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

Chris Junchi Li [◇,⋆] **Yaodong Yu** [◇,⋆] **Nicolas Loizou** [‡] **Gauthier Gidel** [‡]

Yi Ma [◇] **Nicolas Le Roux** [‡,□] **Michael I. Jordan** [◇]

[◇] University of California, Berkeley [‡] Mila, Université de Montréal [□] McGill University

NeurIPS OPT2021 Workshop, December 13, 2021

Introduction

- The general stochastic bilinear minimax optimization problem, also known as the bilinear saddle-point problem,

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \; \underbrace{\mathbf{x}^\top \mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y}}_{\text{coupling term}} + \underbrace{\mathbf{x}^\top \mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}}_{\text{intercept terms}}$$

Introduction

▶ The general stochastic bilinear minimax optimization problem, also known as the bilinear saddle-point problem,

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \underbrace{\mathbf{x}^\top \mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y}}_{\text{coupling term}} + \underbrace{\mathbf{x}^\top \mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}}_{\text{intercept terms}}$$

▶ SEG method composed of an extrapolation step (half-iterates) and an update step (same-sample-and-same-stepsize):

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

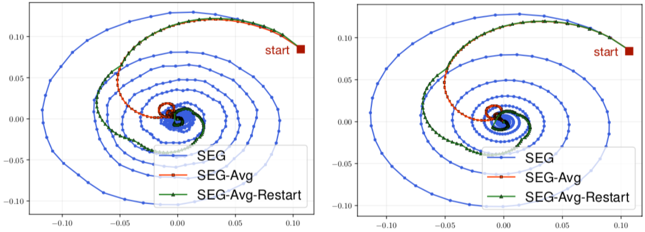
Introduction

▶ The general stochastic bilinear minimax optimization problem, also known as the bilinear saddle-point problem,

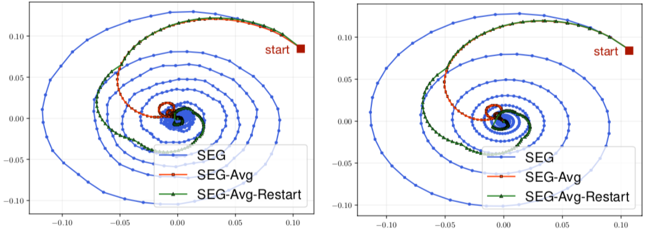
$$\min_{\mathbf{x}} \max_{\mathbf{y}} \underbrace{\mathbf{x}^\top \mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y}}_{\text{coupling term}} + \underbrace{\mathbf{x}^\top \mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}}_{\text{intercept terms}}$$

▶ SEG method composed of an extrapolation step (half-iterates) and an update step (same-sample-and-same-stepsize):

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



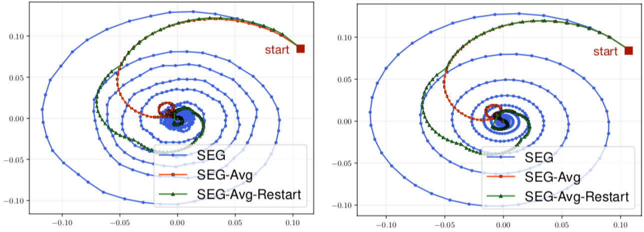
(*a*) General setting. (*b*) Interpolation setting.



(*a*) General setting. (*b*) Interpolation setting.

Contributions

- $1/\sqrt{K}$ convergence rate of SEG with iteration averaging and exponential forgetting by restarting

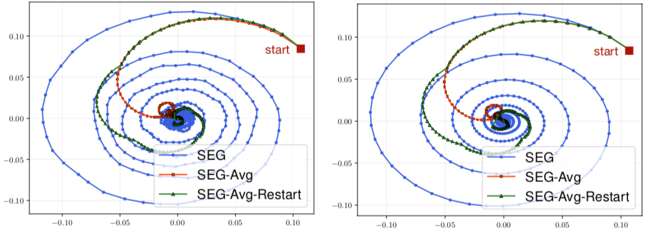


(*a*) General setting.

(*b*) Interpolation setting.

Contributions

- $1/\sqrt{K}$ convergence rate of SEG with iteration averaging and exponential forgetting by restarting
- Achieved sharp convergence rate that generalizes the full-batch version Azizian et al. (2020b) with only access to stochastic estimates



(a) General setting. (b) Interpolation setting.

