Stochastic Extragradient: General Analysis and Improved Rates

Eduard Gorbunov[1,2], Hugo Berard[2], Gauthier Gidel[2,3], Nicolas Loizou[4]

[1]Moscow Institute of Physics and Technology (Russia), [2]Mila & Université de Montréal (Canada), [3]Canada CIFAR AI Chair, [4]Johns Hopkins University (USA)

1. Preliminaries

Problem: unconstrained variational inequality problem (VIP)

$$\text{find } x^* \in \mathbb{R}^d \quad \text{such that} \quad F(x^*) = 0$$

Assumptions:

- Lipschitzness $\quad \|F(x) - F(x')\| \le L\|x - x'\| \quad \forall x, x' \in \mathbb{R}^d$
- Quasi-strong monotonicity $\quad \langle F(x), x - x^*\rangle \ge \mu\|x - x^*\|^2 \quad \forall x \in \mathbb{R}^d$
- $F(x) = \mathbb{E}\left[F_\xi(x)\right]$ or $F(x) = \frac{1}{n}\sum\limits_{i=1}^n F_i(x)$

2. Stochastic Extragradient

Extragradient method (EG) is one of the most popular methods for VIPs

$$x^{k+1} = x^k - \gamma F\left(x^k - \gamma F\left(x^k\right)\right)$$

There are two main options in the stochastic case:

- **Same-Sample Stochastic Extragradient (S-SEG)**

$$x^{k+1} = x^k - \gamma F_{\xi^k}\left(x^k - \gamma F_{\xi^k}\left(x^k\right)\right)$$

- **Independent-Samples Stochastic Extragradient (I-SEG)**

$$x^{k+1} = x^k - \gamma F_{\xi_2^k}\left(x^k - \gamma F_{\xi_1^k}\left(x^k\right)\right)$$

✖ State-of-the-art results [1, 2, 3] are derived via different proof techniques and rely on different assumptions

✖ Some interesting directions are unexplored, e.g., non-uniform sampling

3. Our Contributions

- New theoretical framework for the analysis of SEG
- Our analysis recovers tight guarantees for several known special cases
- New results for known methods, new variants of SEG
- Weak assumptions in the special cases
- Numerical experiments that corroborate our theory

4. General Analysis of SEG

stochastic estimator

Generalized update rule: $$x^{k+1} = x^k - \gamma_{\xi^k} g_{\xi^k}\left(x^k\right)$$

randomness/stochasticity at step *k*

Key assumption: there exist non-negative constants $A, B, C,$ $D_1, D_2 \ge 0,\ \rho \in [0,1]$ and (possibly random) sequence $\{G_k\}$ such that

$$\mathbb{E}_{\xi^k}\left[\gamma_{\xi^k}^2\left\|g_{\xi^k}\left(x^k\right)\right\|^2\right] \le 2AP_k + C\left\|x^k - x^*\right\|^2 + D_1$$

$$P_k \ge \rho\left\|x^k - x^*\right\|^2 + BG_k - D_2$$

$$P_k = \mathbb{E}_{\xi^k}\left[\gamma_{\xi^k}\left\langle g_{\xi^k}\left(x^k\right), x^k - x^*\right\rangle\right]$$

General convergence result: let the key assumption hold with $A \le 1/2$ and $\rho > C \ge 0$. Then the iterates of generalized SEG satisfy

$$\mathbb{E}\left[\left\|x^K - x^*\right\|^2\right] \le (1 + C - \rho)^K\left\|x^0 - x^*\right\|^2 + \frac{D_1 + D_2}{\rho - C}$$

5. Same-Sample SEG

For simplicity, consider a **finite-sum case:** $F(x) = \frac{1}{n}\sum\limits_{i=1}^n F_i(x)$

Assumptions: $$\|F_i(x) - F_i(y)\| \le L_i\|x - y\|$$

$$\langle F_i(x) - F_i(x^*), x - x^*\rangle \ge \mu_i\|x - x^*\|^2$$

allowed to be negative

$$\bar{\mu} = \frac{1}{n}\sum_{i:\mu_i \ge 0}\mu_i + \frac{4}{n}\sum_{i:\mu_i < 0}\mu_i \ge 0$$

S-SEG: $$x^{k+1} = x^k - \gamma_{2,\xi^k}F_{\xi^k}\left(x^k - \gamma_{1,\xi^k}F_{\xi^k}\left(x^k\right)\right)$$

Uniform sampling (US)

$$\mathbb{P}\left[\xi^k = i\right] = \frac{1}{n}$$

$$\gamma_{1,\xi^k} = \gamma$$

Importance sampling (IS)

$$\mathbb{P}\left[\xi^k = i\right] = \frac{L_i}{\sum_{j=1}^n L_j}$$

$$\gamma_{1,\xi^k} = \frac{\gamma\overline{L}}{L_{\xi^k}},\ \overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$$

$$\gamma_{2,\xi^k} = \alpha\gamma_{1,\xi^k},\ \alpha \le 1/4$$

We prove that both options (an much more) fit our framework

6. Independent-Samples SEG

Assumption: $$\mathbb{E}_\xi\left[\|F_\xi(x) - F(x)\|^2\right] \le \delta\|x - x^*\|^2 + \sigma^2$$

- When *δ = 0* it recovers uniformly bounded variance assumption

I-SEG: $$x^{k+1} = x^k - \gamma F_{\xi_2^k}\left(x^k - \gamma F_{\xi_1^k}\left(x^k\right)\right)$$

