Stochastic Extragradient with Flip-Flop Anchoring:
Provable Improvements

Jiseok Chae[1] Chulhee Yun[2] Donghwan Kim[1]

[1]Department of Mathematical Sciences, KAIST

[2]Kim Jaechul Graduate School of AI, KAIST

NeurIPS 2024

Minimax Problems

We consider unconstrained minimax problems with a *finite-sum* structure:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \; f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x}, \boldsymbol{y}).$$

Very versatile, and has many ML applications:

- Generative Adversarial Networks
- Consistency Trajectory Models
- Sharpness-aware Minimization
- Computing Optimal Transport Maps
- ...

Minimax Problems

We consider unconstrained minimax problems with a *finite-sum* structure:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \; f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{x}, \boldsymbol{y}).$$

Denote both min. and max. variables at once by $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y})$.

The *saddle gradient*

$$\boldsymbol{F}(\boldsymbol{z}) = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$$

is more natural than ∇f in minimax problems.

The Extragradient Method

The *gradient descent-ascent* (GDA) method

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \boldsymbol{F}(\boldsymbol{z}_k) \qquad \text{or} \qquad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \eta_k \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \eta_k \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \end{aligned}$$

already does not work for simple convex-concave problems.

The *extragradient* (EG) method (Korpelevich, 1976)

$$\begin{aligned} \boldsymbol{w}_k &= \boldsymbol{z}_k - \eta_k \boldsymbol{F}(\boldsymbol{z}_k) \\ \boldsymbol{z}_{k+1} &= \boldsymbol{z}_k - \eta_k \boldsymbol{F}(\boldsymbol{w}_k) \end{aligned} \qquad \text{or} \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \boldsymbol{F}(\boldsymbol{z}_k - \eta_k \boldsymbol{F}(\boldsymbol{z}_k))$$

on the other hand, **works** on convex-concave problems.

Stochastic Extragradient?

Unlike GDA vs EG, the *stochastic* EG (SEG)

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \boldsymbol{F}_{i(k)}(\boldsymbol{z}_k - \eta_k \boldsymbol{F}_{i(k)}(\boldsymbol{z}_k))$$

does not show a clear advantage in convex-concave problems over GDA.

Even if we additionally assume each f_i are also convex-concave, convergence rates typically look something like:

$$\min_{k=0,1,\ldots,K} \|\boldsymbol{F}\boldsymbol{z}_k\|^2 \leq \mathcal{O}\left(\frac{1}{\mathrm{poly}(K)}\right) + \text{(abs. const.)}$$

* The constant term can be decreased only with strong additional assumptions, such as for example, increasing the batch size every iteration.

For minimization problems...

With-replacement stochastic gradient descent (SGD) works well.

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f_{i(k)}(\boldsymbol{x}_k), \qquad i(k) \sim \mathrm{Unif}(\{1, \ldots, n\})$$

In practice, *shuffling based* SGD is used.

Random reshuffling (RR): in the kth epoch, a permutation $\tau_k : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is chosen randomly, and

$$\begin{aligned} \boldsymbol{x}_i^k &= \boldsymbol{x}_{i-1}^k - \eta_k \nabla f_{\tau_k(i)}(\boldsymbol{x}_{i-1}^k), \qquad i = 1, \ldots, n, \\ \boldsymbol{x}_0^{k+1} &\leftarrow \boldsymbol{x}_n^k. \end{aligned}$$

For minimization problems...

With-replacement stochastic gradient descent (SGD) works well.

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f_{i(k)}(\boldsymbol{x}_k), \qquad i(k) \sim \mathrm{Unif}(\{1, \ldots, n\})$$

In practice, *shuffling based* SGD is used.

