Аппроксимации градиента с помощью оракула нулевого порядка и техники запоминания

Богданов Александр Иванович

Научный руководитель:
к.ф.-м.н. Безносиков А. Н.

Кафедра «Интеллектуальные системы»

03.03.01 — Прикладные математика и физика

Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики

Цель исследования

Проблема: Ставится задача оптимизации с доступом только к зашумленному нулевому оракулу.

Цель: Предложить робастый алгоритм аппроксимации градиента, использующий $\mathcal{O}(1)$ вызовов оракула на каждой итерации.

Решение: Предлагается аппроксимация градиента `JAGUAR`, которая использует технику запоминания.

Постановка задачи

Рассматривается две оптимизационные задачи:

- Нестохастическая

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

  Доступ только к $f_\delta(x) := f(x) + \delta(x)$, где $\delta(x)$ – шум.

- Стохастическая

$$\min_{x \in Q} f(x) := \mathbb{E}_{\xi \sim \pi}\left[f(x, \xi)\right]$$

  Доступ только к $f_\delta(x, \xi) := f(x, \xi) + \delta(x, \xi)$, где $\delta(x, \xi)$ - шум.

Множество $Q \subseteq \mathbb{R}^d$ – произвольное.

Схема аппроксимации:

$$\widetilde{\nabla}_i f_\delta(x) := \frac{f_\delta(x + \tau e_i) - f_\delta(x - \tau e_i)}{2\tau} e_i,$$

где e_i – i-ый базисный вектор, τ – параметр сглаживания.

Алгоритм 1

1: **Вход:** $x, h \in \mathbb{R}^d$
2: Сэмплируем $i \in \overline{1, d}$ равномерно и независимо
3: Считаем $\widetilde{\nabla}_i f_\delta(x) = \frac{f_\delta(x+\tau e_i) - f_\delta(x-\tau e_i)}{2\tau} e_i$
4: $h = h - \langle h, e_i \rangle e_i + \widetilde{\nabla}_i f_\delta(x)$
5: **Выход:** h

Схемы аппроксимации:

$$\widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-) := \frac{f_\delta(x + \tau e_i, \xi^+) - f_\delta(x - \tau e_i, \xi^-)}{2\tau} e_i,$$

где e_i – i-ый базисный вектор, τ – параметр сглаживания. При $\xi^+ \neq \xi^-$ – одноточечная обратная связь (ООС), а при $\xi^+ = \xi^-$ – двухточечная обратная связь (ДОС).

Алгоритм 2

1: **Вход:** $x, h, g \in \mathbb{R}^d$; $\eta \in [0, 1]$
2: Сэмплируем $i \in \overline{1, d}$ равномерно и независимо
3: Сэмплируем ξ: ξ^+ и ξ^- независимо (в ДОС $\xi^+ = \xi^-$)
4: Считаем $\widetilde{\nabla}_i f_\delta(x, \xi^\pm) = \frac{f_\delta(x+\tau e_i, \xi^+) - f_\delta(x - \tau e_i, \xi^-)}{2\tau} e_i$
5: $h = h - \langle h, e_i \rangle e_i + \widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-)$
6: $\rho = h - d \cdot \langle h, e_i \rangle e_i + d \cdot \widetilde{\nabla}_i f_\delta(x, \xi^+, \xi^-)$
7: $g = (1 - \eta) g + \eta \rho$
8: **Выход:** g, h

Обычный алгоритм Франка-Вульфа

Общие допущения:

1. Ограниченность множества Q:

$$\forall x, y \in Q : \|x - y\|^2 \leq D^2.$$

2. Выпуклость множества Q:

$$\forall 0 \leq \alpha \leq 1, \forall x, y \in Q : \alpha x + (1 - \alpha) y \in Q.$$

Алгоритм 3

1: **Вход:** $x_0 \in Q$, γ_k
2: **for** k = 0, 1, 2, ... , N **do**
3: $\quad s^k = \arg\min_{s \in Q} \left\langle s, \nabla f(x^k) \right\rangle$
4: $\quad x^{k+1} = x^k + \gamma_k (s^k - x^k)$
5: **end for**
6: **Выход:** x^{N+1}

Алгоритм Франка-Вульфа с `JAGUAR-d`

Допущения:

1. Функция $f(x)$ L-гладкая на множестве Q:

$$\forall x, y \in Q : \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| .$$

