УДК 519.856, 519.856.3

Параллельные алгоритмы и оценки вероятностей больших уклонений в задачах стохастической выпуклой оптимизации*

П.Е. Двуреченский[1,2], А.В. Гасников[2,3], А.А. Лагуновская[3]

[1]Институт прикладного анализа и стохастики им. К. Вейерштрасса, Моренштрассе, 39, Берлин, Германия, 10117

[2]Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Большой Каретный пер., 19, строение 1, Москва, 127051

[3]Московский физико-технический институт, Институтский пер., 9, Долгопрудный, Московская обл., 141700

E-mails: pavel.dvurechensky@wias-berlin.de (Двуреченский П.Е.), gasnikov.av@mipt.ru (Гасников А.В.), a.lagunovskaya@phystech.edu (Лагуновская А.А.)

Двуреченский П.Е., Гасников А.В., Лагуновская А.А. Параллельные алгоритмы и оценки вероятностей больших уклонений в задачах стохастической выпуклой оптимизации // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. — Новосибирск, 2018. — Т. 21, № 1. — С. 47–53.

В этом коротком сообщении рассматриваются задачи выпуклой стохастической оптимизации при различных предположениях о свойствах стохастических субградиентов. Известно, что если вычислителю доступно значение целевой функции задачи, то можно параллельно вычислить несколько независимых приближений к решению задачи в терминах сходимости по математическому ожиданию. Выбрав приближение с наименьшим значением функции, можно контролировать вероятности больших уклонений невязки по значению функции. В данной работе рассматривается случай, когда значение целевой функции недоступно или требует большого объема вычислений. В предположении субгауссовости распределения стохастических субградиентов, а также в общем случае при умеренном уровне вероятности больших уклонений показано, что параллельное вычисление нескольких приближенных решений с последующим усреднением дает те же оценки вероятностей больших уклонений невязки по функции, что и вычисление одного приближенного решения, но с большим числом итераций. Тем самым в рассматриваемом случае параллельные вычисления позволяют получить решение того же качества, но за меньшее время.

DOI: 10.15372/SJNM20180103

Ключевые слова: *стохастическая выпуклая оптимизация, оценки вероятностей больших уклонений, метод зеркального спуска, параллельные алгоритмы.*

Dvurechensky P., Gasnikov A., Lagunovskaya A. Parallel algorithms and probability of large deviation for stochastic convex optimization problems // Siberian J. Num. Math. / Sib. Branch of Russ. Acad. of Sci. — Novosibirsk, 2018. — Vol. 21, № 1. — P. 47–53.

In this paper, convex stochastic optimization problems under different assumptions on the properties of the available stochastic subgradients are considered. It is known that if a value of the objective function is available, one can obtain, in parallel, several independent approximate solutions in terms of the objective residual expectation. Then, choosing a solution with the minimum function value, one can control the probability of large deviations of the objective residual. On the contrary, in this short paper we address the situation when the objective function value is unavailable or is too expensive to calculate. Under the "light-tail" assumption for stochastic subgradients and in the general case with a moderate probability of large deviations, it is shown that parallelization combined with averaging gives bounds for the probability of large deviations similar to

*Исследование А.В. Гасникова и П.Е. Двуреченского в пункте 3 выполнено в ИППИ РАН за счет гранта Российского научного фонда (проект № 14-50-00150). Исследование А.А. Лагуновской частично поддержано грантом Президента РФ МК-1806.2017.9.

those of a serial method. Thus, in these cases one can benefit from parallel computations and reduce the computational time without any loss in the solution quality.

Keywords: *stochastic convex optimization, probability of large deviation, mirror descent, parallel algorithm.*

1. Введение

Рассмотрим следующий класс задач стохастической оптимизации на выпуклом компактном множестве Q:

$$\min_{x \in Q \subset E} \left\{ f(x) := \mathbb{E}_\xi \left[f(x, \xi) \right] \right\}, \tag{1}$$

где E — конечномерное вещественное векторное пространство, ξ — случайный вектор, $f(x, \xi)$ — замкнутая выпуклая функция от x для почти всех ξ, а ξ имеет распределение, независимое от x. При этих предположениях данная задача является задачей выпуклой оптимизации.

