# Методы выпуклой оптимизации высокого порядка*

***Селиханович Д. О.***[1,2]***, Гасников А. В.***[1,2]***, Воронцова Е. А.***[3]
selihanovich.do@phystech.edu (Селиханович Д.О.), gasnikov.av@mipt.ru (Гасников А.В.), vorontsovaea@gmail.com (Воронцова Е.А.)
[1]Московский физико-технический институт, Институтский пер., 9, Долгопрудный, Московская обл., 141700
[2]Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Большой Каретный пер., 19, строение 1, Москва, 127051
[3]Дальневосточный федеральный университет, Владивосток, Россия

Для выпуклых задач не очень больших размерностей эффективно (до $n \sim 10^3$, иногда даже до $n \sim 10^4$) применяются методы высокого порядка. Принято считать, что это методы второго порядка (использующие вторые производные оптимизируемой функции). В данной работе рассматривается метод второго порядка, который является частным случаем общего метода, предложенного в начале 2018 года в работе Ю.Е. Нестерова [1]. Предлагается на примере задачи логистической регрессии рассмотреть сходимость указанного метода, а также сравнить с другими известными методами, решающими эту задачу - быстрым градиентным методом [2] и реализованными в библиотеке LIBLINEAR [3].

**Ключевые слова**: *выпуклая оптимизация, матрица Гессе, нижние оценки, методы высокого порядка, тензорные методы, быстрый градиентный метод Нестерова, логистическая регрессия, метод сопряжённых градиентов, LBFGS.*

## Введение

Целью данной работы является применение методов высокого порядка [1] в задачах выпуклой оптимизации и исследование их свойств. В качестве тестовой функции рассматривается функция логистических потерь и на основании указанного метода оптимизации строится линейный классификатор, который обучается на данных Covertype Data Set [4]. В качестве критериев сравнения алгоритма с другими были выбраны скорость сходимости по итерациям и качество построенных классификаторов на тестовых данных. Особенностью исследований является применение разных способов решения вспомогательной задачи тензорного шага Нестерова, а именно методов сопряжённых градиентов и LBFGS. Целесообразность применения нового метода объясняется общепринятой практикой использования методов второго порядка в задаче логистической регрессии - Ньютона и IRLS. Рассматриваемый метод сравнивается с быстрым градиентным спуском Нестерова [2] и методом из библиотеки sklearn [3]. В приложении к работе выводится оценка константы Липшица $L_2$ для функции логистических потерь, которая играет важную роль в методе Нестерова второго порядка.

## Постановка задачи

Данные Covertype Data Set [4] представляют собой задачу прогнозирования типа лесного покрова в штате Колорадо по картографическим признакам. Задача мультиклассовой классификации была предобработана в задачу бинарной. Выборка состоит 581012 объ-

Научный руководитель: Гасников А. В. Задачу поставил: Александр Владимирович Гасников. Консультант: Воронцова Е. А.

ектов, наделённых 54 признаками. Из 54 признаков 10 количественные и 44 бинарные. Количественные признаки нормированы на $[0, 1]$.

Предлагается на данной задаче сравнить между собой линейные классификаторы, которые будут отличаться методом нахождения оптимального вектора весов. Так как все рассматриваемые методы работают с гладкими функциями, то в качестве оптимизируемой функции потерь была выбрана логистическая:

$$loss(w, y, X) = \sum_{i=1}^{m} \ln\Big(1 + \exp\big(-y_i\langle X_i, w\rangle\big)\Big) \to \min_{w\in\mathbb{R}^n}, \tag{1}$$

где $y_i \in \{1, -1\}$ - классы объектов, $m$ - число объектов, $X$ - матрица объектов-признаков с константым признаком 1, $n$ - размерность вектора в решающем правиле линейного классификатора:

$$y_{predict}(w, x) = sign\langle w, x\rangle. \tag{2}$$

В работе рассматриваются несколько способов оптимизации функции (1):

— Метод Нестерова второго порядка с разными способами выполнения тензорного шага - с помощью сопряжённых градиентов и LBFGS;
— Быстрый градиентный метод Нестерова;
— Методом, лежащим в основе работы логистической регрессии из библиотеки sklearn с опцией solver = liblinear.

.