2. Функция $f(x)$ выпукла на множестве Q:

$$\forall x, y \in Q : f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

3. Ограниченность оракульного шума:

$$\exists \Delta > 0 \forall x \in Q : |\delta(x)|^2 \leq \Delta^2.$$

Алгоритм Франка-Вульфа с `JAGUAR-d`

Алгоритм 4

1: **Вход:** $x^0 \in Q$, $h^0 = \widetilde{\nabla} f_\delta(x^0)$, γ_k, τ
2: **for** $k = 0, 1, 2, ..., N$ **do**
3: $\quad h^{k+1} = \texttt{JAGUAR-d}\left(x^k, h^k\right)$
4: $\quad s^k = \arg\min\limits_{x \in Q} \left\langle s, h^{k+1} \right\rangle$
5: $\quad x^{k+1} = x^k + \gamma_k(s^k - x^k)$
6: **end for**
7: **Вход:** x^{N+1}

Алгоритм Франка-Вульфа с JAGUAR-d

Теорема (Богданов А., 2023) При шаге оптимизатора:

$$\gamma_k = \frac{4}{k + 8d},$$

получается оценка на сходимость:

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] = \mathcal{O}\left(\frac{d\max\{LD^2, f(x^0) - f(x^*)\}}{N + 8d} + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau}\right).$$

Следствие Пусть ε определяет точность: $\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$:

$$N = \mathcal{O}\left(\frac{d\max\{LD^2, f(x^0) - f(x^*)\}}{\varepsilon}\right),$$

$$\gamma = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD}\right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{dLD^2}\right).$$

Алгоритм Франка-Вульфа с `JAGUAR-s`

Допущения:

1. Функция $f(x, \xi)$ $L(\xi)$-гладкая на множестве Q:
$$\forall\ x, y \in Q : \|\nabla f(x, \xi) - \nabla f(y, \xi)\| \leq L(\xi) \|x - y\| .$$
2. Функция $f(x, \xi)$ выпукла на множестве Q:
$$\forall\ x, y \in Q : f(y, \xi) \geq f(x, \xi) + \langle \nabla f(x, \xi), y - x \rangle .$$
3. Ограниченность оракульного шума:
$$\exists\ \Delta > 0 : \forall\ x \in Q : \mathbb{E}\left[|\delta(x, \xi)|^2\right] \leq \Delta^2$$
4. Ограниченность второго момента градиента:
$$\exists\ \sigma_\nabla^2 : \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^2\right] \leq \sigma_\nabla^2$$
5. Ограниченность второго момента оракула (для OOC):
$$\exists\ \sigma_f^2 : \mathbb{E}\left[\left|f(x, \xi) - f(x)\right|^2\right] \leq \sigma_f^2$$

Алгоритм Франка-Вульфа с `JAGUAR-s`

Алгоритм 5 Стохастический алгоритм Франка-Вульфа с `JAGUAR`

Вход: $x^0 \in Q$, $h^0 = g^0 = \widetilde{\nabla} f_\delta(x^0)$, γ_k, η_k, τ

for $k = 0, 1, 2, ..., N$ **do**

$\quad g^{k+1},\ h^{k+1} = \texttt{JAGUAR-s}\ \left(x^k, h^k, g^k, \eta_k\right)$

$\quad s^k = \underset{x \in Q}{\arg\min} \left\langle s, g^{k+1} \right\rangle$

$\quad x^{k+1} = x^k + \gamma_k(s^k - x^k)$

end for

Вход: x^{N+1}

Алгоритм Франка-Вульфа с JAGUAR-s

Теорема (Богданов А., 2023) При шаге оптимизатора и шаге моментума:

$$\gamma_k = \frac{4}{k + 8d^{3/2}}, \quad \eta_k = \frac{4}{(k + 8d^{3/2})^{2/3}}$$

получается оценка на сходимость:

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] = \mathcal{O}\left(\frac{LD^2 + d\sigma_f D/\tau + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f(x^*))}{(N + 8d^{3/2})^{1/3}} \right.$$
$$\left. + \sqrt{d}LD\tau + \frac{\sqrt{d}\Delta D}{\tau} \right)$$

Следствие Пусть ε определяет точность: $\mathbb{E}\left[f(x^N) - f(x^*)\right] \le \varepsilon$:

$$N = \mathcal{O}\left(\max\left\{ \left[\frac{LD^2 + d\sigma_\nabla D + \sqrt{d}(f(x^0) - f(x^*))}{\varepsilon} \right]^3, \frac{d^{9/2}\sigma_f^3 L^3 D^6}{\varepsilon^6} \right\} \right),$$

$$\gamma = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{d}LD} \right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{dLD^2} \right).$$

Постановка эксперимента

На множестве Q рассматриваются модели LogReg и SVM вида:

