

# Stochastic Extragradient

Dmitry Kovalev

Based on joint work with Konstantin Mishchenko, Egor Shulgin, Peter Richtarik, and Yura Malitsky

# Variational Inequality

Find  $x^* \in \mathcal{K}$  such that

$$g(x) - g(x^*) + \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathbb{R}^d$$

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

# Stochastic Variational Inequality

Find  $x^* \in \mathcal{K}$  such that

$$g(x) - g(x^*) + \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathbb{R}^d$$

$$F(x) = \mathbb{E}_{\xi} [F(x; \xi)]$$

- ▶  $\xi$  is a random variable
- ▶  $F(x; \xi)$  is monotone almost surely

# Examples: Stochastic Convex Minimization

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi} [f(x; \xi)]$$

- ▶  $\mathcal{X} \subset \mathbb{R}^d$  is a convex set
- ▶  $f(x; \xi) : \mathcal{X} \rightarrow \mathbb{R}$  is almost surely convex function

$$F(x; \xi) = \nabla f(x; \xi)$$

# Examples: Stochastic Saddle Point Problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathbb{E}_{\xi} [f(x, y; \xi)]$$

- ▶  $\mathcal{X} \subset \mathbb{R}^{d_x}$ ,  $\mathcal{Y} \subset \mathbb{R}^{d_y}$  are convex sets
- ▶  $f(x, y; \xi) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  is convex in  $x$  and concave in  $y$  almost surely

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

# Extragradient Algorithm

---

**Algorithm 1** Extragradient Method for Variational Inequalities.

---



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

# Stochastic Extragradient Algorithm

---

**Algorithm 1** Extragradient Method for Variational Inequalities.

---

- 1: **Parameters:**  $x^0 \in \mathcal{K}$ , stepsize  $\eta > 0$
- 2: **for**  $t = 0, 1, 2, \dots$  **do**
- 3:      $y^t = \text{prox}_{\eta g}(x^t - \eta F(x^t))$    $F(x^t; \xi_1^t)$
- 4:      $x^{t+1} = \text{prox}_{\eta g}(x^t - \eta F(y^t))$    $F(y^t; \xi_2^t)$
- 5: **end for**

Samples  $\xi_1^t$  and  $\xi_2^t$  are the same or independent?

# Independent Samples: Juditsky et al., 2011

- ▶ Converges under very restrictive uniformly bounded noise assumption
- ▶ Diverges even on bilinear stochastic saddle point problems



# Same Sample: Our Approach

---

**Algorithm 2** Stochastic Extragradient Method for Variational Inequalities.

---

1: **Parameters:**  $x^0 \in \mathcal{K}$ , stepsize  $\eta > 0$   
2: **for**  $t = 0, 1, 2, \dots$  **do**  
3:     **Sample**  $\xi^t$   
4:      $y^t = \text{prox}_{\eta g} (x^t - \eta F(x^t; \xi^t))$   
5:      $x^{t+1} = \text{prox}_{\eta g} (x^t - \eta F(y^t; \xi^t))$   
6: **end for**

---

Requires the noise to be bounded  
at the optimum only!

