OPT2021: 13th Annual Workshop on Optimization for Machine Learning

On the Convergence of Stochastic Extragradient for Bilinear Games using Restarted Iteration Averaging*

Chris Junchi Li† JUNCHILI@BERKELEY.EDU
Yaodong Yu† YYU@EECS.BERKELEY.EDU
University of California, Berkeley

Nicolas Loizou LOIZOUNI@MILA.QUEBEC
Gauthier Gidel GAUTHIER.GIDEL@UMONTREAL.CA
Mila and DIRO, Université de Montréal

Yi Ma YIMA@EECS.BERKELEY.EDU
University of California, Berkeley

Nicolas Le Roux‡ NICOLAS@LE-ROUX.NAME
Mila, Université de Montréal and McGill University

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California, Berkeley

Abstract

We study the stochastic bilinear minimax optimization problem, presenting an analysis of the same-sample Stochastic ExtraGradient (SEG) method with constant step size, and presenting variations of the method that yield favorable convergence. In sharp contrasts with the basic SEG method whose last iterate only contracts to a fixed neighborhood of the Nash equilibrium, SEG augmented with iteration averaging provably converges to the Nash equilibrium under the same standard settings, and such a rate is further improved by incorporating a scheduled restarting procedure. In the interpolation setting where noise vanishes at the Nash equilibrium, we achieve an optimal convergence rate up to tight constants. We present numerical experiments that validate our theoretical findings and demonstrate the effectiveness of the SEG method when equipped with iteration averaging and restarting.

1. Introduction

The *minimax optimization* framework provides solution concepts useful in game theory [33], statistics [4] and online learning [7, 9]. It has recently been prominent in the deep learning community due to its application to generative modeling [17] and robust prediction [29]. There remains, however, a gap between minimax characterizations of solutions and algorithmic frameworks that provably converge to such solutions in practice.

In standard single-objective machine learning applications, a popular optimization algorithm is the stochastic gradient descent (SGD, or one of its variants), where the full gradient is formulated

* Full version available at https://arxiv.org/abs/2107.00464.
† Equal contribution. ‡ Corresponding author.

as an expectation over the data-generating mechanism. In general minimax optimization problems, however, naive use of SGD leads to pathological behavior due to the presence of rotational dynamics [5, 16].

One way to overcome these rotations is to use gradient-based methods specifically designed for the minimax setting (or more generally for the multi-player game setting). A key example of such methods is the celebrated *extragradient method*. Originally introduced by [24], it addresses general minimax optimization problems and yields optimal convergence guarantees in the batch setting [3]. In the stochastic setting, however, it has only been analyzed in special cases, such as the constrained case [23], the bounded-noise case [20], and the interpolatory case [45]. In the current paper, we study the general stochastic bilinear minimax optimization problem, also known as the bilinear saddle-point problem,

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \; \mathbf{x}^\top \mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y} + \mathbf{x}^\top \mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}\,, \tag{1}$$

where the index ξ denotes the randomness associated with stochastic sampling. Following standard practice we assume that the expected coupling matrix $\mathbf{B} = \mathbb{E}[\mathbf{B}_\xi]$ is nonsingular, and that the intercept vectors $\mathbf{g}_\xi^{\mathbf{x}}$ and $\mathbf{g}_\xi^{\mathbf{y}}$ have zero mean: $\mathbb{E}[\mathbf{g}_\xi^{\mathbf{x}}] = \mathbf{0}_n$ and $\mathbb{E}[\mathbf{g}_\xi^{\mathbf{y}}] = \mathbf{0}_m$. Thus the Nash equilibrium point is $[\mathbf{x}^*; \mathbf{y}^*] = [\mathbf{0}_n; \mathbf{0}_m]$. Such assumptions are standard in the literature on bilinear optimization [see, e.g., 31, 45].

