Revisiting Stochastic Extragradient

Dmitry Kovalev
joint work with Konstantin Mishchenko, Egor Shulgin,
Peter Richtárik and Yura Malitsky

ICCOPT
August 5, 2019

Variational Inequality

Find point $x^* \in \mathcal{K}$ satisfying

$$g(x) - g(x^*) + \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathcal{K}, \tag{1}$$

- $\mathcal{K} \subset \mathbb{R}^d$ is a convex set,
- $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semi-continuous convex function,
- $F\colon \mathcal{K} \to \mathbb{R}^d$ is monotone operator, i.e. $\langle F(x) - F(y), x - y \rangle \geq 0$ for all $x, y \in \mathcal{K}$.

Variational Inequality

Find point $x^* \in \mathcal{K}$ satisfying

$$g(x) - g(x^*) + \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathcal{K}, \tag{1}$$

- $\mathcal{K} \subset \mathbb{R}^d$ is a convex set,
- $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semi-continuous convex function,
- $F : \mathcal{K} \to \mathbb{R}^d$ is monotone operator, i.e. $\langle F(x) - F(y), x - y \rangle \geq 0$ for all $x, y \in \mathcal{K}$.

Stochastic setting:

$$F(x) = \mathbb{E}_\xi \left[F(x; \xi) \right]. \tag{2}$$

Examples

- Convex minimization:

$$\min_{x \in \mathcal{X}} f(x), \tag{3}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a convex set, $f : \mathcal{X} \to \mathbb{R}$ is a convex function.

$$F(x) = \nabla f(x).$$

Examples

- Convex minimization:

$$\min_{x \in \mathcal{X}} f(x), \tag{3}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a convex set, $f : \mathcal{X} \to \mathbb{R}$ is a convex function.

$$F(x) = \nabla f(x).$$

- Convex-concave saddle point problem:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y), \tag{4}$$

where $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ are convex sets, $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex in x and concave in y.

$$F(x) = \begin{bmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{bmatrix}.$$

Extragradient Method

Algorithm 1 Extragradient Method for Variational Inequalities.

Parameters: $x^0 \in \mathcal{K}$, stepsize $\eta > 0$
for $t = 0, 1, 2, \dots$ **do**
    $y^t = \mathrm{prox}_{\eta g}\left(x^t - \eta F(x^t)\right)$
    $x^{t+1} = \mathrm{prox}_{\eta g}\left(x^t - \eta F(y^t)\right)$
end for

Stochastic Extragradient Method

Algorithm 2 Stochastic Extragradient Method for Variational Inequalities.

Parameters: $x^0 \in \mathcal{K}$, stepsize $\eta > 0$
for $t = 0, 1, 2, \ldots$ **do**
    Sample ξ^t
    $y^t = \mathrm{prox}_{\eta g}\left(x^t - \eta F(x^t; \xi^t)\right)$
    $x^{t+1} = \mathrm{prox}_{\eta g}\left(x^t - \eta F(y^t; \xi^t)\right)$
end for

Convergence

Theorem (strongly-monotone case)

Assume that g is a μ-strongly convex function, operator $F(\cdot;\xi)$ is almost surely monotone and L-Lipschitz, and that its variance at the optimum x^* is bounded, i.e.

$$\mathbb{E}\|F(x^*;\xi) - F(x^*)\|^2 \leq \sigma^2.$$

Then, for any $\eta \leq 1/(2L)$

$$\mathbb{E}\|x^t - x^*\|^2 \leq (1 - 2\eta\mu/3)^t \|x^0 - x^*\|^2 + 3\eta\sigma^2/\mu.$$

Convergence

Theorem (weakly-monotone case)

