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, Russian Federation
[2] Mila, Université de Montréal, Canada
[3] Canada CIFAR AI Chair
[4] Johns Hopkins University, Baltimore, USA

AISTATS 2022

March 29, 2022

Short Summary of Our Work

- We develop a new theoretical framework for the analysis of SEG

Short Summary of Our Work

- We develop a new theoretical framework for the analysis of SEG
  - Unified assumption on the stochastic estimator, stepsizes, and the problem itself
  - Same-Sample Stochastic Extragradient (S-SEG) and Independent-Samples Stochastic Extragradient (I-SEG) fit the assumption
  - General convergence result under this assumption

Short Summary of Our Work

- We develop a new theoretical framework for the analysis of SEG
  - Unified assumption on the stochastic estimator, stepsizes, and the problem itself
  - Same-Sample Stochastic Extragradient (S-SEG) and Independent-Samples Stochastic Extragradient (I-SEG) fit the assumption
  - General convergence result under this assumption
- Our convergence guarantees give tight rates for several well-known special cases

Short Summary of Our Work

- We develop a new theoretical framework for the analysis of SEG
    - Unified assumption on the stochastic estimator, stepsizes, and the problem itself
    - Same-Sample Stochastic Extragradient (S-SEG) and Independent-Samples Stochastic Extragradient (I-SEG) fit the assumption
    - General convergence result under this assumption
- Our convergence guarantees give tight rates for several well-known special cases
- We obtain new results for known methods and also propose new variants of SEG

Short Summary of Our Work

- We develop a new theoretical framework for the analysis of SEG
  - Unified assumption on the stochastic estimator, stepsizes, and the problem itself
  - Same-Sample Stochastic Extragradient (S-SEG) and Independent-Samples Stochastic Extragradient (I-SEG) fit the assumption
  - General convergence result under this assumption
- Our convergence guarantees give tight rates for several well-known special cases
- We obtain new results for known methods and also propose new variants of SEG
- Weak assumptions in the special cases

Outline

1. Preliminaries

2. Unified Analysis

Problem

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

Problem

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

- Operator $F(x) : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{1}$$

Problem

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

- Operator $F(x) : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{1}$$

- Operator $F(x)$ is μ-quasi strongly monotone, i.e., for $\mu \geq 0$ and for all $x \in \mathbb{R}^d$

$$\langle F(x), x - x^* \rangle \geq \mu\|x - x^*\|^2. \tag{2}$$

Problem

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

- Operator $F(x) : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{1}$$

- Operator $F(x)$ is μ-quasi strongly monotone, i.e., for $\mu \geq 0$ and for all $x \in \mathbb{R}^d$

$$\langle F(x), x - x^* \rangle \geq \mu\|x - x^*\|^2. \tag{2}$$

  We assume that x^* is unique
- Operator $F(x)$ can be
  - expectation $F(x) = \mathbb{E}[F_\xi(x)]$
  - finite-sum $F(x) = \frac{1}{n}\sum\limits_{i=1}^{n} F_i(x)$

Extragradient Method

Extragradient method (EG) [Korpelevich, 1976]:

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

Stochastic Extragradient Method

- Independent-Samples Stochastic Extragradient method (I-SEG) [Nemirovski, 2004]

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

Stochastic Extragradient Method

- Independent-Samples Stochastic Extragradient method (I-SEG) [Nemirovski, 2004]

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

  - ξ_1^k and ξ_2^k are sampled independently

Stochastic Extragradient Method

- Independent-Samples Stochastic Extragradient method (I-SEG) [Nemirovski, 2004]

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

  - ξ_1^k and ξ_2^k are sampled independently

- Same-Sample Stochastic Extragradient method (S-SEG) [Mishchenko et al., 2020]

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

Analysis of Stochastic Extragradient Method

- State-of-the-art theoretical results on SEG [Mishchenko et al., 2020, Beznosikov et al., 2020, Hsieh et al., 2020]

Analysis of Stochastic Extragradient Method

- State-of-the-art theoretical results on SEG [Mishchenko et al., 2020, Beznosikov et al., 2020, Hsieh et al., 2020]
  - are obtained via different proof techniques
  - rely on different assumptions

Analysis of Stochastic Extragradient Method

- State-of-the-art theoretical results on SEG [Mishchenko et al., 2020, Beznosikov et al., 2020, Hsieh et al., 2020]
  - are obtained via different proof techniques
  - rely on different assumptions
- Some interesting directions are unexplored including

