КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 2 С. 309–319
DOI: 10.20537/2076-7633-2022-14-2-309-319

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.8

О связях задач стохастической выпуклой минимизации с задачами минимизации эмпирического риска на шарах в p-нормах

Д. М. Двинских[1,2,a], В. В. Пырэу[1,b], А. В. Гасников[1,2,3,c]

[1]Московский физико-технический институт (национальный исследовательский университет),
Россия, 141701, Московская обл., г. Долгопрудный, Институтский пер., 9
[2]Институт проблем передачи информации РАН им. А. А. Харкевича,
Россия, 127051, г. Москва, Большой Каретный пер., д. 19, стр. 1
[3]Кавказский математический центр Адыгейского государственного университета,
Россия, 385000, Республика Адыгея, г. Майкоп, ул. Первомайская, д. 208

E-mail: [a] dviny.d@yandex.ru, [b] pireyvitalik@phystech.edu, [c] gasnikov@yandex.ru

Получено 01.02.2022.
Принято к публикации 13.02.2022.

В данной работе рассматриваются задачи выпуклой стохастической оптимизации, возникающие в анализе данных (минимизация функции риска), а также в математической статистике (минимизация функции правдоподобия). Такие задачи могут быть решены как онлайн-, так и офлайн-методами (метод Монте-Карло). При офлайн-подходе исходная задача заменяется эмпирической задачей — задачей минимизации эмпирического риска. В современном машинном обучении ключевым является следующий вопрос: какой размер выборки (количество слагаемых в функционале эмпирического риска) нужно взять, чтобы достаточно точное решение эмпирической задачи было решением исходной задачи с заданной точностью. Базируясь на недавних существенных продвижениях в машинном обучении и оптимизации для решения выпуклых стохастических задач на евклидовых шарах (или всем пространстве), мы рассматриваем случай произвольных шаров в p-нормах и исследуем, как влияет выбор параметра p на оценки необходимого числа слагаемых в функции эмпирического риска.

В данной работе рассмотрены как выпуклые задачи оптимизации, так и седловые. Для сильно выпуклых задач были обобщены уже имеющиеся результаты об одинаковых размерах выборки в обоих подходах (онлайн и офлайн) на произвольные нормы. Более того, было показано, что условие сильной выпуклости может быть ослаблено: полученные результаты справедливы для функций, удовлетворяющих условию квадратичного роста. В случае когда данное условие не выполняется, предлагается использовать регуляризацию исходной задачи в произвольной норме. В отличие от выпуклых задач седловые задачи являются намного менее изученными. Для седловых задач размер выборки был получен при условии γ-роста седловой функции по разным группам переменных. Это условие при $\gamma = 1$ есть не что иное, как аналог условия острого минимума в выпуклых задач. В данной статье было показано, что размер выборки в случае острого минимума (седла) почти не зависит от желаемой точности решения исходной задачи.

Ключевые слова: выпуклая оптимизация, стохастическая оптимизация, регуляризация, острый минимум, условие квадратичного роста, метод Монте-Карло

Исследование выполнено за счет гранта Российского научного фонда (проект № 21-71-30005).

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 2 P. 309–319
DOI: 10.20537/2076-7633-2022-14-2-309-319

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.8

On the relations of stochastic convex optimization problems with empirical risk minimization problems on p-norm balls

D. M. Dvinskikh[1,2,a], V. V. Pirau[1,b], A. V. Gasnikov[1,2,3,c]

[1]Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow region, 141701, Russia
[2]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),
19/1 Bol'shoy Karetnyy per., Moscow, 212705, Russia
[3]Caucasus Mathematical Center, Adyghe State University,
208 Pervomaysk st., Maikop, Adyghe, 385000, Russia

E-mail: [a] dviny.d@yandex.ru, [b] pireyvitalik@phystech.edu, [c] gasnikov@yandex.ru

Received 01.02.2022.
Accepted for publication 13.02.2022.

In this paper, we consider convex stochastic optimization problems arising in machine learning applications (e. g., risk minimization) and mathematical statistics (e. g., maximum likelihood estimation). There are two main approaches to solve such kinds of problems, namely the Stochastic Approximation approach (online approach) and the Sample Average Approximation approach, also known as the Monte Carlo approach, (offline approach). In the offline approach, the problem is replaced by its empirical counterpart (the empirical risk minimization problem). The natural question is how to define the problem sample size, i. e., how many realizations should be sampled so that the quite accurate solution of the empirical problem be the solution of the original problem with the desired precision. This issue is one of the main issues in modern machine learning and optimization. In the last decade, a lot of significant advances were made in these areas to solve convex stochastic optimization problems on the Euclidean balls (or the whole space). In this work, we are based on these advances and study the case of arbitrary balls in the ℓ_p-norms. We also explore the question of how the parameter p affects the estimates of the required number of terms as a function of empirical risk.

In this paper, both convex and saddle point optimization problems are considered. For strongly convex problems, the existing results on the same sample sizes in both approaches (online and offline) were generalized to arbitrary norms. Moreover, it was shown that the strong convexity condition can be weakened: the obtained results are valid for functions satisfying the quadratic growth condition. In the case when this condition is not met, it is proposed to use the regularization of the original problem in an arbitrary norm. In contradistinction to convex problems, saddle point problems are much less studied. For saddle point problems, the sample size was obtained under the condition of γ-growth of the objective function. When $\gamma = 1$, this condition is the condition of sharp minimum in convex problems. In this article, it was shown that the sample size in the case of a sharp minimum is almost independent of the desired accuracy of the solution of the original problem.