In this work, we present theoretical results in the general setting of bilinear minimax games for a version of the Stochastic ExtraGradient (SEG) method that incorporates iteration averaging and scheduled restarting. The introduction of stochasticity in the matrix $\mathbf{B}_\xi$ together with an unbounded domain presents technical challenges that have been a major stumbling block in earlier work. Here we show how to surmount these challenges. In fact, convergence results with this type of noise, referred to as *multiplicative noise* [cf. 13], are a key novelty of our analysis. Formally, we introduce the following SEG method composed of an extrapolation step (half-iterates) and an update step:

$$\begin{aligned} \mathbf{x}_{t-1/2} &= \mathbf{x}_{t-1} - \eta_t \left[\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \mathbf{g}_{\xi,t}^{\mathbf{x}}\right] \\ \mathbf{y}_{t-1/2} &= \mathbf{y}_{t-1} + \eta_t \left[\mathbf{B}_{\xi,t}^\top \mathbf{x}_{t-1} + \mathbf{g}_{\xi,t}^{\mathbf{y}}\right] \end{aligned}, \quad \text{and} \quad \begin{aligned} \mathbf{x}_t &= \mathbf{x}_{t-1} - \eta_t \left[\mathbf{B}_{\xi,t}\mathbf{y}_{t-1/2} + \mathbf{g}_{\xi,t}^{\mathbf{x}}\right] \\ \mathbf{y}_t &= \mathbf{y}_{t-1} + \eta_t \left[\mathbf{B}_{\xi,t}^\top \mathbf{x}_{t-1/2} + \mathbf{g}_{\xi,t}^{\mathbf{y}}\right] \end{aligned}. \tag{2}$$

Here and throughout we adopt a ***same-sample-and-step-size*** notation in which the extrapolation and extragradient steps share the same stochastic sample [14, 31] and step size η_t; i.e., the updates in Eq. (2) use the same samples of $\mathbf{B}_\xi$, $\mathbf{g}_\xi^{\mathbf{x}}$ and $\mathbf{g}_\xi^{\mathbf{y}}$. Note that there exist counterexamples [see, e.g., 10, Theorem 1] where the SEG iteration [23] persistently diverges when using independent samples. The same-sample stochastic extra gradient (SEG) method, which has been widely studied in recent literature [14, 31], aims to address this issue. In practice, for the bilinear game problems we consider in this paper as well as other application problems, including generative adversarial networks and adversarial training, it is easy to perform the same-sample SEG updates: in most machine learning applications one can re-use a sample without significant extra cost.

Main contributions. We provide an in-depth study of SEG on bilinear games and we show that, unlike in the minimization-only setting, in the minimax optimization setting the last-iterate SEG algorithm with the same sample and step sizes *cannot* converge in general even when the step sizes are diminishing to zero [Theorems 6 and 7]. This motivates our study of averaging and restarting in order to obtain meaningful convergence rates:

(i) We prove that in the bilinear game setting, under mild assumptions, iteration averaging allows SEG to converge at the rate of $1/\sqrt{K}$ [Theorem 3], K being the number of samples the algorithm has processed. This rate is statistically optimal up to a constant multiplier. Additionally, we can further boost the convergence rate when we combine iteration averaging with scheduled restarting [Theorem 4] when the lower bound of the smallest eigenvalue in the coupling matrix is known to the system. In this case, exponential forgetting of the initialization and an optimal statistical rate are achieved.

(ii) In the special case of the interpolation setting, we are able to show that SEG with iteration averaging and scheduled restarting achieves an accelerated rate of convergence, faster than (last-iterate) SEG [Theorem 5], reducing the dependence of the rate on the condition number to a dependence on its square root. We achieve state-of-the-art rates comparable to the full batch optimal rate [3], with access only to a stochastic estimate of the gradient, improving upon Vaswani et al. [45].

(iii) We provide the first convergence result on SEG with unbounded noise. The only existing result of which we are aware of for the unbounded noise setting is the work of Vaswani et al. [45] in the interpolation setting. Our theoretical results are further validated by experiments on synthetic data.

Organization. The remainder of this paper is organized as follows. §2 details the basic setup and assumptions for our main results. §3 presents our convergence results for SEG with averaging and restarting. §4 concludes our paper. All technical analyses along with auxiliary results are relegated to later sections in the supplementary materials, as well as experiments that validate our theory.