Contributions

- $1/\sqrt{K}$ convergence rate of SEG with iteration averaging and exponential forgetting by restarting
- Achieved sharp convergence rate that generalizes the full-batch version Azizian et al. (2020b) with only access to stochastic estimates
- First convergence result on SEG with unbounded noise

Assumptions

- We first introduce basic setups and assumptions needed for our statement of the dynamics of SEG

- We first introduce basic setups and assumptions needed for our statement of the dynamics of SEG

Assumptions

- **(A1)** Defining $\widehat{\mathbf{M}} \equiv \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi]$ and $\mathbf{M} \equiv \mathbb{E}_\xi \mathbf{M}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top]$. There exists $\sigma_{\mathbf{B}}, \sigma_{\mathbf{B},2} \in [0, \infty)$ such that

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

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

Assumptions

▶ We first introduce basic setups and assumptions needed for our statement of the dynamics of SEG

Assumptions

- **(A1)** Defining $\widehat{\mathbf{M}} \equiv \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi]$ and $\mathbf{M} \equiv \mathbb{E}_\xi \mathbf{M}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top]$. There exists $\sigma_{\mathbf{B}}, \sigma_{\mathbf{B},2} \in [0, \infty)$ such that

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

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

- **(A2)** There exists a $\sigma_{\mathbf{g}} \in [0, \infty)$ such that

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

Assumptions

- We first introduce basic setups and assumptions needed for our statement of the dynamics of SEG

Assumptions

- **(A1)** Defining $\widehat{\mathbf{M}} \equiv \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi]$ and $\mathbf{M} \equiv \mathbb{E}_\xi \mathbf{M}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top]$. There exists $\sigma_{\mathbf{B}}, \sigma_{\mathbf{B},2} \in [0, \infty)$ such that

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

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

- **(A2)** There exists a $\sigma_{\mathbf{g}} \in [0, \infty)$ such that

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

- **(A3)** $\mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] = \mathbf{0}_n$, $\mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{y}}] = \mathbf{0}_m$ and assume independence between the stochastic matrix $\mathbf{B}_\xi$ and the vector $[\mathbf{g}_\xi^{\mathbf{x}}; \mathbf{g}_\xi^{\mathbf{y}}]$

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

Algorithm

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 $\mathsf{Epoch} \geq 1$

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

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

    if $t \in \{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]}$ **then**
        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
    end if
end for
Output: $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$

$$\bar{\mathbf{x}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K}\mathbf{x}_t \qquad \bar{\mathbf{y}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K}\mathbf{y}_t$$

Theorem 1 (SEG Averaged Iterate)

Let Assumptions hold. When the step size η is chosen as $\widehat{\eta}_{\mathbf{M}}(\alpha)$ ($\approx \frac{1}{\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ and $= \frac{1}{\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ when $\mathbf{B}_\xi$ is nonrandom), we have for all $K \geq 1$ the averaged iterate satisfies

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

where $\kappa_\zeta \equiv \frac{\sigma_{\mathbf{B}}^2 + \widehat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}$ is "effective noise condition number"

$$\bar{\mathbf{x}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K}\mathbf{x}_t \qquad \bar{\mathbf{y}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K}\mathbf{y}_t$$

Theorem 1 (SEG Averaged Iterate)

Let Assumptions hold. When the step size η is chosen as $\widehat{\eta}_{\mathbf{M}}(\alpha)$ ($\approx \frac{1}{\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ and $= \frac{1}{\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ when $\mathbf{B}_\xi$ is nonrandom), we have for all $K \geq 1$ the averaged iterate satisfies

$$\mathbb{E}\left[\|\bar{\mathbf{x}}_K\|^2 + \|\bar{\mathbf{y}}_K\|^2\right]$$

$$\leq \frac{16+8\kappa_\zeta}{(1-\alpha)\widehat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2}{(K+1)^2} + \frac{18+12\kappa_\zeta}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\sigma_{\mathbf{g}}^2}{K+1}$$

where $\kappa_\zeta \equiv \frac{\sigma_{\mathbf{B}}^2 + \widehat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}$ is "effective noise condition number"

- Achieve the optimal $O(1/\sqrt{K})$ convergence rate for the averaged iterate

$$\bar{\mathbf{x}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K}\mathbf{x}_t \qquad \bar{\mathbf{y}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K}\mathbf{y}_t$$

Theorem 1 (SEG Averaged Iterate)

Let Assumptions hold. When the step size η is chosen as $\widehat{\eta}_{\mathbf{M}}(\alpha)$ ($\approx \frac{1}{\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ and $= \frac{1}{\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ when $\mathbf{B}_\xi$ is nonrandom), we have for all $K \geq 1$ the averaged iterate satisfies

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

where $\kappa_\zeta \equiv \frac{\sigma_{\mathbf{B}}^2+\widehat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}$ is "effective noise condition number"

- Achieve the optimal $O(1/\sqrt{K})$ convergence rate for the averaged iterate
- Forgets the initialization at a polynomial rate

Theorem 2 (Scheduled Restarting)