Flip-flop sampling (FF) (Rajput et al., 2022) goes one step even further in search for a better sampling scheme: in the kth epoch, a permutation $\tau_k : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is chosen randomly, and

$$\begin{aligned} \boldsymbol{x}_i^k &= \boldsymbol{x}_{i-1}^k - \eta_k \nabla f_{\tau_k(i)}(\boldsymbol{x}_{i-1}^k), && i = 1, \ldots, n, \\ \boldsymbol{x}_i^k &= \boldsymbol{x}_{i-1}^k - \eta_k \nabla f_{\tau_k(2n+1-i)}(\boldsymbol{x}_{i-1}^k), && i = n+1, \ldots, 2n, \\ \boldsymbol{x}_0^{k+1} &\leftarrow \boldsymbol{x}_{2n}^k. \end{aligned}$$

For minimization problems...

With-replacement stochastic gradient descent (SGD) works well.

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f_{i(k)}(\boldsymbol{x}_k), \qquad i(k) \sim \mathrm{Unif}(\{1, \ldots, n\})$$

In practice, *shuffling based* SGD is used.

In terms of convergence rates,

- RR is in general faster than with-replacement SGD.

  (Ahn et al., 2020; Mishchenko et al., 2020)

- If all f_i are quadratic functions then FF is even faster, thanks to the stochastic error term being smaller. (Rajput et al., 2022)

Our contributions

- Stochastic EG indeed does not work on convex-concave problems. Shuffling does not resolve the problem.
  - An explicit counterexample with divergent iterates
- On top of FF, adding a simple *anchoring* step

$$z_0^{k+1} \leftarrow \frac{z_{2n}^k + z_0^k}{2}$$

  reduces the stochastic error by an order of magnitude (w.r.t. stepsize), finally allowing a convergence rate of $\tilde{\mathcal{O}}(1/k^{1/3})$.
- The reduced error also benefits the convergence on strongly-convex-strongly-concave problems, enjoying a rate of $\tilde{\mathcal{O}}(1/nk^4)$.
  - Without anchoring (*i.e.*, with-replacement sampling or RR only), the convergence rate is at best $\Omega(1/nk^3)$.

Algorithm

Stochastic Extragradient with Flip-Flop Anchoring (SEG-FFA)

For each $k = 0, 1, \ldots:$ `# epoch level outer loop`

$\tau_k \sim \mathrm{Unif}(\mathfrak{S}_n)$ `# sample random permutation`

For each $i = 1, \ldots, n:$ `# flip`

$$\boldsymbol{z}_i^k = \boldsymbol{z}_{i-1}^k - \eta_k \boldsymbol{F}_{\tau_k(i)} \left(\boldsymbol{z}_{i-1}^k - \frac{\eta_k}{2} \boldsymbol{F}_{\tau_k(i)}(\boldsymbol{z}_{i-1}^k) \right)$$

For each $i = n+1, \ldots, 2n:$ `# flop`

$$\boldsymbol{z}_i^k = \boldsymbol{z}_{i-1}^k - \eta_k \boldsymbol{F}_{\tau_k(2n+1-i)} \left(\boldsymbol{z}_{i-1}^k - \frac{\eta_k}{2} \boldsymbol{F}_{\tau_k(2n+1-i)}(\boldsymbol{z}_{i-1}^k) \right)$$

$$\boldsymbol{z}_0^{k+1} \leftarrow \frac{\boldsymbol{z}_{2n}^k + \boldsymbol{z}_0^k}{2}$$ `# anchoring`

Thank you for your attention.

Visit us at the Poster Session!

References I

Ahn, Kwangjun, Chulhee Yun, and Suvrit Sra (2020). "SGD with shuffling: optimal rates without component convexity and large epoch requirements". In: *Advances in Neural Information Processing Systems* 33, pp. 17526–17535.

Korpelevich, Galina M (1976). "The extragradient method for finding saddle points and other problems". In: *Matecon* 12, pp. 747–756.

Mishchenko, Konstantin, Ahmed Khaled, and Peter Richtárik (2020). "Random reshuffling: Simple analysis with vast improvements". In: *Advances in Neural Information Processing Systems* 33, pp. 17309–17320.

Rajput, Shashank, Kangwook Lee, and Dimitris Papailiopoulos (2022). "Permutation-Based SGD: Is Random Optimal?" In: *International Conference on Learning Representations*.