КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2021 Т. 13 № 6 С. 1137–1147
DOI: 10.20537/2076-7633-2021-13-6-1137-1147

Ки&М

ЧИСЛЕННЫЕ МЕТОДЫ И ОСНОВЫ ИХ РЕАЛИЗАЦИИ

УДК: 519.85

Метод эллипсоидов для задач выпуклой стохастической оптимизации малой размерности[1]

Е. Л. Гладин[1,2,3,a], К. Э. Зайнуллина[1,b]

[1]Национальный исследовательский университет «Московский физико-технический институт»,
Россия, 141701, Московская область, г. Долгопрудный, Институтский пер., д. 9
[2]Институт проблем передачи информации РАН,
Россия, 127051, г. Москва, Б. Каретный пер., д. 9
[3]Сколковский институт науки и технологий,
Россия, 121205, г. Москва, Большой бульвар, д. 30, с. 1

E-mail: [a] gladin.el@phystech.edu, [b] zaynullina.ke@phystech.edu

Получено 09.11.2020, после доработки — 15.11.2021.
Принято к публикации 16.11.2021.

В статье рассматривается задача минимизации математического ожидания выпуклой функции. Задачи такого вида повсеместны в машинном обучении, а также часто возникают в ряде других приложений. На практике для их решения обычно используются процедуры типа стохастического градиентного спуска (SGD). В нашей работе предлагается решать такие задачи с использованием метода эллипсоидов с мини-батчингом. Алгоритм имеет линейную скорость сходимости и может оказаться эффективнее SGD в ряде задач. Это подтверждается в наших экспериментах, исходный код которых находится в открытом доступе. Для получения линейной скорости сходимости метода не требуется ни гладкость, ни сильная выпуклость целевой функции. Таким образом, сложность алгоритма не зависит от обусловленности задачи. В работе доказывается, что метод эллипсоидов с наперед заданной вероятностью находит решение с желаемой точностью при использовании мини-батчей, размер которых пропорционален точности в степени −2. Это позволяет выполнять алгоритм параллельно на большом числе процессоров, тогда как возможности для батч-параллелизации процедур типа стохастического градиентного спуска весьма ограничены. Несмотря на быструю сходимость, общее количество вычислений градиента для метода эллипсоидов может получиться больше, чем для SGD, который неплохо сходится и при маленьком размере батча. Количество итераций метода эллипсоидов квадратично зависит от размерности задачи, поэтому метод подойдет для относительно небольших размерностей.

Ключевые слова: стохастическая оптимизация, выпуклая оптимизация, метод эллипсоидов, мини-батчинг

Авторы выражают благодарность Гасникову Александру Владимировичу за идею для написания работы.

[1] Статья была подготовлена в ходе проектной смены «Современные методы теории информации, оптимизации и управления» («Сириус», 2–23 августа 2020 г.) и научно-образовательной школы-конференции «Управление. Информация. Оптимизация» («Сириус», 23–29 августа 2020 г.).

COMPUTER RESEARCH AND MODELING
2021 VOL. 13 NO. 6 P. 1137–1147
DOI: 10.20537/2076-7633-2021-13-6-1137-1147

Ки&М

NUMERICAL METHODS AND THE BASIS FOR THEIR APPLICATION

UDC: 519.85

Ellipsoid method for convex stochastic optimization in small dimension

E. L. Gladin[1,2,3,a], K. E. Zainullina[1,b]

[1]National Research University Moscow Institute of Physics and Technology,
9, Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia
[2]Institute for Information Transmission Problems RAS,
9, B. Karetny lane, Moscow, 127051, Russia
[3]Skolkovo Institute of Science and Technology,
30/1, Bolshoy Boulevard, Moscow, 121205, Russia

E-mail: [a] gladin.el@phystech.edu, [b] zaynullina.ke@phystech.edu

Received 09.11.2020, after completion — 15.11.2021.
Accepted for publication 16.11.2021.

The article considers minimization of the expectation of convex function. Problems of this type often arise in machine learning and a variety of other applications. In practice, stochastic gradient descent (SGD) and similar procedures are usually used to solve such problems. We propose to use the ellipsoid method with mini-batching, which converges linearly and can be more efficient than SGD for a class of problems. This is verified by our experiments, which are publicly available. The algorithm does not require neither smoothness nor strong convexity of the objective to achieve linear convergence. Thus, its complexity does not depend on the conditional number of the problem. We prove that the method arrives at an approximate solution with given probability when using mini-batches of size proportional to the desired accuracy to the power -2. This enables efficient parallel execution of the algorithm, whereas possibilities for batch parallelization of SGD are rather limited. Despite fast convergence, ellipsoid method can result in a greater total number of calls to oracle than SGD, which works decently with small batches. Complexity is quadratic in dimension of the problem, hence the method is suitable for relatively small dimensionalities.