Following the same setup as in Theorem 1, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies:

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

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

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

Theorem 2 (Scheduled Restarting)

Following the same setup as in Theorem 1, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies:

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

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

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

- Key: halt the restarting procedure once the last iterate reaches stationarity in squared Euclidean metric

Theorem 2 (Scheduled Restarting)

Following the same setup as in Theorem 1, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies:

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

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

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

- Key: halt the restarting procedure once the last iterate reaches stationarity in squared Euclidean metric
- Achieve the optimal $O(1/\sqrt{K})$ convergence rate for the averaged iterate

Theorem 2 (Scheduled Restarting)

Following the same setup as in Theorem 1, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies:

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

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

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

- Key: halt the restarting procedure once the last iterate reaches stationarity in squared Euclidean metric
- Achieve the optimal $O(1/\sqrt{K})$ convergence rate for the averaged iterate
- With the help of restarting, forgets the initialization at an exponential rate

Theorem 3 (Interpolation Setting)

Let Assumptions hold and $\sigma_{\mathbf{g}} = 0$. For the same setup as above, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies

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

where $C(\alpha)$ is defined as

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

Theorem 3 (Interpolation Setting)

Let Assumptions hold and $\sigma_{\mathbf{g}} = 0$. For the same setup as above, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies

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

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 contraction rate (in terms of the 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 terms in variance parameters of $\mathbf{B}_\xi$

An Example: Interpolation Setting

Theorem 3 (Interpolation Setting)

Let Assumptions hold and $\sigma_{\mathbf{g}} = 0$. For the same setup as above, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies

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

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 contraction rate (in terms of the 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 terms in variance parameters of $\mathbf{B}_\xi$
- Does *not* require an explicit Polyak- or Nesterov-type momentum update rule; in the case of nonrandom $\mathbf{B}_\xi$, this rate matches the lower bound (Ibrahim et al., 2020; Zhang et al., 2019)

Theorem 3 (Interpolation Setting)

Let Assumptions hold and $\sigma_{\mathbf{g}} = 0$. For the same setup as above, the output $\widehat{\mathbf{x}}_K, \widehat{\mathbf{y}}_K$ satisfies

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

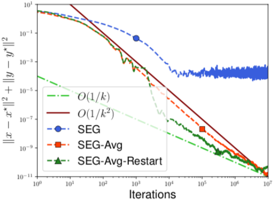
where $C(\alpha)$ is defined as

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

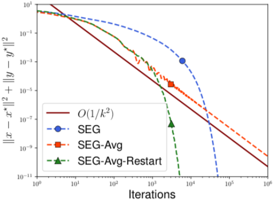
- The contraction rate (in terms of the exponent) to the Nash equilibrium $-\frac{\eta_{\mathbf{M}}^2}{4} \cdot \left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right)$ improves to $-\frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$ plus higher-order terms in variance parameters of $\mathbf{B}_\xi$
- Does *not* require an explicit Polyak- or Nesterov-type momentum update rule; in the case of nonrandom $\mathbf{B}_\xi$, this rate matches the lower bound (Ibrahim et al., 2020; Zhang et al., 2019)
- Previous algorithm achieving this optimal rate to our best knowledge is Azizian et al. (2020b) without an explicit $1/e$-prefactor

Numerical Experiments

- **Comparing SEG, SEG-Avg, and SEG-Avg-Restart**



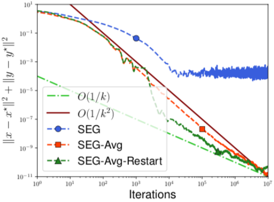
(*a*) General setting.



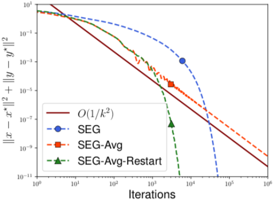
(*b*) Interpolation setting.

Numerical Experiments

- **Comparing SEG, SEG-Avg, and SEG-Avg-Restart**

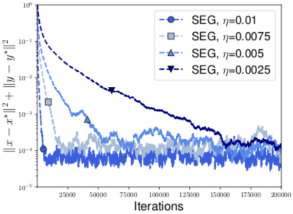


(a) General setting.

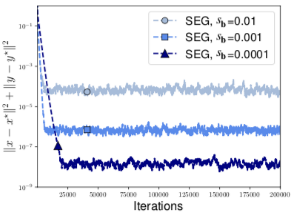


(b) Interpolation setting.

- **SEG (w/o Averaging) with Different Step Sizes and Noise Magnitudes**



(a) Different step size η.



(b) Different noise $\mathrm{std}_{\mathbf{g}}$.

Thanks!

Full version of this work: `https://arxiv.org/abs/2107.00464`