Notation. Throughout this paper we use the following notation. For two real symmetric matrices, $\mathbf{B}_1, \mathbf{B}_2$, we denote $\mathbf{B}_1 \preceq \mathbf{B}_2$ when $\mathbf{v}^\top \mathbf{B}_1 \mathbf{v} \leq \mathbf{v}^\top \mathbf{B}_2 \mathbf{v}$ holds for all vectors $\mathbf{v}$. Let $\lambda_{\max}(\mathbf{B})$ (resp. $\lambda_{\min}(\mathbf{B})$) be the largest (resp. smallest) eigenvalue of a generic (real symmetric) matrix $\mathbf{B}$. Let $\|\mathbf{B}\|_{op}$ denotes the operator norm of $\mathbf{B}$. Let $\mathcal{F}_t$ be the filtration generated by the stochastic samples, $\mathbf{B}_{\xi,s}, \mathbf{g}_{\xi,s}$, $s = 1, \ldots, t$, in the bilinear game. Let $\max(a, b)$ or $a \vee b$ denote the maximum value of $a, b \in \mathbb{R}$, and let $\min(a; b)$ or $a \wedge b$ denote the minimum. For two real sequences, (a_n) and (b_n), we write $a_n = O(b_n)$ to mean that $|a_n| \leq C b_n$ for a positive, numerical constant C, for all $n \geq 1$, and let $a_n = \tilde{O}(b_n)$ mean that $|a_n| \leq C b_n$ where C hides a logarithmic factor in relevant parameters. We also denote $\widehat{\mathbf{M}}_\xi \equiv \mathbf{B}_\xi^\top \mathbf{B}_\xi$ and $\mathbf{M}_\xi \equiv \mathbf{B}_\xi \mathbf{B}_\xi^\top$ for brevity, each being positive semi-definite for each realization of ξ. Finally, let $[n] = \{1, \ldots, n\}$ for n being a natural number.

2. Setup for our Main Results

In this section, we introduce the basic setup and assumptions needed for our statement of the convergence of the stochastic extragradient (SEG) algorithm. We first make the following assumptions on $\mathbf{B}_\xi$. Let us recall that $\widehat{\mathbf{M}} \equiv \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi]$ and $\mathbf{M} \equiv \mathbb{E}_\xi \mathbf{M}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top]$.

Assumption 1 (Assumption on $\mathbf{B}_\xi$) *Denote $\mathbf{B} = \mathbb{E}_\xi[\mathbf{B}_\xi]$ for $\mathbf{B} \in \mathbb{R}^{n \times m}$ and impose the following regularity conditions: $\lambda_{\max}(\mathbf{B}^\top \mathbf{B}) > 0$ and $\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}}) > 0$. We assume that there exist $\sigma_{\mathbf{B}}, \sigma_{\mathbf{B},2} \in [0, \infty)$ such that*

$$\max\left(\|\mathbb{E}_\xi[(\mathbf{B}_\xi - \mathbf{B})^\top (\mathbf{B}_\xi - \mathbf{B})]\|_{op}; \|\mathbb{E}_\xi\left[(\mathbf{B}_\xi - \mathbf{B})(\mathbf{B}_\xi - \mathbf{B})^\top\right] \|_{op} \right) \leq \sigma_{\mathbf{B}}^2 \tag{3}$$

and

$$\max\left(\|\mathbb{E}_\xi[\mathbf{B}_\xi^\top\mathbf{B}_\xi-\widehat{\mathbf{M}}]^2\|_{op};\|\mathbb{E}_\xi[\mathbf{B}_\xi\mathbf{B}_\xi^\top-\mathbf{M}]^2\|_{op}\right)\le\sigma_{\mathbf{B},2}^2. \tag{4}$$

The assumption of $n\ge m$ (i.e. $\mathbf{B}$ is tall) is without loss of generality; we can convert the SEG iterates with a wide coupling matrix to that of its transpose. Note also $\sigma_{\mathbf{B}}=0$ corresponds to the nonrandom $\mathbf{B}_\xi=\mathbf{B}$ case. The stochasticity introduced in $\mathbf{B}_\xi$ allows us to conclude the first convergence result under the unbounded noise condition.[1] Next we impose an assumption on the intercept vector $\mathbf{g}_\xi$.