## Базовый алгоритм

В работе [1] рассматривается следующий итерационный процесс для решения задачи выпуклой оптимизации $\min\limits_{x\in\mathbb{E}} f(x)$:

$$x_{t+1} = T_{p,M}(x_t), t \geqslant 0, \tag{3}$$

где функция $f(x) \in \mathcal{F}_p$ - класс выпуклых и $p$-раз гладких функций, а $T_{p,M}(x_t) \in \arg\min\limits_{y\in\mathbb{E}} \Omega_{x_t,p,M}(y)$,

$$\Omega_{x,p,M}(y) = f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(x)[y - x]^i + \frac{M}{(p-1)!} \frac{||y - x||^{p+1}}{p+1} \tag{4}$$

при $M \geqslant L_p$, $L_p$ - константа Липшица для $p$-й производной функции $f$. Здесь $D^i$ обозначает $i$-ю производную функции $f(x)$.

Предлагается рассмотреть случай $p = 2$ для евклидовой нормы на примере функции $loss(w, y, X)$:

$$\Omega_{x,2,M}(y) = f(x) + \langle\nabla f(x), y - x\rangle + \big\langle\nabla^2 f(x)(y - x), y - x\big\rangle + \frac{M}{3}||y - x||_2^3 \to \min_{y\in\mathbb{R}^n}. \tag{5}$$

В терминах оптимизируемого вектора параметров $w$ и обучающей выборки $(X_i, y_i)_{i=1}^m$ процесс при $t \geqslant 0$ примет вид:

$$w_{t+1} = T_{2,M}(w_t) = \arg\min_{w\in R^n} \Big[\langle\nabla_w loss(w_t, y, X), w - w_t\rangle + \\ + \big\langle\nabla_w^2 loss(w_t, y, X)(w - w_t), w - w_t\big\rangle + \frac{M}{3}||w - w_t||_2^3\Big]. \tag{6}$$

Для данной задачи градиент и гессиан функции потерь равны:

$$\nabla_w loss(w_t, y, X) = -\sum_{i=1}^{m} \frac{y_i X_i}{1 + \exp\big(y_i \langle X_i, w\rangle\big)}, \tag{7}$$

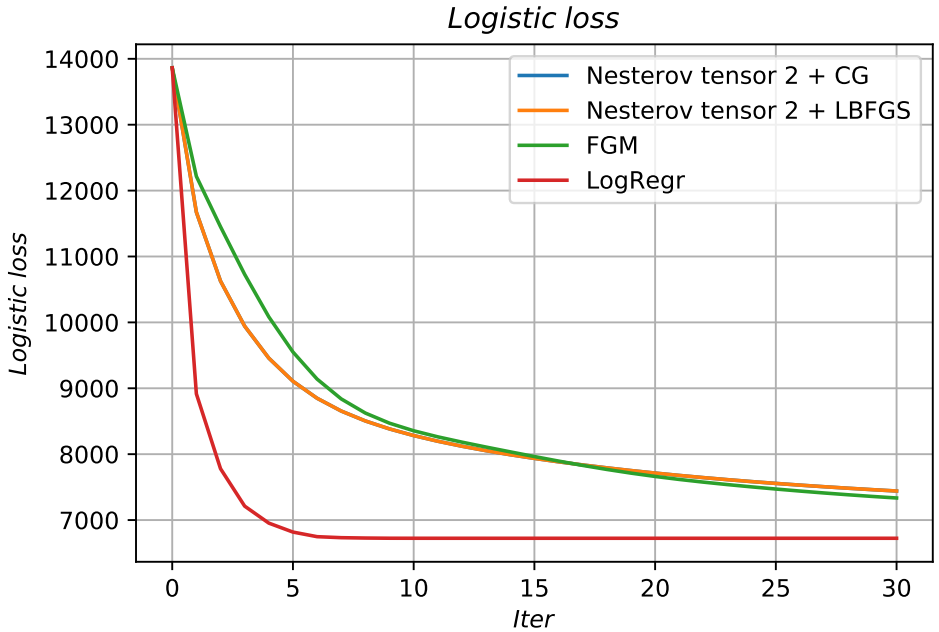
$$\nabla_w^2 loss(w_t, y, X) = \sum_{i=1}^{m} \frac{\exp\big(y_i \langle X_i, w\rangle\big)}{\Big(1 + \exp\big(y_i \langle X_i, w\rangle\big)\Big)^2} X_i X_i^T. \tag{8}$$

В приложении показывается, что в качестве константы $M \geqslant L_2$ можно взять величину $\frac{1}{10}\sum\limits_{i=1}^{m} ||X_i||_2^3$. Задачу (6) предлагается решать с помощью реализованных в модуле optimize библиотеки scipy методов сопряжённых градиентов и LBFGS на основе книги [11].

