K(x_t, z_t; y_t)$ and $\nabla \tilde{K}_t := \nabla \tilde{K}(x_t, z_t; y_t)$ for $t \in \mathbb{N}$.

Next, we state some basic assumptions in this paper.

Assumption III.1. *The vectors $G_x(x, y, \xi_1)$ and $G_y(x, y, \xi_2)$ are unbiased stochastic estimators of $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$, respectively, i.e. $\mathbb{E}[\nabla \tilde{K}(x, z; y)] = \nabla K(x, z; y)$.*

Assumption III.2. *The function f is differentiable and there exists a positive constant $L > 0$ such that for all $x_1, x_2 \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$,*

$$\begin{aligned} \|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| &\le L[\|x_1 - x_2\| + \|y_1 - y_2\|], \\ \|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| &\le L[\|x_1 - x_2\| + \|y_1 - y_2\|]. \end{aligned}$$

Algorithm 1 Probabilistic Variance-Reduced Smoothed Gradient Descent-Ascent (PVR-SGDA)

Initialize: (x_0, y_0, z_0), step sizes $\eta_x > 0$, $\eta_y > 0$, $\rho > 0$, number of epochs T.
for $t = 0, \ldots, T-1$ **do**
  $v_t = \begin{cases} \nabla_x K_t & \text{with prob. } p \\ v_{t-1} + \nabla_x \tilde{K}_t - \nabla_x \tilde{K}_{t-1} & \text{with prob. } 1-p \end{cases}$
  $w_t = \begin{cases} \nabla_y K_t & \text{with prob. } p \\ w_{t-1} + \nabla_y \tilde{K}_t - \nabla_y \tilde{K}_{t-1} & \text{with prob. } 1-p \end{cases}$
  $x_{t+1} = P_{\mathcal{X}}(x_t - \eta_x v_t)$
  $y_{t+1} = P_{\mathcal{Y}}(y_t + \eta_y w_t)$
  $z_{t+1} = z_t + \rho(x_{t+1} - z_t)$
end for

Remark III.1. *Assumption III.1 naturally holds when the samples ξ_1 and ξ_2 are chosen independently from an identical distribution. Assumption III.2 is standard in minimax optimization. These two assumptions are commonly adopted in existing theoretical studies [11], [12], [15].*

With the above assumptions, we assume $r > L$ in the rest of this paper, and then the regularized function K owns the following important property.

Lemma III.1. *The function $K(\cdot, z; y)$ is strongly convex with modular $r - L$ and $\nabla_x K(\cdot, z; y)$ is Lipschitz continuous with constant $L + r$.*

IV. Convergence analysis

This section concentrates on the convergence analysis of the proposed algorithm[1]. To introduce the main result, we first define the stationarity measure that we are interested in.

Definition IV.1. *Let $\epsilon \geq 0$ be given. The point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is said to be an ϵ-game-stationary point (ϵ-GS) if*

$$\begin{aligned} \operatorname{dist}(0, \nabla_x f(x, y) + \partial \mathbf{1}_{\mathcal{X}}(x)) \leq \epsilon, \\ \operatorname{dist}(0, -\nabla_y f(x, y) + \partial \mathbf{1}_{\mathcal{Y}}(y)) \leq \epsilon. \end{aligned}$$

The notion of game stationarity is a natural extension of that of first-order stationarity in a minimization problem. In addition, we define the following functions:

- (Dual function) $d(y, z) := \min_{x \in \mathcal{X}} K(x, z; y)$;
- (Proximal function) $P(z) := \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} K(x, z; y)$;
- (Primal solution) $x(y, z) := \operatorname{argmin}_{x \in \mathcal{X}} K(x, z; y)$;
- (Dual PGD step) $y_+(z) := P_{\mathcal{Y}}(y + \eta_y \nabla_y K(x(y, z), z; y)$.