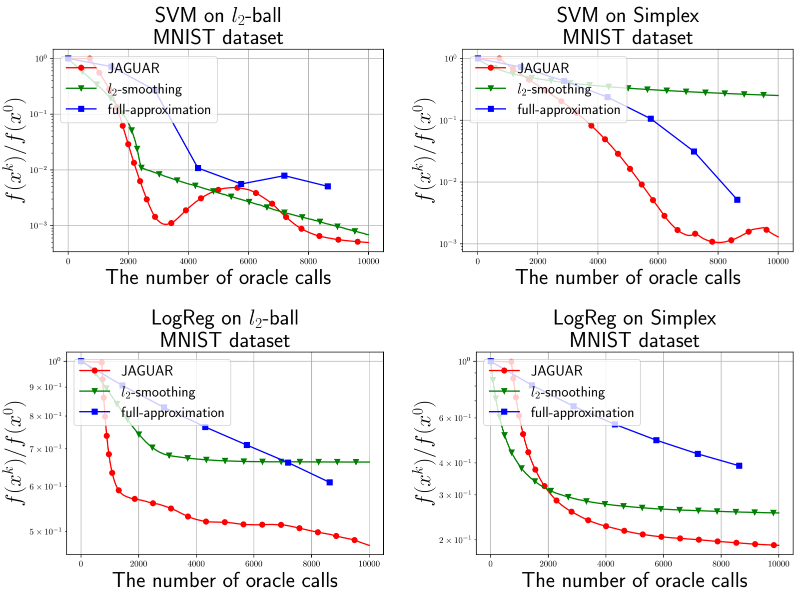
$$\min_{w \in Q} \left\{ f(w) = \frac{1}{m} \sum_{k=1}^{m} \log\left(1 + \exp\left[-y_k (Xw)_k\right]\right) + \frac{1}{2C} \|w\|^2 \right\};$$

$$\min_{w \in Q, b \in \mathbb{R}} \left\{ f(w, b) = \frac{1}{m} \sum_{k=1}^{m} (1 - y_k[(Xw)_k - b])_+ + \frac{1}{2C} \|w\|^2 \right\}.$$

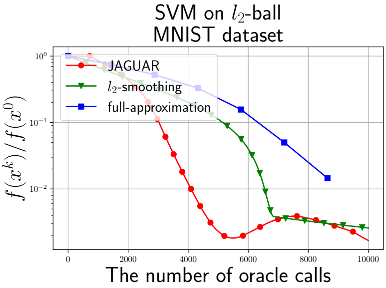
Эксперимент проводится с регуляризационным членом $C = 10$, на множествах: симплексе Δ_d и l_2-шаре; датасете MNIST.
В качестве базовых оценок градиента рассматриваются *l_2-сглаживание* и *полная аппроксимация*.
Рассматривался шум:

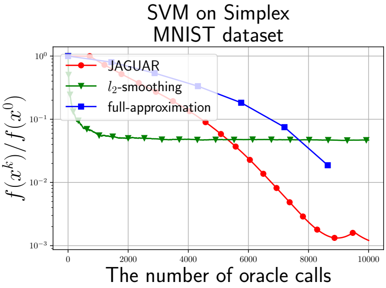
- Стохастический: $f_\delta(x) = \text{round}(f(x), 5)$;
- Детерминированный: $f_\delta(x, \xi) = f(x) + \xi; \xi \sim \mathcal{N}(0, 0.1)$.

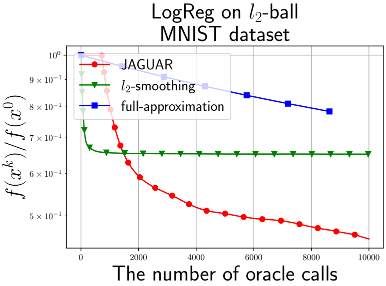
Эксперимент для нестохастического случая

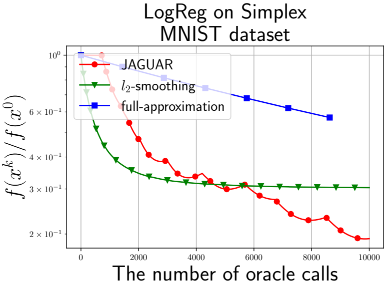


Эксперимент для стохастического случая









Выносится на защиту

1. Предложен робастый алгоритм аппроксимация градиента `JAGUAR`, использующий $\mathcal{O}(1)$ вызовов оракула на каждой итерации.
2. Доказаны теоретические оценки сходимости использования данного метода для нестохастического и стохастического случаев в алгоритме Франка-Вульфа. Показано теоретическое превосходство над *l_2-сглаживанием* и *полной аппроксимаций*.
3. Проведены вычислительные эксперименты, в которых сравнивается `JAGUAR`-аппроксимация с l_2-сглаживанием и полной аппроксимаций на различных задачах минимизации. Показано практическое превосходство.

Литература

- Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. Naval research logistics quarterly, 3(1-2):95–110, 1956.
- Darina Dvinskikh, Vladislav Tominin, Iaroslav Tominin, and Alexander Gasnikov. Noisy zeroth-order optimization for non-smooth saddle point problems. In International Conference on Mathematical Optimization Theory and Operations Research, pages 18–33. Springer, 2022.
- Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 3468–3477. PMLR, 2019.