Использование гладкости высших порядков в безградиентных методах с неточным оракулом для задач выпуклой стохастической оптимизации

Новицкий Василий Геннадьевич
студент 2 курса магистратуры ФПМИ

Московский физико-технический институт
Кафедра «Интеллектуальные системы» ВЦ РАН
Научный руководитель д.ф.-м. н. А.В. Гасников

16 июня 2021

Проблема

- изучается проблема выпуклой стохастической минимизации функции безградиентными методами

$$f(x) \to \min_{x \in Q}$$

Особенности

$$f(x) \to \min_{x \in Q}$$

1. тип минимизации — стохастический градиентный
2. вместо градиента на каждой итерации известны значения функции с аддитивным шумом (с конечной дисперсией)
$$\widetilde{f}(x_k, \xi_k) = f(x_k) + \xi_k,\ \mathbb{E}[\xi_k^2] \leq \sigma^2$$
3. высокий порядок гладкости функции f по x (липшицевость гессиана или производных большего порядка)

Актуальность проблемы

- Akhavan A., Pontil M., Tsybakov A.B. Exploiting higher-order smoothness in derivative-free optimization and continuous bandits (NeurIPS, 2020)
- Yan Zhang, Yi Zhou, Kaiyi Ji, Michael M. Zavlanos. Boosting One-point Derivative-free Online Optimization Via Residual Feedback (arXiv: 2020)
- F. Bach and V. Perchet. Highly-smooth zero-th order online optimization (PMLR, 2016)
- Shamir O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback (Journal of Machine Learning Research, 2017)

Результаты

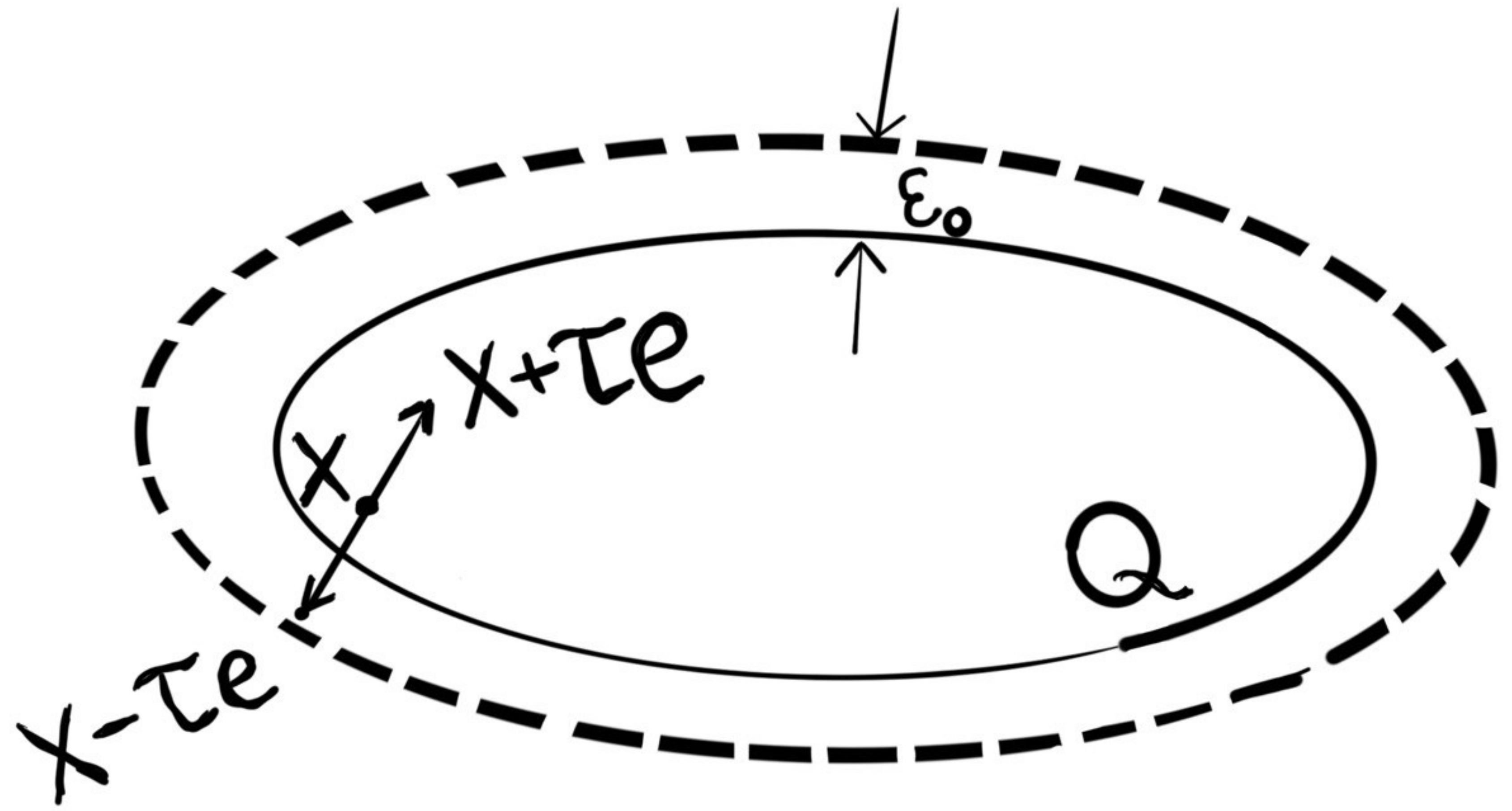
- получены новые верхние оценки скорости сходимости для задачи стохастической минимизации при условии повышенной гладкости целевой функции (одноточечный оракул)
- получены новые верхние оценки для седловой задачи при условии повышенной гладкости целевой функции (одноточечный оракул)
- диплом победителя конференции МФТИ
- Improved Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandit (V.Novitskii, A.Gasnikov), Optimization Letters [Under Review], https://arxiv.org/abs/2101.03821
- One-Point Gradient-Free Methods for Smooth and Non-Smooth Saddle-Point Problems (A.Beznosikov, V.Novitskii, A.Gasnikov), MOTOR 2021 [accepted], https://arxiv.org/pdf/2103.00321.pdf