Inspired by the works [18], [19], we propose a novel potential function as follows:

$$\Phi_t := V_t + \frac{\gamma}{2p}(\|\nabla_x K_t - v_t\|^2 + \|\nabla_y K_t - w_t\|^2), \tag{5}$$

[1]The full paper including detailed theoretical analysis is present at https://anonymous.4open.science/r/ICASSP_2025_NC_C-F070.

where $V_t := K_t - 2d(y_t, z_t) + 2P(z_t)$ and γ is constant parameter to be determined. The first term V_t in (5) can be rewritten as

$$V_t = \underbrace{K_t - d(y_t, z_t)}_{\text{Primal Descent}} + \underbrace{P(z_t) - d(y_t, z_t)}_{\text{Dual Ascent}} + \underbrace{P(z_t)}_{\text{Proximal Descent}}.$$

The potential function closely links the proximal function P to the updates in the proposed algorithm on K, bridged by an ascent step on the dual function d. The second term in (5) accounts for the error in the gradient estimate.

The following proposition quantifies the change of Φ_t after one round of updates.

Proposition IV.1. *Suppose Assumptions III.1 and III.2 hold. Without loss of generality, we set $L \geq 1$. Let $2L \leq r \leq 4L$, $\gamma \geq 4 + \frac{2}{L}$, and $p \in (0, 1]$. The step-sizes satisfy*

$$\begin{aligned} \eta_x &\leq \frac{p}{p\,(1 + 24L + 2L^2) + 80\gamma L^2}, \\ \eta_y &\leq \min\left\{ \frac{p}{2p\,(1 + 9L) + 10\gamma L^2}, \frac{1}{2L(1+\omega)^2} \right\}, \\ \rho &\leq \frac{4p}{1200p + 9r\gamma}. \end{aligned}$$

then for any $t \geq 0$,

$$\begin{aligned} &\mathbb{E}[\Phi_t - \Phi_{t+1}] \\ &\geq c_x \mathbb{E}[\|x_{t+1} - x_t\|^2] + c_y \mathbb{E}[\|y_t - y_+^t(z_t)\|^2] \\ &\quad + c_z \mathbb{E}[\|z_t - z_{t+1}\|^2] - 24 r \rho \kappa \mathbb{E}[\|y_t - y_+^t(z_t)\|] \\ &\quad + c_v \mathbb{E}[\|\nabla_x K_t - v_t\|^2] + c_w \mathbb{E}[\|\nabla_y K_t - w_t\|^2], \end{aligned}$$

where $c_x = \frac{1}{2\eta_x}$, $c_y = \frac{1}{4\eta_y}$, $c_z = \frac{r}{6\rho}$, $c_v = c_w = \frac{\gamma}{4}$.

Using Proposition IV.1, we establish the main theorem concerning the iteration complexity of Algorithm 1 for solving (1) based on the following connection between various iterate gaps and the GS measure in the following lemma.

Lemma IV.1. *Let $\epsilon \geq 0$ be given. Suppose that*

$$\begin{aligned} \max\Bigg\{ &\frac{\|x_t - x_{t+1}\|}{\eta_x}, \frac{\|y_t - y_+^t(z_t)\|}{\eta_y}, \frac{\|z_{t+1} - z_t\|}{\rho}, \\ &\|\nabla_x K_t - v_t\|, \|\nabla_y K_t - w_t\| \Bigg\} \leq \epsilon. \end{aligned}$$

Then, there exists a $\beta > 0$ such that (x_{t+1}, y_{t+1}) is a $\beta\epsilon$-GS.

Theorem IV.1. *Under the setting of Proposition IV.1 and $\rho = \mathcal{O}(T^{-\frac{1}{2}})$, for any $T > 0$, there exists a $t \in \{1, \ldots, T\}$ such that (x^{t+1}, y^{t+1}) is an $\mathcal{O}(T^{-\frac{1}{4}})$-GS in expectation, i.e.,*

$$\begin{aligned} \mathbb{E}[\operatorname{dist}(0, \nabla_x f(x_{t+1}, y_{t+1}) + \partial \mathbf{1}_{\mathcal{X}}(x_{t+1}))] \leq \epsilon, \\ \mathbb{E}[\operatorname{dist}(0, -\nabla_y f(x_{t+1}, y_{t+1}) + \partial \mathbf{1}_{\mathcal{Y}}(y_{t+1}))] \leq \epsilon, \end{aligned} \tag{6}$$

where $\epsilon = \mathcal{O}(T^{-\frac{1}{4}})$.

