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[◇] Johns Hopkins University[◁] Mila[†] Université de Montréal[‡]
McGill University[§] Microsoft Research[□] (*Equal contribution)

Problem Setup

- **Bilinear saddle-point problem**

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

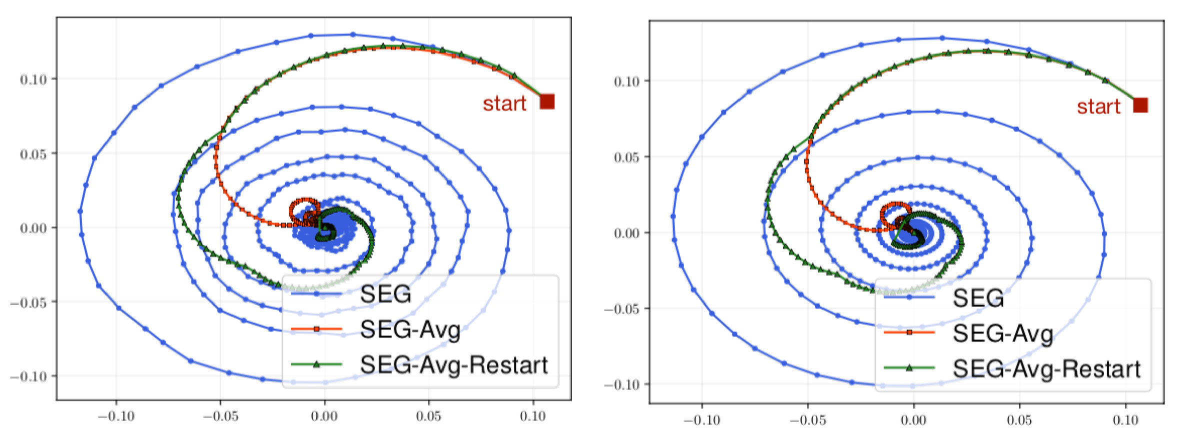
$$\min_{\mathbf{x}} \max_{\mathbf{y}} \ \mathbf{x}^\top \mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y} + \mathbf{x}^\top \mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}. \quad (1)$$

- ξ denotes the randomness associated with stochastic sampling.
- $\mathbf{g}_\xi^{\mathbf{x}}$ and $\mathbf{g}_\xi^{\mathbf{y}}$ have zero mean.
- Nash equilibrium point is $[\mathbf{x}^\star, \mathbf{y}^\star] = [\mathbf{0}, \mathbf{0}]$.

- **Stochastic Extragradient Method (SEG)**

SEG method composed of an extrapolation step (half-iterates) and an update step:

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



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

- **Contributions**
  - $1/\sqrt{K}$ convergence rate of SEG with iteration averaging and exponential forgetting by restarting.
  - Under the interpolation setting, achieved sharp convergence rate comparable with 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

- **(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

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

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

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

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

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 $\mathsf{Epoch} \ge 1$

for $t = 1, 2, \dots, K$ **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: $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$

Theoretical Results

- **Theorem 1 (SEG Averaged Iterate)**

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

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

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

- **Theorem 2 (Scheduled Restarting)**

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

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

where $\hat{K} \equiv K - K_{\mathsf{complexity}}$ is equals to

$$\frac{\text{logarithmic factor}}{\sqrt{(1-\alpha)\hat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} - C_1}.$$

- Key: halt the restarting procedure once the last iterate reaches stationarity in squared Euclidean metric.
- 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 that forgets the initialization at an exponential rate instead of the polynomial rate that is obtained without restarting.

Theoretical Results (Interpolation Setting)

- **Theorem 3 (Interpolation Setting)**

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

$$\begin{aligned}
&\mathbb{E}[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2] \\
&\le e^{-\frac{K}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} + C_2} \cdot \left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right]
\end{aligned} \quad (3)$$