Assumption 2 (Assumption on $\mathbf{g}_\xi$) *There exists a $\sigma_{\mathbf{g}}\in[0,\infty)$ such that*

$$\mathbb{E}_\xi\left[\|\mathbf{g}_\xi^{\mathbf{x}}\|^2+\|\mathbf{g}_\xi^{\mathbf{y}}\|^2\right]\le\sigma_{\mathbf{g}}^2<\infty.$$

Furthermore, we let $\mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}]=\mathbf{0}_n$, $\mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{y}}]=\mathbf{0}_m$ and assume independence between the stochastic matrix $\mathbf{B}_\xi$ and the vector $[\mathbf{g}_\xi^{\mathbf{x}};\mathbf{g}_\xi^{\mathbf{y}}]$.

We remark that the independence assumption in Assumption 2 significantly simplifies our analysis.[2] In particular, it ensures $\mathbb{E}[\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}]=\mathbf{0}_n$ and $\mathbb{E}[\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}]=\mathbf{0}_m$, so the Nash equilibrium is the equilibrium point that the last-iterate SEG oscillates around. The independence structure of $\mathbf{B}_\xi$ and $[\mathbf{g}_\xi^{\mathbf{x}};\mathbf{g}_\xi^{\mathbf{y}}]$ in Assumption 2 is crucial for our analysis, which is satisfied in certain statistical models. Specially, when one of the $\mathbf{B}_\xi$ and $[\mathbf{g}_\xi^{\mathbf{x}};\mathbf{g}_\xi^{\mathbf{y}}]$ is nonrandom this is always satisfied. Our analysis can be further generalized to more relaxed assumptions on zero correlation between $[\mathbf{g}_\xi^{\mathbf{x}};\mathbf{g}_\xi^{\mathbf{y}}]$ and the first three moments of $\mathbf{B}_\xi$, with a second-moment condition similar to $\mathbb{E}_\xi[\|\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\|^2+\|\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\|^2]\le C(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}}))\sigma_{\mathbf{g}}^2$. We defer the full development of this extension to future work. With Assumptions 1 and 2 at hand, we are ready to state our main results on the convergence of SEG variants.

3. SEG with Averaging and Restarting

Recall that in contrast to SGD theory in convex optimization, the last iterate of SEG does *not* converge to an arbitrarily small neighborhood of the Nash equilibrium even for the case of a converging step size [20]. We accordingly turn to an analysis of the *averaged iterate* of $\mathbf{x}_t$ and $\mathbf{y}_t$, $t=0,1,\dots,K$, denoted as

$$\overline{\mathbf{x}}_K\equiv\frac{1}{K+1}\sum_{t=0}^K\mathbf{x}_t,\qquad\overline{\mathbf{y}}_K\equiv\frac{1}{K+1}\sum_{t=0}^K\mathbf{y}_t. \tag{5}$$

For simplicity we focus on the case in which $\mathbf{B}_\xi,\mathbf{B}$ are square matrices. Let us define $\eta_{\mathbf{M}}$ as follows, which is the maximal step size that the SEG algorithm analysis takes:

$$\eta_{\mathbf{M}}\equiv\frac{1}{\sqrt{\lambda_{\max}\left(\mathbf{M}^{-1/2}[\mathbb{E}_\xi\mathbf{M}_\xi^2]\mathbf{M}^{-1/2}\right)\vee\lambda_{\max}\left(\widehat{\mathbf{M}}^{-1/2}[\mathbb{E}_\xi\widehat{\mathbf{M}}_\xi^2]\widehat{\mathbf{M}}^{-1/2}\right)}}. \tag{6}$$

[1] As a comparison, Hsieh et al. [20] only provides a proof for the bounded noise case.

[2] In practice, such independence can be *approximately* achieved via the following decoupling argument: we formulate the random Jacobian-vector product and the random intercept using two independent random samples, separately. Note an approximate knowledge of the Nash equilibrium is required in this decoupling argument.

