Проект по статье "The Power of First-Order Smooth Optimization for Black-Box Non-Smooth Problems"

by Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii1, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu

Шевцова Маргарита
(МФТИ)

1 декабря 2023 г.

Содержание

Введение: о статье

"The Power of First-Order Smooth Optimization for Black-Box Non-Smooth Problems"by Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii1, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu

Постановка задачи

Задача

$$\min_{x \in Q \subseteq R^d} f(x) \qquad (1)$$

для оракула нулевого порядка. Оракул возвращает $f(x)$ в точке запроса x, возможно с некоторым шумом, ограниченным $\Delta > 0$.

$\gamma > 0$ – малое число
$Q_\gamma := Q + B_2^d(\gamma)$
Предположим, что

1. множество Q выпуклое
2. функция f выпуклая на множестве Q_γ
3. функция f Липшецево с константой M , то есть $|f(y) - f(x)| \leq M||y - x||_p$ на Q_γ, где $p \in [1, 2]$ и $|| \cdot ||_p$ —p-норма. Если $p = 2$, то обозначим как M_2 константу Липшица.

Обзор результатов

Случайное сглаживание негладкой целевой функции

Случайное сглаживание негладкой целевой функции f

$f_\gamma(x) = \mathbb{E}_u f(x+u)$, где $u \sim RB_2^d(\gamma)$, u случайный вектор с равномерным распределением в $B_2^d(\gamma)$

Theorem 2.1 (свойства f_γ)

Для всех $x, y \in Q$ верно

1. $f(x) \leq f_\sigma(x) \leq f(x) + \sigma M_2$
2. $f_\sigma(x)$ Липшицева с константой M:

$$|f_\sigma(y) - f_\sigma(x)| \leq M||y-x||_p;$$

3. градиент $f_\sigma(x)$ Липшицев с константой $L = \frac{\sqrt{d}M}{\sigma}$:
$||\nabla f_\sigma(y) - \nabla f_\sigma(x)||_q \leq L||y-x||_p$, где q такое, что $1/p + 1/q = 1$.

Несмещенный стохастический градиент для f_γ

Несмещенный стохастический градиент для f_γ [67]

$$\nabla f_\gamma(x, e) = d\frac{f(x + \gamma e) - f(x - \gamma e)}{2\gamma}e \qquad (2),$$

где $e \sim RS_d^2(1)$ случайный вектор с равномерным распределением на сфере.

Theorem 2.2 (свойства $\nabla f_\gamma(x, e)$)

Для всех $x \in Q$

1. $\nabla f_\gamma(x, e)$ – несмещенная оценка $\nabla f_\gamma(x)$: $3\mathbb{E}_e[\nabla f_\gamma(x, e)] = \nabla f_\gamma(x)$;
2. $\nabla f_\gamma(x, e)$ имеет ограниченную дисперсию:

$$\mathbb{E}_e[||\nabla f_\gamma(x, e)||_q^2] \leq \kappa(p, d) \cdot (dM_2^2 + \frac{d^2\Delta^2}{\gamma^2}), \text{где } 1/p + 1/q = 1$$

и $\kappa(p, d) = O(\sqrt{\mathbb{E}_e||e||_q^4}) = O(\min\{q, \ln d\}d^{\frac{2}{q-1}})$.

Алгоритм решения гладкой задачи

Алгоритм решения гладкой задачи

$A(L, \sigma^2)$ – алгоритма с batch, который решает (1)

1. в предположении, что f гладкая и $||\nabla f(y) - \nabla f(x)||_q \leq L||y - x||_p, x, y \in Q_\gamma$
2. используя стохастического оракула 1-порядка, который зависит от случайной величины ν и возвращает в точке x несмещенный стохастический градиент $\nabla_x f(x, \nu)$ с ограниченной дисперсией: $\mathbb{E}_\nu[||\nabla_x f(x, \nu) - \nabla f(x)||_q^2] \leq \sigma^2$.

$N(L, \varepsilon)$ последовательных итераций
$T(L, \sigma^2, \varepsilon)$ стохастических вызовов оракула 1-порядка.
Значит, в $A(L, 2)$ можно использовать batch-parallelization со средним размером batch $B(L, \sigma^2, \varepsilon) = T(L, \sigma^2, \varepsilon)/N(L, \varepsilon)$.

Метод

Стохастический оракул 1-порядка

Зависит от случайной величины ν и возвращает в точке x несмещенный стохастический градиент $\nabla_x f(x, \nu)$ с ограниченной дисперсией: $\mathbb{E}_\nu[||\nabla_x f(x, \nu) - \nabla f(x)||_q^2] \leq \sigma^2$.

Метод

Применяем $A(L, \sigma^2)$ к сглаженной задаче

$$\min_{x \in Q \subseteq R^d} f_\gamma(x) \qquad (6)$$

$$\text{с } \gamma = \varepsilon/(2M_2) \qquad (7)$$

и $\nu = e$, $\nabla_x f(x, \nu) = \nabla f_\gamma(x, e)$, где $\varepsilon > 0$ точность, с которой хотим решить задачу (1) в терминах матожидания ошибки.