Keywords: stochastic optimization, convex optimization, ellipsoid method, mini-batching

Citation: *Computer Research and Modeling*, 2021, vol. 13, no. 6, pp. 1137–1147 (Russian).

Authors express gratitude to Alexander Gasnikov for the idea of the paper.

Введение

В приложениях часто возникает задача оптимизации математического ожидания некоторой функции. В качестве примера можно рассмотреть машинное обучение, в котором подбор параметров модели производится путем минимизации математического ожидания функции потерь.

Пусть данные представлены в виде пар (ψ, y), где y — метка, то есть величина, которую необходимо научиться предсказывать на основе вектора признаков ψ для данного объекта. Среди многообразия задач машинного обучения часто встречается проблема классификации, в которой y принимает значения из конечного множества классов, и регрессии, где метка принадлежит некоторому подмножеству числовой оси. Данные имеют некоторое распределение с неизвестной функцией плотности $P(\psi, y)$. Обозначим через $\widehat{y} = h(\psi; \theta)$ предсказание модели с вектором весов θ на объекте с вектором признаков ψ. Качество предсказания на паре (ψ, y), как правило, определяется значением некоторой функции потерь (лосс-функции) $\ell(\widehat{y}, y)$. Задача машинного обучения, таким образом, состоит в минимизации математического ожидания функции потерь [Bottou, Curtis, Nocedal, 2018]:

$$\min_{\theta} \Big\{ L(\theta) := \mathbb{E}_P \, \ell(h(\psi; \theta), y) \Big\}. \tag{1}$$

Поскольку реальное распределение данных P неизвестно, на практике часто осуществляется минимизация эмпирического риска [Deisenroth, Faisal, Ong, 2020]:

$$\widetilde{\theta}_{ERM} = \arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \ell(h(\psi_i; \theta), y_i), \tag{2}$$

где $\{(\psi_i, y_i)\}_{i=1}^{m}$ — элементы выборки.

Вместо того чтобы работать с готовой выборкой, можно решать задачу в онлайн-режиме, получая обучающие примеры последовательно [Hazan, 2016]. Для этого существует множество процедур, одной из самых известных является стохастический градиентный спуск (SGD) (см., например, [Bottou, 2010]). На каждой итерации этого метода вычисляется градиент лосс-функции для очередной пары (ψ_i, y_i) и осуществляется шаг в направлении против этого градиента:

$$\theta \leftarrow \theta - \alpha_t \cdot \nabla_\theta \ell(h(\psi_i; \theta), y_i),$$

где ∇_θ обозначает градиент (в недифференцируемом случае — субградиент) по переменным θ, α_t — размер шага на итерации t. В общем случае, когда рассматривается задача минимизации математического ожидания произвольной функции

$$\min_{x \in Q} \Big\{ f(x) := \mathbb{E}_\xi f(x, \xi) \Big\}, \tag{3}$$

где ξ — случайная величина, на итерациях SGD используется *стохастический субградиент* $\nabla_x f(x, \xi)$, обладающий свойством несмещенности: $\mathbb{E}_\xi \nabla_x f(x, \xi) = \nabla f(x)$. При выводе оценок сложности методов решения задачи (3) часто предполагают, что функция f выпуклая, а стохастический субградиент удовлетворяет некоторым условиям (например, ограниченность дисперсии). В одной из простейших вариаций этого метода производится усреднение по *батчу* — набору реализаций стохастического субградиента:

$$x \leftarrow x - \alpha_t \cdot \overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right), \quad \text{где} \quad \overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) := \frac{1}{r} \sum_{i=1}^{r} \nabla_x f(x, \xi^l). \tag{4}$$

Процедура (4) называется mini-batch SGD (стохастический градиентный спуск с мини-батчингом). В контексте машинного обучения батч представляет собой набор объектов из обучающей

выборки. Стохастический градиентный спуск стал широко использоваться для обучения нейронных сетей [Bottou, 2012], что привело к появлению множества разновидностей этого метода, включая Adagrad [Duchi, Hazan, Singer, 2011], Adadelta [Zeiler, 2012], RMSprop [Tieleman, Hinton, 2012], Adam [Kingma, Ba, 2014], Nadam [Dozat, 2016] и многие другие. Хотя эти методы хорошо показывают себя во многих практических задачах, существуют примеры выпуклых задач, на которых некоторые из них не сходятся к оптимуму [Reddi, Kale, Kumar, 2018].