Keywords: convex optimization, stochastic optimization, regularization, empirical risk minimization, stochastic approximation, sample average approximation, quadratic growth condition, sharp minimum

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 2, pp. 309–319 (Russian).
This research was funded by Russian Science Foundation (project 21-71-30005).

Введение

Подавляющее число задач математической статистики [Spokoiny, Dickhaus, 2015; Shapiro, Dentcheva, Ruszczynski, 2021] и машинного обучения [Shalev-Shwartz, Ben-David, 2014; Bach, 2021] в конечном итоге сводится к задачам стохастической оптимизации: минимизации функции риска, представляющей собой математического ожидание функции потерь. Данные задачи можно решать в онлайн-режиме [Nemirovski et al., 2009; Agarwal et al., 2012] (методами типа стохастического градиентного спуска), когда решение (например, оцениваемый параметр) корректируется по мере поступления новых данных (выборки) и в офлайн-режиме (методом Монте-Карло), когда исходная задача подменяется задачей минимизации функции эмпирического риска [Shapiro, Nemirovski, 2005; Shalev-Shwartz et al., 2009; Shalev-Shwartz, Ben-David, 2014; Bach, 2021] (выборочного среднего функции потерь). Офлайн-подход в последние годы стал достаточно популярным в связи с ростом размерностей задач и необходимостью использовать распределенные вычисления [Gorbunov et al., 2020]. Офлайн-подход прекрасно позволяет хранить разные части данных (выборки) на разных устройстах. Если онлайн-подход для задач выпуклой стохастической оптимизации достаточно хорошо проработан [Nemirovski et al., 2009; Agarwal et al., 2012; Woodworth, Srebro, 2021], то в офлайн-подходе теоретически обоснованных результатов поменьше [Li, Liu, 2021]. В частности, если задача стохастической оптимизации рассматривается на неевклидовом шаре, то офлайн-подход не позволяет учитывать такую специфику (за исключением работ [Dvinskikh, 2021b; Dvinskikh, 2021a], в которых рассматривался один частный случай — задача на шаре в 1-норме, без оценок вероятностей больших отклонений), в отличие от онлайн-подхода. В настоящей работе устраняется отмеченный недостаток офлайн-подхода.

Основные результаты

Рассмотрим задачу стохастической оптимизации

$$\min_{x \in X} F(x) := \mathbb{E}_{\xi} f(x, \xi). \tag{1}$$

Как правило, под множеством X будем понимать шар $B_p^d(R)$ радиусом R с центром в точке 0 в p-норме, $p \geqslant 1$, в пространстве $\mathbb{R}^d$.

Предположение 1 (липшицевость). *Для всех $x \in X$ и всех ξ выполняется*

$$|f(y, \xi) - f(x, \xi)| \leqslant M\|y - x\|_p.$$

Предположение 2 (гладкость). *Для всех $x \in X$ и всех ξ выполняется*

$$\left\|\nabla_x f(y, \xi) - \nabla_x f(x, \xi)\right\|_q \leqslant L\|y - x\|_p,$$

где $\frac{1}{p} + \frac{1}{q} = 1$.

Задача заключается в определении числа сэмплов (объема выборки) N, т. е. независимых одинаково распределенных реализаций случайной величины ξ, которое будет достаточно, чтобы некоторый алгоритм (подход) A позволял по $\{\xi^k\}_{k=1}^N$ определить такой $x\left(\{\xi^k\}_{k=1}^N\right)$, что

$$\mathbb{P}\left(F\left(x\left(\{\xi^k\}_{k=1}^N\right)\right) - \min_{x \in X} F(x) \leqslant \varepsilon\right) \geqslant 1 - \sigma. \tag{2}$$

Естественно ожидать, что N зависит от M, L, R, d, ε, σ. Как будет видно в дальнейшем, существенной зависимости от L в общем случае нет.

Важным местом в приведенном определении является наличие некоторого подхода (алгоритма), обозначенного через A, выдающего $x\left(\{\xi^k\}_{k=1}^N\right)$. В действительности оценка параметра N должна также зависеть и от A. Принципиально различаются два подхода к тому, как получать $x\left(\{\xi^k\}_{k=1}^N\right)$.

Первый подход — *онлайн* (в западной литературе часто используется название Stochastic Approximation). Базируется на процедурах типа (проекции π_X) *стохастического градиентного спуска*

$$x^{k+1} = \pi_X\left\{x^k - h\nabla_x f(x^k, \xi^k)\right\}, \quad k = 1, \ldots, N,$$

и вариациях этого метода [Поляк, 1990; Polyak, Juditsky, 1992; Nemirovski et al., 2009; Shapiro, Dentcheva, Ruszczynski, 2021]. Отметим, что, как правило, в таких процедурах выдается не последняя точка, а среднее по траектории [Поляк, 1990]. Большими преимуществами такого подхода являются простота получения искомой оценки, возможность адаптивной корректировки оцениваемого вектора параметров x по мере поступления новых данных (выборки). В действительности именно такие подходы приводят к наилучшим оценкам для параметра N в случае, когда F — выпуклая функция [Nemirovski et al., 2009; Agarwal et al., 2012; Shapiro, Dentcheva, Ruszczynski, 2021].

