Исследование метода тяжелого шарика с усреднением

Малиновский Григорий Станиславович
Научный руководитель: Поляк Б.Т.

Московский физико-технический институт

ВЦ РАН, Москва, 2019

Цель работы

Задача

Требуется создать эффективный алгоритм для обучения моделей, в особенности нейронных сетей. А также теоретически обосновать свойства алгоритмов типа "тяжелый шарик".

Мотивация

В настоящее время самыми популярными методами оптимизации для обучения моделей являются ускоренные градиентные методы, а также их адаптации. Однако они нуждаются в теоретической подкреплении.

Предложение

В данной работе основной задачей является исследование сходимости среднего по Чезаро метода "тяжелого шарика". А также обобщение некоторых результатов для невыпуклой оптимизации.

Литература

- Some methods of speeding up the convergence of iteration methods, BT Polyak, USSR Computational Mathematics and Mathematical Physics 4 (5), 1-17
- On Cezari convergence of gradient methods for approximating saddle points of convex-concave functions AS NEMIROVSKII, DB IUDIN DOKLADY AKADEMII NAUK SSSR 239 (5), 1056-1059
- Acceleration of stochastic approximation by averaging, Boris T Polyak, Anatoli B Juditsky, SIAM Journal on Control and Optimization 30 (4), 838-855
- Global convergence of the heavy-ball method for convex optimization, E Ghadimi, HR Feyzmahdavian, M Johansson, 2015 European Control Conference (ECC), 310-315

Постановка задачи

$$x_* = \arg\min_{x \in \mathbb{R}^d} f(x)$$

Рассматриваются следующие классы:

- $\mathcal{F}_L^{1,1}$ — выпуклая, непрерывно дифференцируемая и имеющая Липшецев градиент с константой L
- $\mathcal{F}_{\mu,L}^{1,1}$ — еще и строго выпуклая с константой μ
- невыпуклые непрерывно дифференцируемая и имеющая Липшецев градиент с различными ограничениями

Метод тяжелого шарика

Input: размер шага α, параметр импульса $\mu \in [0, 1)$, начальная стартовая $\mathbf{x}_0 \in \mathbb{R}^d$, нам доступны запросы к оракулу для получения градиентов $f : \mathbb{R}^d \to \mathbb{R}$.
function: $\mathrm{HB}(\mathbf{x}_0, \alpha, \mu)$
Initialize: $\mathbf{x}_1 = \mathbf{x}_0$
For: $t = 1, 2, \ldots$ **do**

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k) + \mu(\mathbf{x}_k - \mathbf{x}_{k-1})$$

End for
End function

Метод тяжелого шарика

Запишем метод тяжелого шарика в виде блочной матрицы, введем $2n$-мерный вектор $\mathbf{z}_k = \{\mathbf{x}_k, \mathbf{x}_{k-1}\}$, тогда итерационный процесс имеет вид:

$$\mathbf{z}_{k+1} = \mathbf{F}\mathbf{z}_k + o(\mathbf{z}_k),$$

где матрица $2n \times 2n$ имеет вид:

$$\mathbf{F} = \begin{pmatrix} (1+\beta)\mathbf{I} - \alpha\mathbf{B} & -\alpha\mathbf{B} \\ \mathbf{I} & 0 \end{pmatrix},$$

где $\mathbf{B} = \nabla^2 f(\mathbf{x}_*)$

Усреднение по Чезаро

- $\hat{\mathbf{x}}_{k+1} = \frac{\mathbf{x}_1 + \ldots \mathbf{x}_{k+1}}{k+1} = \frac{k}{k+1}\hat{\mathbf{x}}_k + \frac{1}{k+1}\mathbf{x}_{k+1}$ — итеративная формула пересчета
- Следствие теоремы Штольца: если числовая последовательность a_k сходится к a_*, то $\frac{a_1 + \cdots + a_{k+1}}{k+1}$ сходится к a_*
- Лемма: пусть матрица $\mathbf{F}$ стабильна и $\mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k$, тогда $\hat{\mathbf{x}}_{k+1} = \frac{\mathbf{x}_1 + \ldots \mathbf{x}_{k+1}}{k+1}$ сходится как $O(\frac{\lambda_{max}(\mathbf{F}(\mathbf{I} - \mathbf{F})^{-1})}{k})$