В стохастической оптимизации часто рассматриваются два подхода к определению точности приближенного решения $\widetilde{\theta}$ задачи (1) [Dvurechensky, Gasnikov, Lagunovskaya, 2018]. Первый из них связан с математическим ожиданием точности по функции. В этом случае $\widetilde{\theta}$ называется ε-решением (1) для $\varepsilon > 0$, если $\mathbb{E}L(\widetilde{\theta}) - L_* \leqslant \varepsilon$, где L_* — оптимальное значение в задаче (1), математическое ожидание берется по случайности, возникающей в стохастическом алгоритме. Второй подход позволяет ограничить вероятность того, что значение функции в приближенном решении сильно отклоняется от оптимального. В этом случае $\widetilde{\theta}$ называется (ε, β)-решением (1) для $\varepsilon > 0$, $\beta \in (0, 1)$, если $\mathbb{P}\{L(\widetilde{\theta}) - L_* > \varepsilon\} \leqslant \beta$.

Одним из показателей эффективности метода решения задачи (1) является верхняя оценка на число итераций, гарантирующих желаемую точность решения. Если функция $L(\theta)$ является выпуклой и непрерывной на компактном множестве, SGD позволяет получить ε-решение задачи (1) за $N = O\left(\frac{1}{\varepsilon^2}\right)$ итераций [Zinkevich, 2003]. Некоторые условия гладкости позволяют улучшить эту оценку до $O\left(\frac{1}{\varepsilon}\right)$ [Murata, 1998]. Если же мы хотим, чтобы $\widetilde{\theta}_{ERM}$ из (2) было ε-решением задачи (1), то потребуется [Shapiro, Dentcheva, Ruszczynski, 2014] размер выборки $m \sim \frac{n}{\varepsilon^2}$, где n — размерность θ. Впрочем, этот результат можно также привести к $m \sim \frac{1}{\varepsilon^2}$, если регуляризовать целевую функцию в (2), добавив к ней слагаемое $\sim \varepsilon\|\theta\|_2^2$. Тем не менее найти точное решение (2) — часто сложная, а порой и невыполнимая задача.

Вот уже несколько десятков лет разрабатываются способы производить вычисления эффективнее с помощью распределенных систем [Bertsekas, Tsitsiklis, 1989]. Параллельные и распределенные алгоритмы позволяют решать задачи в разы быстрее, а в последние годы благодаря продолжающемуся развитию искусственного интеллекта и вычислительной техники они стали особенно популярны. В частности, процедуры типа (4) можно осуществлять на нескольких процессорах. Рассмотрим модель, в которой параллельно выполняются r вычислений стохастического субградиента. Такой модели соответствует централизованная система с быстрой коммуникацией между вычислительными узлами. В случае замкнутой выпуклой функции поиск (ε, β)-решения (1) с помощью процедур типа SGD удается производить лишь на $\Theta(\ln(\beta^{-1}))$-машинах за счет параллельного вычисления независимых решений [Dvurechensky, Gasnikov, Lagunovskaya, 2018], что не дает большого преимущества. Для случая дифференцируемой функции с непрерывным по Липшицу градиентом существуют методы, допускающие батч-параллелизацию на $\Theta\left(\frac{1}{\varepsilon^{3/2}}\right)$-процессорах. Таким образом, число (параллельных) итераций ускоренного метода стохастического градиентного спуска будет всего $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ [Cotter et al., 2011]. Указанная асимптотическая зависимость от ε соответствует нижней оценке сложности стохастических градиентных методов [Woodworth et al., 2018; Woodworth et al., 2021], то есть является оптимальной.

Существует класс алгоритмов, называемых методами секущей плоскости и позволяющих минимизировать непрерывные выпуклые функции с линейной скоростью, то есть их сложность зависит от желаемой точности как $N = O(\ln \varepsilon^{-1})$. Примерами таких алгоритмов служат метод центров тяжести, метод эллипсоидов, метод Вайды (см., например, [Bubeck, 2015]) и другие. Одним из их недостатков является рост вычислительной сложности с размерностью переменных, что делает их неприменимыми для задач большой размерности. Возможно, это одна из причин, по которым этим алгоритмам уделяется меньше внимания, чем градиентным методам. В частности, они не находят широкого применения в стохастической оптимизации. Тем не менее

если бы был найден способ использовать такие методы для задач типа (3), то их благоприятная асимптотическая сложность $N(\varepsilon)$ могла бы стать хорошим преимуществом в случае малой размерности переменных и высокой требуемой точности. В настоящей статье предлагается такой вариант метода эллипсоидов для решения задачи (3). Использование мини-батчинга позволяет достичь линейной скорости сходимости при параллелизации на $\widetilde{O}\left(\frac{1}{\varepsilon^2}\right)$ машинах. Число итераций метода квадратично зависит от размерности задачи. Таким образом, предлагаемый алгоритм подойдет для задач небольшой размерности (несколько десятков) в случае, когда есть достаточное количество вычислительных узлов и требуется высокая точность решения. Эффективность предлагаемого метода для таких постановок проиллюстрирована численным экспериментом, в котором он применяется для обучения модели логистической регрессии.