Let g be a convex function, $F(\cdot;\xi)$ be monotone and L-Lipschitz almost surely. Then, the iterates of Algorithm 2 with stepsize $\eta = \mathcal{O}(1/(\sqrt{t}L))$ satisfy for any set $\mathcal{X}$

$$\sup_{x\in\mathcal{X}} \left\{ g(\hat{x}^t) - g(x) + \left\langle F(x), \hat{x}^t - x \right\rangle \right\} \le \frac{1}{\sqrt{t}L} \sup_{x\in\mathcal{X}} \left\{ \frac{L^2}{2}\|x^0 - x\|^2 + \sigma_x^2 \right\}.$$

where $\hat{x}^t = \frac{1}{t}\sum_{k=0}^t y^k$ and $\sigma_x^2 \stackrel{def}{=} \mathbb{E}\|F(x) - F(x;\xi)\|^2$, i.e. σ_x^2 is the variance of F at point x.

Bilinear Min-Max Problem

$$\min_x \max_y f(x,y) = x^\top \mathbf{B} y + a^\top x + b^\top y, \tag{5}$$

where $\mathbf{B}$ is a full rank square matrix.

Algorithm 3 The extragradient method for min-max problems.

Require: Stepsizes η_1, η_2, initial vectors x^0, y^0

1: **for** $t = 0, 1, \ldots$ **do**

2: $\quad u^t = x^t - \eta_1 \nabla_x f(x^t, y^t)$

3: $\quad v^t = y^t + \eta_1 \nabla_y f(x^t, y^t)$

4: $\quad x^{t+1} = x^t - \eta_2 \nabla_x f(u^t, v^t)$

5: $\quad y^{t+1} = y^t + \eta_2 \nabla_y f(u^t, v^t)$

6: **end for**

Bilinear Min-Max Problem

Theorem

Let f be bilinear with a full-rank matrix $\mathbf{B}$ and apply Algorithm 3 to it. Choose any η_1 and η_2 such that $\eta_2 < 1/\sigma_{\max}(\mathbf{B})$ and $\eta_1\eta_2 < 2/\sigma_{\max}(\mathbf{B})^2$, then the rate is

$$\|x^t - x^*\|^2 + \|y^t - y^*\|^2 \le \rho^{2t}(\|x^0 - x^*\|^2 + \|y^0 - y^*\|^2),$$

where $\rho \overset{def}{=} \max\{(1-\eta_1\eta_2\sigma_{\max}(\mathbf{B})^2)^2+\eta_2^2\sigma_{\max}(\mathbf{B})^2, (1-\eta_1\eta_2\sigma_{\min}(\mathbf{B})^2)^2+\eta_2^2\sigma_{\min}(\mathbf{B})^2\}.$

Bilinear Min-Max Problem

Corollary

Under the same assumption as in Theorem 3, consider two choices of stepsizes:

1. *if* $\eta_1 = \eta_2 = 1/(\sqrt{2}\sigma_{\max}(\mathbf{B}))$ *we get*

$$\|x^t - x^*\|^2 + \|y^t - y^*\|^2 \leq \left(1 - \sigma_{\min}(\mathbf{B})^2/6\sigma_{\max}(\mathbf{B})^2\right)^{2t} \left(\|x^0 - x^*\|^2 + \|y^0 - y^*\|^2\right),$$

2. *if* $\sigma_{\min}(\mathbf{B}) > 0$, *and* $\eta_1 = \kappa/(\sqrt{2}\sigma_{\max}(\mathbf{B})^2)$, $\eta_2 = 1/(\sqrt{2}\kappa\sigma_{\max}(\mathbf{B})^2)$ *with* $\kappa \overset{def}{=} \sigma^2_{\min(\mathbf{B})}/\sigma^2_{\max(\mathbf{B})}$, *then the rate is*

$$\|x^t - x^*\|^2 + \|y^t - y^*\|^2 \leq \left(1 - \sigma_{\min}(\mathbf{B})^2/4\sigma_{\max}(\mathbf{B})^2\right)^{2t} \left(\|x^0 - x^*\|^2 + \|y^0 - y^*\|^2\right).$$

Non-convex minimization

$$\min_x \mathbb{E}_\xi f(x;\xi), \tag{6}$$

where $f(\cdot;\xi)$ is smooth but potentially non-convex function.

