КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 2 С. 381–391
DOI: 10.20537/2076-7633-2023-15-2-381-391

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.8

Сравнение оценок онлайн- и офлайн-подходов для седловой задачи в билинейной форме

С. Н. Скорик[1,2,a], В. В. Пырэу[1,b], С. А. Седов[3,c], Д. М. Двинских[1,4,d]

[1]Московский физико-технический институт (национальный исследовательский университет),
Россия, 141701, Московская обл., г. Долгопрудный, Институтский пер., 9
[2]Институт системного программирования им. В. П. Иванникова РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25
[3]Высшая школа экономики (национальный исследовательский университет),
Россия, 109028, г. Москва, Покровский б-р, д. 11, стр. 1
[4]Институт проблем передачи информации РАН им. А. А. Харкевича,
Россия, 212705, г. Москва, Большой Каретный переулок, д. 19, стр. 1

E-mail: [a] skorik.sn@phystech.edu, [b] pireyvitalik@phystech.edu, [c] sasedov@edu.hse.ru, [d] dviny.d@yandex.ru

Получено 19.02.2023.
Принято к публикации 23.02.2023.

Стохастическая оптимизация является актуальным направлением исследования в связи со значительными успехами в области машинного обучения и их применениями для решения повседневных задач. В данной работе рассматриваются два принципиально различных метода решения задачи стохастической оптимизации — онлайн- и офлайн-алгоритмы. Соответствующие алгоритмы имеют свои качественные преимущества перед друг другом. Так, для офлайн-алгоритмов требуется решать вспомогательную задачу с высокой точностью. Однако это можно делать распределенно, и это открывает принципиальные возможности, как, например, построение двойственной задачи. Несмотря на это, и онлайн-, и офлайн-алгоритмы преследуют общую цель — решение задачи стохастической оптимизации с заданной точностью. Это находит отражение в сравнении вычислительной сложности описанных алгоритмов, что демонстрируется в данной работе.

Сравнение описанных методов проводится для двух типов стохастических задач — выпуклой оптимизации и седел. Для задач стохастической выпуклой оптимизации существующие решения позволяют довольно подробно сравнить онлайн- и офлайн-алгоритмы. В частности, для сильно выпуклых задач вычислительная сложность алгоритмов одинаковая, причем условие сильной выпуклости может быть ослаблено до условия γ-роста целевой функции. С этой точки зрения седловые задачи являются гораздо менее изученными. Тем не менее существующие решения позволяют наметить основные направления исследования. Так, значительные продвижения сделаны для билинейных седловых задач с помощью онлайн-алгоритмов. Оффлайн-алгоритмы представлены всего одним исследованием. В данной работе на этом примере демонстрируется аналогичная с выпуклой оптимизацией схожесть обоих алгоритмов. Также был проработан вопрос точности решения вспомогательной задачи для седел. С другой стороны, седловая задача стохастической оптимизации обобщает выпуклую, то есть является ее логичным продолжением. Это проявляется в том, что существующие результаты из выпуклой оптимизации можно перенести на седла. В данной работе такой перенос осуществляется для результатов онлайн-алгоритма в выпуклом случае, когда целевая функция удовлетворяет условию γ-роста.

Ключевые слова: стохастическая оптимизация, выпуклая оптимизация, выпукло-вогнутая оптимизация, острый минимум, условие квадратичного роста

Исследование выполнено за счет гранта Российского научного фонда (проект № 21-71-30005), https://rscf.ru/project/21-71-30005/.

COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 2 P. 381–391
DOI: 10.20537/2076-7633-2023-15-2-381-391

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.8

Comparsion of stochastic approximation and sample average approximation for saddle point problem with bilinear coupling term

S. N. Skorik[1,2,a], V. V. Pirau[1,b], S. A. Sedov[3,c], D. M. Dvinskikh[1,4,d]

[1]Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow region, 141701, Russia
[2]Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25 A. Solzhenitsyn st., Moscow, 109004, Russia
[3]Higher School of Economics,
11/1 Pokrovsky boul., Moscow, 109028, Russia
[4]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),
19/1 Bol'shoy Karetnyy per., Moscow, 212705, Russia

E-mail: [a] skorik.sn@phystech.edu, [b] pireyvitalik@phystech.edu, [c] sasedov@edu.hse.ru, [d] dviny.d@yandex.ru

Received 19.02.2023.
Accepted for publication 23.02.2023.

Stochastic optimization is a current area of research due to significant advances in machine learning and their applications to everyday problems. In this paper, we consider two fundamentally different methods for solving the problem of stochastic optimization — online and offline algorithms. The corresponding algorithms have their qualitative advantages over each other. So, for offline algorithms, it is required to solve an auxiliary problem with high accuracy. However, this can be done in a distributed manner, and this opens up fundamental possibilities such as, for example, the construction of a dual problem. Despite this, both online and offline algorithms pursue a common goal — solving the stochastic optimization problem with a given accuracy. This is reflected in the comparison of the computational complexity of the described algorithms, which is demonstrated in this paper.