We introduce the following variants:

$$\hat{\eta}_{\mathbf{M}}(\alpha) \equiv \frac{\eta_{\mathbf{M}}}{\sqrt{2}} \wedge \frac{\alpha\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{2\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}, \qquad \text{and} \quad \bar{\eta}_{\mathbf{M}}(\alpha) \equiv \eta_{\mathbf{M}} \wedge \frac{\alpha\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{2\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}, \tag{7}$$

which reduce to $1/\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$ and $1/\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$ when $\mathbf{B}_\xi$ is nonrandom. We state our first main result on SEG with iteration averaging, Theorem 3:

Theorem 3 (SEG Averaged Iterate) *Let Assumptions 1 and 2 hold with $n = m$. Prescribing an $\alpha \in (0,1)$, when the step size η is chosen as $\hat{\eta}_{\mathbf{M}}(\alpha)$ as defined in Eq. (7), we have for all $K \geq 1$ the following convergence bound for the averaged iterate:*

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \leq \frac{16 + 8\kappa_\zeta}{(1-\alpha)\hat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{(K+1)^2} + \frac{18 + 12\kappa_\zeta}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\sigma_{\mathbf{g}}^2}{K+1}, \tag{8}$$

where $\kappa_\zeta \equiv \frac{\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}$ denotes an effective noise condition number of problem Eq. (1).

Measured by the Euclidean metric, Theorem 3 indicates an $O(1/\sqrt{K})$ leading-order convergence rate for the averaged iterate of SEG in the general stochastic setting, which is known to be statistically optimal up to a constant multiplier. Nevertheless, the iteration slowly forgets initial conditions at a polynomial rate, and this result can be improved if we utilize a restarting scheme and take advantage of the knowledge of the smallest eigenvalue of $\mathbf{B}\mathbf{B}^\top$. Indeed, in the following result, we boost the convergence rate shown in Eq. (8), when the smallest eigenvalue $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ is available to the system, via a novel restarting procedure at specific times. The rationale behind this analysis is akin to that used in boosting sublinear convergence in convex optimization to linear convergence when the designer has (an estimate of) the strong convexity parameter.

We now develop this argument in detail. We continue to assume the case of square matrices $\mathbf{B}_\xi, \mathbf{B}$. In Algorithm 1 we run SEG with averaging and restart the iteration at chosen timestamps, $\{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]} \subseteq [K]$, initializing at the averaged iterate of the previous epoch. The principle behind our choice of parameters in this algorithm is that we trigger the restarting when the expected squared Euclidean metric $\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right]$ decreases by a factor of $1/e^2$, and we halt the restarting procedure once the last iterate reaches stationarity in squared Euclidean metric in the sense of Theorem 6:[3]

$$\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 \approx \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}.$$

Given these choices, summarized in Algorithm 1, we obtain the following theorem:

Theorem 4 (SEG with Averaging and Restarting) *Let Assumptions 1 and 2 hold with $n = m$. For any prescribed $\alpha \in (0,1)$, choose the step size $\hat{\eta}_{\mathbf{M}}(\alpha)$ as in Eq. (7) and assume a proper restarting schedule. For all $K \geq K_{\text{complexity}} + 1$ we have the following convergence bound for the output $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$ of Algorithm 1:*

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2\right] \leq \left[1 + \underbrace{\frac{O(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}}_{\textit{higher-order term } O(\kappa_\zeta)}\right] \cdot \frac{18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{1}{K - K_{\text{complexity}} + 1}, \tag{9}$$

[3]The choice of the discount factor $1/e^2$ is to be consistent with our optimal choice in the interpolation setting, where in the $\sigma_{\mathbf{B}} = 0$ case the total complexity is minimized to $e\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})/\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$.