Второй подход — *офлайн*, который также можно называть подходом на основе метода Монте-Карло (в западной литературе часто используется название Sample Average Approximation) [Shapiro, Nemirovski, 2005; Shalev-Shwartz et al., 2009; Shapiro, Dentcheva, Ruszczynski, 2021]. В основе подхода — замена функционала задачи (1) на выборочное среднее:

$$\min_{x\in X} \widehat{F}(x) := \frac{1}{N}\sum_{k=1}^{N} f(x, \xi^k). \tag{3}$$

Решение (приближенное) задачи (3) понимается как $x\left(\{\xi^k\}_{k=1}^N\right)$ при офлайн-подходе. Очевидным недостатком подхода является необходимость достаточно точно решать задачу (3). Впрочем, в ряде случаев это может быть и достоинством, если, например, $f(x, \xi)$ обладает дорогим прямым оракулом, выдающим $\nabla_x f(x, \xi)$, но дешевым двойственным, выдающим градиент сопряженной по x функции [Dvinskikh, 2021a; Dvinskikh, 2021b]. Другим недостатком является более скромная теория, которая приводит в целом к худшим оценкам N в выпуклом случае [Shapiro, Nemirovski, 2005; Shalev-Shwartz et al., 2009; Feldman, Vondrak, 2019; Klochkov, Zhivotovskiy, 2021; Li, Liu, 2021]. Причем (сильную) выпуклость требуется понимать теперь как (сильную) выпуклость $f(x, \xi)$ по x, (сильной) выпуклости только F уже недостаточно для конечности N [Sekhari, Sridharan, Kale, 2021]. Впрочем, как будет видно в дальнейшем, это условие можно заметно ослабить — в большей степени, чем при онлайн-подходе. Отличительным достоинством офлайн-подхода является возможность организации распределенных вычислений [Gorbunov et al., 2020] при решении задачи (3), что представляется принципиально важным для многих современных приложений, приходящих, например, из обучения глубоких нейронных сетей [Huang et al., 2019].

Далее в статье постараемся сравнить подробнее оба подхода. Для этого потребуется обобщить некоторые результаты, связанные с офлайн-подходом.

Выпуклый случай

Для возможности сравнения двух подходов (онлайн и офлайн) предположим, что $f(x, \xi)$ удовлетворяет предположению 1, а $F(x) = \mathbb{E}_\xi f(x, \xi)$ — выпуклая функция при $x \in B_p^d(R)$.

Из результатов [Nemirovski et al., 2009] следует, что в онлайн-подходе

- при $N \leqslant d$

$$N = O\left(\kappa_p(d)\frac{M^2R^2}{\varepsilon^{\max\{2,\,p\}}}\ln\left(\frac{1}{\sigma}\right)\right), \tag{4}$$

где $\kappa_p(d) = O(1)$, при $p \geqslant 2$, при $p \in [1, 2]$ функция $\kappa_p(d)$ убывает от $O(\ln d)$ при $d = 1$ до $O(1)$ при $p = 2$;

- при $N \geqslant d$

$$N = O\left(d^{1-2/\max\{2,\,p\}}\frac{M^2R^2}{\varepsilon^2}\ln\left(\frac{1}{\sigma}\right)\right). \tag{5}$$

Причем данные оценки с точностью до логарифмических множителей не могут быть улучшены в общем случае, в том числе даже при дополнительном предположении 2 [Немировский, Юдин, 1979; Agarwal et al., 2012].

Из результатов [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021] следует, что в офлайн-подходе

$$N = O\left(\frac{M^2R^2}{(\varepsilon-\delta)^2}\left(d\ln\left(\frac{MR}{\varepsilon-\delta}\right)+\ln\left(\frac{1}{\sigma}\right)\right)\right), \tag{6}$$

где δ — точность решения задачи (3). Причем данная оценка с точностью до логарифмических множителей не может быть в общем случае улучшена, в том числе даже при дополнительном предположении 2 [Feldman, 2016].

Сопоставляя оценки, которые можно получить при онлайн-подходе (4), (5) с оценкой офлайн-подхода, получаем, что, за исключением случая $p = \infty$, онлайн-подход доминирует над офлайн-подходом. В частности, при $p = 2$ имеем $N_{\text{офлайн}} \simeq d \cdot N_{\text{онлайн}}$.

На самом деле приведенные выше результаты можно обобщить и на случай, когда $M = M(\xi)$ в предположении 1 не равномерно ограничена по ξ, а ограниченным является лишь второй момент $\mathbb{E}_\xi\left[M(\xi)^2\right]$ [Gorbunov et al., 2021; Shapiro, Dentcheva, Ruszczynski, 2021].

В заключение заметим, что оценка офлайн-подхода (6) может быть получена и без предположения выпуклости функции F [Shapiro, Dentcheva, Ruszczynski, 2021]. То есть выпуклость при офлайн-подходе в общем случае ничего не дает. Ситуация существенно меняется в сильно выпуклом случае.