Свойства функции

$$f(x) \to \min_{x \in Q},$$

- где $f : U_{\varepsilon_0}(Q) \to \mathbb{R}$ — выпуклая функция,
- $Q \subset \mathbb{R}^n$ — выпуклое замкнутое множество
- метрика евклидова
- $\|\nabla f(x) - \nabla f(y)\| \leq G\|x - y\|$

  для всех $x, y \in U_{\varepsilon_0}(Q)$



$$\widetilde{g}_k = \frac{f(x_k + \tau_k e_k) + \xi_k^+ - f(x_k - \tau_k e_k) - \xi_k^-}{2\tau_k} n e_k$$

Гладкость высоких порядков

Определение: функция $f(x)$ удовлетворяет условию Гельдера с параметром β $[f \in \mathcal{F}_\beta(L)]$, если

$$\left| f(z) - \sum_{0 \le |m| \le l} \frac{1}{m!} D^m f(x)(z-x)^m \right| \le L \|z-x\|^\beta,$$

- здесь l – максимальное целое число, строго меньшее, чем β, сумма ведется по мультииндексу $m = (m_1, \dots, m_n) \in \mathbb{N}^n$, $m! = m_1! \cdot \dots \cdot m_n!$, $L > 0$.

Мы исследуем функции, удовлетворяющие условию Гельдера с параметром $\beta > 2$.