Эта задача рассматривается в крупномасштабной постановке в предположении большой размерности пространства E. Наша основная цель — приближенно решить эту задачу с использованием некоторого алгоритма первого порядка, используя только значения функции и субградиенты или их аппроксимации. Обычно в литературе по стохастической оптимизации [1, 2] рассматриваются две меры качества приближенного решения $\bar{x}$. Первая — математическое ожидание невязки целевой функции. В этом случае $\bar{x}$ является ε-решением (1) для $\varepsilon > 0$, если и только если $\mathbb{E}f(\bar{x}) - f_* \leq \varepsilon$, где f_* — оптимальное значение в (1), и математическое ожидание берется относительно всех случайностей в алгоритмическом процессе. Вторая — граница вероятности большого уклонения невязки целевой функции. В этом случае $\bar{x}$ является (ε, σ)-решением (1) для $\varepsilon > 0$, $\sigma \in (0, 1)$, если и только если $\mathbb{P}\{f(\bar{x}) - f_* > \varepsilon\} \leq \sigma$. В данной статье главным образом рассматривается вторая мера качества.

Известно, что если значение целевой функции доступно (например, в рандомизированных методах [3]), то можно параллельно получить логарифмическое по σ^{-1} число независимых ε-решений. Тогда решение с минимальным значением функции является (ε, σ)-решением. Тем не менее значение целевой функции может быть недоступно или его вычисление или аппроксимация могут быть слишком дорогими. Последнее легко представить себе, поскольку для $\xi \in \mathbb{R}^p$ вычислительные затраты для вычисления $\bar{f}$ такого, что $|\bar{f} - f(x)| \leq \delta$, могут достичь $O(\delta^{-p})$ вычислений $f(x, \xi)$ при различных ξ. Наша цель — предложить метод, позволяющий получить (ε, σ)-решение на основе ряда параллельно вычисленных ε-решений без вычисления значения функции $f(x)$.

Наш подход основан на алгоритме стохастического зеркального спуска [2]. Оказывается, что при некоторых слабых предположениях ε-решение (1), полученное методом стохастического зеркального спуска, также является $(\tilde{\varepsilon}, \sigma)$-решением. Мы используем этот факт и параллельно вычисляем логарифмическое по σ^{-1} число независимых ε-решений, усредняем их и доказываем, что это среднее является (ε, σ)-решением. Таким образом, мы можем извлечь пользу из распараллеливания и сократить время вычисления без потери качества решения.

2. Метод стохастического зеркального спуска

В данном пункте приводится описание метода стохастического зеркального спуска (СЗС) [2, 3] и свойств его сходимости в терминах математического ожидания невязки целевой функции, а также в терминах вероятности больших уклонений этой невязки. Эти результаты по сходимости служат основой нашего подхода к построению (ε, σ)-решения путем распараллеливания. Возьмем некоторую норму $\|\cdot\|$ на E и обозначим сопряженную норму в двойственном пространстве E^* как $\|\cdot\|_*$. Предположим, что в любой точке $x \in Q$ стохастический субградиент $\nabla_x f(x, \xi)$ от $f(x)$ имеется и удовлетворяет

$$\mathbb{E}_\xi\left[\nabla_x f(x,\xi)\right] \in \partial f(x), \quad \mathbb{E}_\xi\left[\left\|\nabla_x f(x,\xi)\right\|_*^2\right] \le M^2 \tag{2}$$

для некоторой постоянной $M > 0$. Здесь $\partial f(x)$ обозначает субдифференциал f в точке x. Выберем функцию близости $d(x)$, $x \in Q$, которая является 1-строго выпуклой в $\|\cdot\|$. Пусть $x^0 = \arg\min_{x\in Q} d(x)$. Без потери общности предположим, что $d(x^0) = 0$. Алгоритм использует дивергенцию Брегмана $V_z(x) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle$. Пусть x_* — решение (1), R — число такое, что $V_{x^0}(x_*) \le R^2$, а $\bar{R}$ — число такое, что $\max_{x\in Q} V_x(x_*) \le \bar{R}$.