Сильно выпуклый случай. Условие квадратичного роста

Отмеченный в предыдущем разделе зазор в оценках N в онлайн- и офлайн-подходе в выпуклом случае исчезает в сильно выпуклом случае [Shalev-Shwartz et al., 2009].

Для простоты сначала предположим, что $f(x, \xi)$ — μ-сильно выпуклая в p-норме функция по x при $x \in X$ (X — выпуклое множество) и при всех ξ, т. е. для всех $x, y \in X$

$$f(y, \xi) \geqslant f(x, \xi) + \langle \nabla_x f(x, \xi), y - x\rangle + \frac{\mu}{2}\|y - x\|_p^2. \tag{7}$$

Также будем предполагать, что $f(x, \xi)$ удовлетворяет предположению 1 и является неотрицательной функцией своих аргументов $f(x, \xi) \geqslant 0$. Отметим, что для онлайн-подхода μ-сильную выпуклость в p-норме $f(x, \xi)$ можно ослабить до μ-сильной выпуклости в p-норме $F(x) = \mathbb{E}_\xi f(x, \xi)$, а условие неотрицательности $f(x, \xi)$ можно опустить совсем.

Из результатов [Juditsky, Nemirovski, 2011; Juditsky, Nesterov, 2014; Harvey et al., 2019] следует, что в онлайн-подходе

$$N = O\left(\kappa_p(d)\frac{M^2}{\mu\varepsilon}\ln\left(\frac{\ln\left(\frac{M^2}{\mu\varepsilon}\right)}{\sigma}\right)\right), \tag{8}$$

где $\kappa_p(d)$ было определено в формуле (4). Данная оценка (8) с точностью до логарифмических множителей не может быть улучшена в общем случае, в том числе даже при дополнительном предположении 2 [Немировский, Юдин, 1979].

Из результатов работ [Shalev-Shwartz et al., 2009; Feldman, Vondrak, 2019; Klochkov, Zhivotovskiy, 2021; Li, Liu, 2021], в которых рассматривался случай $p = 2$, следует, что в офлайн-подходе

$$N = O\left(\frac{M^2}{\mu\varepsilon}\left(\ln\left(\frac{M^2}{\mu\varepsilon}\right) + \ln\ln\left(\frac{1}{\sigma}\right)\right)\ln\left(\frac{1}{\sigma}\right)\right). \tag{9}$$

При этом требуется решить задачу (3) с точностью $\delta = O(\mu\varepsilon^2)$. Оценки на N и δ с точностью до логарифмических множителей не могут быть в общем случае улучшены, в том числе даже при дополнительном предположении 2 [Немировский, Юдин, 1979; Shalev-Shwartz et al., 2009].

В данной работе устанавливается следующий результат.

Теорема 1. *Пусть $f(x, \xi) \geqslant 0$ удовлетворяет условию* (7) *на выпуклом множестве X и удовлетворяет предположению* 1*, где $p \in [1, \infty]$. Пусть задача* (3)*, с N, определяемым по формуле* (9)*, решена с точностью по функции $\delta = O(\mu\varepsilon^2)$ с вероятностью $1 - \frac{\sigma}{2}$, т. е. получен такой $x\left(\{\xi^k\}_{k=1}^N\right)$, что*

$$\mathbb{P}\left(\widehat{F}\left(x\left(\{\xi^k\}_{k=1}^N\right)\right) - \min_{x\in X}\widehat{F}(x) \leqslant \delta\right) \geqslant 1 - \frac{\sigma}{2}.$$

Тогда $x\left(\{\xi^k\}_{k=1}^N\right)$ будет ε-решением по функции задачи (1) *с вероятностью $1 - \sigma$ (см.* (2))*, т. е.*

$$\mathbb{P}\left(F\left(x\left(\{\xi^k\}_{k=1}^N\right)\right) - \min_{x\in X}F(x) \leqslant \varepsilon\right) \geqslant 1 - \sigma.$$

Следствие 1 (условие квадратичного роста). *В условиях теоремы* 1 *можно ослабить условие сильной выпуклости* (7) *до условия выпуклости $f(x, \xi)$ по x и условия квадратичного роста функций $\widehat{F}(x)$ из* (3) *и $F(x)$ из* (1):

для всех $x \in X$ (и всех $\{\xi^k\}_{k=1}^N$, см. (3))

$$\widehat{F}(x) - \widehat{F}(\widehat{x}_*) \geqslant \frac{\mu}{2}\|x - \widehat{x}_*\|_p^2, \tag{10}$$

где $\widehat{x}_*$ — проекция x на множество решений задачи (3);

$$F(x) - F(x_*) \geqslant \frac{\mu}{2}\|x - x_*\|_p^2, \tag{11}$$

где x_* — проекция x на множество решений задачи (1).

При $p = 2$ это следствие было установлено в работе [Li, Liu, 2021]. Также в данной работе приведены другие обобщения приведенной теоремы при $p = 2$, в частности на случай, когда можно совсем отказаться от условий выпуклости, заменив их намного более слабым условием Поляка – Лоясиевича, которому должна удовлетворять функция F, а не $\widehat{F}$:

для всех $x \in X$

$$F(x) - F(x_*) \leqslant \frac{1}{2\mu}\|\nabla F(x)\|_2^2, \tag{12}$$

где x_* — проекция (в 2-норме) x на множество решений задачи (1).