The comparison of the described methods is carried out for two types of stochastic problems — convex optimization and saddles. For problems of stochastic convex optimization, the existing solutions make it possible to compare online and offline algorithms in some detail. In particular, for strongly convex problems, the computational complexity of the algorithms is the same, and the condition of strong convexity can be weakened to the condition of γ-growth of the objective function. From this point of view, saddle point problems are much less studied. Nevertheless, existing solutions allow us to outline the main directions of research. Thus, significant progress has been made for bilinear saddle point problems using online algorithms. Offline algorithms are represented by just one study. In this paper, this example demonstrates the similarity of both algorithms with convex optimization. The issue of the accuracy of solving the auxiliary problem for saddles was also worked out. On the other hand, the saddle point problem of stochastic optimization generalizes the convex one, that is, it is its logical continuation. This is manifested in the fact that existing results from convex optimization can be transferred to saddles. In this paper, such a transfer is carried out for the results of the online algorithm in the convex case, when the objective function satisfies the γ-growth condition.

Keywords: stochastic optimization, stochastic approximation, sample average approximation, decentralization

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 2, pp. 381–391 (Russian).

The research was supported by Russian Science Foundation (project No. 21-71-30005), https://rscf.ru/en/project/21-71-30005/.

Введение

Задача стохастической оптимизации возникает в машинном обучении в качестве минимизации функции риска. Алгоритмы, решающие эту задачу, можно условно разделить на два типа. Первый — онлайн-алгоритм (в английской литературе Stochastic Approximation), работа которого основана на стохастическом градиентном спуске, а также его вариациях [Поляк, 1990; Nemirovski et al., 2009]. Алгоритмы такого типа являются наиболее распространенными. Второй — офлайн-алгоритм (в английской литературе Sample Average Approximation). Особенностями этого типа алгоритмов являются переход и решение вспомогательной задачи. Тем не менее в определенных постановках такой переход может быть эффективнее вычисления прямого оракула. Это достигается за счет принципиальной возможности офлайн-алгоритма к построению двойственной задачи, в которой вычисление сопряженных градиентов может быть гораздо выгоднее вычисления прямых [Dvinskikh, 2021]. Также вспомогательную задачу можно решать распределенно, что особенно актуально в случае обучения глубоких нейронных сетей [Huang et al., 2019]. В данной работе описанные алгоритмы сравниваются с точки зрения их вычислительной сложности для задач выпуклой и седловой стохастической оптимизации. Также седловая задача рассматривается как продолжение выпуклой: результаты выпуклой оптимизации для офлайн-алгоритмов переносятся на случай седла [Dvinskikh et al., 2021]. В данной работе такой перенос был проведен для онлайн-алгоритмов.

Обзор существующих решений

Выпуклая оптимизация

Рассматривается задача стохастической оптимизации

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

и ее эмпирическая вариация

$$\min_{x\in\mathcal{X}} \widehat{F}(x) := \frac{1}{N}\sum_{k=1}^{N} f(x, \xi^k) \tag{2}$$

на некотором множестве $\mathcal{X}$. В течение всей работы мы будем пользоваться следующими определениями.

Определение 1 (липшицевость). Функция f называется *M-липшицевой* на $\mathcal{X}$ по норме p, если для любых $x_1, x_2 \in \mathcal{X}$ и всех ξ выполнено

$$|f(x_1, \xi) - f(x_2, \xi)| \leqslant M\|x_1 - x_2\|_p. \tag{3}$$

Определение 2 (гладкость). Функция f называется *L-гладкой* на $\mathcal{X}$ по норме p, если для любых $x_1, x_2 \in \mathcal{X}$ и всех ξ выполнено

$$\|\nabla_x f(x_1, \xi) - \nabla_x f(x_2, \xi)\|_q \leqslant M\|x_1 - x_2\|_p, \tag{4}$$

где q такое, что $\frac{1}{q} + \frac{1}{p} = 1$.

Определение 3 (сильная выпуклость). Функция f называется *μ-сильно выпуклой* на $\mathcal{X}$ по норме p, если для любых $x_1, x_2 \in \mathcal{X}$ и всех ξ выполнено

$$f(x_1, \xi) \geqslant f(x_2, \xi) + \langle \nabla_x f(x_2, \xi),\, x_1 - x_2\rangle + \frac{\mu}{2}\|x_1 - x_2\|_p^2. \tag{5}$$

В частности, при $\mu = 0$ функция f называется *выпуклой*.

Определение 4 (условие квадратичного роста). Функция $F(x)$ удовлетворяет условию квадратичного роста по норме p, если для всех $x \in \mathcal{X}$ справедливо

$$F(x) - F(x_*) \geqslant \frac{\mu}{2}\|x - x_*\|_p^2, \tag{6}$$

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

Под решением задачи будем понимать некоторый алгоритм $\mathcal{A}$, который для заданных значений ε, δ, получив на вход выборку $\{\xi_k\}_{k=1}^N$, где $N = N(\varepsilon, \delta)$, предоставит такой $\widehat{x} = \mathcal{A}(\{\xi_k\}_{k=1}^N)$, что

$$\mathbb{P}\left\{F(\widehat{x}) - \min_{x \in \mathcal{X}} F(x) \leqslant \varepsilon\right\} \geqslant 1 - \delta.$$