Итерации стохастического зеркального спуска [2, 3] проводятся следующим образом, начиная с $x^0 \in Q$:

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h\nabla_x f\left(x^k, \xi^k\right)\right), \quad \mathrm{Mirr}_{x^k}(v) := \arg\min_{x\in Q}\left\{\left\langle v, x - x^k\right\rangle + V_{x^k}(x)\right\}, \ v \in E^*, \tag{3}$$

где $h > 0$ — размер шага, $\left\{\xi^k\right\}_{k\ge 0}$ — выборка независимых одинаково распределенных ξ. Основное свойство СЗС-шага [3] следующее:

$$2V_{x^{k+1}}(x) \le 2V_{x^k}(x) + 2h\left\langle \nabla_x f\left(x^k, \xi^k\right), x - x^k\right\rangle + h^2\left\|\nabla_x f\left(x^k, \xi^k\right)\right\|_*^2 \quad \forall\, x \in Q.$$

Теперь, используя выпуклость $f(x)$, для любых $\nabla f(x^k) \in \partial f(x^k)$ и $x \in Q$ мы имеем

$$f\left(x^k\right) - f(x) \le \left\langle \nabla f\left(x^k\right), x^k - x\right\rangle \le \left\langle \nabla f(x^k) - \nabla_x f\left(x^k, \xi^k\right), x^k - x\right\rangle + \frac{1}{h}\left(V_{x^k}(x) - V_{x^{k+1}}(x)\right) + \frac{h}{2}\left\|\nabla_x f\left(x^k, \xi^k\right)\right\|_*^2.$$

Возьмем условное математическое ожидание относительно $\xi^1, \ldots, \xi^{k-1}$ и используем (2). Тогда получим

$$f\left(x^k\right) - f(x) \le \frac{1}{h}\left(V_{x^k}(x) - \mathbb{E}\left[V_{x^{k+1}}(x) \mid \xi^1, \ldots, \xi^{k-1}\right]\right) + \frac{h}{2}\underbrace{\mathbb{E}\left[\left\|\nabla_x f\left(x^k, \xi^k\right)\right\|_*^2 \mid \xi^1, \ldots, \xi^{k-1}\right]}_{\overset{(2)}{\le} M^2}.$$

Поскольку $\left\{\xi^k\right\}_{k\ge 0}$ является выборкой независимых одинаково распределенных ξ, то, взяв полное математическое ожидание от обеих сторон этих неравенств для $k = 0, \ldots, N-1$, сложив их и приняв $x = x_*$, получим, вследствие выпуклости $f(x)$:

$$\mathbb{E}\left[f\left(\bar{x}^N\right)\right] - f_* \le \mathbb{E}\,\frac{1}{N}\sum_{k=0}^{N-1} f\left(x^k\right) - f_* \le \frac{1}{hN}V_{x^0}(x_*) + \frac{M^2 h}{2} \le \sqrt{\frac{2M^2R^2}{N}},$$

где

$$R \text{ — такое, что } V_{x^0}(x_*) \le R^2, \quad \bar{x}^N := \frac{1}{N}\sum_{k=0}^{N-1} x^k, \quad h = \frac{R}{M}\sqrt{\frac{2}{N}}. \tag{4}$$

Выбрав некоторую точность $\varepsilon > 0$ и

$$N = \left\lceil \frac{2M^2R^2}{\varepsilon^2} \right\rceil, \tag{5}$$

мы получим, что $\bar{x}^N$ удовлетворяет $\mathbb{E}\left[f\left(\bar{x}^N\right)\right] - f_* \le \varepsilon$ и, следовательно, это ε-решение. Отметим, что эта граница для N является оптимальной [3] с точностью до постоянного множителя для класса задач выпуклого стохастического программирования (1) с почти всюду ограниченными стохастическими субградиентами.