## Вычислительный эксперимент

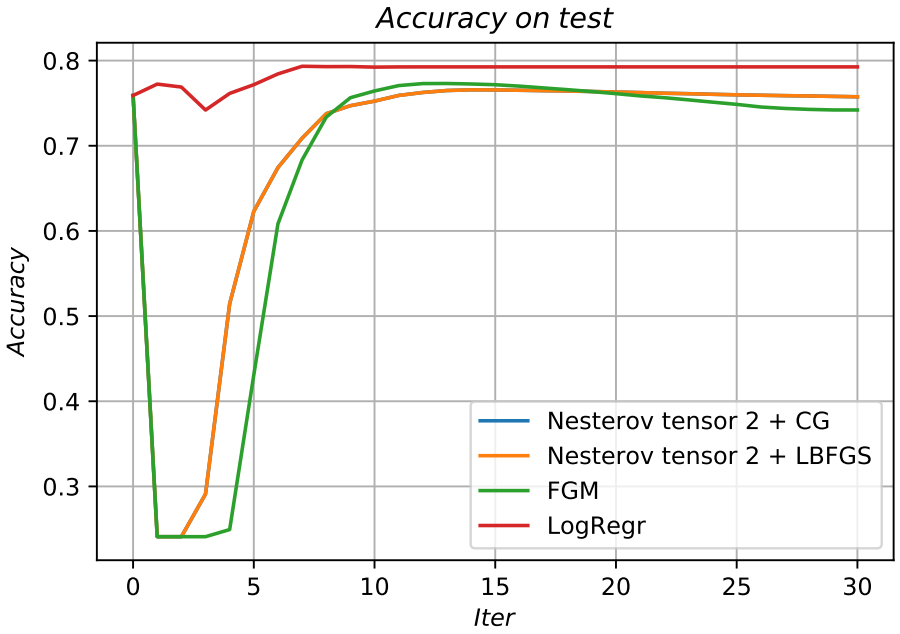
В качестве обучения линейных классификаторов было выбрано 20000 объектов, в тест - 30000. Распределение классов было следующим: 1-ый класс в train - 13978, в test - 7229, 2-й класс в train - 6022, в test = 22771. Отметим особенности первого графика. Во-первых, реализации тензорного шага Нестерова обоими методами - сопряжёнными градиентами и LBFGS - дали практически идентичный результат при использовании параметров по умолчанию. Это может объясняться тем, что обе процедуры схожи в том смысле, что сводятся к одномерному поиску. При этом метод второго порядка работает не намного лучше быстрого градиентного спуска и значительно уступает по скорости сходимости методу из библиотеки LIBLINEAR.

**Рис. 1.** Зависимость функции потерь от номера итерации




Как и можно было ожидать из первого графика, качество классификации на тестовых данных у построенных классификаторов оказалось наилучшем для Logistic Regression, реализованного в библиотеке sklearn.

**Рис. 2.** Зависимость точности классификации от номера итерации




## Выводы

Хотя в целом результаты применения рассматриваемого метода оказались хуже стандартной реализации логистической регрессии из sklearn, работа показывает, что методы высоких порядков могут использоваться в задачах машинного обучения. Важно заметить, что с повышением порядка метода увеличиваются вычислительные расходы по времени и памяти. В дальнейшем планируется подбирать константы Липшица адаптивно. При этом для указанных методом можно выбирать разные алгоритмы для выполнения тензорного шага, что порождает целый класс новых методов оптимизаций. Возможна модификация работы с рассмотрением случая $p = 3$, а также других гладких функций потерь.

## Литература

[1] NESTEROV Yu. Implementable tensor methods in unconstrained convex optimization // CORE Discussion Papers 2018005. 2018. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

[2] Родоманов Антон Олегович, Кропотов Дмитрий Александрович, Ветров Дмитрий Петрович. Анализ быстрого градиентного метода Нестерова для задач машинного обучения с $L_1$-регуляризацией // e-print. 2014. URL: http://www.machinelearning.ru/wiki/images/0/03/Rodomanov_FGM.pdf.