Существует два принципиально различных вида алгоритма $\mathcal{A}$, причем оба типа алгоритмов имеют свои качественные преимущества друг перед другом, поэтому выбор конкретного из них определяется особенностями решаемой задачи (1). Сравним подробнее приведенные виды алгоритмов с точки зрения их вычислительной сложности. Для этого рассмотрим задачу (1) на ограниченном константой B множестве $\mathcal{X} \subset \mathbb{R}^d$ в предположениях M-липшицевости по норме p функции $f(x, \xi)$ и выпуклости целевой функции $F(x)$. Из работ [Nemirovski et al., 2009; Shapiro, Nemirovski, 2005] следует соотношение на размер N между онлайн- и офлайн-алгоритмами.

$$N_{\text{офлайн}} = d^{2/\max\{2,p\}} N_{\text{онлайн}}. \tag{7}$$

Из (7) следует, что при сделанных предположениях офлайн-алгоритм требует большего размера N. Частично это объясняется тем, что приведенная оценка для офлайн-алгоритма может быть получена и при более слабых предположениях, а именно без предположения выпуклости целевой функции F [Shapiro, Dentcheva, Ruszczynski, 2021].

Ситуация изменится в сильно выпуклом случае [Shalev-Shwartz et al., 2009]. Для этого дополнительно предположим, что функция $f(x, \xi)$ является μ-сильно выпуклой в норме $p = 2$, а также $f(x, \xi) \geqslant 0$. Из работ [Shalev-Shwartz et al., 2009; Zhivotovskiy, Klochkov, 2021; Li, Liu, 2021] следует, что для офлайн-алгоритма

$$N = \widetilde{O}\left(\frac{M^2}{\mu\varepsilon} \ln\left(\frac{1}{\delta}\right)\right). \tag{8}$$

При этом вспомогательную задачу (2) необходимо решить с точностью $\sigma = O\left(\mu\varepsilon^2\right)$.

В работе [Li, Liu, 2021] для $p = 2$ было установлено, что оценку (8) можно получить и при более слабых предположениях, чем сильная выпуклость функции $f(x, \xi)$. А именно, достаточно потребовать выпуклость функции $f(x, \xi)$ и условия квадратичного роста (определение 4) функций $\widehat{F}(x)$ и $F(x)$ из (2), (1) соответственно. В работах [Dvinskikh et al., 2021; Shapiro, Dentcheva, Ruszczynski, 2021] этот результат был продолжен для шаров $\mathcal{X} = \mathcal{B}_p^d(R)$ на общий случай $p \in [1, \infty)$ соответствующим обобщением условия квадратичного роста (определение 5). Дополнительно в [Dvinskikh et al., 2021] аналогичные оценки были перенесены для выпукло-вогнутой оптимизации. В данной работе мы производим соответствующие выкладки для онлайн-алгоритмов.

Определение 5 (условие γ-роста). Функция $F(x)$ удовлетворяет условию γ-роста ($\gamma \geqslant 1$) по норме p на $\mathcal{X}$, если для всех $x \in \mathcal{X}$ справедливо

$$F(x) - F(x_*) \geqslant \mu_\gamma \|x - x_*\|_p^\gamma, \tag{9}$$

где x_* — проекция x на множество решений задачи (1). В случае $\gamma = 1$ говорят, что функция $F(x)$ удовлетворяет условию острого минимума.

Ослабим соответствующее предположение μ-сильной выпуклости функции $f(x, \xi)$ в случае онлайн-алгоритма. Для этого введем соответствующие предположения.

Предположение 1 (ограниченный шум). *Обозначим стохастический оракул* $g_k = \nabla f(x_k, \xi_k)$ *и шум* $\varepsilon_k = g_k - \nabla f(x_k)$. *Стохастический оракул* $g(x_k)$ *предполагается несмещенным с ограниченным шумом* σ^2 *для любого* k *в двойственной норме* $\|\cdot\|_*$, *то есть справедливо*

$$\begin{cases} \mathbb{E}[g(x_k) \mid x_k, g_k] = \nabla f(x_k), \\ \mathbb{E}\|\varepsilon_k\|_*^2 \leqslant \sigma^2 < \infty \end{cases} \quad \forall k = \overline{1, N}.$$

Предположение 2 (легкие хвосты). *В обозначениях предположения* 1 *существует такая константа* $\sigma < \infty$, *что*

$$\mathbb{E}\left[\exp\left\{\|\xi_k\|_*^2\sigma^{-2}\right\} \mid x_k, g_k\right] \leqslant \exp(1), \quad k = \overline{1, N}.$$

Тогда в предположении M-липшицевости и выпуклости функции $f(x, \xi)$, условия γ-роста функции $F(x)$ ($\gamma \geqslant 2$) по норме p, а также выполнении предположений 1, 2 из работы [Judistky, Nesterov, 2014] для онлайн-алгоритма

$$N = O\left(\frac{(M^2 + \sigma^2)C_d}{\mu_\gamma^{2/\gamma}\varepsilon^{2(\gamma-1)/\gamma}\mu_d}\ln\left(\frac{1}{\delta}\right)\right), \tag{10}$$