Постановка задачи и полученные результаты

Рассмотрим задачу

$$\min_{x\in Q}\left\{f(x) := \mathbb{E}_{\xi} f(x,\,\xi)\right\}, \tag{5}$$

где $Q \subseteq \mathbb{R}^n$ — выпуклое компактное множество с непустой внутренностью, размерность n относительно небольшая (до ста), $f(x)$ — непрерывная выпуклая функция. Обозначим $D := \sup\limits_{x,\,y\in Q} \|x - y\|$, $B := \sup\limits_{x,\,y\in Q} |f(x) - f(y)|$, ρ — радиус некоторого евклидова шара, содержащегося в Q. Здесь и далее $\|\cdot\|$ означает евклидову норму. Будем считать, что стохастический субградиент $\nabla_x f(x,\,\xi)$ удовлетворяет для некоторого $\sigma > 0$ условию

$$\mathbb{E}_{\xi} \exp\left(\frac{\|\nabla_x f(x,\,\xi) - g\|^2}{\sigma^2}\right) \leqslant \exp(1), \tag{6}$$

где g — некоторый субградиент f в точке x (обозначение: $g \in \partial f(x)$). В настоящей статье предлагается использовать метод эллипсоидов с мини-батчингом и доказывается, что он позволяет найти $(\varepsilon,\,\beta)$-решение задачи (5) для $\varepsilon > 0$, $\beta \in (0,\,1)$ после

$$N = \left\lceil 2n^2 \ln\left(\frac{DB}{\rho\varepsilon}\right)\right\rceil$$

итераций при размере батча

$$r = \widetilde{O}\left(\frac{\sigma^2 D^2}{\varepsilon^2}\right),$$

где $\lceil\cdot\rceil$ — потолок числа, $\widetilde{O}(\cdot)$ означает $O(\cdot)$ с точностью до логарифмического по ε^{-1} и β^{-1} множителя.

Вывод оценки сложности метода

Нам потребуется следующее определение.

Определение 1. Пусть $\delta \geqslant 0$, $Q \subseteq \mathbb{R}^n$ — выпуклое множество, $f\colon Q \to \mathbb{R}$ — выпуклая функция. Вектор $g \in \mathbb{R}^n$ называется δ-субградиентом f в точке $x \in Q$, если

$$f(y) \geqslant f(x) + \langle g,\, y - x\rangle - \delta \quad \forall y \in Q.$$

Множество δ-субградиентов f в точке x обозначается как $\partial_\delta f(x)$.

Заметим, что δ-субградиент при $\delta = 0$ совпадает с обычным субградиентом. Приведенная ниже теорема устанавливает, что если в методе эллипсоидов (см., например, [Bubeck, 2015]) вместо субградиента использовать δ-субградиент (алгоритм 1), то неточность δ не будет накапливаться от шага к шагу.

Алгоритм 1. Метод эллипсоидов с δ-субградиентом для задачи $\min\limits_{x\in Q} f(x)$

Вход: Число итераций $N \geqslant 1$, $\delta \geqslant 0$, шар $\mathcal{B}_R \supseteq Q$, его центр c и радиус R.

$\mathcal{E}_0 := \mathcal{B}_R$, $H_0 := R^2 I_n$, $c_0 := c$.
for $k = 0, \ldots, N-1$ **do**
  if $c_k \in Q$ **then**
    $w_k := w \in \partial_\delta f(x)$
    if $w_k = 0$ **then**
      return c_k
    end if
  else
    $w_k := w$, где $w \neq 0$ таков, что $Q \subset \{x \in \mathcal{E}_k : \langle w, x - c_k\rangle \leqslant 0\}$
  end if
  $c_{k+1} := c_k - \frac{1}{n+1}\frac{H_k w_k}{\sqrt{w_k^T H_k w_k}}$
  $H_{k+1} := \frac{n^2}{n^2-1}\left(H_k - \frac{2}{n+1}\frac{H_k w_k w_k^T H_k}{w_k^T H_k w_k}\right)$
  $\mathcal{E}_{k+1} := \{x : (x - c_{k+1})^T H_{k+1}^{-1}(x - c_{k+1}) \leqslant 1\}$
end for