Предположение 3 (условие на шум). *Существует такая константа $B > 0$, что для всех $k = 2, \ldots, N$ выполняется*

$$\mathbb{E}_\xi\left[\|\nabla_x f(x_*, \xi)\|_2^k\right] \leqslant B^{k-2}k!\mathbb{E}_\xi\left[\|\nabla_x f(x_*, \xi)\|_2^2\right],$$

где x_* — решение задачи (1).

А именно, в [Li, Liu, 2021] показано, что если дополнительно (к условию Поляка – Лоясиевича для F) для $f(x, \xi) \geqslant 0$ выполняются предположения 1, 2, 3, то при достаточно большом N с вероятностью $1 - \sigma$ справедлива оценка

$$F\left(x\left(\{\xi^k\}_{k=1}^N\right)\right) - F(x_*) = O\left(\frac{(B^2 + \mu^2)\ln^2\left(\frac{1}{\sigma}\right)}{\mu N^2} + \frac{LF(x_*)\ln\left(\frac{1}{\sigma}\right)}{\mu N}\right).$$

В перепараметризованном случае $F(x_*) \simeq 0$ получаем, что $N \sim \frac{1}{\sqrt{\mu\varepsilon}}$, что сильно лучше оценки (9), но может быть хуже оценки, которую можно получить в перепараметризованном случае для онлайн-подхода $N \sim \frac{L}{\mu}\ln\left(\frac{\Delta F}{\varepsilon}\right)$ (см., например, [Woodworth, Srebro, 2021]).

Интересно было попробовать обобщить и эти результаты на случай $p \in [1, \infty]$. Насколько нам известно, это пока еще не сделано.

Регуляризация

Из предыдущих разделов следует, что в случае выпуклой задачи выгодно сделать ее сильно выпуклой с помощью *регуляризации* (см., например, [Shalev-Shwartz et al., 2009; Dvinskikh, 2021b; Dvinskikh, 2021a]). Причем «эффект» от такой регуляризации будет значительно выше, чем это имеет место в обычной оптимизации [Немировский, Юдин, 1979; Гасников, 2021].

Под регуляризацией понимается замена исходной задачи (1) на задачу с $f(x, \xi) := f(x, \xi) + \mu V(x, x^0)$, где $V(x, x^0)$ — 1-сильно выпуклая по x на X в p-норме ($p \in [1, 2]$) функция, такая, что (см. обозначения в разделе «Выпуклый случай») $V(x, x^0) \leqslant \kappa_p(d)\|x - x^0\|_p^2 = O\left(\|x - x^0\|_p^2 \ln d\right)$. Можно показать, что такие функции существуют [Ben-Tal, Nemirovski, 2022] и уже вполне успешно применялись в рассматриваемом здесь контексте [Dvinskikh, 2021b; Dvinskikh, 2021a] при $p = 1$. В данной работе рассматривается общий случай $p \in [1, 2]$.

Ключевое наблюдение (см., например, замечание 4.1 в [Гасников, 2021]) заключается в том, что если $x\left(\{\xi^k\}_{k=1}^N\right)$ — $\left(\frac{\varepsilon}{2}, \sigma\right)$-решение регуляризованной задачи в смысле (2) с $\mu \leqslant \frac{\varepsilon}{2V(x_*, x^0)}$, где x_* — такое решение задачи (1), которое наиболее близко к x^0 (в смысле минимальности $V(x_*, x^0)$), то $x\left(\{\xi^k\}_{k=1}^N\right)$ будет (ε, σ)-решением исходной задачи (3) в смысле (2).

Выбирая «на пределе» $\mu = \frac{\varepsilon}{2\kappa_p(d)R^2}$, получим (с точностью до логаримических множителей) из формул раздела «Сильно выпуклый случай. . . » формулы раздела «Выпуклый случай», только без лишнего d-множителя в офлайн-случае (6).

Таким образом, регуляризация решает отмеченную проблему нестыковки оценок онлайн- и офлайн-подходов в выпуклом случае ($f(x, \xi)$ — выпуклая функция от x). Впервые приблизительно такая конструкция была предложена в данном контексте при $p = 2$ в работе [Shalev-Shwartz et al., 2009] (см. также ее изложение, вошедшее в классический учебник по машинному обучению [Shalev-Shwartz, Ben-David, 2014]), а для $p = 1$ близкая конструкция была описана в работе [Dvinskikh, 2021b]. Описанный выше подход обобщает схему из [Dvinskikh, 2021b] на случай $p \in [1, 2]$.

Промежуточная выпуклость. Острый минимум

Условие квадратичного роста можно обобщить. Введем, следуя Шапиро – Немировскому (см., например, [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021]), *условие γ-роста* ($\gamma \geqslant 1$):

для всех $x \in X_{2\varepsilon} = \{x \in X\colon\ F(x) \leqslant F(x_*) + 2\varepsilon\}$

$$F(x) - F(x_*) \geqslant \mu_\gamma\|x - x_*\|_p^\gamma, \tag{13}$$

где x_* — проекция (в p-норме) x на множество решений задачи (1).