где C_d и μ_d соответствуют константам квадратичного роста и сильной выпуклости прокс-функции $d(x)$ на единичном шаре в p-норме. Стоит отметить, что исходно в работе [Judistky, Nesterov, 2014] вместо γ-роста функции $F(x)$ использовалась ее равномерная выпуклость, что является более сильным условием. Тем не менее его можно ослабить соответствующим образом без изменения структуры и логики получения результатов. Для $\gamma = 1$ аналогичный результат был получен в [Juditsky, 1993]. Предложенные рассуждения в [Judistky, Nesterov, 2014] обобщаются и на случай $\gamma \in [1, 2]$. Рассмотрим функцию $F(x)$, удовлетворяющую условию квадратичного роста ($\gamma = 2$). В норме $p = 2$ соответствующая прокс-функция $d(x) = \frac{1}{2}\|x\|_2^2$, то есть $\mu_d = 1$, $C_d = \frac{1}{2}$. Подставляя соответствующие μ_d, C_d, γ, получим

$$N = O\left(\frac{M^2 + \sigma^2}{\mu\varepsilon}\ln\left(\frac{1}{\delta}\right)\right). \tag{11}$$

В сделанных предположениях оценки (11), (8) не могут быть улучшены с точностью до логарифмических множителей.

Выпукло-вогнутая оптимизация

Рассмотрим стохастическую седловую задачу

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\Phi(x, y) := \mathbb{E}_\xi[\Phi(x, y, \xi)]. \tag{12}$$

Здесь $\mathcal{X}$, $\mathcal{Y}$ — выпуклые компакты, ξ является случайной величиной. Заметим, что седловая задача (12) обобщает выпуклую постановку (1) при $\Phi(x, y) \equiv \Phi(x)$, то есть является ее логичным продолжением. В связи с этим отдельный интерес представляют сравнение оценок из предыдущего раздела с результатами для задачи (12), а также перенос возможных идей или решений на случай седла.

Задача (12) рассматривается при следующих предположениях.

Предположение 3 (липшицевость). $\Phi(x, y, \xi)$ *является* M_x*-липшицевой на* $\mathcal{X}$ *относительно нормы* $\|\cdot\|_{p_x}$ *по переменной* x *и* M_y*-липшицевой на* $\mathcal{Y}$ *относительно* $\|\cdot\|_{p_y}$ *по* y.

Предположение 4 (гладкость). $\Phi(x, y, \xi)$ *является* L_x*-,* L_y*-,* L_{xy}*-гладкой по каждому из градиентов* ∇_x, ∇_y *и по каждой переменной* x, y *относительно норм* $\|\cdot\|_{p_x}$, $\|\cdot\|_{p_y}$. *То есть* $\nabla_x\Phi(x, y, \xi)$ *является* L_x*-липшицевой по* x *и* L_{xy}*-липшицевой по* y *и аналогично* $\nabla_y\Phi(x, y, \xi)$.

Предположение 5 (сильная выпуклость – сильная вогнутость). $\Phi(\cdot, y, \xi)$ *является* μ_x*-сильно выпуклой относительно* $\|\cdot\|_{p_x}$*-нормы для почти всех* ξ. $\Phi(x, \cdot, \xi)$ *является* μ_y*-сильно вогнутой относительно* $\|\cdot\|_{p_y}$*-нормы для почти всех* ξ.

Предположение 6 (ограниченный шум). *Существуют константы* $\sigma_x, \sigma_y \geqslant 0$ *такие, что для выбранных* $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\xi_x, \xi_y \sim \mathcal{D}_\xi$, *для которых выполнено* $\mathbb{E}_\xi[\nabla_x\Phi(x, y, \xi_x)] = \nabla_x\Phi(x, y)$ *и* $\mathbb{E}_\xi[\nabla_y\Phi(x, y, \xi_y)] = \nabla_y\Phi(x, y)$, *справедливо*

$$\begin{cases} \mathbb{E}_\xi\left[\|\nabla_x\Phi(x, y, \xi_x) - \nabla_x\Phi(x, y)\|_{p_x}^2\right] \leqslant \sigma_x^2, \\ \mathbb{E}_\xi\left[\|\nabla_x\Phi(x, y, \xi_y) - \nabla_y\Phi(x, y)\|_{p_y}^2\right] \leqslant \sigma_y^2. \end{cases} \tag{13}$$

Пусть (x^*, y^*) есть истинное решение задачи (12). Результатом работы алгоритма $\mathcal{A}\left(\{\xi_k\}_{k=1}^N\right)$ является пара $(\widehat{x}, \widehat{y})$, которая оценивается с помощью метрики сходимости по аргументу $d^2(\widehat{x}, \widehat{y}) := \mathbb{E}\left[\left\|\widehat{x} - x^*\right\|_{p_x}^2 + \left\|\widehat{y} - y^*\right\|_{p_y}^2\right]$ и по функции $\Delta^s(\widehat{x}, \widehat{y}) := \mathbb{E}_\xi\left[\max\limits_{y\in\mathcal{Y}} \Phi(\widehat{x}, y) - \min\limits_{x\in\mathcal{X}} \Phi(x, \widehat{y})\right]$.