Algorithm 1 Iteration Averaged SEG with Scheduled Restarting

Require: Initialization $\mathbf{x}_0$, step sizes η_t, total number of iterates K, restarting timestamps $\{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]} \subseteq [K]$ with the total number of epoches $\mathsf{Epoch} \geq 1$, time index $s \leftarrow 0$

1: **for** $t = 1, 2, \ldots, K$ **do**
2: $s \leftarrow s + 1$
3: Update $\mathbf{x}_t$, $\mathbf{y}_t$ via Eq. (2)
4: Update $\hat{\mathbf{x}}_t$, $\hat{\mathbf{y}}_t$ via

$$\hat{\mathbf{x}}_t \leftarrow \frac{s-1}{s}\hat{\mathbf{x}}_{t-1} + \frac{1}{s}\mathbf{x}_t \quad \text{and} \quad \hat{\mathbf{y}}_t \leftarrow \frac{s-1}{s}\hat{\mathbf{y}}_{t-1} + \frac{1}{s}\mathbf{y}_t$$

5: **if** $t \in \{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]}$ **then**
6: Overload $\mathbf{x}_t \leftarrow \hat{\mathbf{x}}_t$, $\mathbf{y}_t \leftarrow \hat{\mathbf{y}}_t$, and set $s \leftarrow 0$ //restarting procedure is triggered
7: **end if**
8: **end for**
9: **Output:** $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$

where $K_{\mathrm{complexity}}$ *is the fixed* burn-in complexity *defined as*

$$\frac{\textit{logarithmic factor}}{\frac{1}{e}\sqrt{(1-\alpha)\hat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} - O\left(\hat{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\right)}. \quad (10)$$

Here we not only achieve the optimal $O(1/\sqrt{K})$ convergence rate for the averaged iterate, but the proper restarting schedule allows us to achieve a convergence rate bound for iteration-averaged SEG in Eq. (9) that forgets the initialization at an exponential rate instead of the polynomial rate that is obtained without restarting [cf. Theorem 4].

Finally, we consider the interpolation setting, where the noise vanishes at the Nash equilibrium. That is, $\mathbf{g}_\xi^{\mathbf{x}} = \mathbf{0}_n$ and $\mathbf{g}_\xi^{\mathbf{y}} = \mathbf{0}_m$, which is also equivalent to $\sigma_{\mathbf{g}} = 0$ in Assumption 2. In that setting, we prove that SEG with iteration averaging achieves an accelerated linear convergence rate. Set the (constant) interval length of restarting timestamps $K_{\mathrm{thres}}(\alpha)$ as

$$\frac{2}{\frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} - O\left(\bar{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\right)}. \quad (11)$$

We present an analysis of this algorithm in the following theorem, which can be seen as a corollary of Theorem 4 but benefits from a refined analysis where tight constant prefactor sits in each term of the bound:

Theorem 5 (Interpolation Setting) *Let Assumptions 1 and 2 hold with $n = m$ and $\sigma_{\mathbf{g}} = 0$. For any prescribed $\alpha \in (0,1)$ choosing the step size $\eta = \bar{\eta}_{\mathbf{M}}(\alpha)$ as in Eq. (7) and the restarting timestamps $\mathcal{T}_i = i \cdot K_{\mathrm{thres}}(\alpha)$ where $K_{\mathrm{thres}}(\alpha)$ was defined as in Eq. (11), we conclude for all $K \geq 1$ that is divisible by $K_{\mathrm{thres}}(\alpha)$ the following convergence rate for the output $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$ of Algorithm 1:*

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2\right] \leq e^{-\frac{K}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} + C(\alpha)}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right], \quad (12)$$

where $C(\alpha)$ is defined as

$$C(\alpha) = O\left(K\bar{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\right).$$

The idea behind Theorem 5 is, in plain words, to trigger restarting whenever the last-iterate SEG has travelled through a full cycle, giving insights on the design of $K_{\text{thres}}(\alpha)$ in the restarting mechanism. Compared with Eq. (13) in Theorem 6 with $\sigma_{\mathbf{g}}$ equal to zero, the contraction rate (in terms of its exponent) to the Nash equilibrium $-\frac{\eta_{\mathbf{M}}^2}{4} \cdot \left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right)$ improves to $-\frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$ plus higher-order moment terms involving $\mathbf{B}_\xi$. It is worth mentioning that Algorithm 1 achieves this accelerated convergence rate in Eq. (12) via simple restarting and does *not* require an explicit Polyak- or Nesterov-type momentum update rule [36]. In the case of nonrandom $\mathbf{B}_\xi$, this rate matches the lower bound [21, 46], and the only algorithm that achieves this optimal rate to our best knowledge is Azizian et al. [3] without an explicit $1/e$-prefactor on the right hand of Eq. (12).