Оказывается, что при дополнительных предположениях можно доказать наличие границ вероятности больших уклонений для $f\left(\bar{x}^N\right) - f_*$.

Предложение [2, 5, 6]. *Предположим, что верно одно из следующих предположений*:
а) $\|\nabla_x f(x,\xi)\|_* \le M$ *для почти всех* ξ;
б) $\mathbb{E}_\xi\left(\exp\left(\left\|\nabla_x f(x,\xi)\right\|_*^2 / M^2\right)\right) \le \exp(1)$ *и* $\ln\sigma^{-1} \ll N$;
в) *существует некоторая* $\alpha > 2$ *такая, что для всех* $t \ge 0$ $\mathbb{P}\left(\frac{\|\nabla f(x,\xi)\|_*^2}{M^2} \ge t\right) \le \frac{1}{(t+1)^\alpha}$ *и* $\sigma^{-1/(\alpha-1)} \ll N$.
Тогда точка $\bar{x}^N$, *сгенерированная при помощи СЗС* (3), (4) *после* $N \ge 0$ *шагов, удовлетворяет*

$$\mathbb{P}\left\{f\left(\bar{x}^N\right) - f_* \le \frac{C_1 M}{\sqrt{N}}\left(R + C_2\bar{R}\sqrt{\ln(1/\sigma)}\right)\right\} \ge 1 - \sigma, \tag{6}$$

где R — *такое, что* $V_{x^0}(x_*) \le R^2$, $\bar{R}$ — *такое, что* $\max_{x\in Q} V_x(x_*) \le \bar{R}$, *и в случае а)* $C_1 = \sqrt{2}$, $C_2 = 2\sqrt{2}$; *в случае* б) $C_1 = C_2 = 2\sqrt{2}$; *а в случае* в) $C_1 = C_1(\alpha)$, $C_2 = 1$.

Следствие. *Пусть любое из трех предположений* а)–в) *предложения верно. Возьмем* $N = \left\lceil \frac{CM^2\bar{R}^2}{\varepsilon^2} \right\rceil$, *где постоянная* C *зависит от* C_1, C_2. *Тогда точка* $\bar{x}^N$, *сгенерированная при помощи СЗС* (3), (4), *удовлетворяет для любого* $c \ge 0$ *следующему соотношению*:

$$\mathbb{P}\left\{f\left(\bar{x}^N\right) - f_* \ge c\right\} \le \mathbb{P}\left\{\eta \ge c\right\}, \tag{7}$$

где $\eta \in \mathrm{N}\left(\varepsilon, \varepsilon^2\right)$ — *нормальная случайная переменная со средним* ε *и дисперсией* ε^2.

3. Распараллеливание и границы вероятности больших уклонений

В данном пункте сначала обсудим известный способ получения (ε,σ)-решения с использованием ряда ε-решений, вычисленных параллельно. Затем предложим новый метод его получения без вычисления значения целевой функции $f(x)$, сформулируем и докажем основной результат.

Предположим, что $\bar{x}^N$ — $\varepsilon/2$-решение, полученное с использованием СЗС (3), (4) при $N = \left\lceil \frac{8M^2R^2}{\varepsilon^2} \right\rceil$. Используя неравенство Маркова [5], получим

$$\mathbb{P}\Big\{f\big(\bar{x}^N\big) - f_* \geq \varepsilon\Big\} \leq \frac{\mathbb{E}\left[f\left(\bar{x}^N\right)\right] - f_*}{\varepsilon} \leq \frac{1}{2}.$$

Если параллельно вычислить $K = \left\lceil \log_2\left(\sigma^{-1}\right)\right\rceil$ независимых СЗС $\varepsilon/2$-решений $\left\{\bar{x}^{N,i}\right\}_{i=1}^K$ и выбрать $\bar{x}^N_{\min}$, минимизирующее $f\left(\bar{x}^{N,i}\right)$, то, выполнив всего

$$\left\lceil \frac{8M^2R^2}{\varepsilon^2} \right\rceil \left\lceil \log_2\left(\sigma^{-1}\right)\right\rceil$$

вычислений стохастического субградиента, мы получим

$$\mathbb{P}\Big\{f\left(\bar{x}^N_{\min}\right) - f_* \geq \varepsilon\Big\} \leq \sigma.$$

Таким образом, $\bar{x}^N_{\min}$ является (ε, σ)-решением (1). Здесь важным моментом является возможность вычисления значения функции $f(x)$.