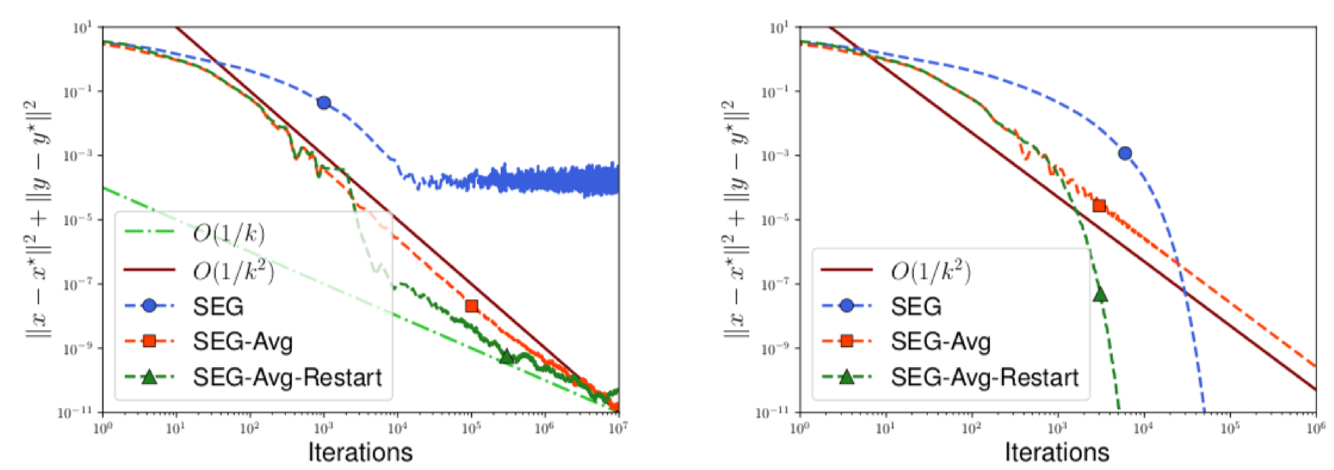
where C_2 is defined as

$$C_2 = O\left(K\bar{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\right).$$

- 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).
- The only algorithm that achieves this optimal rate to our best knowledge is Azizian et al. (2020b) without an explicit $1/e$-prefactor on the right hand of Eq. (3).

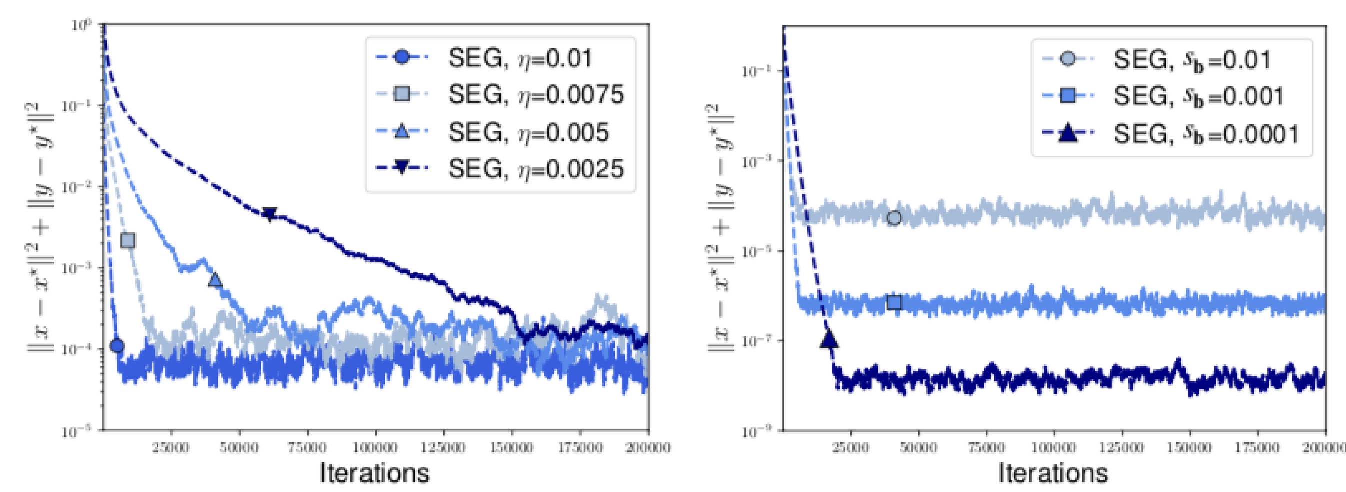
Numerical Experiments

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



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

- **Different Step Sizes and Noise Magnitudes**



(a) Different step size η. (b) Different noise $\mathrm{std}_{\mathbf{g}}$.

Paper Link

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