

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)

26 ноября 2023 г.

1. Введение

- Классическая задача символической динамики
- Постановка задачи

2. Способ сглаживания

- Случайное сглаживание негладкой целевой функции
- Несмещенный стохастический градиент для f_γ

3. Метод

4. Области применения сглаживания

5. Эксперименты

- Варианты приближения градиента
- LAD regression
- Support Vector Machine
- Reinforcement Learning

6. Заключение

"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

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

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

- ❶ множество Q выпуклое
- ❷ функция f выпуклая на множестве Q_γ
- ❸ функция f Липшицево 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.



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]

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

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

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:

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

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

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 $B(L, \sigma^2, \varepsilon) = T(L, \sigma^2, \varepsilon)/N(L, \varepsilon)$.

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

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

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

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 $\sigma^2 \leq 2\kappa(p, d)dM_2^2$.

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

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

//

Минимизация эмпирического риска

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

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

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

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

- ❶ $\{\lg(n)\}$ всюду плотен на $[0, 1)$
- ❷ $\{\lg(n!)\}$ всюду плотен на $[0, 1)$

¹Точные многочлены для некоторых частных случаев есть в

Central finite difference (Central)

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

Forward finite difference (Forward)

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

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

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

негладкая функция

задача оптимизации с конечной суммой

Данные

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

Альтернативы

- 1 Central Coordinate: считаем аппроксимацию градиента по каждой координате. Поэтому $2d$ вызовов оракула на каждый шаг.
- 2 Forward Coordinate: считаем аппроксимацию градиента по каждой координате. Поэтому $d + 1$ вызовов оракула на каждый шаг.

LAD Regression: сравнение методов

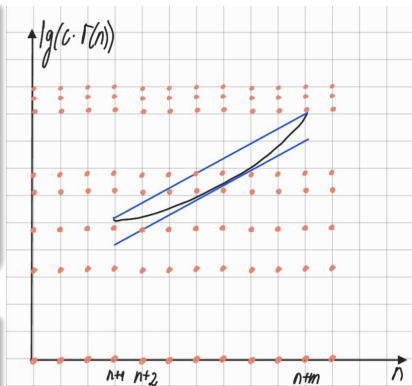
Least absolute deviation (LAD) regression

$$\min_{w \in R^d} \{f(w) = \frac{1}{n} \sum_{k=1}^m |x_k^T w - y_k|\}$$

негладкая функция

Данные

dataset "abalone scale" from the LibSVM



Критические позиции для подслова в факториальном слове

позиции, в которых подслово отличается от хотя бы одного из своих слов-ограничителей.

Лемма

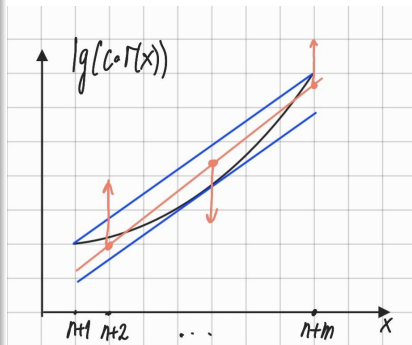
Любое квазиарифметическое слово длины больше 2 отличается от своих слов-ограничителей на не более чем на 1 в одну сторону по модулю 9 в каждом знаке.

Лемма

При фиксированном m при достаточно большом n все точки разделения, лежащие в параллелограмме ограничения для подслова $\alpha[n+1] \dots \alpha[n+m]$ факториального слова, лежат на одной прямой (критическая прямая).

(Критерий арифметичности подслова)

При фиксированном m и достаточно большом n не арифметичность подслова $\alpha[n+1] \dots \alpha[n+m]$ факториального слова равносильна на критической прямой для подслова $\alpha[n+1] \dots \alpha[n+m]$ есть такие 3 различные точки разделения с координатами (x_1, y_1) , (x_2, y_2) , (x_3, y_3) такие, что $x_1, x_2, x_3 \in [n+1, n+m]$ и $\lg(c \cdot \Gamma(x_1)) \geq y_1$, $\lg(c \cdot \Gamma(x_2)) < y_2$, $\lg(c \cdot \Gamma(x_3)) \geq y_3$.



- 1 В факториальном слове все арифметические подслова встречаются бесконечное число раз.
- 2 Каждое квазиарифметическое подслово является близким к арифметическому или подозрительным.
- 3 Количество различных подслов длины m , встречающихся в факториальном слове бесконечное число раз, не превосходит многочлена 3 степени от m .
- 4 Мера c , при которых квазиарифметическое подслово встречается в факториальном слове, равна нулю.
- 5 Для любого подозрительного или близкого к арифметическому подслова существуют такие c_1, c_2 , что в последовательности первых цифр $c_1 \cdot \Gamma(c_2 \cdot n)$ оно встречается бесконечное число раз.

Спасибо за внимание!