Теперь предложим метод, не основанный на предположении о наличии значения функции $f(x)$. Это предположение может быть неверным [1] во многих реальных задачах стохастического программирования, например в методе максимального правдоподобия, используемом в математической статистике.

Теорема. *Пусть любое из трех предположений* а)–в) *предложения верно. Пусть* $K = \left\lceil 2\ln\left(\sigma^{-1}\right)\right\rceil$ *и* $\left\{\bar{x}^{N,i}\right\}_{i=1}^K$ *— независимые точки, полученные с использованием СЗС* (3), (4) *при* $N = \left\lceil \frac{4CM^2\bar{R}^2}{\varepsilon^2} \right\rceil$. *Тогда точка* $\bar{x}^K = \frac{1}{K}\sum_{i=1}^K \bar{x}^{N,i}$ *является* (ε, σ)*-решением* (1).

Доказательство. Используя следствие при $c = \varepsilon$, мы получим для всех $i = 1, \ldots, K$:

$$\mathbb{P}\Big\{f\big(\bar{x}^{N,i}\big) - f_* \geq \varepsilon\Big\} \leq \mathbb{P}\big\{\eta_i \geq \varepsilon\big\}, \quad \eta_i \in \mathrm{N}\left(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{4}\right). \tag{8}$$

Вследствие выпуклости $f(x)$, поскольку $\left\{\bar{x}^{N,i}\right\}_{i=1}^K$ независимы и одинаково распределены и $\{\eta_i\}_{i=1}^K$ независимы и одинаково распределены, мы имеем

$$\begin{aligned}\mathbb{P}\Big\{f\big(\bar{x}^K\big) - f_* \geq \varepsilon\Big\} &\leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=i}^K \Big(f\big(\bar{x}^{N,i}\big) - f_*\Big) \geq \varepsilon\right\} \overset{(8)}{\leq} \mathbb{P}\left\{\frac{1}{K}\sum_{i=1}^K \eta_i \geq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\right\} \\ &= \mathbb{P}\left\{\left(\frac{1}{K}\sum_{i=1}^K \eta_i\right) - \frac{\varepsilon}{2} \geq \frac{\varepsilon}{2}\right\} = \mathbb{P}\Big\{\eta - \frac{\varepsilon}{2} \geq \frac{\varepsilon}{2}\sqrt{K}\Big\} \leq \sigma,\end{aligned}$$

где $\eta \in \mathrm{N}\left(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{4}\right)$. Здесь мы использовали хорошо известные свойства суммы независимых нормальных случайных переменных [4]:

$$\mathrm{N}\left(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{4}\right) + \cdots + \mathrm{N}\left(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{4}\right) \overset{d}{=} \mathrm{N}\left(\frac{K\varepsilon}{2}, \frac{K\varepsilon^2}{4}\right), \qquad \frac{1}{K}\,\mathrm{N}\left(\frac{K\varepsilon}{2}, \frac{K\varepsilon^2}{4}\right) \overset{d}{=} \mathrm{N}\left(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{4K}\right).$$

Здесь $K = \left\lceil 2\ln\left(\sigma^{-1}\right)\right\rceil$ и для $\eta \in \mathrm{N}\left(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{4}\right)$, $\mathbb{P}\left\{\eta - \frac{\varepsilon}{2} \geq \frac{\varepsilon}{2}\sqrt{2\ln\left(\sigma^{-1}\right)}\right\} \leq \sigma$. Таким образом, точка $\bar{x}^K$ является (ε, σ)-решением (1). □

