О стохастическом экстраградиентном методе для вариационных неравенств

Дмитрий Ковалев

Московский физико-технический институт

Физтех-школа прикладной математики и информатики

Кафедра интеллектуальных систем

Научный руководитель: д.ф-м.н. Гасников А.В.

Вариационное неравенство

Найти вектор $x^* \in \mathbb{R}^d$, удовлетворяющий

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

- $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ – μ-сильно выпуклая замкнутая функция
- $F : \mathbb{R}^d \to \mathbb{R}^d$ – монотонный L-липшицев оператор

Стохастическое вариационное неравенство

Найти вектор $x^* \in \mathbb{R}^d$, удовлетворяющий

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

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

- ξ – случайный вектор
- $F(x;\xi)$ – почти наверное монотонный L-липшицев оператор

Пример: стохастическая минимизация

$$\min_{x \in \mathbb{R}^d} \mathbb{E}_\xi [f(x;\xi)] + g(x)$$

- $f(x;\xi) : \mathbb{R}^d \to \mathbb{R}$ – почти наверное выпуклая L-гладкая функция

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

Пример: стохастическая седловая задача

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \mathbb{E}_\xi [f(x, y; \xi)] + g_x(x) - g_y(y)$$

- $g_x(x) \colon \mathbb{R}^{d_x} \to \mathbb{R}$ и $g_y(y) \colon \mathbb{R}^{d_x} \to \mathbb{R}$ –
  μ-сильно выпуклые функции
- $f(x, y; \xi) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ – почти наверное
  выпуклая по x и вогнутая по y
  L-гладкая функция

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

Экстраградиентный алгоритм

промежуточная точка

Algorithm 1 Extragradient Method for Variational Inequalities.

1: **Parameters:** $x^0 \in \mathcal{K}$, stepsize $\eta > 0$
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: $\quad y^t = \mathrm{prox}_{\eta g}\left(x^t - \eta F(x^t)\right)$ ← градиентный шаг из x^t
4: $\quad x^{t+1} = \mathrm{prox}_{\eta g}\left(x^t - \eta F(y^t)\right)$
5: **end for**

градиентный шаг из x^t с градиентом взятым в y^t

Стохастический экстраградиентный алгоритм

Algorithm 1 Extragradient Method for Variational Inequalities.

Parameters: $x^0 \in \mathcal{K}$, stepsize $\eta > 0$
for $t = 0, 1, 2, \ldots$ **do**
  $y^t = \mathrm{prox}_{\eta g}\left(x^t - \eta F(x^t)\right)$ ← $F(x^t; \xi_1^t)$
  $x^{t+1} = \mathrm{prox}_{\eta g}\left(x^t - \eta F(y^t)\right)$ ← $F(y^t; \xi_2^t)$
end for

Векторы ξ_1^t и ξ_2^t равны или выбраны независимо?

Экстраградиент с независимыми сэмплами: Judistsky et al., 2011

- требует равномерную ограниченность шума на всем пространстве для сходимости
- расходится на простой билинейной седловой задаче когда область определения неограничена

Предлагаемый подход: экстраградиент с одинаковыми сэмплами

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

требует ограниченность шума только в оптимуме!

Эксперимент: билинейная седловая задача

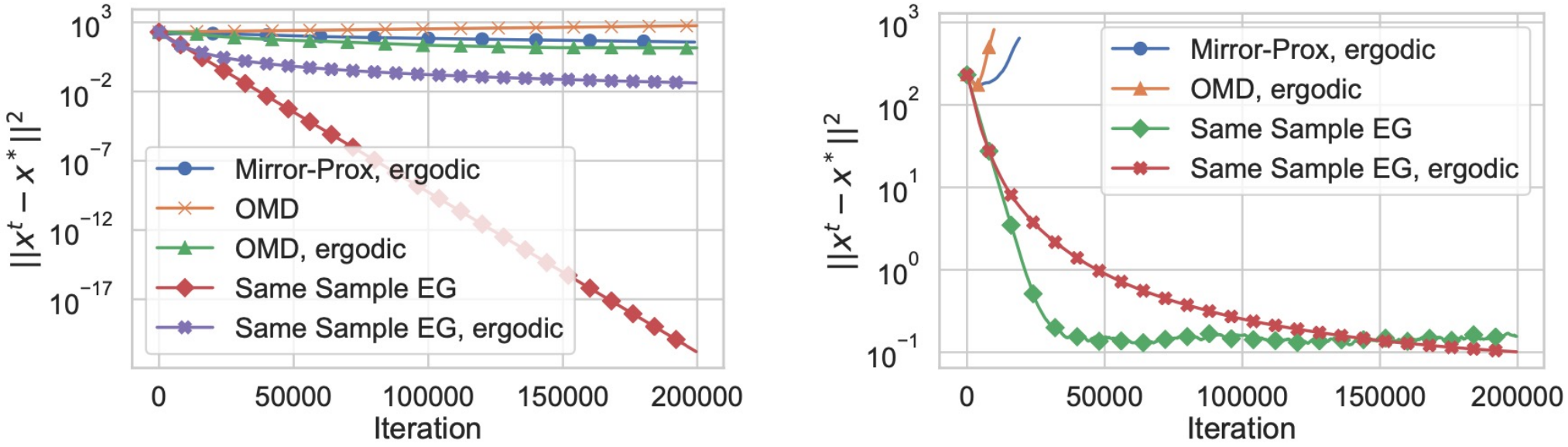

Figure 1: Left: comparison of using independent samples and averaging as suggested by [Juditsky et al., 2011] and the same sample as proposed in this work. The problem here is the sum of randomly sampled matrices $\min_x \max_y \sum_{i=1}^n x^\top \mathbf{B}_i y$. Since at point (x^*, y^*) the noise 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.

Эксперимент: генерация смеси гауссиан

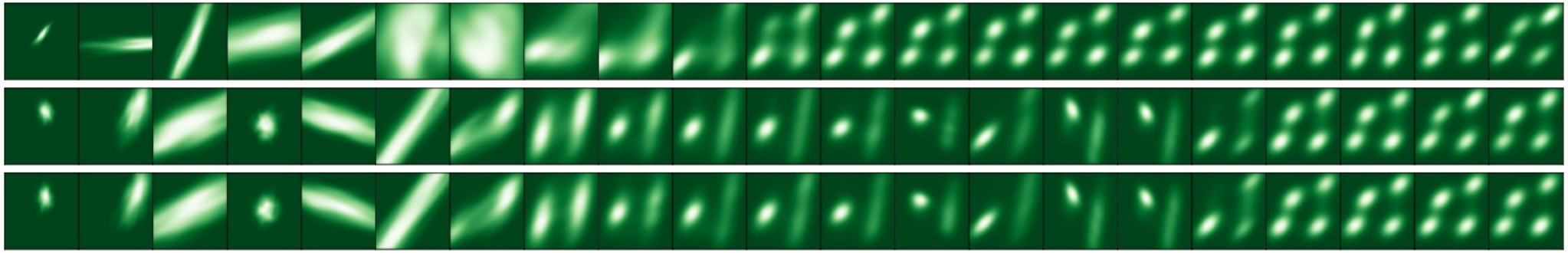

Figure 2: Top line: extragradient with the same sample. Middle line: gradient descent-ascent. Bottom line: extragradient with different samples. Since the same seed was used for all methods, the former two methods performed extremely similarly, although when zooming it should be clear that their results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

Эксперимент: GAN, CelebA

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.