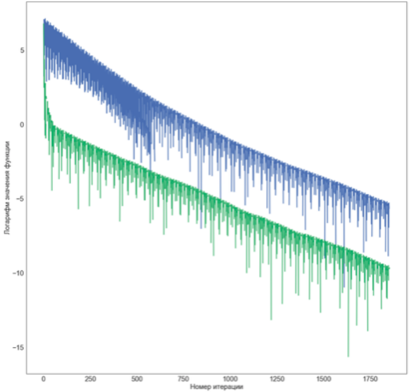
Базовые результаты

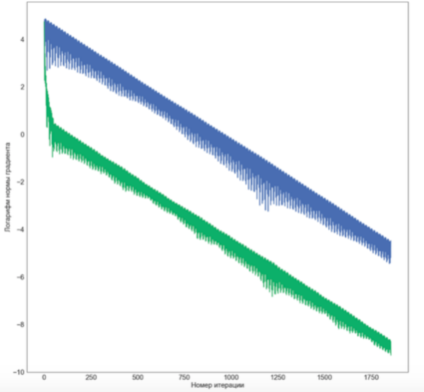
- $\mathcal{F}^{1,1}_{\mu,L}$ — глобальная линейная сходимость
- $\mathcal{F}^{1,1}_{L}$ — глобальная сублинейная сходимость последовательности и средних $O(1/k)$
- глобальная линейная сходимость при условии Поляка-Лоясиевича

Ожидаемые результаты

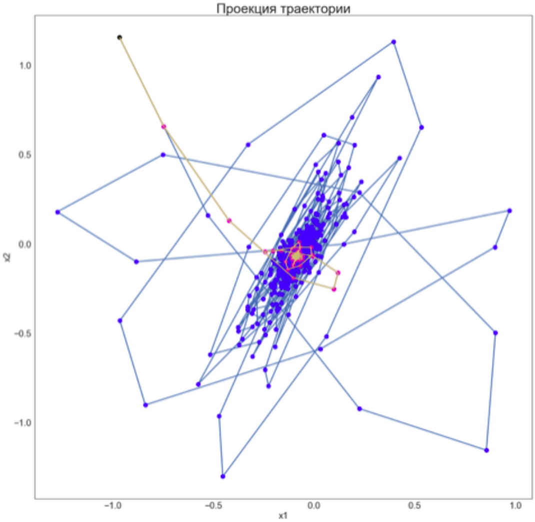
- $\mathcal{F}_{\mu,L}^{1,1}$ — глобальная(локальная) линейная сходимость средних с лучшей константой
- $\mathcal{F}_{L}^{1,1}$ — глобальная(локальная) ускоренная сходимость средних $O(1/k^2)$
- глобальная линейная сходимость при условии Поляка-Лоясиевича для средних