Оценки количества итераций и вызовов оракула

Чтобы добиться ε-точности в матожидании, алгоритму требуется:

$A(L, \sigma^2)$ для гладкой задачи

$N(L, \varepsilon)$ последовательных итераций
$T(L, \sigma^2, \varepsilon)$ стохастических вызовов оракула 1-порядка.
Значит, в $A(L, 2)$ можно использовать batch-parallelization со средним размером batch $B(L, \sigma^2, \varepsilon) = T(L, \sigma^2, \varepsilon)/N(L, \varepsilon)$.

$A(L, \sigma^2)$ после адаптации для негладкой задачи

$N(\frac{2\sqrt{d}MM_2}{\varepsilon}, \varepsilon)$ последовательных итераций
$2T(\frac{2\sqrt{d}MM_2}{\varepsilon}, 2\kappa(p, d)dM_2^2, \varepsilon)$ вызовов оракула 1-порядка.

По Theorem 2.1 и формуле для γ, $L \leq \frac{2\sqrt{d}MM_2}{\varepsilon}$. По Theorem 2.2, если достаточно малое, то $\sigma^2 \leq 2\kappa(p, d)dM_2^2$.

Стохастическая оптимизация

Если оракул нулевого порядка возвращает несмещенное стохастическое значение $f(x, \xi)$ с шумом ($\mathbb{E}_\xi f(x, \xi) = f(x)$), то с помощью оракула по 2 точкам получаем такой аналог (2):

$$\nabla f_\gamma(x, \xi, e) = d \frac{f(x + \gamma e, \xi) - f(x - \gamma e, \xi)}{2\gamma} e.$$

Негладкая выпуклая задача стохастической оптимизации

$$\min_{w \in R^d} \{f(w) := \mathbb{E}_\xi f(x, \xi)\} \qquad (10)$$

//

Задачи оптимизации с конечной суммой

Минимизация эмпирического риска

Задача (10), но ξ имеет равномерное распределение по $1 \dots m$

$$\min_{w \in R^d} \{f(w) := \mathbb{E}_\xi f(x, \xi) = \frac{1}{m} \sum_{k=1}^{m} f_k(x)\} \qquad (10)$$

Оракул 0-порядка возвращает значения $\{f(x_i, \xi)\}_{i=1}^2$ для точек x_1, x_2.

Сильно выпуклые задачи оптимизации

Теорема: оценка на количество арифметических подслов

Количество различных арифметических подслов длины m хотя бы $(m-1)m(m+1)/6$ и не больше $3(m-1)m(m+1)$ [1]

Кузнечик Кронеккера

1. $\{\lg(n)\}$ всюду плотен на $[0,1)$
2. $\{\lg(n!)\}$ всюду плотен на $[0,1)$

[1] Точные многочлены для некоторых частных случаев есть в

Варианты приближения градиента

Central finite difference (Central)

$$\nabla f_\gamma(x, e) = d\frac{f(x + \gamma e) - f(x - \gamma e)}{2\gamma}e \qquad (2)$$

Forward finite difference (Forward)

$$\nabla f_\gamma(x, e) = d\frac{f(x + \gamma e) - f(x)}{\gamma}e. \qquad (14)$$

As a result, coordinate steps are more accurate approximation of the gradient but also, they are more expensive computationally.

LAD Regression

Least absolute deviation (LAD) Regression

$$\min_{w \in R^d} \{f(w) = \frac{1}{m} \sum_{k=1}^{m} |x_i^T w - y_i|\} \qquad \text{негладкая функция}$$

задача оптимизации с конечной суммой

Данные

Датасет "abalone scale"из LibSVM, всего 8 признаков.

Альтернативы

1. Central Coordinate: считаем апроксимацию градиента по каждой координате. Поэтому $2d$ вызова оракула на каждый шаг.
2. Forward Coordinate: считаем апроксимацию градиента по каждой координате. Поэтому $d + 1$ вызов оракула на каждый шаг.