Подведем итог: при некоторых слабых предположениях, но без вычисления значения целевой функции, мы предлагаем способ получения (ε, σ)-решения с использованием ряда ε-решений, вычисленных параллельно. Этот подход позволяет сократить время вычисления без потери качества решения. В то же самое время мы получили ответ на вопрос Ю. Нестерова [7]. Этот вопрос можно сформулировать следующим образом: когда качество решения задачи (1), полученное одним мудрым старцем, который думал $\Theta\left(M^2\bar{R}^2 \ln\left(\sigma^{-1}\right)/\varepsilon^2\right)$ дней, такое же, что и полученное $\Theta\left(\ln\left(\sigma^{-1}\right)\right)$ экспертами, каждый из которых думал $\Theta\left(M^2\bar{R}^2/\varepsilon^2\right)$ дней? Наш ответ состоит в том, что качество одинаково при любом из трех предположений а)–в) предложения.

Литература

1. **Shapiro A., Dentcheva D., and Ruszczynski A.** Lectures on Stochastic Programming: Modeling and Theory. — MPS-SIAM Series on Optimization, 2014.
2. **Nemirovski A., Juditsky A., Lan, G., and Shapiro A.** Robust stochastic approximation approach to stochastic programming // SIAM J. Optim. — 2009. — Vol. 19. — P. 1574–1609.
3. **Ben-Tal A., Nemirovski A.** Lectures on Modern Convex Optimization.— 2013. — http://www2.isye.gatech.edu/ nemirovs/Lect_ModConvOpt.pdf.
4. **Durrett R.** Probability: Theory and Examples. — Cambridge University Press, 2010.
5. **Guigues V., Juditsky A., and Nemirovski A.** Non-asymptotic confidence bounds for the optimal value of a stochastic program // Optimization Methods and Software. — 2017. — Vol. 32 (5). — P. 1033–1058.
6. **Гасников А.В.** Эффективные численные методы поиска равновесий в больших транспортных сетях: Дис. ... доктора физ.-мат.наук: 05.13.18. — Москва, 2016. — Перевод: Gasnikov A.V. Searching Equilibriums in Large Transport networks. — 2016. — (Preprint / Cornell University Library; arXiv:1607.03142). — (Doctoral Thesis).
7. **Nesterov Yu., Vial J.-Ph.** Confidence level solutions for stochastic programming // Automatica. — 2008. — Vol. 44, № 6. — P. 1559–1568.

Поступила в редакцию 24 января 2017 г.,
в окончательном варианте 7 июля 2017 г.

Литература в транслитерации

1. **Shapiro A., Dentcheva D., and Ruszczynski A.** Lectures on Stochastic Programming: Modeling and Theory. — MPS-SIAM Series on Optimization, 2014.
2. **Nemirovski A., Juditsky A., Lan, G., and Shapiro A.** Robust stochastic approximation approach to stochastic programming // SIAM J. Optim. — 2009. — Vol. 19. — P. 1574–1609.
3. **Ben-Tal A., Nemirovski A.** Lectures on Modern Convex Optimization.— 2013. — http://www2.isye.gatech.edu/ nemirovs/Lect_ModConvOpt.pdf.
4. **Durrett R.** Probability: Theory and Examples. — Cambridge University Press, 2010.
5. **Guigues V., Juditsky A., and Nemirovski A.** Non-asymptotic confidence bounds for the optimal value of a stochastic program // Optimization Methods and Software. — 2017. — Vol. 32 (5). — P. 1033–1058.
6. **Gasnikov A.V.** Effektivnye chislennye metody poiska ravnovesiy v bol'shih transportnyh setyah: Dis. ... doktora fiz.-mat.nauk: 05.13.18. — Moskva, 2016. — Perevod: Gasnikov A.V. Searching

Equilibriums in Large Transport networks. — 2016. — (Preprint / Cornell University Library; arXiv:1607.03142). — (Doctoral Thesis).

7. **Nesterov Yu., Vial J.-Ph.** Confidence level solutions for stochastic programming // Automatica. — 2008. — Vol. 44, № 6. — P. 1559–1568.