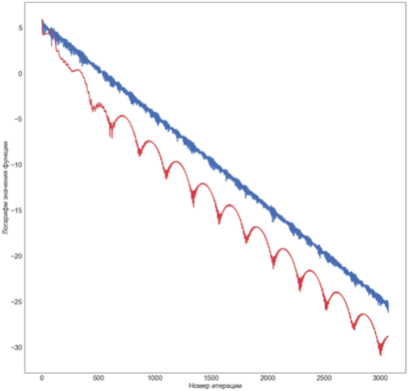
Квадратичная функция

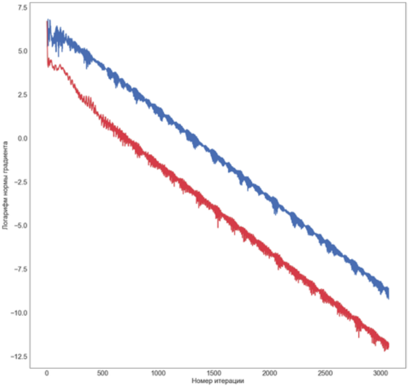




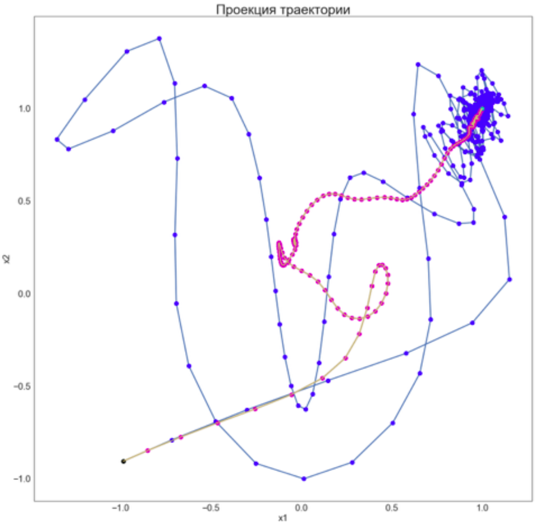
Квадратичная функция



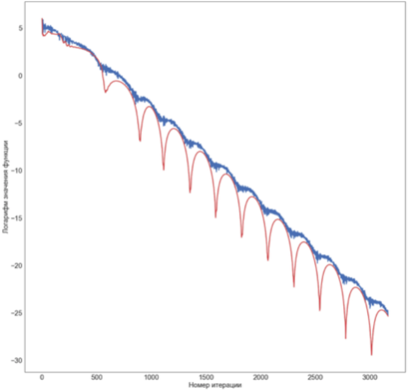

5
0
−5
−10
−15
−20
−25
−30
Логарифм значения функции
0
500
1000
1500
2000
2500
3000
Номер итерации

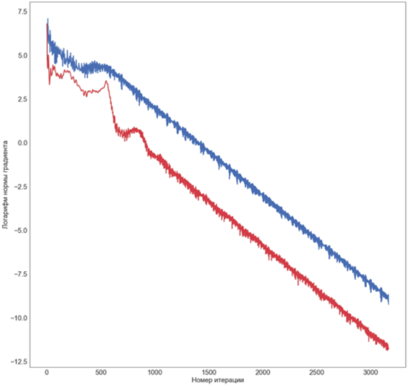

7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
Логарифм нормы градиента
0
500
1000
1500
2000
2500
3000
Номер итерации

Функция Розенброка

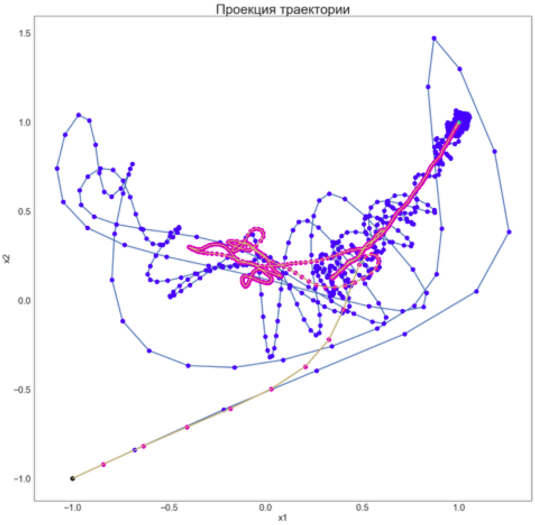


Функция Нестерова





Функция Нестерова



Заключение

- Был предложен усовершенственный метод "тяжелого шарика"для оптмизации функции ошибки
- Были доказаны оценки сходимости и проанализированы свойства
- Были проведены эксперементы на пробных функциях