Ослабим также предположение 1. А именно, предположим, что для любых $x, y \in X$ субгауссовская дисперсия $f(y, \xi) - f(x, \xi) - (F(y) - F(x))$ ограничена сверху $\lambda^2\|y - x\|_p^2$, т. е.

$$\mathbb{E}_\xi \left[\exp\left(t \cdot (f(y, \xi) - f(x, \xi) - (F(y) - F(x)))\right)\right] \leqslant \exp\left(\frac{t^2\lambda^2\|y - x\|_p^2}{2}\right). \tag{14}$$

Заметим, что если выполняется предположение 1, то $\lambda^2 \leqslant 2M^2$.

Если $f(x, \xi)$ — выпуклая по x функция (на самом деле это условие можно ослабить [Shapiro, Dentcheva, Ruszczynski, 2021]), то при сделанных предположениях[1]

$$N = O\left(\frac{\lambda^2}{\mu_\gamma^{2/\gamma}\varepsilon^{2(\gamma-1)/\gamma}}\left(d\ln\left(\frac{MR_\varepsilon}{\varepsilon}\right) + \ln\left(\frac{1}{\sigma}\right)\right)\right), \tag{15}$$

где $\delta = \frac{\varepsilon}{2}$ — точность решения задачи (3). Причем данная оценка (15) с точностью до логарифмических множителей не может быть в общем случае улучшена [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021]. В цитированных работах оценка (15) была доказана, насколько удалось понять обозначения, для случая $p = 2$. Однако в [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021] общий случай $p \in [1, \infty)$ получается дословным повторением всех рассуждений, что также было нам подтверждено в ходе личной беседы одним из авторов [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021], Александром Шапиро.

Формула (15) особенно интересна в случае «острого минимума» $\gamma = 1$. Она не зависит от ε.

Тот факт, что формула (15) не может быть улучшена, хорошо поясняет пример из книги [Shapiro, Dentcheva, Ruszczynski, 2021], в котором $p = 2$: $f(x, \xi) = \|x\|_2^\gamma - \gamma\sigma\langle\xi, x\rangle$, $\xi \in \mathcal{N}(0, I_d)$ — стандартное нормальное распределение (с нулевым математическим ожиданием и единичной корреляционной матрицей), $X = B_2^d(1)$. В этом случае N не может быть меньше, чем $\frac{d\sigma^2}{\varepsilon^{2(\gamma-1)/\gamma}}$. Однако мы привели здесь этот пример, чтобы показать, что предположение 1 и условие (14) могут довольно сильно отличаться. А именно, для этого примера предположение 1 выполняется лишь в «среднем» с $\mathbb{E}_\xi\left[M(\xi)^2\right] = \gamma^2\sigma^2 d$, при том что условие (14) выполняется с $\lambda = \gamma^2\sigma^2$.

В связи со всем написанным выше в этом разделе и написанным ранее в разделе «Выпуклый случай» может показаться, что оценка (15) при $\gamma \to \infty$ (вырожденный случай) противоречит нижней оценке (6) [Feldman, 2016]. Ведь оценка N сверху (15) получается лучше в плане возможности использования параметра λ вместо M. Но при этом предположение 1 является более узким, чем условие (14). На самом деле никакого противоречия нет. Обе оценки точные в своих классах функций $f(x, \xi)$. Возникающий здесь парадокс с описанным примером связан с тем, что $\gamma \to \infty$ влечет за собой то, что $\lambda \to \infty$ и $M \to \infty$. Поэтому данный пример в пределе $\gamma \to \infty$ не отражает точное поведение оценки (15).

Результат, аналогичный (15) (с заменой $d\lambda^2$ на M^2 и R_ϵ на R при $\gamma = 1$) при условии (13) может быть получен и для онлайн-методов типа рестартованного стохастического градиентного спуска при $\gamma \geqslant 2$ [Juditsky, Nesterov, 2014]. Результаты работы [Juditsky, Nesterov, 2014] переносится и на случай $\gamma \in [1, 2]$. Случай $\gamma = 1$ был исследован в работе [Juditsky, 1993].

[1] Если M в предположении 1 зависит от ξ, то под M в формуле (15) следует понимать $\mathbb{E}_\xi M(\xi)$ [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021]. Параметр R_ε в этой формуле отвечает диаметру множества $X_{2\varepsilon}$ в p-норме. В частности, при $\gamma = 1$ параметр $R_\varepsilon \leqslant \frac{4\varepsilon}{\mu_1}$. Таким образом, в случае «острого минимума» ($\gamma = 1$) N не зависит от ε [Shapiro, Dentcheva, Ruszczynski, 2021].

Седловые задачи. Промежуточная выпукло-вогнутость. Острый минимум

К сожалению, такой богатой теории, которая уже создана для задач (выпуклой) оптимизации, для седловых задач нам не известно. Из всех приведенных выше результатов на данный момент удалось перенести только результат (15). Рассуждения практически дословно повторяют выкладки из работ [Shapiro, Nemirovski, 2005; Shapiro, Dentcheva, Ruszczynski, 2021]. Далее излагается соответствующая теория.