Как мы используем гладкость высоких порядков

Сглаживающее ядро: $K : [-1,1] \to \mathbb{R}$,

$\mathbb{E}[K(r)] = 0$, $\mathbb{E}[rK(r)] = 1$, $\mathbb{E}[r^j K(r)] = 0$, $j = 2, \ldots, l$, $\mathbb{E}\left[|r|^\beta |K(r)|\right] \le \infty$

$r \sim U[-1,1]$

Стохастическая аппроксимация градиента:

$$g_k = \frac{f(x_k + r_k\delta_k) - f(x_k - r_k\delta_k)}{2\|\delta_k\|} nK(r_k)e_k$$

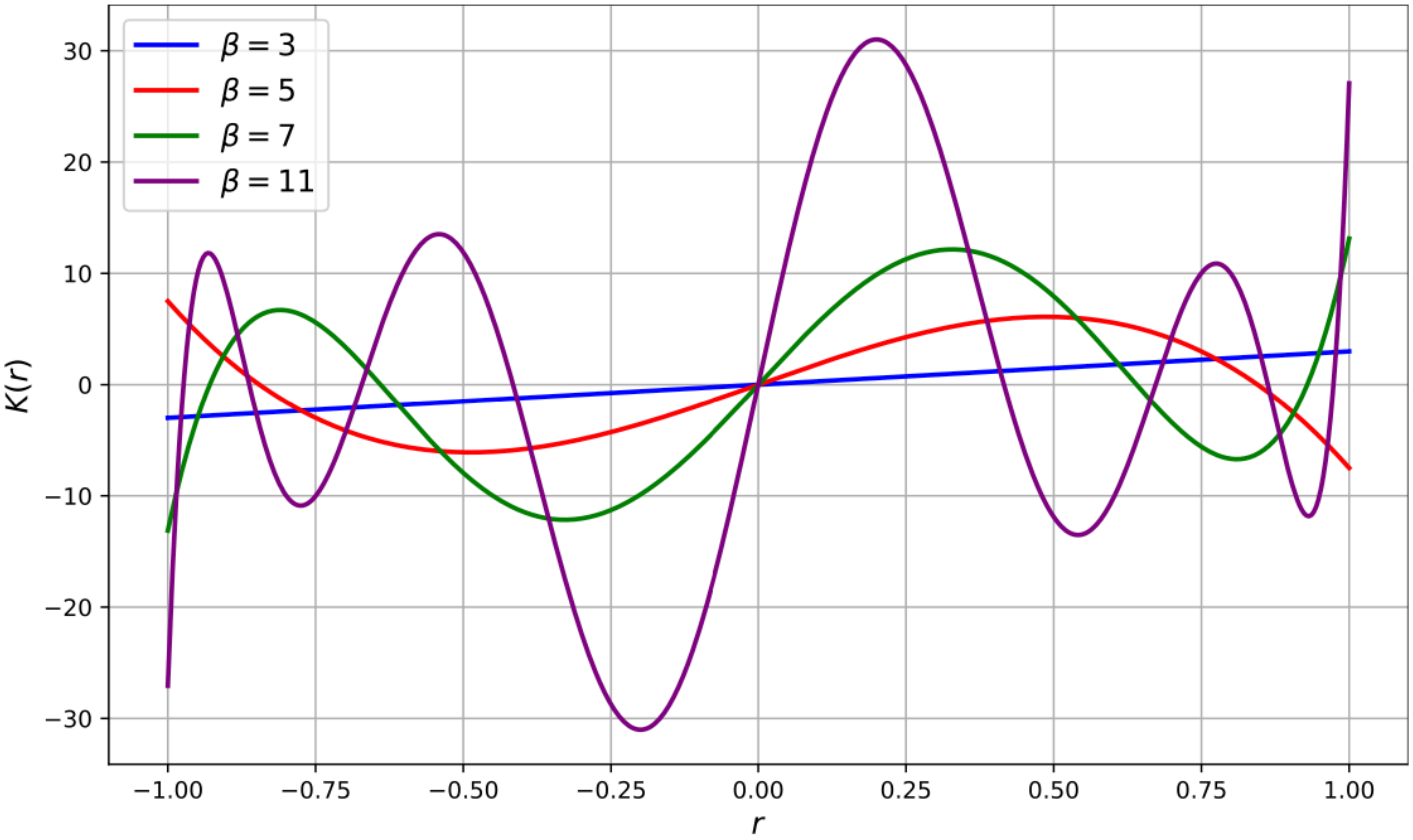
e_k — распределен равномерно на Евклидовой сфере

r_k — распределен равномерно на [-1; 1]