[3] LIBLINEAR: A library for large linear classification / Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh [и др.] // Journal of machine learning research. 2008. Т. 9, № Aug. С. 1871–1874.

[4] Blackard Jock A, Dean Denis J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables // Computers and electronics in agriculture. 1999. Т. 24, № 3. С. 131–151.

[5] Современные численные методы оптимизации. Метод универсального градиентного спуска: учебное пособие / А. В. Гасников. – М. : МФТИ, 2018. – 166с. – Изд. 2-е, доп.

[6] The global rate of convergence for optimal tensor methods in smooth convex optimization / Alexander Gasnikov, Dmitry Kovalev, Ahmed Mohhamed [и др.] // arXiv preprint arXiv:1809.00382. 2018.

[7] Nesterov Yu. Accelerating the cubic regularization of Newton's method on convex problems // Mathematical Programming. 2008. Т. 112, № 1. С. 159–181.

[8] Monteiro Renato DC, Svaiter Benar Fux. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods // SIAM Journal on Optimization. 2013. Т. 23, № 2. С. 1092–1125.

[9] Долгополик М.В. Оптимальный градиентный метод минимизации выпуклых функций // e-print. 2016. URL: http://www.apmath.spbu.ru/cnsa/pdf/2016/DolgopolikMV_10November2016.pdf.

[10] Bubeck Sébastien [и др.]. Convex optimization: Algorithms and complexity // Foundations and Trends® in Machine Learning. 2015. Т. 8, № 3-4. С. 231–357.

[11] Wright S., Nocedal J. Numerical optimization 2nd // Springer Science. 2006.

## Приложение. Оценка констант Липшица для функции логистических потерь.

Напомним определение константы Липшица для $p$-й производной функции $f$:

$$||D^p f(x) - D^p f(y)|| \leqslant L_p||x-y||,\ x,y \in \operatorname{dom} f, p \geqslant 1. \tag{9}$$

Для $p = 1 \to D^1 f(x) = \nabla f(x)$, для $p = 2 \to D^2 f(x) = \nabla^2 f(x)$. Будем рассматривать случай евклидовой нормы.

Оценим разность $||\nabla_w loss(w_1, y, X) - \nabla_w loss(w_2, y, X)||_2$ согласно (7), пользуясь неравенством треугольника для нормы и равенством $|y_i| = 1\ \forall i \in \overline{1,m}$:

$$||\nabla_w loss(w_1,y,X) - \nabla_w loss(w_2,y,X)||_2 \leqslant \sum_{i=1}^{m} \Big| \frac{1}{1+\exp\big(y_i\langle X_i, w_1\rangle\big)} - \frac{1}{1+\exp\big(y_i\langle X_i, w_2\rangle\big)} \Big| ||X_i||_2. \tag{10}$$

Заметим, что согласно формуле Лагранжа конечных приращений:

$$\Big|\frac{1}{1+e^x} - \frac{1}{1+e^y}\Big| \leqslant \max_{x\in\mathbb{R}} \Big|\big(\frac{1}{1+e^x}\big)'\Big| |x-y| = \max_{z\in\mathbb{R}} \Big|\frac{e^z}{(1+e^z)^2}\Big| |x-y| = \frac{1}{4}|x-y|. \tag{11}$$

Это позволяет оценить величину под знаком суммы при помощи (11) и неравенства Коши-Буняковского-Шварца:

$$\Big| \frac{1}{1+\exp\big(y_i\langle X_i, w_1\rangle\big)} - \frac{1}{1+\exp\big(y_i\langle X_i, w_2\rangle\big)} \Big| \leqslant \frac{1}{4}\big|\langle X_i, w_1 - w_2\rangle\big| \leqslant \frac{1}{4}||w_1 - w_2||_2 ||X_i||_2. \tag{12}$$

Суммируя, получим оценку:

$$||\nabla_w loss(w_1,y,X) - \nabla_w loss(w_2,y,X)||_2 \leqslant \frac{1}{4}\sum_{i=1}^{m} ||X_i||_2^2 ||w_1 - w_2||_2 = M_1||w_1 - w_2||_2, \tag{13}$$

где $M_1 = \frac{1}{4}\sum\limits_{i=1}^{m} ||X_i||_2^2 \geqslant L_1$ - оценка на константу Липшица $L_1$.