Analysis of Stochastic Extragradient Method

- State-of-the-art theoretical results on SEG [Mishchenko et al., 2020, Beznosikov et al., 2020, Hsieh et al., 2020]
  - are obtained via different proof techniques
  - rely on different assumptions
- Some interesting directions are unexplored including
  - non-uniform sampling
  - different stepsizes for extrapolation and update without strong assumptions on them

Analysis of Stochastic Extragradient Method

- State-of-the-art theoretical results on SEG [Mishchenko et al., 2020, Beznosikov et al., 2020, Hsieh et al., 2020]
  - are obtained via different proof techniques
  - rely on different assumptions
- Some interesting directions are unexplored including
  - non-uniform sampling
  - different stepsizes for extrapolation and update without strong assumptions on them

A single unifying framework allowing to tighten known results and to obtain new ones is required

Generalized Update Rule

$$x^{k+1} = x^k - \gamma_{\xi^k} g_{\xi^k}(x^k), \tag{3}$$

- $g_{\xi^k}(x^k)$ – some stochastic operator evaluated at point x^k
- ξ^k – the randomness/stochasticity appearing at iteration k (e.g., the sample used at step k)
- γ_{ξ^k} – the stepsize that is allowed to depend on ξ^k

Key Assumption

Assumption 1

We assume that there exist non-negative constants $A, B, C, D_1, D_2 \geq 0$, $\rho \in [0, 1]$, and (possibly random) non-negative sequence $\{G_k\}_{k\geq 0}$ such that

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

Key Assumption

Assumption 1

We assume that there exist non-negative constants $A, B, C, D_1, D_2 \geq 0$, $\rho \in [0,1]$, and (possibly random) non-negative sequence $\{G_k\}_{k\geq 0}$ such that

$$\mathbb{E}_{\xi^k}\left[\gamma_{\xi^k}^2\|g_{\xi^k}(x^k)\|^2\right] \leq 2AP_k + C\|x^k - x^*\|^2 + D_1, \tag{4}$$

$$P_k \geq \rho\|x^k - x^*\|^2 + BG_k - D_2, \tag{5}$$

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

General Convergence Result

Theorem 1

Let Assumption 1 hold with $A \leq 1/2$ and $\rho > C \geq 0$. Then, the iterates of SEG given by (3) satisfy

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

General Convergence Result

Theorem 1

Let Assumption 1 hold with $A \leq 1/2$ and $\rho > C \geq 0$. Then, the iterates of SEG given by (3) satisfy

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

In one theorem, we either recover the best-known results for SEG or improve them

Achieved Results

- Better rates for S-SEG:

Achieved Results

- Better rates for S-SEG:
    - We improve the result by Mishchenko et al. [2020] for S-SEG with uniform sampling

Achieved Results

- Better rates for S-SEG:
  - We improve the result by Mishchenko et al. [2020] for S-SEG with uniform sampling
  - We derive better rates for other sampling strategies including importance sampling and mini-batching without replacement

Achieved Results

- Better rates for S-SEG:
  - We improve the result by Mishchenko et al. [2020] for S-SEG with uniform sampling
  - We derive better rates for other sampling strategies including importance sampling and mini-batching without replacement
- Better rates for I-SEG:

Achieved Results

- Better rates for S-SEG:
  - We improve the result by Mishchenko et al. [2020] for S-SEG with uniform sampling
  - We derive better rates for other sampling strategies including importance sampling and mini-batching without replacement
- Better rates for I-SEG:
  - We improve the result by Hsieh et al. [2020]

Achieved Results

- Better rates for S-SEG:
    - We improve the result by Mishchenko et al. [2020] for S-SEG with uniform sampling
    - We derive better rates for other sampling strategies including importance sampling and mini-batching without replacement
- Better rates for I-SEG:
    - We improve the result by Hsieh et al. [2020]
    - We generalize the result by Beznosikov et al. [2020]

In the Paper We Also Have

- Results for S-SEG in the case of Arbitrary Sampling
- Results for the case when $\mu = 0$
- Numerical experiments corroborating our theorerical findings
- Link to the code:
  `https://github.com/hugobb/Stochastic-Extragradient`

References I

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

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.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

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.

References II

A. 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.