Как мы используем гладкость высоких порядков

$$
\begin{aligned}
K_\beta(r) &= 3r, & \beta &\in [2,3],\\
K_\beta(r) &= \frac{15r}{4}(5-7r^2), & \beta &\in (3,5],\\
K_\beta(r) &= \frac{105r}{64}(99r^4-126r^2+35), & \beta &\in (5,7].
\end{aligned}
$$

Как мы используем гладкость высоких порядков



Шум

$$f(x) \to \min_{x \in Q}$$

- значения функции на каждой итерации известны с аддитивным шумом (с, вообще говоря, ненулевым мат. ожиданием и конечной дисперсией)

$$\widetilde{f}(x_k, \xi_k) = f(x_k) + \xi_k,\ \mathbb{E}[\xi_k^2] \le \sigma^2$$

$$\widetilde{g_k} = \frac{f(x_k + \tau_k r_k e_k) + \xi_k^+ - f(x_k - \tau_k r_k e_k) - \xi_k^-}{2\tau_k} n K(r_k) e_k$$

e_k — распределен равномерно на Евклидовой сфере

r_k — распределен равномерно на [-1; 1]

Предположения на шум

$$\widetilde{g_k} = \frac{f(x_k + \tau_k r_k e_k) + \xi_k^+ - f(x_k - \tau_k r_k e_k) - \xi_k^-}{2\tau_k} nK(r_k) e_k$$

e_k — распределен равномерно на Евклидовой сфере

r_k — распределен равномерно на [-1; 1]

1. $\mathbb{E}[\xi_k^+] \leq \sigma^2, \mathbb{E}[\xi_k^-] \leq \sigma^2$
2. случайные величины ξ_k^+ and ξ_k^- независимы с e_k и r_k
3. случайные величины e_k и r_k независимы

так как $\xi_k^+ \neq \xi_k^-$, то оракул одноточечный

Алгоритм (метод проекции градиента)