Выход: $x^N = \arg\min\limits_{x\in\{c_0,\ldots,c_N\}\cap Q} f(x)$

Теорема 1 ([Гладин и др., 2020]). *Пусть Q — компактное выпуклое множество, которое содержится в некотором евклидовом шаре радиусом R и включает некоторый евклидов шар радиусом ρ, $f : Q \to \mathbb{R}$ — непрерывная выпуклая функция, число $B > 0$ таково, что $|f(x) - f(x')| \leqslant B\ \forall x, x' \in Q$. После $N \geqslant 2n^2 \ln\frac{R}{\rho}$ итераций метод эллипсоидов с δ-субградиентом* (*алгоритм* 1) *возвращает такую точку $x^N \in Q$, что*

$$f(x^N) - \min_{x\in Q} f(x_*) \leqslant \frac{BR}{\rho}\exp\left(-\frac{N}{2n^2}\right) + \delta.$$

Далее приведены две леммы, которые позволяют связать δ-субградиент со стохастическим субградиентом, усредненным по батчу:

$$\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) := \frac{1}{r}\sum_{i=1}^r \nabla_x f(x, \xi^l), \quad \xi^l \text{ — независимые реализации случайной величины } \xi.$$

Лемма 1. *Пусть $g \in \partial f(x)$, и пусть для вектора $\widetilde{g}$ выполнено $\|\widetilde{g} - g\| \leqslant \varepsilon$, тогда $\widetilde{g} \in \partial_\delta f(x)$ для $\delta = \varepsilon D$, где $D := \sup\limits_{x,y\in Q}\|x - y\|$.*

Доказательство. По неравенству Коши – Буняковского, для любого $y \in Q$ справедливо

$$\langle \widetilde{g} - g, y - x\rangle \leqslant \|\widetilde{g} - g\| \cdot \|y - x\| \leqslant \|\widetilde{g} - g\| \cdot D \leqslant \varepsilon D. \tag{7}$$

В силу выпуклости f для $g \in \partial f(x)$ имеем

$$f(y) \geqslant f(x) + \langle g, y - x\rangle \quad \forall y \in Q. \tag{8}$$

Сложив (7) и (8), получим

$$f(y) \geqslant f(x) + \langle \widetilde{g}, y - x\rangle - \varepsilon D \quad \forall y \in Q, \tag{9}$$

т. е. $\widetilde{g} \in \partial_\delta f(x)$, $\delta = \varepsilon D$.

Лемма 2. *Пусть стохастический субградиент* $\nabla_x f(x, \xi)$ *удовлетворяет условию* $\mathbb{E}_\xi \exp\left(\frac{\|\nabla_x f(x,\xi)-g\|^2}{\sigma^2}\right) \leqslant \exp(1)$, *где* $g \in \partial f(x)$, *тогда для любого* $\beta \in (0, 1)$

$$\mathbb{P}\left\{\left\|\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) - g\right\| < \left[\sqrt{2} + \sqrt{6\ln\beta^{-1}}\right] \cdot \frac{\sigma}{\sqrt{r}}\right\} \geqslant 1 - \beta \quad \forall x \in Q.$$

Доказательство. Согласно теореме 2.1 (ii) из [Juditsky, Nemirovski, 2008], для любого $\gamma \geqslant 0$

$$\mathbb{P}\left\{\|S_r\| \geqslant \left[\sqrt{2} + \sqrt{2}\gamma\right] \cdot \sqrt{\sum_{i=1}^r \sigma_i^2}\right\} \leqslant \exp\left(-\frac{\gamma^2}{3}\right),$$

где S_r — сумма независимых случайных векторов $\{\zeta_i\}_{i=1}^r$ с нулевым математическим ожиданием, для которых выполнено $\mathbb{E}\exp\left(\frac{\|\zeta_i\|^2}{\sigma_i^2}\right) \leqslant \exp(1)$. В нашем случае $\zeta_i = \nabla_x f(x, \xi^i) - g$ и $\|S_r\| = r \times \left\|\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) - g\right\|$, поэтому

$$\mathbb{P}\left\{\left\|\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) - g\right\| \geqslant \left[\sqrt{2} + \sqrt{2}\gamma\right] \cdot \sqrt{\frac{\sigma^2}{r}}\right\} \leqslant \exp\left(-\frac{\gamma^2}{3}\right). \tag{10}$$

Обозначим $\beta = \exp\left(-\frac{\gamma^2}{3}\right)$, тогда $\gamma = \sqrt{3\ln\beta^{-1}}$ и неравенство (10) приобретает вид

$$\mathbb{P}\left\{\left\|\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) - g\right\| \geqslant \left[\sqrt{2} + \sqrt{6\ln\beta^{-1}}\right] \cdot \frac{\sigma}{\sqrt{r}}\right\} \leqslant \beta.$$

