Nesterov Meets Optimism: Rate-Optimal Optimistic-Gradient-Based Method for Stochastic Bilinearly-Coupled Minimax Optimization

Chris Junchi Li*
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720
JUNCHILI@BERKELEY.EDU

Angela Yuan*
Department of Computer Sciences
University of California, Los Angeles
Los Angeles, CA 90095
HZYUAN@CS.UCLA.EDU

Gauthier Gidel
DIRO, Université de Montréal and Mila
2920 Chemin de la Tour
Montréal, Québec H3T 1J4, Canada
GAUTHIER.GIDEL@UMONTREAL.CA

Michael I. Jordan
Department of Electrical Engineering and Computer Sciences & Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
JORDAN@CS.BERKELEY.EDU

Abstract

We provide a novel first-order optimization algorithm for bilinearly-coupled strongly-convex-concave minimax optimization called the AcceleratedGradient OptimisticGradient (AG-OG). The main idea of our algorithm is to leverage the structure of the considered minimax problem and operates Nesterov's acceleration on the individual part and optimistic gradient on the coupling part of the objective. We motivate our method by showing that its continuous-time dynamics corresponds to an organic combination of the dynamics of optimistic gradient and of Nesterov's acceleration. By discretizing the dynamics we conclude polynomial convergence behavior in discrete time. Further enhancement of AG-OG with proper restarting allows us to achieve rate-optimal (up to a constant) convergence rates with respect to the conditioning of the coupling and individual parts, which results in the first single-call algorithm achieving improved convergence in the deterministic setting and rate-optimality in the stochastic setting under bilinearly coupled minimax problem sets.

1. Introduction

Optimization is the workhorse of machine learning (ML) and artificial intelligence research; indeed, many basic ML learning tasks can be cast as a minimization problem. In an increasing number of applications, however, such as generative adversarial networks (GANs) [12], robust/adversarial optimization [3, 22], Markov games (MGs) [26], and reinforcement learning (RL) [7, 8, 27], the

. *Equal contribution. Full version available at https://arxiv.org/abs/2210.17550

© C.J. Li*, A. Yuan*, G. Gidel & M. I. Jordan.

goal is to solve instead a minimax problem of the form:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}). \tag{1}$$

When $\mathcal{L}(x,y) : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is a smooth function that is convex in x and concave in y, we refer to this problem as a *convex-concave saddle-point problem*. In this work, we focus on designing sharp or optimal deterministic and stochastic first-order algorithms for solving convex-concave saddle-point problems of the form (1).

Unlike in the minimization setting, there are no general convergence guarantees when simple gradient methods are used on convex-concave objectives. Indeed, there are examples showing the divergence of gradient descent ascent (GDA) on bilinear objectives [10, 20]. This has led to the development of extrapolation-based methods that include the extra-gradient (EG) method [17] and the optimistic gradient descent ascent (OGDA) method [25], both of which can be shown to converge in the convex-concave setting. While the EG algorithm summons an additional gradient oracle at each step, the OGDA algorithm can be seen as a single-call version of EG [10, 14]. In this paper, we design algorithms based on the idea of the OGDA iteration, aiming to improve its performance while retaining the convergence guarantee.

We focus on a specific instance of the general minimax problem, which we call the *bilinearly coupled strongly convex-strongly concave saddle point problem* (bi-SC-SC), formulated as follows:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}) \equiv f(\boldsymbol{x}) + \boldsymbol{x}^\top\mathbf{B}\boldsymbol{y} - g(\boldsymbol{y}), \tag{2}$$

where f is L_f-smooth and μ_f-strongly convex, and g is L_g-smooth and μ_g-strongly convex. Here $\boldsymbol{x}^\top\mathbf{B}\boldsymbol{y}$ is called the bilinear coupling term and is L_H-smooth where $L_H \equiv \sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$. Moreover, we consider throughout this paper the unconstrained problem where $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$ except specified in some application instances. When considering $L := \max(L_H, L_f, L_g)$ and $\mu = \min(\mu_f, \mu_g)$, the standard OGDA algorithm can be shown to yield a coarse-grained complexity of $\frac{L_f\vee L_g\vee L_H}{\mu_f\wedge\mu_g}\log\left(\frac{1}{\epsilon}\right)$ when applied to Problem (2) [10, 23]. In fact this rate is optimal *in the coarse sense* , as has been shown by Azizian et al. [2] that the minimal complexity is $\Omega(\frac{L}{\mu}\log(\frac{1}{\epsilon}))$. Nevertheless, the above convergence rate has a dependency on the parameters L_f, L_g, L_H and also μ_f, μ_g as a whole. Moreover, when reducing to an individual optimization problem $\min_{\boldsymbol{x}}\max_{\boldsymbol{y}} f(\boldsymbol{x}) - g(\boldsymbol{y})$, it does not achieve accelerated rate. This motivates and illustrates the difficulty of deriving *fine-grained* convergence rates that depend on the condition numbers of f, g and H separately and for which a notion of acceleration is possible.

Providing fine-grained rates in the *unbalanced* strongly-convex-strongly-concave setting where $\mu_x \neq \mu_y$ is of particular importance when the constants μ_x, μ_y, L_H, L_x and L_y are drastically different. For instance when μ_g is significantly larger than μ_f, decoupling the dependencies on L_f, μ_f and L_g, μ_g from the coarsened smoothness and strong convexity parameters $L_f\vee L_g$, $\mu_f\wedge\mu_g$ would significantly improve the theoretical complexity. In this paper, we focus on accelerating OGDA when constrained to the specific problem (2), and show that it enjoys a *fine-grained*, *accelerated* convergence rate in the sense that it has sharp dependency on the aforementioned Lipschitz constants in an individual fashion. More precisely, one can reduce the complexity due to the condition number of f and g by a square root. Overall, the best rate that can be achieved in this setting is given by the lower bound $\Omega\left(\sqrt{\frac{L_f}{\mu_f}+\frac{L_g}{\mu_g}}+\sqrt{\frac{L_H^2}{\mu_f\mu_g}}\right)\log\left(\frac{1}{\epsilon}\right)$, established by Zhang et al. [32]. With the goal of

matching such a first-order complexity, we are devoted to the problem of: *Designing rate-optimal single-call algorithms for deterministic and stochastic bilinearly coupled saddle point problems*.

