Ускоренные методы нулевого порядка в гладкой выпуклой стохастической оптимизации

Фанис Адикович Хафизов

Московский физико-технический институт

Курс: Автоматизация научных исследований
(практика, В. В. Стрижов)/Группа 105

Эксперт: к.ф.-м.н А. Н. Безносиков

Консультант: А. И. Богданов

2024

Цель исследования

Цель:

- Создать ускоренный безградиентный метод решения задачи безусловной гладкой стохастической оптимизации
- Доказать сходимость в случае стохастического шума
- Поставить численный эксперимент и сравнить с методом Нестерова и градиентным спуском в случаях детерминированного и стохастического шума

Литература

- Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, Eric Moulines. First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities. 2023.
- Andrey Veprikov, Alexander Bogdanov, Vladislav Minashkin, Aleksandr Beznosikov, Alexander Gasnikov. New Aspects of Black Box Conditional Gradient: Variance Reduction and One Point Feedback. 2024.
- Eduard Gorbunov, Pavel Dvurechensky, Alexander Gasnikov. An Accelerated Method for Derivative-Free Smooth Stochastic Convex Optimization. 2020.

Постановка задачи

Дано

Доступ к оракулу нулевого порядка $f_\delta(x) = f(x, \xi) + \delta(x, \xi)$.

Требуется

Построить алгоритм, решающий задачу

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi \sim \pi}[f(x, \xi)],$$

обращаясь к оракулу f_δ.

Предположения

A (1. Сильная выпуклость)

$$\exists \mu > 0 : \forall x, y \in \mathbb{R}^d \hookrightarrow \frac{\mu}{2}\|x - y\|^2 \leqslant f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \tag{1}$$

A (2. Гладкость)

$$\exists L(\xi) > 0 : \forall x, y \in \mathbb{R}^d \hookrightarrow \|\nabla f(x, \xi) - \nabla f(y, \xi)\| \leqslant L(\xi)\|x - y\|. \tag{2}$$

$$L^2 := \mathbb{E}\left[L(\xi)^2\right]. \tag{3}$$

Предположения

A (3. Ограниченность оракульного шума)

$$\exists \Delta > 0 : \forall x \in \mathbb{R}^d \hookrightarrow \mathbb{E}\left[|\delta(x,\xi)|^2\right] \leqslant \Delta^2. \tag{4}$$

A (4. Ограниченность второго момента градиента)

$$\exists \sigma_\nabla^2 : \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f(x)\|^2\right] \leqslant \sigma_\nabla^2. \tag{5}$$

A (5. Ограниченность второго момента функции)

$$\exists \sigma_f^2 : \mathbb{E}\left[\|f(x,\xi) - f(x)\|^2\right] \leqslant \sigma_f^2. \tag{6}$$

Предложенный метод

Algorithm Accelerated Stochastic Gradient Descent (OPF)

Require: stepsize $\gamma > 0$, momentums θ, η, β, p, number of iterations N, approximation parameter $\tau > 0$.

Initialization: choose $x^0 = x_f^0$

for $k = 0, 1, \ldots, N-1$ **do**

  $x_g^k = \theta x_f^k + (1-\theta)x^k$

  Sample $i \sim U\{1, \ldots, d\}$

  Sample 2 realizations of ξ: ξ_k^- and ξ_k^+ independently

  $g^k = d\frac{f_\delta(x_g^k + \tau e_i, \xi_k^+) - f_\delta(x_g^k - \tau e_i, \xi_k^-)}{2\tau} e_i$

  $x_f^{k+1} = x_g^k - p\gamma g^k$

  $x^{k+1} = \eta x_f^{k+1} + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k$

end for

Теорема о сходимости

Теорема

Предположим А 1-5. Тогда ускоренный стохастический градиентный спуск (Algorithm 1) имеет скорость сходимости:

$$
\begin{aligned}
&\mathbb{E}\left[\|x^N - x^*\|^2 + \frac{6}{\mu}(f(x_f^N) - f(x^*))\right] \leqslant \\
&\leqslant \exp\left(-N\sqrt{\frac{p^2\mu\gamma}{3}}\right)\left(\|x^0 - x^*\|^2 + \frac{6}{\mu}(f(x_f^0) - f(x^*))\right) + \\
&+ \frac{3\sqrt{3\gamma}}{\mu^{3/2}}\left(\left(1 + \sqrt{\frac{3}{\gamma\mu}}\right)\left(\frac{dL^2\tau^2}{2} + \frac{2d\Delta^2}{\tau^2}\right) + \right. \\
&\left. + \frac{dL^2\tau^2}{2} + \frac{4d\sigma_f^2}{\tau^2} + 4\sigma_\nabla^2 d + \frac{\Delta^2}{2\tau^2}\right),
\end{aligned}
$$