Для детерминированной версии задачи (12) в предположениях 3, 4, 5 по норме $p = 2$ существует нижняя оценка сложности алгоритма [Zhang, Hong, Zhang, 2019]

$$\Omega\left(\left(\sqrt{\frac{L_x}{\mu_x}} + \frac{L_{xy}}{\sqrt{\mu_x\mu_y}} + \sqrt{\frac{L_y}{\mu_y}}\right)\ln\left(\frac{1}{\varepsilon}\right)\right). \tag{14}$$

И для случая билинейного седла $\Phi(x, y) = f(x) + x^\top\mathbf{A}y - g(y)$ разработаны оптимальные онлайн-алгоритмы [Kovalev, Gasnikov, Richtárik, 2021; Jin, Sidford, Tian, 2022; Thekumparampil, He, Oh, 2022].

Для стохастического случая существующая теория более скромная. Случай $\mu_x = \mu_y$ и $L_x = L_y$ хорошо известен в литературе, и для него мы можем объединить пространства $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ и переменные $\Phi(z, \xi) := \Phi(x, y, \xi) = [\nabla_x\Phi(x, y, \xi) - \nabla_y\Phi(x, y, \xi)]$, рассматривая вместо задачи (12) стохастическое вариационное неравенство:

$$\text{найти } z^* \text{ такое, что } \langle\Phi(z^*), z^* - z\rangle \geqslant 0 \ \forall z \in \mathcal{Z}, \tag{15}$$

где $\Phi(z) = \mathbb{E}_{\xi\sim\mathcal{D}}[\Phi(z, \xi)]$. Для такой постановки в предположении 5 ($\mu_x = \mu_y = 0$) известно, что $z^* \in \mathcal{Z}$ является решением задачи (15) тогда и только тогда, когда z^* является решением задачи (12). Для задач (15) существует обзор онлайн-алгоритмов [Beznosikov et al., 2022]. Из него, в частности, следует (теорема 9), что при сделанных предположениях 3, 5, 6 ($\sigma_x^2 = \sigma_y^2 = \sigma^2$) для онлайн-алгоритма

$$N = O\left(\frac{L}{\mu}\ln\left(\frac{LR_0^2}{\mu\varepsilon}\right) + \frac{\sigma^2}{\mu^2\varepsilon}\right), \tag{16}$$

где $L = M_x = M_y$, $\mu = \mu_x = \mu_y$ и $R_0^2 = \left\|z^0 - z^*\right\|^2$. Данная оценка выполняется для сходимости по аргументу.

В случае различных констант сильной выпуклости допущения делаются относительно вида целевой функции $\Phi(x, y)$. Так, распространенным видом для исследования является билинейный класс функций $\Phi(x, y) = f(x) + x^\top\mathbf{A}y - g(y)$. Интерес к такому классу связан с его широким

применением в различных задачах машинного обучения, таких как обучение с подкреплением, регуляризированная задача минимизации эмпирического риска и другие. На этот счет существует сразу несколько исследований для онлайн-алгоритмов [Metelev et al., 2022; Li et al., 2022; Du et al., 2022], которые имеют свои особенности. Так, работа [Metelev et al., 2022] обобщает (12) на децентрализованную постановку, а [Li et al., 2022; Du et al., 2022] используют стохастические версии онлайн-алгоритмов спуска-подъема (в зарубежной литературе — gradient descent ascent). Специфичный вид функционала переопределяет введенные предположения. Так, предположения 3, 4, 5 в данном случае накладываются на функции $f(x)$ и $g(y)$. При этом предположение 6, задающее стохастическое условие, может интерпретироваться по-разному. Например, в работе [Metelev et al., 2022] билинейное слагаемое $x^\top \mathbf{A} y$ являлось детерминированным и различалась стохастичность $f(x) = f(x, \xi_x)$, $g(y) = g(y, \xi_y)$. А в работах [Li et al., 2022; Du et al., 2022] билинейное слагаемое $H(x, y) = x^\top \mathbf{A} y = x^\top \mathbf{A}_\eta y = H(x, y, \eta)$ было стохастичным, но стохастика $F(z) := f(x) + g(y) = F(z, \xi)$ объединялась. Обозначим константы из предположения 6: σ_f^2, σ_g^2, σ_H^2, σ_F^2 соответственно. Из [Metelev et al., 2022] в предположениях 4, 5, 6 для онлайн-алгоритма

$$N = O\left(\left(N_{\text{det}} + N_{\text{stoch}}\right) \log\left(\frac{1}{\varepsilon}\right)\right), \tag{17}$$

где

$$N_{\text{det}} = \max\left\{\sqrt{\frac{L_x}{\mu_x}},\ \frac{L_{xy}}{\sqrt{\mu_x \mu_y}},\ \sqrt{\frac{L_y}{\mu_y}}\right\} \tag{18}$$

и

$$N_{\text{stoch}} = \frac{L_x}{L_{xy}\sqrt{\mu_x \mu_y}}\left(\left(\frac{1}{L_x} + \sqrt{\frac{\mu_x}{\mu_y}}\frac{1}{L_{xy}}\right)\frac{\sigma_f^2}{\varepsilon} + \left(\frac{1}{L_y} + \sqrt{\frac{\mu_y}{\mu_x}}\frac{1}{L_{xy}}\right)\frac{\sigma_g^2}{\varepsilon}\right). \tag{19}$$

А для [Li et al., 2022; Du et al., 2022] в тех же предположениях

$$N = O\left(\left(\sqrt{\frac{L_x}{\mu_x} \vee \frac{L_y}{\mu_y}} + \frac{L_{xy}}{\sqrt{\mu_x \mu_y}}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma_F^2 + \sigma_H^2}{\mu_x^2 \varepsilon^2}\right). \tag{20}$$

Стоит отметить, что формулы (18), (19) являются одним из возможных вариантов (полный список можно найти в теореме 4.2 оригинальной статьи). Сравнивая формулы (17), (20), можно заметить, что они обе оптимальны по части детерминированного слагаемого, однако стохастическая часть в (17) лучше. В сравнении с оценкой (11) стохастическая часть в (17) содержит дополнительный множитель $\log \frac{1}{\varepsilon}$. Авторы статьи объясняют его синхронизацией в организации распределенных вычислений.