Assumption (bounded variance)

There exists a constant $\sigma > 0$ such that for all x it holds

$$\mathbb{E}\|\nabla f(x;\xi) - \nabla f(x)\|^2 \leq \sigma^2.$$

Non-convex minimization

Theorem

Choose $\eta \le \frac{1}{4L}$ and apply extragradient to (6). Then, its iterates satisfy

$$\mathbb{E}\|\nabla f(\hat{x}^t)\|^2 \le \frac{5}{\eta t}(f(x^0) - f^*) + 11\eta L\sigma^2,$$

where $\hat{x}^t$ is sampled uniformly from $\{x^0, \ldots, x^{t-1}\}$ and $f^* = \inf_x f(x)$.

Corollary

If we choose $\eta = \Theta\left(1/(L\sqrt{t})\right)$, then the rate is $O\left((f(x^0)-f^*)/\sqrt{t} + \sigma^2/\sqrt{t}\right)$, which is the same as the rate of SGD under our assumptions.

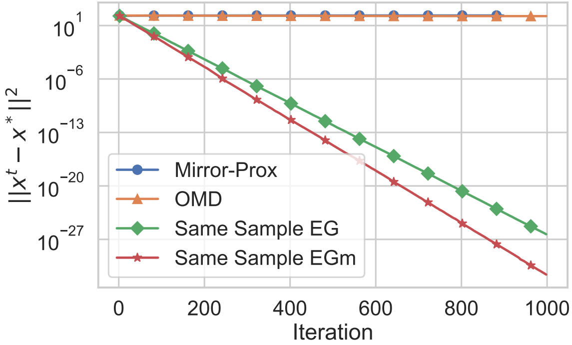


Figure: Comparison of using independent samples and averaging as suggested by [Juditsky et al., 2011] and the same sample as proposed in this work. The problem here is the sum of randomly sampled matrices $\min_x \max_y \sum_{i=1}^n x^\top \mathbf{B}_i y$. Since at point (x^*, y^*) the noise is equal 0, the convergence of Algorithm 3 is linear unlike the slow rate of [Juditsky et al., 2011]. 'EGm' is the version with negative momentum equal $\beta = -0.3$.

Experiments: Generating Mixture of Gaussians

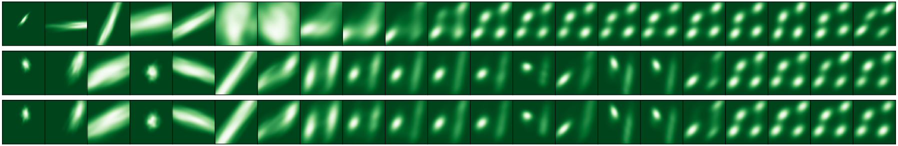


Figure: Top line: extragradient with the same sample. Middle line: gradient descent-ascent. Bottom line: extragradient with different samples. Since the same seed was used for all methods, the former two methods performed extremely similarly, although when zooming it should be clear that their results are slightly different.