We end this section with some remarks. For the results in this section, we can forgo fully optimizing the prefactor over α and simply set a step size η as in Eq. (7). Both the analyses of Theorems 3 and 4 adopt a step size of $\eta_{\mathbf{M}}/\sqrt{2}$, capped by some α-dependent threshold, due to the fact that our analysis relies heavily on the last-iterate convergence to stationarity. In the meantime, Theorem 5 does not rely on such an argument and accommodates the larger (thresholded) $\eta_{\mathbf{M}}$ as the step size. Lastly, we emphasize that the knowledge of $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ is required for the algorithm to achieve the accelerated rate. Considerations regarding such knowledge are related to the topic of adaptivity of stochastic gradient algorithms [see, e.g., 25].

4. Conclusions

We have presented an analysis of the classical Stochastic ExtraGradient (SEG) method for stochastic bilinear minimax optimization. Despite that the last iterate only contracts to a fixed neighborhood of the Nash equilibrium and the diameter of the neighborhood is independent of the step size, we show that SEG accompanied by iteration averaging converges to Nash equilibria at a sublinear rate. Moreover, the forgetting of the initialization is optimal when we use a scheduled restarting procedure in both the general and interpolation settings. Numerical experiments further validate this use of iteration averaging and restarting in the SEG setting.

Further directions for research include justification of the optimality of our convergence result, improvement of the convergence of SEG for nonlinear convex-concave optimization problems with relaxed assumptions, and connection to the Hamiltonian viewpoint for bilinear minimax optimization.

Acknowledgements

We would like to acknowledge support from the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764. Gauthier Gidel and Nicolas Le Roux are supported by Canada CIFAR AI Chairs. Nicolas Loizou acknowledges support by the IVADO Postdoctoral Funding Program. Yi Ma acknowledges support from ONR grant N00014-20-1-2002 and the joint Simons Foundation-NSF DMS grant number 2031899.

References

[1] Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *International Conference on Artificial Intelligence and Statistics*, pages 486–495. PMLR, 2019.

[2] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020.

[3] 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.

[4] Francis Bach. The "η-trick" or the effectiveness of reweighted least-squares. https://francisbach.com/the-%ce%b7-trick-or-the-effectiveness-of-reweighted-least-squares/, 2019.

[5] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

[6] Hugo Berard, Gauthier Gidel, Amjad Almahairi, Pascal Vincent, and Simon Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

[7] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[8] Radu Ioan Bot, Panayotis Mertikopoulos, Mathias Staudigl, and Phan Tu Vuong. Forward-backward-forward methods with variance reduction for stochastic variational inequalities. *arXiv preprint arXiv:1902.03355*, 2019.

[9] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[10] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, volume 32, pages 393–403, 2019.

[11] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 9236–9246, 2018.

[12] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[13] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

[14] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[15] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

[16] Ian Goodfellow. NIPS2016 Tutorial: Generative Adversarial Networks. *arXiv preprint arXiv:1701.00160*, 2016.

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[18] Robert Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

[19] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[20] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pages 16223–16234, 2020.

[21] Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

[22] Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

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

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

[25] Lihua Lei and Michael I Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020.

[26] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[27] Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic Hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

[28] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[30] Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

[31] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

[32] 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.

[33] Oskar Morgenstern and John Von Neumann. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[34] Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[35] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4): 1574–1609, 2009.

[36] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

[37] Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[38] James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, pages 1–46, 2021.

[39] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

[40] Mark Schmidt. Convergence rate of stochastic gradient with constant step size. *Technical Report, University of British Columbia*, 2014.

[41] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. On the convergence of the stochastic heavy ball method. *arXiv preprint arXiv:2006.07867*, 2020.

[42] Lloyd N Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, 1997.

[43] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[44] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[45] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32, pages 3732–3745, 2019.

[46] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.