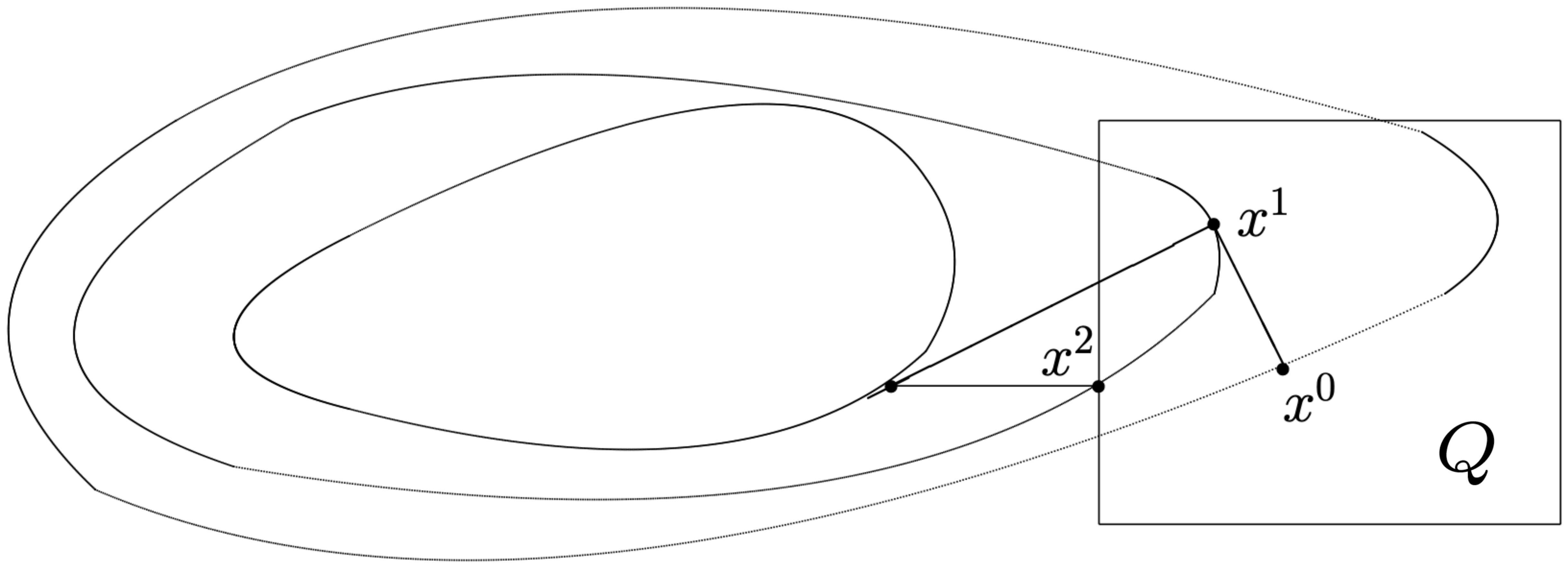
for $k = 1, \ldots, N$ **do**

1. $\widetilde{g}_k = \dfrac{f(x_k + \tau_k r_k e_k) + \xi_k^+ - f(x_k - \tau_k r_k e_k) - \xi_k^-}{2\tau_k} nK(r_k) e_k$

2. $x_{k+1} := \mathrm{Proj}_Q(x_k - \alpha_k \widetilde{g}_k)$

end

Return: $\overline{x}_N = \dfrac{1}{N} \sum_{k=1}^{N} x_k$



Задача

$$f(x) \to \min_{x \in Q}$$

- $\|\nabla f(x) - \nabla f(y)\| \leq G\|x - y\|$ для всех $x,\ y \in U_{\varepsilon_0}(Q)$
- функция удовлетворяет условию Гельдера с параметром $\beta > 2$ в Q
- значения функции доступны с шумом, удовлетворяющим условиям:

1. $\mathbb{E}[\xi_k^+] \leq \sigma^2, \mathbb{E}[\xi_k^-] \leq \sigma^2$
2. случайные величины ξ_k^+ and ξ_k^- независимы с e_k и r_k
3. случайные величины e_k и r_k независимы ,

Основной вопрос

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \le \varepsilon$$

Как ε зависит от параметров:

- размерность пространства n
- число итераций N
- параметр гладкости β
- константа сильной выпуклости γ (для случая сильно выпуклой функции)

Теорема 1. (Новицкий, 2020)

Для γ-сильно выпуклой функции f алгоритм с параметрами

$$\tau_k = \left(\frac{3\kappa\sigma^2 n}{2(\beta-1)(\kappa_\beta L)^2} \right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \; \alpha_k = \frac{2}{\gamma k}, \; k = 1, \ldots, N$$

дает оценку ошибки оптимизации:

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*) \right] \leq \mathcal{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}} \right).$$

Теорема 1. (Новицкий, 2020)

Для γ-сильно выпуклой функции ошибка оптимизации $\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leq \varepsilon$ достигается после $N(\varepsilon)$ итераций, где

$$N(\varepsilon) = \mathcal{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}\right).$$

Теорема 2. (Новицкий, 2020)

Для выпуклой функции f ошибка оптимизации $\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leq \varepsilon$ достигается после

$$N(\varepsilon) = \mathcal{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$$

итераций алгоритма для γ-сильно выпуклой регуляризованной функции $f_\gamma(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|_2^2$, где $\gamma \leq \frac{\varepsilon}{2\|x_0 - x^*\|^2}$.