- ξ_1^k, ξ_2^k are sampled independently

We prove that I-SEG in this setup fit our framework as well

7. Comparison with SOTA Results

Our general theorem implies the results highlighted in green

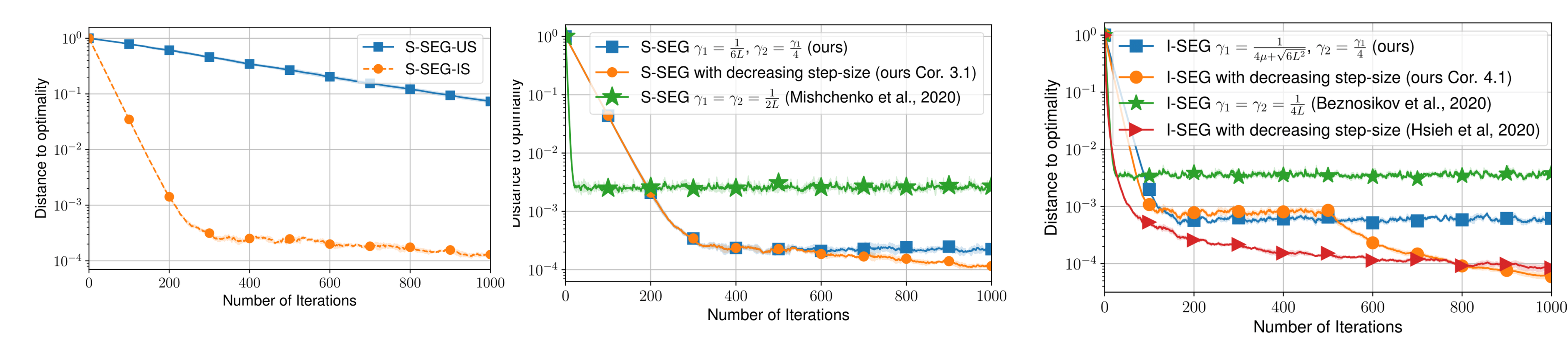
Method	Citation	Convergence Rate for Constant Stepsize	Convergence Rate for Diminishing Stepsize
S-SEG-US	(Mishchenko et al., 2020)	$(1-\gamma\mu_{\min})^K R_0^2 + \frac{\gamma\sigma_{\text{US}*}^2}{\mu_{\min}}$	$\frac{L_{\max}R_0^2}{\mu_{\min}}\exp\left(-\frac{\mu_{\min}K}{L_{\max}}\right) + \frac{\sigma_{\text{US}*}^2}{\mu_{\min}^2 K}$
	This paper	$(1-\gamma\bar{\mu})^K R_0^2 + \frac{\gamma\sigma_{\text{US}*}^2}{\bar{\mu}}$	$\frac{L_{\max}R_0^2}{\bar{\mu}}\exp\left(-\frac{\bar{\mu}K}{L_{\max}}\right) + \frac{\sigma_{\text{US}*}^2}{\bar{\mu}^2 K}$
S-SEG-IS	This paper	$(1-\gamma\bar{\mu})^K R_0^2 + \frac{\gamma\sigma_{\text{IS}*}^2}{\bar{\mu}}$	$\frac{\overline{L}R_0^2}{\bar{\mu}}\exp\left(-\frac{\bar{\mu}K}{\overline{L}}\right) + \frac{\sigma_{\text{IS}*}^2}{\bar{\mu}^2 K}$
I-SEG	(Hsieh et al., 2020)	$(1-\gamma_1\gamma_2\mu^2)^K R_0^2 + \frac{C\sigma^2}{\mu^2 b}$, $C = \gamma_1 L(1+\gamma_1 L) + \frac{\gamma_2}{\gamma_1}$	$\frac{L^2\sigma^2}{\mu^4 b K^{1/3}}$
	(Beznosikov et al., 2020)	$(1-\gamma\mu)^K R_0^2 + \frac{\gamma\sigma^2}{\mu b}$	$R_0^2\exp\left(-\frac{\mu K}{L}\right) + \frac{\sigma^2}{\mu^2 b K}$
	This paper	$(1-\gamma\mu)^K R_0^2 + \frac{\gamma\sigma^2}{\mu b}$	$\kappa R_0^2\exp\left(-\frac{K}{\kappa}\right) + \frac{\sigma^2}{\mu^2 b K}$, $\kappa = \max\left\{\frac{\delta}{\mu^2 b}, \frac{L+\sqrt{\delta/b}}{\mu}\right\}$

Notation: $\mu_{\min} = \min_{i\in[n]}\mu_i \quad L_{\max} = \max_{i\in[n]} L_i$

$$\sigma_{\text{US}*}^2 = \frac{1}{n}\sum_{i=1}^n\|F_i(x^*)\|^2 \qquad \sigma_{\text{IS}*}^2 = \frac{1}{n}\sum_{i=1}^n\frac{\overline{L}}{L_i}\|F_i(x^*)\|^2$$

8. Numerical Experiments

We tested the methods on qudratic unconstrained games. The first experiment shows the benefits of importance sampling compared to the uniform sampling, the last two experiments compare our results with theoretical SOTA.



References

[1] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtarik, and Y. Malitsky. **Revisiting stochastic extragradient**. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of Proceedings of Machine Learning Research, pages 4573–4582. PMLR, 26–28 Aug 2020.

[2] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. **Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling**. *Advances in Neural Information Processing Systems*, 33, 2020.

[3] A. Beznosikov, V. Samokhin, and A. Gasnikov. **Distributed saddle-point problems: Lower bounds, optimal algorithms and federated GANs**. arXiv preprint ArXiv:2010.13112, 2020.