Совсем недавно в работе [Zhang, Aybat, Gürbüzbalaban, 2021] был предложен онлайн-алгоритм, работающий с различными константами сильной выпуклости целевой функции $\Phi(x, y) = f(x) + H(x, y) - g(y)$, у которой $H(x, y)$ не обязательно принадлежит билинейному классу. В ней $f(x)$, $g(y)$ предполагаются собственными, замкнутыми, μ_x-, μ_y-сильно выпуклыми функциями соответственно. Слагаемое $H(x, y)$ удовлетворяет предположениям 3, 5 (с $\mu_x = \mu_y = 0$), 6. При сделанных предположениях

$$N = O\left(\left[\frac{L_{xx}}{\mu_x} + \frac{L_{xy}}{\sqrt{\mu_x \mu_y}} + \frac{L_{yy}}{\mu_y} + \left\{\left(1 + \sqrt{\frac{\mu_x}{\mu_y}}\right)\frac{\sigma_x^2}{\mu_x} + \left(1 + \frac{L_{xy}}{L_{yx}} + \sqrt{\frac{\mu_y}{\mu_x}}\right)\frac{\sigma_y^2}{\mu_y}\right\}\frac{1}{\varepsilon}\right]\ln\left(\frac{1}{\varepsilon}\right)\right). \tag{21}$$

Несмотря на то что оценка (21) не является оптимальной ни по детерминированной, ни по стохастической части, это один из единственных конкурентных результатов для выпукло-вогнутой оптимизации в таких предположениях общности. Описанные онлайн-алгоритмы работают для сходимости по аргументу в евклидовой норме.

Для офлайн-алгоритмов на данный момент нам известна всего лишь одна основная работа [Zhang et al., 2021], которая исследует задачу (12). Соответствующая ей вспомогательная задача —

$$\min_{x\in X}\max_{y\in Y}\widehat{\Phi}_n(x,\,y) := \frac{1}{n}\sum_{i=1}^{n}\Phi(x,\,y,\,\xi_i). \tag{22}$$

В данной задаче предположение 3 ослаблено: функция $\Phi(x, y, \xi)$ неравномерно ограничена по ξ, то есть $M_x(y) = M_x(\xi, y)$ и $M_y(x) = M_y(\xi, x)$. Вместо этого ограничены вторые моменты $M_x^2 = \mathbf{E}_\xi\left[\sup_{y\in Y} M_x^2(\xi, y)\right]$ и аналогично M_y^2. Также такое ограничение является офлайн-альтернативой предположению 6, поскольку в этом случае алгоритм не работает со стохастическими оракулами. С учетом этих замечаний из работы [Zhang et al., 2021] в предположениях 3, 5 для сходимости по аргументу офлайн-алгоритма

$$N = O\left(\frac{1}{\mu_{\min}}\left(\frac{M_x^2}{\varepsilon\mu_x} + \frac{M_y^2}{\varepsilon\mu_y}\right)\right) \tag{23}$$

и в предположениях 3, 4, 5 для сходимости по функции

$$N = O\left(\sqrt{1 + \frac{L_{xy}^2}{\mu_x\mu_y}}\left(\frac{M_x^2}{\varepsilon\mu_x} + \frac{M_y^2}{\varepsilon\mu_y}\right)\right). \tag{24}$$

Учитывая общий смысл констант M_x^2 и σ_f^2, M_y^2 и σ_g^2 и сравнивая формулы (17) и (23), можно заметить их концептуальную схожесть. Это же мы наблюдали и для соответствующих оценок (11), (8) в выпуклом случае. Такое сравнение указывает на общую природу офлайн- и онлайн-алгоритмов. Сравнивая стохастические части (17) с (11) и (23) с (8), мы также наблюдаем общие паттерны, которые демонстрируют идею того, что стохастические седловые задачи (12) являются продолжением (1).

Еще одной особенностью взаимосвязи задач (1) и (12) является возможность переноса результатов из одной области в другую. Это частично было сделано в [Dvinskikh et al., 2021] для острого минимума в случае офлайн-алгоритма. В нынешней работе оценка (10) будет перенесена на выпукло-вогнутую оптимизацию. Также был проработан вопрос точности решения вспомогательной задачи (22).