Based on the convergence result in Theorem IV.1, we find that the proposed algorithm requires an iteration complexity of $\mathcal{O}(\epsilon^{-4})$ to achieve an ϵ-GS. Compared to the deterministic algorithms in [17]–[19], the proposed stochastic algorithm achieves the same iteration complexity while using sampled

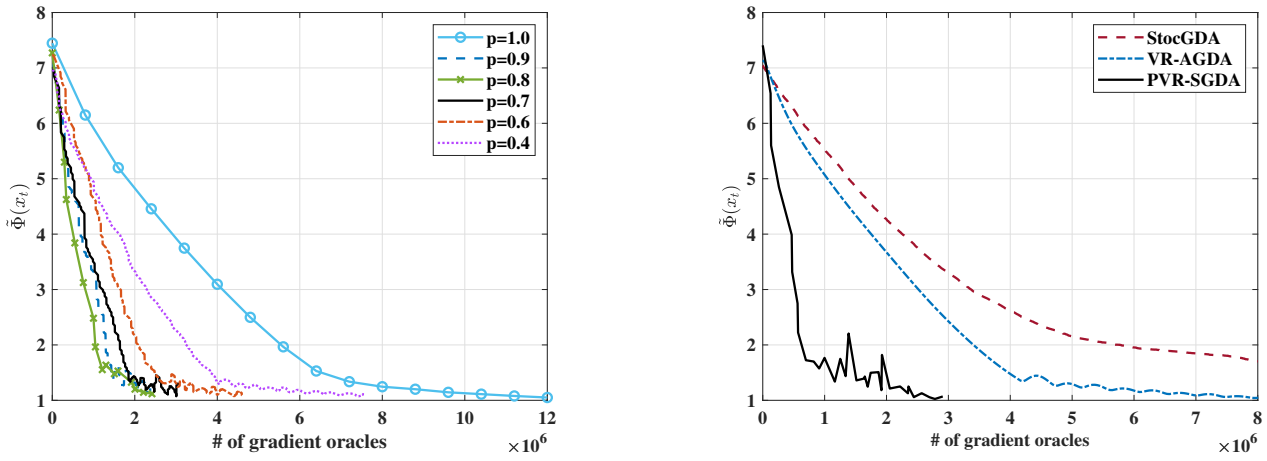

Fig. 1. (a) Convergence of PVR-SGDA algorithm with different p. (b) Performance for different algorithms.

gradients, thereby reducing gradient complexity. Furthermore, the upper bounds on the step sizes in Proposition IV.1 indicate that as the probability p increases, the ranges for the step sizes η_x, η_y, and ρ expand, potentially leading to faster convergence. On the other hand, as p increases, the number of gradient calls in PVR-SGDA also rises for each update step. Therefore, a trade-off exists between the choice of p and the overall convergence efficiency.

V. Numerical Results

In this section, we apply the proposed algorithm to robust logistic regression to demonstrate its practical efficacy. For a dataset $\{(a_i, b_i)\}_{i=1}^n$, where $a_i \in \mathbb{R}^d$ is the feature and $b_i \in \{-1, 1\}$ is the label, the nonconvex-regularized problem is formulated as follows:

$$\min_{x\in\mathbb{R}^d} \max_{y\in\Delta_n} f(x,y) = \sum_{i=1}^{n} y_i \log\left(1+\exp\left(-b_i a_i^\top x\right)\right) + g(x),$$

where y_i is the i-th component of variable y and Δ_n denotes the simplex in $\mathbb{R}^n$. The nonconvex regularization g has the form $g(x) := \lambda \sum_{i=1}^d \alpha x_i^2/(1+\alpha x_i^2)$. Following the settings in [15], [20], we set $\lambda = 0.001$ and $\alpha = 10$ in our experiment. We conduct the experiment on the public dataset a9a, where $d = 123$ and $n = 32561$. To measure the convergence performance of algorithms, we evaluate the function value $\tilde{\Phi}(x) = \max_{y\in\Delta_n} f(x,y)$ with respect to the number of gradient oracles.