LAD Regression: сравнение методов

Рис. 1: Gradient Descent

Рис. 2: Nesterov

LAD Regression: сравнение методов

Рис. 3: Gradient Descent

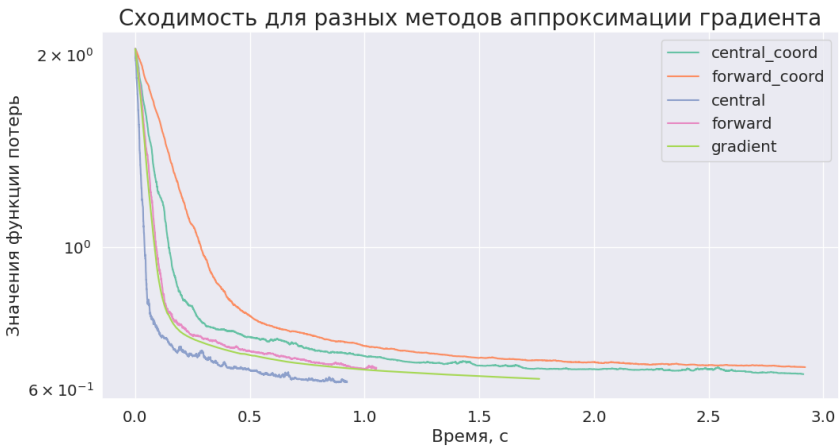
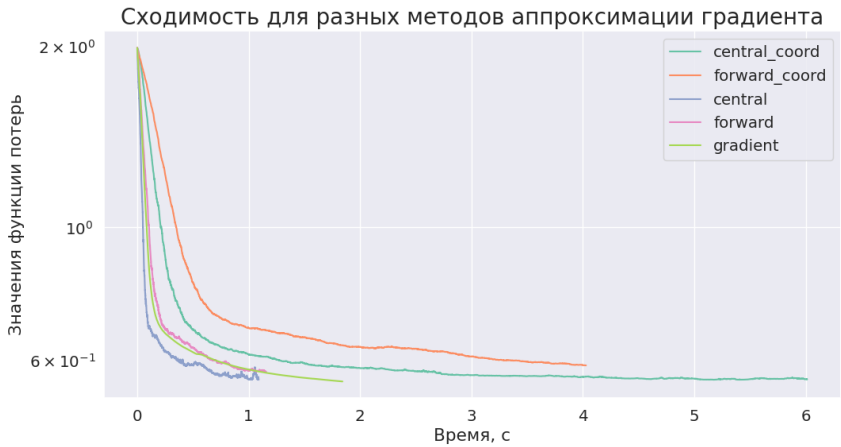


Рис. 4: Nesterov



LAD Regression: сравнение методов

Рис. 5: Gradient Descent

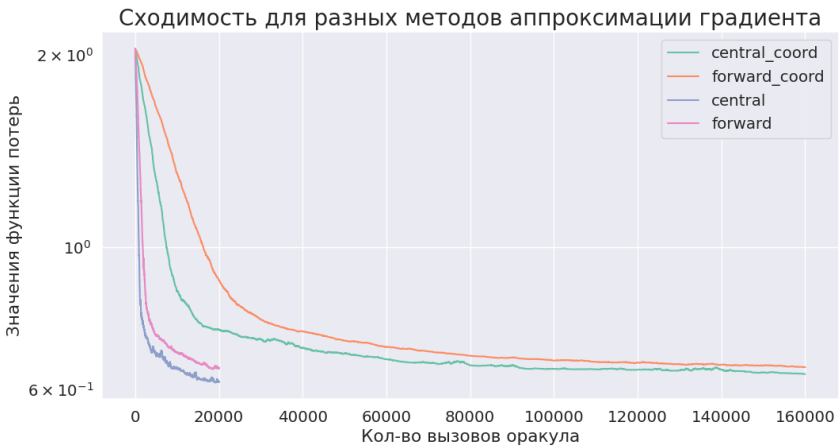
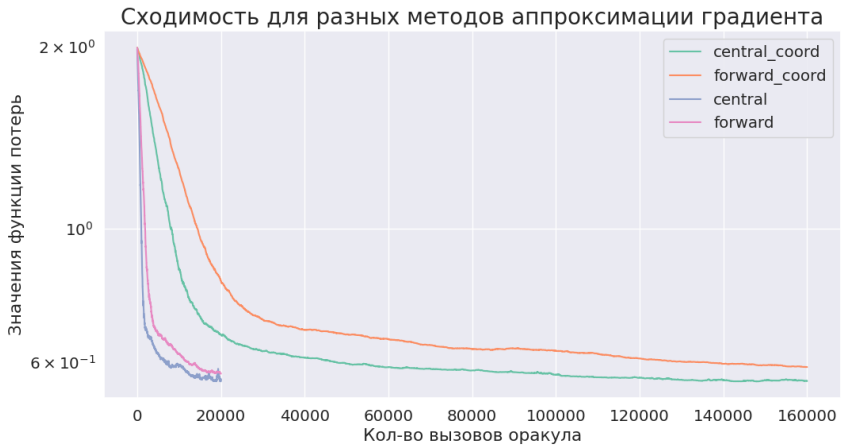


Рис. 6: Nesterov



Support Vector Machine

Support Vector Machine

$$\min_{w \in R^d} \{f(w) = \frac{\mu}{2}||w||^2 + \frac{1}{m}\sum_{k=1}^{m}(1 - y_i \cdot x_i^T w)_+\},$$

Данные

LibSVM basic dataset "a9a"

Выводы

1.

Спасибо за внимание!