Тогда

$$\mathbb{P}\left\{\left\|\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) - g\right\| < \left[\sqrt{2} + \sqrt{6\ln\beta^{-1}}\right] \cdot \frac{\sigma}{\sqrt{r}}\right\} =$$
$$= 1 - \mathbb{P}\left\{\left\|\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) - g\right\| \geqslant \left[\sqrt{2} + \sqrt{6\ln\beta^{-1}}\right] \cdot \frac{\sigma}{\sqrt{r}}\right\} \geqslant 1 - \beta.$$

Следствие 1. *Пусть величина* $\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right)$ *была вычислена* N *раз,* $\widetilde{g}_i$ *— ее значение на шаге* $i \in \overline{1, N}$ *в точке* x^i. *Для любого* $\beta \in (0, 1)$ *справедливо*

$$\mathbb{P}\left(\bigcap_{i=1}^N \{\widetilde{g}_i \in \partial_\delta f(x^i)\}\right) \geqslant 1 - \beta N, \quad \text{где} \quad \delta = \left[\sqrt{2} + \sqrt{6\ln\beta^{-1}}\right] \cdot \frac{\sigma D}{\sqrt{r}}.$$

Доказательство. Согласно леммам 1 и 2, для любых $x \in Q$, $\beta \in (0, 1)$

$$\mathbb{P}\left\{\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) \in \partial_\delta f(x)\right\} \geqslant 1 - \beta, \quad \text{где} \quad \delta = \left[\sqrt{2} + \sqrt{6\ln\beta^{-1}}\right] \cdot \frac{\sigma D}{\sqrt{r}}.$$

Это эквивалентно

$$\mathbb{P}\left\{\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right) \notin \partial_\delta f(x)\right\} \leqslant \beta.$$

Значит, вероятность того, что хотя бы на одном из N шагов величина $\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right)$ не будет являться δ-субградиентом, не превышает βN. Следовательно, вероятность того, что на всех N шагах будет вычислен δ-субградиент, составляет не менее $1 - \beta N$.

Приведенное следствие позволяет использовать в строчке 4 алгоритма 1 в качестве δ-субградиента величину $\overset{r}{\nabla}_x f\left(x, \{\xi^l\}_{l=1}^r\right)$. Будем называть такой вариант алгоритма методом эллипсоидов с мини-батчингом.

Теорема 2. *Пусть $Q \subseteq \mathbb{R}^n$ — выпуклое компактное множество с непустой внутренностью, $f(x)$ — непрерывная выпуклая функция. Метод эллипсоидов с мини-батчингом для задачи* (5) *возвращает (ε, β)-решение для $\varepsilon > 0$, $\beta \in (0, 1)$ после*

$$N = \left\lceil 2n^2 \ln\left(\frac{DB}{\rho\varepsilon}\right)\right\rceil$$

итераций при размере батча

$$r = \widetilde{O}\left(\frac{\sigma^2 D^2}{\varepsilon^2}\right),$$

где $D := \sup\limits_{x,\, y \in Q} \|x - y\|$, $B := \sup\limits_{x,\, y \in Q} |f(x) - f(y)|$, ρ — радиус некоторого евклидова шара, содержащегося в Q.

Доказательство. Согласно теореме 1 ε-решение задачи (5) может быть получено за

$$N = \left\lceil 2n^2 \ln\left(\frac{DB}{\rho\varepsilon}\right)\right\rceil$$

итераций метода эллипсоидов с $\frac{\varepsilon}{2}$-субградиентом (алгоритм 1). Воспользуемся следствием 1 для определения размера батча r, необходимого для того, чтобы усредненный по батчу стохастический субградиент являлся $\frac{\varepsilon}{2}$-субградиентом на каждой из N итераций с вероятностью не менее $1 - \beta$:

$$\frac{\varepsilon}{2} = \left[\sqrt{2} + \sqrt{6 \ln \frac{N}{\beta}}\right] \cdot \frac{\sigma D}{\sqrt{r}} \Longrightarrow r = \widetilde{O}\left(\frac{\sigma^2 D^2}{\varepsilon^2}\right).$$

Численный эксперимент

Рассмотрим модель логистической регрессии для задачи классификации:

$$\widehat{p}_x(w) = \frac{1}{1 + e^{-\langle w,\, x\rangle}},$$

где x — вектор признаков для объекта обучающей выборки (включая константный признак), w — веса модели. В качестве функции потерь выступает кросс-энтропия:

$$f_x(w) = y \ln \widehat{p}_x(w) + (1 - y) \ln(1 - \widehat{p}_x(w)),$$

где $y \in \{0, 1\}$ — класс объекта обучающей выборки. Задача оптимизации имеет вид

$$\min_{w \in Q} \{f(w) := \mathbb{E}_x f_x(w)\},$$

где в качестве Q можно взять евклидов шар с достаточно большим радиусом. Таким образом, целевая функция является выпуклой и непрерывной, и минимизация осуществляется на компактном множестве с непустой внутренностью, что соответствует предположениям, в которых выведена оценка сложности предлагаемого метода.