Рассматривается стохастическая седловая задача

$$\min_{x\in X}\max_{y\in Y} F(x,\,y) := \mathbb{E}_{\xi} f(x,\,y,\,\xi). \tag{16}$$

Множества $X \subset \mathbb{R}^{d_x}$ и $Y \subset \mathbb{R}^{d_y}$ предполагаются выпуклыми компактами, функция $f(x,\,y,\,\xi)$ — выпуклая по x при $x \in X$ и вогнутая по y при $y \in Y$ для всех ξ. Также будем считать, что по каждой группе переменных x и y на X и на Y функция $f(x,\,y,\,\xi)$ удовлетворяет

- предположению 1 с параметрами соответственно $M_x(\xi)$ в p_x-норме ($p_x \in [1,\,2]$) и $M_y(\xi)$ в p_y-норме ($p_y \in [1,\,2]$), причем $\mathbb{E}M_x(\xi) = M_x < \infty$ и $\mathbb{E}M_y(\xi) = M_y < \infty$;
- условию (14) с параметрами соответственно λ_x, p_x и λ_y, p_y;
- условию γ_x-роста в p_x-норме с константой $\mu_{\gamma,x}$ и γ_y-роста в p_y-норме с константой $\mu_{\gamma,y}$:

$$F(x,\,y) - F(x_*(y),\,y) \geqslant \mu_{\gamma,x}\|x - x_*(y)\|_{p_x}^{\gamma_x}, \tag{17}$$

где $x_*(y)$ — проекция (в p_x-норме) x на множество решений задачи $\min\limits_{x\in X} F(x,\,y)$, и

$$F(x,\,y_*(x)) - F(x,\,y) \geqslant \mu_{\gamma,y}\|y - y_*(x)\|_{p_y}^{\gamma_y}, \tag{18}$$

где $y_*(x)$ — проекция (в p_y-норме) y на множество решений задачи $\max\limits_{y\in Y} F(x,\,y)$.

Введем эмпирическую функцию

$$\widehat{F}(x,\,y) = \frac{1}{N}\sum_{k=1}^{N} f(x,\,y,\,\xi^k).$$

Пусть удалось найти такие $(\widehat{x},\,\widehat{y})$, что

$$\widehat{F}(\widehat{x},\,\widehat{y}) - \widehat{F}(\widehat{x}_*(\widehat{y}),\,\widehat{y}) \leqslant \frac{\varepsilon}{4},$$
$$\widehat{F}(\widehat{x},\,\widehat{y}_*(\widehat{x})) - \widehat{F}(\widehat{x},\,\widehat{y}) \leqslant \frac{\varepsilon}{4},$$

где $\widehat{x}_*(y)$, $\widehat{y}_*(x)$ определяются по $\widehat{F}$ аналогично тому, как $x_*(y)$, $y_*(x)$ определялись по F.

Тогда если

$$N = O\left(N_x + N_y\right),$$
$$N_x = O\left(\frac{\lambda_x^2}{\mu_{\gamma,x}^{2/\gamma_x}\varepsilon^{2(\gamma_x-1)/\gamma_x}}\left(d_x\ln\left(\frac{M_xR_x}{\varepsilon}\right) + \ln\left(\frac{1}{\sigma}\right)\right)\right),$$
$$N_y = O\left(\frac{\lambda_y^2}{\mu_{\gamma,y}^{2/\gamma_y}\varepsilon^{2(\gamma_y-1)/\gamma_y}}\left(d_y\ln\left(\frac{M_yR_y}{\varepsilon}\right) + \ln\left(\frac{1}{\sigma}\right)\right)\right),$$

то с вероятностью $1 - \sigma$

$$F(\widehat{x}, \widehat{y}) - F(x_*(\widehat{y}), \widehat{y}) \leqslant \frac{\varepsilon}{2},$$
$$F(\widehat{x}, y_*(\widehat{x})) - F(\widehat{x}, \widehat{y}) \leqslant \frac{\varepsilon}{2},$$

где R_x — диаметр X в p_x-норме, R_y — диаметр Y в p_y-норме (можно уточнить эти оценки и использовать диаметры соответствующих множеств Лебега, подобно тому, как это делалось выше, см. сноску 1). Следовательно,

$$0 \leqslant \max_{y \in Y} F(\widehat{x}, y) - \min_{x \in X} F(x, \widehat{y}) = F(\widehat{x}, y_*(\widehat{x})) - F(x_*(\widehat{y}), \widehat{y}) \leqslant \varepsilon.$$

Приведенные выше оценки могут быть получены и в онлайн-подходе (с заменой $d\lambda^2$ на M^2). Немного в более общем контексте это недавно было показано в работе [Dvinskikh, 2022].

Авторы выражают благодарность Александру Шапиро и Анатолию Юдицкому за ценные советы.

Статья приурочена к 60-летию Анатолия Борисовича Юдицкого, внесшего значительный вклад в развитие методов стохастического градиентного спуска.

Список литературы (References)

Гасников А. В. Современные численные методы оптимизации. Метод универсального градиентного спуска. — М.: МЦНМО, 2021.
Gasnikov A. V. Sovremennye chislennye metody optimizatsii. Metod universal'nogo gradientnogo spuska [Universal gradient method]. — MCCME, 2021 (in Russian).

Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. — М.: Наука, 1979.
Nemirovsky A. S., Yudin D. B. Slozhnost' zadach i effektivnost' metodov optimizatsii [Problem Complexity and Optimization Method Efficiency]. — M.: Nauka, 1979 (in Russian).

Поляк Б. Т. Новый метод типа стохастической аппроксимации // Автоматика и телемеханика. — 1990. — С. 98–107.
Polyak B. T. Novyi metod tipa stokhasticheskoi approksimatsii [A new method of stochastic approximation type] // Autom. Remote Control. — 1990. — Vol. 51, No. 7. — P. 937–946.

Agarwal A., Bartlett P., Ravikumar P., Wainwright M. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization // Advances in Neural Information Processing Systems. — 2009. — Vol. 22.

Bach F. Learning Theory from First Principles Draft. — 2021.

Ben-Tal A., Nemirovski A. Lectures on modern convex optimization: analysis, algorithms, and engineering applications. — SIAM, 2022. — https://www2.isye.gatech.edu/~nemirovs/LMCOLN2022WithSol.pdf

Boucheron S., Lugosi G., Massart P. Concentration inequalities: A nonasymptotic theory of independence. — Oxford university press, 2013.

Dvinskikh D. Decentralized algorithms for Wasserstein barycenters. — arXiv preprint. — 2021. — https://arxiv.org/pdf/2105.01587 — Дис. — Humboldt Universitaet zu Berlin (Germany), 2021.

Dvinskikh D. et al. Gradient-free optimization for non-smooth minimax problems with maximum value of adversarial noise. — arXiv preprint. — 2022. — https://arxiv.org/pdf/2202.06114

Dvinskikh D. Stochastic approximation versus sample average approximation for Wasserstein barycenters // Optimization Methods And Software. — 2021. — P. 1–33.

Feldman V. Generalization of erm in stochastic convex optimization: The dimension strikes back // Advances In Neural Information Processing Systems. — 2016. — Vol. 29. — P. 3576–3584.

Feldman V., Vondrak J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate // Conference On Learning Theory, PMLR, 2019. — P. 1270–1279.

Gorbunov E., Danilova M., Shibaev I., Dvurechensky P., Gasnikov A. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. — arXiv preprint. — 2021. — https://arxiv.org/pdf/2106.05958

Gorbunov E., Rogozin A., Beznosikov A., Dvinskikh D., Gasnikov A. Recent theoretical advances in decentralized distributed convex optimization. — arXiv preprint. — 2020. — https://arxiv.org/pdf/2011.13259

Harvey N., Liaw C., Plan Y., Randhawa S. Tight analyses for non-smooth stochastic gradient descent // Conference On Learning Theory, PMLR, 2019. — P. 1579–1613.

Huang Y., Cheng Y., Bapna A., Firat O., Chen D., Chen M., Lee H., Ngiam J., Le Q., Wu Y., and others. Gpipe: Efficient training of giant neural networks using pipeline parallelism // Advances In Neural Information Processing Systems. — 2019. — Vol. 32. — P. 103–112.

Juditsky A. A stochastic estimation algorithm with observation averaging // IEEE transactions on automatic control. — 1993. — Vol. 38, No. 5. — P. 794–798.

Juditsky A., Nemirovski A. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods // Optimization For Machine Learning. — 2011. — Vol. 30, No. 9. — P. 121–148.

Juditsky A., Nesterov Yu. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stochastic Systems. — 2014. — Vol. 4, No. 1. — P. 44–80.

Klochkov Y., Zhivotovskiy N. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$ // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — https://arxiv.org/pdf/2103.12024

Li S., Liu Y. Improved learning rates for stochastic optimization: two theoretical viewpoints. — arXiv preprint. — 2021. — https://arxiv.org/pdf/2107.08686

Nemirovski A., Juditsky A., Lan G., Shapiro A. Robust stochastic approximation approach to stochastic programming // SIAM Journal On Optimization. — 2009. — Vol. 19, No. 4. — P. 1574–1609.

Polyak B., Juditsky A. Acceleration of stochastic approximation by averaging // SIAM Journal On Control And Optimization. — 1992. — Vol. 30, No. 4. — P. 838–855.

Robbins H., Monro S. A stochastic approximation method // The annals of mathematical statistics. — 1951. — Vol. 2. — P. 400–407.

Sekhari A., Sridharan K., Kale S. SGD: the role of implicit regularization, batch-size and multiple-epochs // Advances In Neural Information Processing Systems. — 2021. — Vol. 34.

Shalev-Shwartz S., Ben-David S. Understanding machine learning: From theory to algorithms. — Cambridge university press, 2014.

Shalev-Shwartz S., Shamir O., Srebro N., Sridharan K. Stochastic Convex Optimization // COLT. — 2009. — Vol. 2, No. 4.

Shapiro A., Dentcheva D., Ruszczynski A. Lectures on stochastic programming: modeling and theory. — SIAM, 2021.

Shapiro A., Nemirovski A. On complexity of stochastic programming problems // Continuous Optimization. — Boston: Springer, 2005. — P. 111–146.

Spokoiny V., Dickhaus T. Basics of modern mathematical statistics. — Heidelberg: Springer, 2015.

Woodworth B., Srebro N. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — https://arxiv.org/pdf/2106.02720