где $\gamma \in (0, \frac{3}{4L}], p \simeq (2(1+\gamma L)(4d+1))^{-1}, \beta \simeq \sqrt{p^2\mu\gamma}, \eta \simeq \sqrt{\frac{1}{\mu\gamma}},$
$\theta \simeq \frac{p\eta^{-1}-1}{\beta p\eta^{-1}-1}.$

Оценка на количество оракульных вызовов

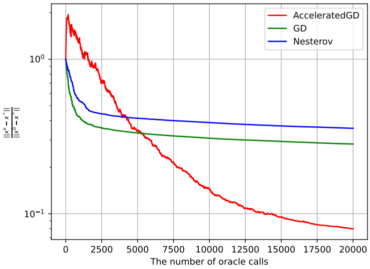
Следствие

В предположениях теоремы 1 и выборе $\gamma = \frac{3}{4L}$ *для достижения* ε*-точности решения* ($\mathbb{E}\|x^N - x^*\|^2 \leqslant \varepsilon$) *требуется* $\mathcal{O}\left(d\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon-\sigma}\right)$ *обращений к оракулу, где*
$\sigma = \frac{9}{2\mu^2}\left(2dL^2\tau^2 + \frac{\Delta^2}{2\tau^2}(12d+1) + \frac{4d\sigma_f^2}{\tau^2} + 4d\sigma_\nabla^2\right)$.

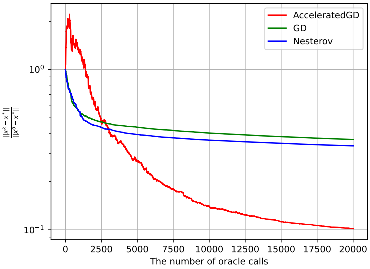
Вычислительный эксперимент. Квадратичная задача

$$\min_{x \in \mathbb{R}^d} f(x) = x^T A x + b^T x + c$$

$$A \in \mathbb{S}_d : \mu \preceq A \preceq L, \mu = 1, L = 1000, d = 100$$



Детерминированный шум, $\Delta = 10^{-6}$, $\tau = 10^{-4}$

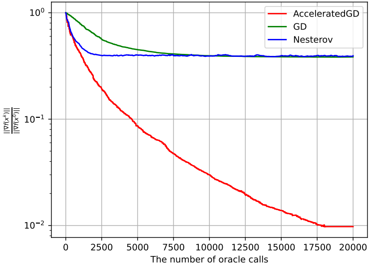


Стохастический шум, $\Delta = 10^{-6}$, $\tau = 10^{-4}$

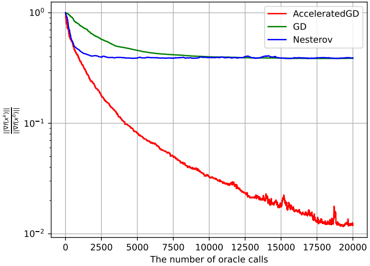
Вычислительный эксперимент. Задача логистической регрессии

Датасет mushrooms, $d = 112$, $m = 8124$, $y_i \in \{0, 1\}$, $\lambda = 0,1$.

$$\min_{w \in R^d} f(w) = \frac{1}{m} \sum_{k=1}^{m} \log(1 + \exp(-y_k \cdot (Xw)_k)) + \lambda \|w\|_2^2$$



Детерминированный шум, $\Delta = 10^{-6}$, $\tau = 10^{-4}$



Стохастический шум, $\Delta = 10^{-6}$, $\tau = 10^{-4}$

Результаты

- Предложен метод нулевого порядка, решающий поставленную задачу
- Получена скорость сходимости на описанном классе функций в случае стохастического шума
- Показано превосходство предложенного метода над методом Нестерова и градиентным спуском

Будущая работа

Поставить эксперимент на более сложной задаче.