Основные результаты

Оценка точности оптимизации эмпирической седловой функции

Теорема 1. *Пусть функция* $\Phi(x, y, \xi)$*, определенная в* (12)*, удовлетворяет предположениям* (3)*,* (4) *и* (5)*, то есть она является сильно выпуклой – сильно вогнутой, имеет место липшицевость функции и ее градиента. Тогда если решение* $(\widehat{x}_{\varepsilon'}, \widehat{y}_{\varepsilon'})$ *обеспечивает точность* $\varepsilon' = \Delta_n^s(\widehat{x}_{\varepsilon'}, \widehat{y}_{\varepsilon'})$*, то*

$$\varepsilon = \Delta^s(\widehat{x}_{\varepsilon'}, \widehat{y}_{\varepsilon}) \leqslant C\sqrt{\varepsilon'} + B, \tag{25}$$

$$C = \max\left\{\frac{2M_x^s}{\sqrt{\mu_x}}, \frac{2M_y^s}{\sqrt{\mu_y}}\right\}, \quad \Delta^s(\widehat{x}, \widehat{y}) \leqslant \frac{2\sqrt{2}}{n}\cdot\sqrt{\frac{L_{xy}^2}{\mu_x\mu_y} + 1}\cdot\left(\frac{(M_x^s)^2}{\mu_x} + \frac{(M_y^s)^2}{\mu_y}\right) = B. \tag{26}$$

Таким образом, мы должны решать эмпирическую задачу с точностью $\varepsilon' = O\left(\varepsilon^2\right)$*, чтобы обеспечить точность* ε*. Однако на практике* μ_x *и* μ_y *зачастую порядка* ε*. Введем* $\mu_{\min} = \min\{\mu_x, \mu_y\}$*,*

тогда $\varepsilon' = O\left(\mu_{\min} \cdot \varepsilon^2\right)$*, то есть для метрики SGM зависимость порядка третьей степени. Теорема является обобщением для седлового случая аналогичных результатов теоремы* 2.1.3 *из* [Dvinskikh, 2021]*, оценивающих точность приближенного решения эмпирической функции в выпуклой оптимизационной задаче. Ограничение на* $\Delta^s(\widehat{x}, \widehat{y})$ *— результат теоремы* 3 *из* [Zhang et al., 2021].

Условие острого минимума

Для задачи (12) будет удобнее обозначить

$$z = (x,\, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y},$$
$$\Phi(x,\, y) = \Phi(z).$$

На пространстве $\mathcal{Z}$ можно ввести некоторую метрику, использующую метрики $\mathcal{X}$, $\mathcal{Y}$ (например, сумму). Подобно работе [Judistky, Nesterov, 2014], для точки $x_0 \in \mathcal{X}$ обозначим через $\mathcal{X}_R(x_0) = \mathcal{X} \cap B_R(x_0)$, где $B_R(x_0)$ — шар радиусом R с центром x_0. Аналогично введем множества $\mathcal{Y}_R(y_0)$ и $\mathcal{Z}_R(z_0)$.

В рассматриваемом онлайн-методе будем считать, что выбраны начальная точка $z_0 = (x_0, y_0) \in \mathcal{Z}$ и радиусы R_x, R_y. Заметим, что $\mathcal{X}_{R_x} \times \mathcal{Y}_{R_y} \subset \mathcal{Z}_R(z_0)$ для некоторого R, зависящего от метрики на $\mathcal{Z}$. В случае метрики-суммы $\rho(z_1, z_2) = \rho_x(x_1, x_2) + \rho_y(y_1, y_2)$, где ρ_x, ρ_y — метрики на $\mathcal{X}$, $\mathcal{Y}$, $R = R_x + R_y$.

Будем далее считать, что в задаче (12) оптимизация ведется не на пространствах $\mathcal{X}$, $\mathcal{Y}$, а на пространствах $\mathcal{X}_{R_x}(x_0)$, $\mathcal{Y}_{R_y}(y_0)$.

Вслед за условием γ-роста для выпуклых задач можно обобщить условие (5). Сформулируем условие ρ-роста ($\rho > 1$) для седловых задач:

$$\Phi(x,\, y^*) - \Phi(x^*,\, y) \geqslant \mu(\Phi) \left\| z - z^* \right\|^{\rho}. \tag{27}$$

Теорема 2. *Рассмотрим* (12) *в следующих предположениях*:

- *предположение* 1 *выполнено для функций* $\Phi(x, \cdot)$, $\Phi(\cdot, y)$ *с константами* L_x, L_y;
- *функция* $\Phi(x, y)$ *удовлетворяет условию* ρ*-роста* (27);
- *стохастические оракулы* $g_k^x = \nabla_x \Phi(x_k, y_k, \xi)$, $g_k^y = \nabla_y \Phi(x_k, y_k, \xi)$ *удовлетворяют предположению* 1;
- *соответствующие стохастическим оракулам* g_k^x, g_k^y *ошибки* ε_k^x, ε_k^y *удовлетворяют предположению* 2.

Пусть $z^* = (x^*, y^*)$ *— решение задачи* (12)*. Тогда для достижения точности* $\|z_k - z\| \leqslant \varepsilon$ *онлайн-алгоритму достаточен размер выборки* $N = O\left(\mu(\Phi)^{-2/\rho} \varepsilon^{-2(\rho-1)/\rho}\right)$.