В ходе экспериментов было проведено сравнение метода эллипсоидов и стохастического градиентного спуска с использованием модели логистической регрессии и кросс-энтропии, выступающей в качестве функции потерь, при размере датасета $\sim 500\,000$ объектов и размерности пространства признаков 55. Согласно выведенным оценкам сложности метод эллипсоидов требует большой размер батча. В эксперименте он был выбран равным 2^{13}, поскольку при меньших значениях кривая сходимости была зашумленной. Для SGD использовался размер батча 16,

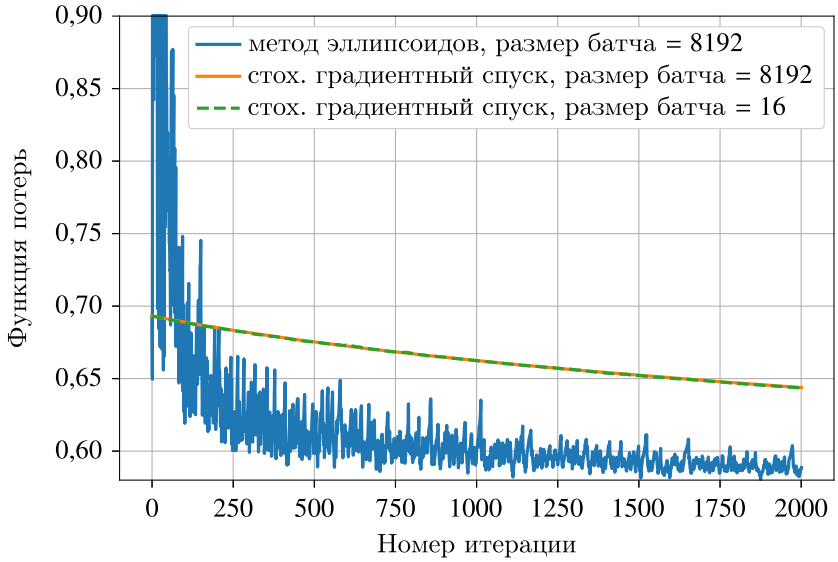

Рис. 1. Сходимость метода эллипсоидов и стохастического градиентного спуска при обучении логистической регрессии

а также 2^{13} для сравнения методов. После каждой итерации обучения модели на отложенной тестовой выборке рассчитывалось значение функции потерь. Данная зависимость функции потерь от количества итераций для каждого из методов отражена на рис. 1.

Как видно из графика, метод эллипсоидов сходится существенно быстрее стохастического градиентного спуска, хотя кривая функции потерь флуктуирует сильнее. Заметим, что метод эллипсоидов требует большого размера батча и будет эффективен только в случае, когда есть возможность осуществлять вычисления параллельно. Несмотря на быструю сходимость, общее количество вычислений стохастического субградиента для метода эллипсоидов может получиться больше, чем для SGD, который неплохо сходится и при маленьком размере батча.

Реализацию метода эллипсоидов на PyTorch и эксперименты можно найти в GitHub-репозитории: https://github.com/Karina1997/math-optimization-ellipsoids-method.

Заключение

В начале данной работы была выдвинута гипотеза о том, что методы секущей плоскости могут оказаться полезны в стохастической оптимизации. Такие методы эффективны для минимизации выпуклых функций, зависящих от относительно небольшого числа переменных (несколько десятков), причем гладкость задачи не требуется. Одним из примеров таких алгоритмов является метод эллипсоидов. Известно, что если на его итерациях использовать δ-субградиенты (неточные субградиенты) целевой функции, то ошибка не накапливается от шага к шагу. Это свойство позволяет использовать в методе эллипсоидов стохастический субградиент, усредненный по батчу, который с некоторой вероятностью является δ-субградиентом (см. леммы 1 и 2). Таким образом, если использовать батч достаточно большого размера r, то с высокой вероятностью стохастический субградиент на всех итерациях метода будет являться δ-субградиентом, где δ зависит от r как $\delta \sim \frac{1}{\sqrt{r}}$. В связи с этим предлагаемый алгоритм целесообразно использовать в случае, когда имеется возможность параллельно вычислять много стохастических субградиентов — например, если имеется система с большим количеством вычислительных узлов и быстрыми коммуникациями.