Аналогичным способом получим оценку на константу Липшица $L_2$:

$$||\nabla_w^2 loss(w_1,y,X) - \nabla_w^2 loss(w_2,y,X)||_2 \leqslant$$
$$\leqslant \sum_{i=1}^{m} \Big| \frac{\exp\big(y_i\langle X_i, w_1\rangle\big)}{\big(1+\exp\big(y_i\langle X_i, w_1\rangle\big)\big)^2} - \frac{\exp\big(y_i\langle X_i, w_2\rangle\big)}{\big(1+\exp\big(y_i\langle X_i, w_2\rangle\big)\big)^2} \Big| ||X_i X_i^T||_2. \tag{14}$$

По определению нормы оператора:

$$|||X_iX_i^T||_2 = \sup_{z\neq 0}\frac{||X_iX_i^Tz||_2}{||z||_2} = \sup_{z\neq 0}\sqrt{\frac{\langle X_i(X_i^Tz), X_i(X_i^Tz)\rangle}{\langle z,z\rangle}} = \sup_{z\neq 0}\frac{|X_i^Tz|||X_i||_2}{||z||_2} \leqslant$$
$$\leqslant \sup_{z\neq 0}\frac{||z||_2||X_i||_2^2}{||z||_2} = ||X_i||_2^2, \qquad (15)$$

где мы воспользовались ассоциативностью матричного умножения и неравенством Коши-Буняковского-Шварца. С другой стороны, предполагая $X_i \neq 0_n$, получим

$$|||X_iX_i^T||_2 = \sup_{z\neq 0}\frac{||X_iX_i^Tz||_2}{||z||_2} \geqslant \frac{||X_iX_i^TX_i||_2}{||X_i||_2} = \frac{||X_i(X_i^TX_i)||_2}{||X_i||_2} = \frac{|X_i^TX_i|||X_i||_2}{||X_i||_2} = ||X_i||^2. \quad (16)$$

Очевидно, что полученная оценка $|||X_iX_i^T||_2 \geqslant ||X_i||^2$ в случае $X_i \neq 0_n$ остаётся справедливой и в случае $X_i = 0_n$. Таким образом, мы заключаем, что всегда имеет место равенство $||X_iX_i^T||_2 = ||X_i||_2^2$. Величину под знаком суммы в (14) также оценим через формулу Лагранжа конечных приращений:

$$\Big|\frac{e^x}{(1+e^x)^2} - \frac{e^y}{(1+e^y)^2}\Big| \leqslant \max_{z\in\mathbb{R}}\Big|\Big(\frac{e^z}{(1+e^z)^2}\Big)'\Big||x-y| =$$
$$= \max_{z\in\mathbb{R}}\Big|\frac{e^z(1-e^z)}{(1+e^z)^3}\Big||x-y| = \frac{1}{6\sqrt{3}}|x-y| \qquad (17)$$

Отсюда:

$$\Big|\frac{\exp\big(y_i\langle X_i, w_1\rangle\big)}{\big(1+\exp\big(y_i\langle X_i, w_1\rangle\big)\big)^2} - \frac{\exp\big(y_i\langle X_i, w_2\rangle\big)}{\big(1+\exp\big(y_i\langle X_i, w_2\rangle\big)\big)^2}\Big| \leqslant$$
$$\leqslant \frac{1}{6\sqrt{3}}\big|\langle X_i, w_1-w_2\rangle\big| \leqslant \frac{1}{6\sqrt{3}}||X_i||_2||w_1-w_2||_2. \qquad (18)$$

Суммируя, получим оценку:

$$||\nabla_w^2 loss(w_1,y,X) - \nabla_w^2 loss(w_2,y,X)||_2 \leqslant \frac{1}{6\sqrt{3}}\sum_{i=1}^m ||X_i||_2^3||w_1-w_2||_2 \leqslant$$
$$\leqslant \frac{1}{10}\sum_{i=1}^m ||X_i||_2^3||w_1-w_2||_2 = M_2||w_1-w_2||_2, \qquad (19)$$

где $M_2 = \frac{1}{10}\sum\limits_{i=1}^m ||X_i||_2^3 \geqslant L_2$ - оценка на константу Липшица $L_2$.