Заключение

Решение стохастической седловой задачи (12) является актуальным направлением исследования: существующие передовые работы были опубликованы в течение последних двух-трех лет. Несмотря на это, данная область по-прежнему содержит большое количество открытых проблем. В нынешней работе эти проблемы были подсвечены через призму сравнения существующих решений с аналогичными результатами в выпуклой стохастической оптимизации (1). Так,

в частности, большой зазор остается в исследовании общности задачи (12): передовые онлайн-алгоритмы работают в ограничениях вида целевой функции $\Phi(x, y)$, евклидовости норм $\|\cdot\|_{p_x} = \|\cdot\|_{p_y} = \|\cdot\|_2$ и метрики сходимости по аргументу $d^2(\widehat{x}, \widehat{y})$. Также офлайн-алгоритмы для (12) представлены единственным конкурентным исследованием [Zhang et al., 2021], которое работает в высокой степени общности постановки седловой задачи, но теряет в детерминированном слагаемом оценки вычислительной сложности (23), (24).

В данной работе удается устранить некоторые зазоры, присутствующие в сравнении алгоритмов для задач (1) и (12). В частности, это удается сделать, взглянув на задачу (12) как на продолжение задачи (1). Эта идея уже использовалась в работе [Dvinskikh et al., 2021] для переноса результатов офлайн-алгоритма задачи (1) в случае острого минимума на выпукло-вогнутую постановку. В нынешней работе такой перенос был осуществлен для онлайн-алгоритма (10). Также был проработан вопрос точности решения вспомогательной седловой задачи (22).

Список литературы (References)

Поляк Б. Т. Новый метод типа стохастической аппроксимации // Автоматика и телемеханика. — 1990. — № 7. — С. 98–107.

Polyak B. T. A new method of stochastic approximation type // Autom. Remote Control. — 1990. — Vol. 51, No. 7. — P. 937–946. (Original Russian paper: *Polyak B. T.* Novyi metod tipa stokhasticheskoi approksimatsii // Avtomatika i telemekhanika. — 1990. — No. 7. — P. 98–107.)

Beznosikov A. et al. Smooth monotone stochastic variational inequalities and saddle point problems–survey // arXiv preprint arXiv:2208.13592. — 2022.

Du S. S. et al. Optimal extragradient-based bilinearly-coupled saddle-point optimization // arXiv preprint arXiv:2206.08573. — 2022.

Dvinskikh D. Decentralized algorithms for wasserstein barycenters. — Germany: Humboldt Universitaet zu Berlin, 2021.

Dvinskikh D. et al. On the relations of stochastic convex optimization problems with empirical risk minimization problems on *p*-norm balls // Computer research and modeling. — 2022. — Vol. 14, No. 2. — P. 309–319.

Huang Y. et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism // Advances in neural information processing systems. — 2019. — Vol. 32. — P. 103–112.

Jin Y., Sidford A., Tian K. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods // Conference on Learning Theory. — PMLR, 2022. — P. 4362–4415.

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

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

Kovalev D., Gasnikov A., Richtárik P. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling // arXiv preprint arXiv:2112.15199. — 2021.

Li C. J. et al. Nesterov meets optimism: Rate-optimal optimistic-gradient-based method for stochastic bilinearly-coupled minimax optimization // arXiv preprint arXiv:2210.17550. — 2022.

Li S., Liu Y. Improved learning rates for stochastic optimization: Two theoretical viewpoints // arXiv preprint arXiv:2107.08686. — 2021.

Metelev D. et al. Decentralized saddle-point problems with different constants of strong convexity and strong concavity // arXiv preprint arXiv:2206.00090. — 2022.

Nemirovski A. et al. Robust stochastic approximation approach to stochastic programming // SIAM Journal on optimization. — 2009. — Vol. 19, No. 4. — P. 1574–1609.

Nesterov Y. Accelerating the cubic regularization of Newton's method on convex problems // Mathematical Programming. — 2008. — Vol. 112, No. 1. — P. 159–181.

Nesterov Y. Primal-dual subgradient methods for convex problems // Mathematical programming. — 2009. — Vol. 120, No. 1. — P. 221–259.

Shalev-Shwartz S., Shamir O., Srebro N., Sridharan K. Stochastic convex optimization // COLT. — 2009. — Vol. 2. — P. 5.

Shapiro A., Dentcheva D., Ruszczynski A. Lectures on stochastic programming: modeling and theory. — Society for Industrial and Applied Mathematics, 2021.

Shapiro A., Nemirovski A. On complexity of stochastic programming problems // Continuous optimization: Current trends and modern applications. — 2005. — P. 111–146.

Thekumparampil K. K., He N., Oh S. Lifted primal-dual method for bilinearly coupled smooth minimax optimization // International Conference on Artificial Intelligence and Statistics. — PMLR, 2022. — P. 4281–4308.

Zhang J. et al. Generalization bounds for stochastic saddle point problems // International Conference on Artificial Intelligence and Statistics. — PMLR, 2021. — P. 568–576.

Zhang J., Hong M., Zhang S. On lower iteration complexity bounds for the saddle point problems // arXiv preprint arXiv:1912.07481. — 2019.

Zhang X., Aybat N. S., Gürbüzbalaban M. Robust accelerated primal-dual methods for computing saddle points // arXiv preprint arXiv:2111.12743. — 2021.

Zhivotovskiy N., Klochkov Y. Stability and deviation optimal risk bounds with convergence rate $O\left(\frac{1}{n}\right)$ // NeurIPS Proceedings. — 2022.