# Experiments: Bilinear Saddle Point Problem

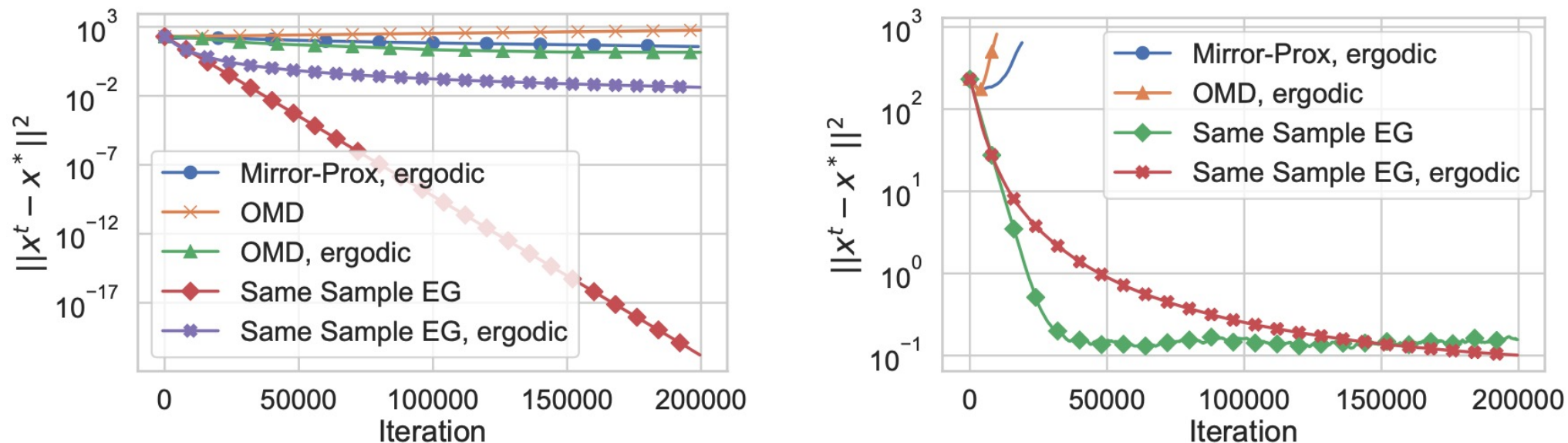


Figure 1: Left: comparison of using independent samples and averaging as suggested by [Juditsky et al., 2011] and the same sample as proposed in this work. The problem here is the sum of randomly sampled matrices  $\min_x \max_y \sum_{i=1}^n x^\top \mathbf{B}_i y$ . Since at point  $(x^*, y^*)$  the noise is equal 0, the convergence of Algorithm 1 is linear unlike the slow rates of [Juditsky et al., 2011] and [Gidel et al., 2019a]. 'EGm' is the version with negative momentum [Gidel et al., 2019b] equal  $\beta = -0.3$ . Right: bilinear example with linear terms.

# Experiments: Generating Mixture of Gaussians

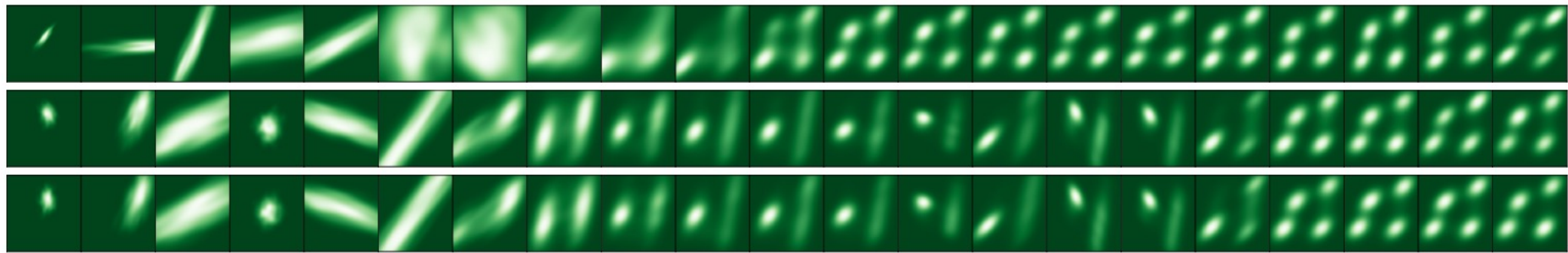


Figure 2: Top line: extragradient with the same sample. Middle line: gradient descent-ascent. Bottom line: extragradient with different samples. Since the same seed was used for all methods, the former two methods performed extremely similarly, although when zooming it should be clear that their results are slightly different.

# Experiments: GAN, CelebA dataset



Figure 9: Adam (top) and ExtraAdam (bottom) results of training self attention GAN for two epochs. The results of training with the three best performing stepsizes,  $10^{-3}$ ,  $2 \cdot 10^{-3}$ ,  $4 \cdot 10^{-3}$ , are provided for each method (from the left to the right). Best seen in color by zooming on a computer screen.