Основные результаты

Зависимость количества итераций $N(\varepsilon)$ от n (размерность пространства), γ (сильная выпуклость), β (параметр гладкости)

	сильно выпуклая	выпуклая
нижняя оценка, Akhavan, Pontil, Tsybakov (2020)	$\mathcal{O}\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)$	$\mathcal{O}\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)$
Новицкий (2020)	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Akhavan, Pontil, Tsybakov (2020)	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{2}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Bach, Perchet (2016)	$\mathcal{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\gamma\varepsilon^{\frac{\beta+1}{\beta-1}}}\right)$	$\mathcal{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Gasnikov and al. (2015), $\beta = 2$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\gamma\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^3}\right)$

$\gamma \sim \varepsilon$

Седловая задача

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \varphi(x, y)$$

Изменения:

1. $\widetilde{g}_k := \frac{n}{2\tau_k}(\widetilde{\varphi}_k^+ - \widetilde{\varphi}_k^-) \begin{pmatrix} (\mathbf{e}_k)_x \\ -(\mathbf{e}_k)_y \end{pmatrix} K(r_k)$

2. Функция ошибки:

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \leq \max_{y \in \mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x \in \mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \leq \varepsilon$$

Теорема 3 (Новицкий, 2021)

Для γ-сильно-выпуклой-сильно-вогнутой φ алгоритм с параметрами

$$\tau_k = \left(\frac{3\kappa\sigma^2 n}{2(\beta-1)(\kappa_\beta L)^2} \right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \ \alpha_k = \frac{2}{\gamma k}, \ k = 1, \ldots, N$$

дает оценку ошибки оптимизации:

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \leq \max_{y \in \mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x \in \mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \leq \mathcal{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\gamma N^{\frac{\beta-1}{\beta}}} \right).$$

Теорема 3 (Новицкий, 2021)

Для γ-сильно-выпуклой-сильно-вогнутой функции φ ошибка оптимизации

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \le \max_{y \in \mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x \in \mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \le \varepsilon$$

достигается после $N(\varepsilon)$ итераций, где

$$N(\varepsilon) = \mathcal{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\gamma\varepsilon)^{\frac{\beta}{\beta-1}}}\right).$$

Теорема 4 (Новицкий, 2021)

Для выпукло-вогнутой функции φ ошибка оптимизации

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \le \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \le \varepsilon$$

достигается после

$$N(\varepsilon) = \mathcal{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$$

итераций алгоритма для γ-сильно-выпуклой-сильно-вогнутой регуляризованной функции $\varphi_\gamma(x, y) := \varphi(x, y) + \frac{\gamma}{2}\|x - x_0\|^2 - \frac{\gamma}{2}\|y - y_0\|^2$, где $\gamma \le \frac{\varepsilon}{\|x_0 - x^*\|^2 + \|y_0 - y^*\|^2}$.