В работе приводится не только теоретическое обоснование применения метода эллипсоидов к задачам стохастической оптимизации, но и численный эксперимент, в котором указанный метод используется для обучения модели логистической регрессии. Для сравнения эта же задача решается стохастическим градиентным спуском с мини-батчингом. Обнаружено, что при большом размере батча метод эллипсоидов сходится быстрее. Этот результат иллюстрирует тот факт, что при достаточной параллелизации зависимость числа итераций предлагаемого алгоритма от требуемой точности ε асимптотически лучше, чем у SGD. Отметим, что число итераций метода эллипсоидов квадратично зависит от размерности переменных n. С этой точки зрения существуют более оптимальные методы секущей плоскости, поэтому одно из дальнейших направлений работы — исследование таких алгоритмов в применении к задачам выпуклой стохастической оптимизации.

Список литературы (References)

Гладин Е. Л., Курузов И. А., Стонякин Ф. С., Пасечнюк Д. А., Алкуса М. С., Гасников А. В. Решение сильно выпукло-вогнутых композитных седловых задач с небольшой размерностью одной из групп переменных // arXiv preprint arXiv:2010.02280. — 2020.
Gladin E. Reshenie sil'no vypuklo-vognutyh kompozitnyh sedlovyh zadach s nebol'shoj razmernost'yu odnoj iz grupp peremennyh [Solving strongly convex-concave composite saddle point problems with a small dimension of one of the variables] // arXiv preprint arXiv:2010.02280. — 2020 (in Russian).

Bertsekas D., Tsitsiklis J. Parallel and distributed computation: numerical methods. — 1989.

Bottou L. Large-Scale machine learning with stochastic gradient descent / Proceedings of COMPSTAT'2010. — 2010. — P. 177–186.

Bottou L. Stochastic gradient descent tricks // Neural Networks: Tricks of the Trade. — 2012. — P. 421–436.

Bottou L., Curtis F., Nocedal J. Optimization methods for large-scale machine learning // Siam Review. — 2018. — Vol. 60, № 2. — P. 223–311.

Bubeck S. Convex optimization: algorithms and complexity // Foundations and Trends in Machine Learning. — 2015. — Vol. 8, № 3–4. — P. 231–357.

Cotter A., Shamir O., Srebro N., Sridharan K. Better mini-batch algorithms via accelerated gradient methods / Advances in Neural Information Processing Systems. — 2011.

Deisenroth M. P., Faisal A. A., Ong C. S. Mathematics for machine learning. — Cambridge University Press, 2020.

Dozat T. Incorporating nesterov momentum into adam. — ICLR Workshop, 2016.

Duchi J., Hazan E., Singer Y. Adaptive subgradient methods for online learning and stochastic optimization // Journal of machine learning research. — 2011. — Vol. 12, no. 7.

Dvurechensky P. E., Gasnikov A. V., Lagunovskaya A. A. Parallel algorithms and probability of large deviation for stochastic convex optimization problems // Numerical Analysis and Applications. — 2018. — Vol. 11, no. 1. — P. 33–37.

Hazan E. Introduction to Online Convex Optimization // Foundations and Trends in Optimization. — 2016. — Vol. 2, no. 3–4. — P. 157–325.

Juditsky A., Nemirovski A. Large deviations of vector-valued martingales in 2-smooth normed spaces // arXiv preprint arXiv:0809.0813. — 2008.

Kingma D. P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

Murata N. A statistical study of on-line learning // Online Learning and Neural Networks. — 1998. — P. 63–92.

Reddi S. J., Kale S., Kumar S. On the convergence of Adam & Beyond // International Conference on Learning Representations. — 2018. — Vol. 24.

Shapiro A., Dentcheva D., Ruszczynski A. Lecture on stochastic programming. Modeling and theory / MPS-SIAM series on Optimization. — 2014.

Tieleman T., Hinton G. Lecture 6.5-RmsProp: Divide the gradient by a running average of its recent magnitude // COURSERA: Neural networks for machine learning. — 2012. — Vol. 4, no. 2. — P. 26–31.

Wilmott P. Machine learning: an applied mathematics introduction. — Panda Ohana Publishing, 2019.

Woodworth B. et al. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization / Advances in Neural Information Processing Systems. — 2018. — P. 8496–8506.

Woodworth B. E., Bullins B., Shamir O., Srebro N. The min-max complexity of distributed stochastic convex optimization with intermittent communication / Proceedings of Thirty Fourth Conference on Learning Theory. — 2021. — Vol. 134. — P. 4386–4437.

Zeiler M. D. Adadelta: an adaptive learning rate method. // arXiv preprint arXiv:1212.5701. — 2012.

Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent / Proceedings of the 20th International Conference on Machine Learning. — 2003. — P. 928–936.