Experiments: Adam vs ExtraAdam

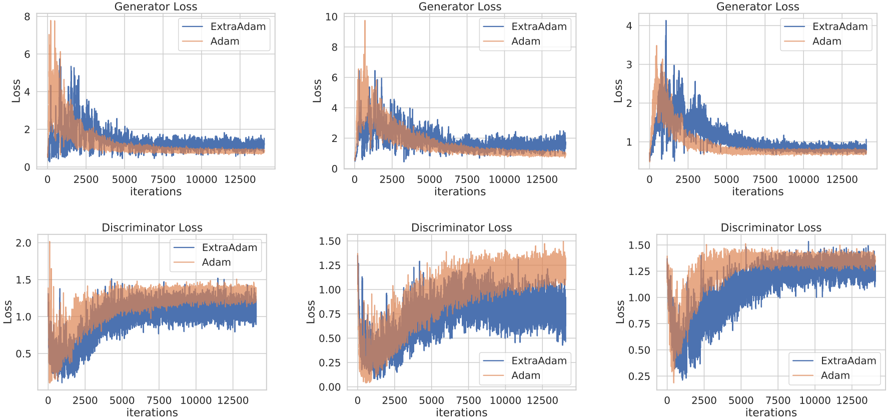
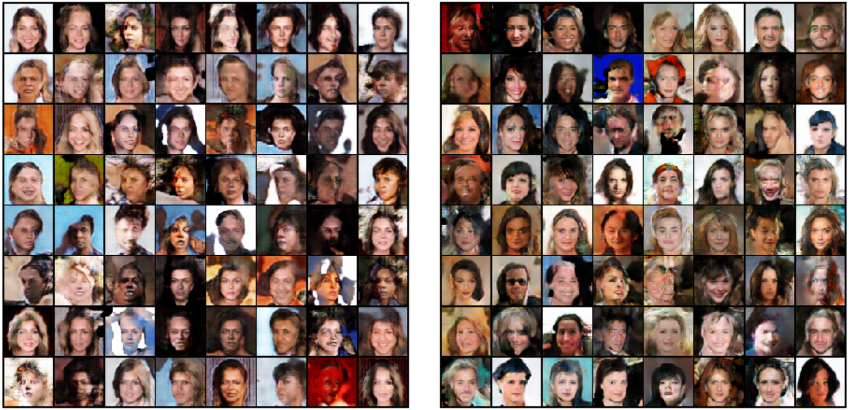


Figure: Adam and ExtraAdam results of training conditional GAN for two epochs.

Experiments: Adam vs ExtraAdam



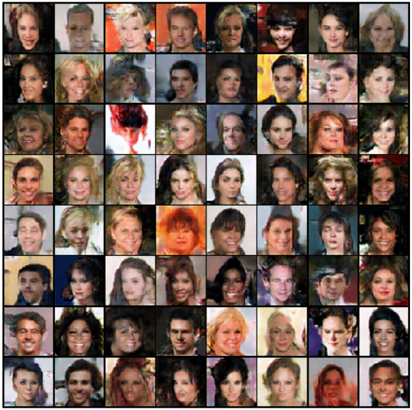
(a) Adam

(b) ExtraAdam

Figure: Adam and ExtraAdam results of training self attention GAN for two epochs.

Experiments: Adam vs ExtraAdam

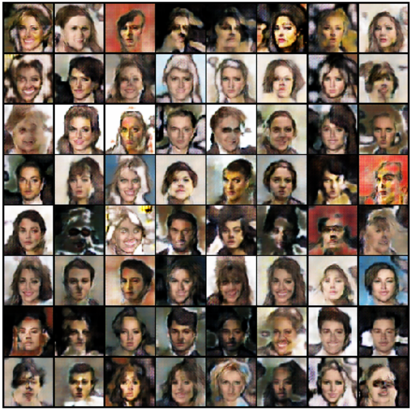
(a) Adam



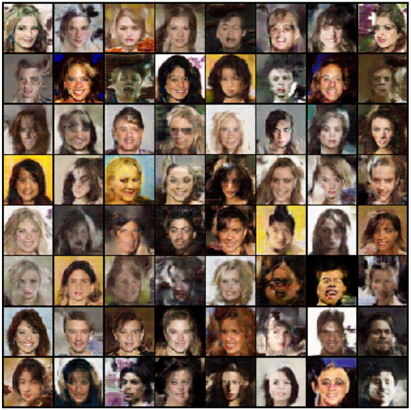
(b) ExtraAdam

Figure: Adam and ExtraAdam results of training self attention GAN for two epochs.

Experiments: Adam vs ExtraAdam



(a) Adam



(b) ExtraAdam

Figure: Adam and ExtraAdam results of training self attention GAN for two epochs.

References

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011).
Solving variational inequalities with stochastic mirror-prox algorithm.
Stochastic Systems, 1(1):17–58.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. (2019).
Revisiting stochastic extragradient.
arXiv preprint arXiv:1905.11373.