1.1. Contributions

We list our contributions in this subsection. First, we present a novel algorithm that blends acceleration dynamics based on the single-call OGDA algorithm for the adversarial part and Nesterov's acceleration for the individual part. We refer to this blend as the Accelerated Gradient-Optimistic Gradient (AG-OG) algorithm. Second, we illustrate how the dynamics of AG-OG can be seen as the summation of the OGDA dynamics and Nesterov's acceleration dynamics, which simplifies the understanding of acceleration on OGDA and provides insights into the convergence proof. Additionally, equipped with a scheduled restarting technique, we derive an "accelerated optimistic gradient with Nesterov's acceleration and restarting" (AVATAR) method that achieves an upper bound that matches the lower bound on Problem (2). Finally, when it comes to stochastic settings, we present a stochastic version of AVATAR and establish its convergence at an optimal rate. Overall, our work provides novel acceleration schemes for deterministic/stochastic OGDA under bilinearly-coupled saddle point problems (2) (Thms 3.4 and 4.2). Our algorithm is simple, shows high interpretability in discrete and continuous dynamics, and is the first single-call algorithm with a rate-optimal convergence rate with known sharpest bias term in the stochastic settings.

1.2. Related Work

Deterministic Case. Many works have studied the linear convergence rate of gradient-based methods for games in the context of strongly monotone operators (which is implied by strong convex-concavity) [23] and several works [6, 29–31, 33] have slowly bridged the gap with the lower bound provided for *unbalanced* strongly-convex-strongly-concave objective. There has been a series of papers along this direction [6, 21, 23, 29, 30], and only very recently have optimal results that reach the lower bound been presented Jin et al. [15], Kovalev et al. [18], Thekumparampil et al. [28]. These three works proposed improved methods leveraging convex duality. However, it could be challenging to extend the proposed methods to the non-bilinear coupling term case or the stochastic case. In particular, none of these previous works provide result in the non-finite-sum stochastic case.

Stochastic Case. There exists a rich literature on stochastic variational inequalities with application to solving stochastic minimax problems [1, 4, 5, 14, 16, 34]. However, only a few works have proposed fined-grained bounds suited to the (Bil-)SC-SC setting. To the best of our knowledge, most fined-grained bounds have been proposed in the finite-sum setting [15, 24]. Two closely related works are Li et al. [19], who provide a convergence rate for extra-gradient in the purely bilinear setting and Du et al. [9], who study an accelerated version of extra-gradient, dubbed as AcceleratedGradient-ExtraGradient (AG-EG) in the Bil-SC-SC setting. Our work provides results in the same vein as Du et al. [9] but instead employs the optimistic gradient instead of extra-gradient to handle the bilinear coupling part. Optimistic-gradient-based methods have been considered extensively in literature due to its requiring fewer gradient oracle calls per iteration than standard extra-gradient and can be applied to the online learning setting [11]. Note that, in general, EG and OG methods often share some similarities in their analyses, but also acknowledge sigfinicant differences [11, §3.1], [13, §2]. More specifically in our case, using a *single-call algorithm* that reuses

previously calculated gradients alters the recursion (Eq. (31)). Although the main part of the proof follows the streamline of estimating Nesterov's acceleration terms first, an additional squared error norm involving the previous iterates is present, intrinsically implying an additional iterative rule (Eq. (32)) underneath the original iterative rule essential for proving bounded iterates. In addition, due to the accumulated error across iterates, the maximum stepsize allowed in single-call algorithms is forced to be smaller.

References

[1] Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

[2] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, pages 1705–1715. PMLR, 2020.

[3] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. *arXiv preprint arXiv:2002.04017*, 2020.

[4] Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems–survey. *arXiv preprint arXiv:2208.13592*, 2022.

[5] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Michael B Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. *arXiv preprint arXiv:2011.06572*, 2020.

[7] 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 *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

[8] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

[9] Simon S Du, Gauthier Gidel, Michael I Jordan, and Chris Junchi Li. Optimal extragradient-based stochastic bilinearly-coupled saddle-point optimization. *arXiv preprint arXiv:2206.08573*, 2022.

[10] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[11] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 2020.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in neural information processing systems*, 2022.

[14] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. *arXiv preprint arXiv:2202.04640*, 2022.

[16] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[17] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

[18] Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *arXiv preprint arXiv:2112.15199*, 2021.

[19] Chris Junchi Li, Yaodong Yu, Nicolas Loizou, Gauthier Gidel, Yi Ma, Nicolas Le Roux, and Michael Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pages 9793–9826. PMLR, 2022.

[20] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

[21] Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[23] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[24] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

[25] Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[26] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

[27] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[28] Kiran K Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 4281–4308. PMLR, 2022.

[29] Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.

[30] Guangzeng Xie, Yuze Han, and Zhihua Zhang. Dippa: An improved method for bilinear saddle point problems. *arXiv preprint arXiv:2103.08270*, 2021.

[31] Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020.

[32] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, pages 1–35, 2021.

[33] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

[34] Renbo Zhao. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022.