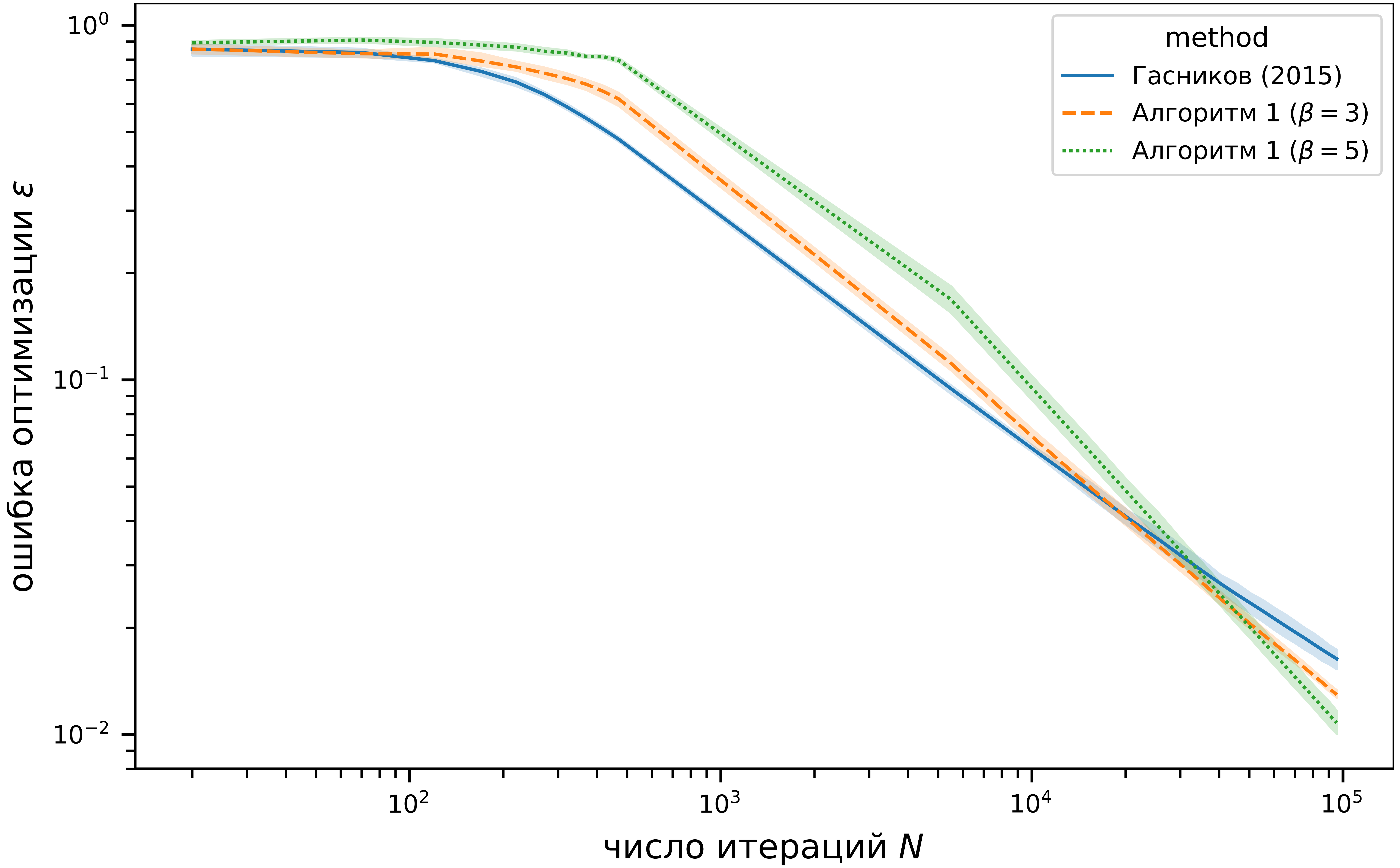
Эксперимент

$$f(x) = \frac{1}{2}x^T A x + \frac{1}{10}\sum_{k=1}^{50} |x_k|^4 \to \min_{x \in B_1(0)}$$

собственные значения матрицы: 0.25, 1, 4.

	теория
Гасников, $\beta = 2$, 2015	$\varepsilon \sim N^{-0.5}$
Алгоритм, $\beta = 3$, 2020	$\varepsilon \sim N^{-2/3}$
Алгоритм, $\beta = 5$, 2020	$\varepsilon \sim N^{-4/5}$

Эксперимент



	теория	эксперимент
Гасников, $\beta = 2$, 2015	$\varepsilon \sim N^{-0.5}$	$\varepsilon \sim N^{-0.61}$
Алгоритм, $\beta = 3$, 2020	$\varepsilon \sim N^{-2/3}$	$\varepsilon \sim N^{-0.73}$
Алгоритм, $\beta = 5$, 2020	$\varepsilon \sim N^{-4/5}$	$\varepsilon \sim N^{-0.91}$

Результаты

- получены новые верхние оценки для задачи стохастической минимизации при условии повышенной гладкости целевой функции (одноточечный оракул)
- получены новые верхние оценки для седловой задачи при условии повышенной гладкости целевой функции (одноточечный оракул)
- диплом победителя конференции МФТИ
- Improved Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandit (V. Novitskii, A. Gasnikov), Optimization Letters [Under Review], https://arxiv.org/abs/2101.03821
- One-Point Gradient-Free Methods for Smooth and Non-Smooth Saddle-Point Problems (A. Beznosikov, V. Novitskii, A. Gasnikov), MOTOR 2021 [accepted], https://arxiv.org/pdf/2103.00321.pdf