Figure 1(a) illustrates the convergence trajectories of the PVR-SGDA algorithm for different values of the probability p. It can be observed that p significantly impacts the convergence behavior of the proposed PVR-SGDA algorithm. As p increases, the convergence rate initially accelerates but then decreases as p approaches 1. When $p = 1$, where the proposed algorithm reduces to the deterministic smoothed GDA algorithm as described in [17], the convergence slows down due to the increased gradient computation burden at each iteration. Thus, the trade-off on the probability can also be observed from the numerical experiments.

In addition, we compare the proposed PVR-SGDA algorithm with several existing algorithms, including the popular StocGDA algorithm [8] and the SVRG-based variance-reduced AGDA (VR-AGDA) algorithm [21]. It is important to note that VR-AGDA is designed for minimax problems that satisfy the one-sided Polyak-Łojasiewicz inequality, and its theoretical analysis does not directly apply to the NC-C minimax problems considered here. Therefore, we only compare the numerical performance of VR-AGDA in this context. The convergence trajectories with respect to the number of gradient oracle calls are presented in Figure 1(b). The fluctuations at the tail of the PVR-SGDA convergence curve are influenced by the choice of probability p, which can be seen from Fig. 1(a). From these trajectories, we observe that the proposed PVR-SGDA algorithm converges faster than the other baseline algorithms, thereby validating the effectiveness of our approach.

VI. Conclusion

This paper presents a single-loop variance-reduced algorithm for solving NC-C minimax problems. By incorporating a probability-based variance reduction step, the proposed algorithm reduces the need for full gradient calculations, which is typical in deterministic algorithms, and achieves convergence that is robust to gradient variance. Utilizing the Moreau-Yosida smoothing technique, the algorithm achieves the best-known complexity of $\mathcal{O}(\epsilon^{-4})$ among stochastic algorithms that solve NC-C minimax problems.

References

[1] Y. Huang, Z. Meng, J. Sun, and W. Ren, "A unified distributed method for constrained networked optimization via saddle-point dynamics," *IEEE Transactions on Automatic Control*, vol. 69, no. 3, pp. 1818–1825, 2024.

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton University Press, 2009.

[3] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

[4] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*. Informs, 2019, pp. 130–166.

[5] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 157–163.

[6] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song, "SBEED: Convergent reinforcement learning with nonlinear function approximation," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 10–15 Jul 2018, pp. 1125–1134.

[7] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

[8] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 2020, pp. 6083–6093.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[10] J. Yang, S. Zhang, N. Kiyavash, and N. He, "A catalyst framework for minimax optimization," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 5667–5678.

[11] R. I. Boţ and A. Böhm, "Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems," *SIAM Journal on Optimization*, vol. 33, no. 3, pp. 1884–1913, 2023.

[12] X. Zhang, N. S. Aybat, and M. Gurbuzbalaban, "SAPD+: An accelerated stochastic method for nonconvex-concave minimax problems," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 21 668–21 681.

[13] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[14] A. Cutkosky and F. Orabona, "Momentum-based variance reduction in non-convex SGD," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[15] L. Luo, H. Ye, Z. Huang, and T. Zhang, "Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 20 566–20 577.

[16] C. Song, S. J. Wright, and J. Diakonikolas, "Variance reduction via primal-dual accelerated dual averaging for nonsmooth convex finite-sums," in *International Conference on Machine Learning*, vol. 139. PMLR, 18–24 Jul 2021, pp. 9824–9834.

[17] J. Zhang, P. Xiao, R. Sun, and Z. Luo, "A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 7377–7389.

[18] J. Li, L. Zhu, and A. M.-C. So, "Nonsmooth nonconvex-nonconcave minimax optimization: Primal-dual balancing and iteration complexity analysis," *arXiv:2209.10825*, 2022.

[19] T. Zheng, L. Zhu, A. M.-C. So, J. Blanchet, and J. Li, "Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 54 075–54 110.

[20] W. Xian, F. Huang, Y. Zhang, and H. Huang, "A faster decentralized algorithm for nonconvex minimax problems," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 25 865–25 877.

[21] J. Yang, N. Kiyavash, and N. He, "Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1153–1165.