Градиентные и прямые методы с неточным оракулом для задач стохастической онлайн оптимизации. Промежуточный метод

Гасников А.В. (МФТИ, ВШЭ, ИППИ),

Двуреченский П.Е. (МФТИ, ИППИ)

Аннотация

В работе обзорно приводятся накопившиеся у авторов новые результаты по градиентным и прямым методам стохастической выпуклой оптимизации с неточным оракулом. Стоит отметить, что приведенные в статье оценки содержат все основные известные авторам результаты в этой области с одной стороны и демонстрируют эти результаты максимально компактно с другой. Кроме того, приведенные в статье оценки непрерывны. То есть основная их отличительная черта это овыпукление ранее известных оценок. В статье также упоминается приложение полученных результатов к онлайн постановке.

Рассматривается задача стохастической выпуклой оптимизации в гильбертовом пространстве

$$f(x) := E_\xi\left[f(x,\xi)\right] \to \min_{x\in Q}. \tag{1}$$

Норму (евклидову), порожденную скалярным произведение, будем обозначать 2-нормой. Относительно множества Q предполагается, что оно может быть вложено в шар (в 2-норме) конечного радиуса в этом пространстве. Функция $f(x)$ предполагается μ_2-сильно выпуклой в 2-норме. Далее будем считать, что в гильбертовом пространстве задана такая норма $\|\ \|$, что единичный шар в этой норме содержится внутри единичного шара в 2-норме. Считаем также, что задана прокс-структура относительно этой нормы [1]. Прокс-диаметр множества Q считаем равным R. Мы будем добавлять нижний индекс 2, если проск-структура предполагается евклидовой.

Предположение 1 (см. [2], [3]). *(δ, L, D)-оракул выдает (на запрос, в котором указывается только одна точка x) такие $\left(F(x,\xi), G(x,\xi)\right)$ (с.в. ξ независимо разыгрывается из одного и того же распределения, фигурирующего в постановке (1)), что для всех $x\in Q$ ограничена дисперсия*

$$E_\xi\left[\left\|G(x,\xi)-E_\xi\left[G(x,\xi)\right]\right\|_*^2\right] \le D,$$

и для любых $x, y\in Q$

$$0 \le E_\xi\left[f(y,\xi)\right] - E_\xi\left[F(x,\xi)\right] - \left\langle E_\xi\left[G(x,\xi)\right], y-x\right\rangle \le \frac{L}{2}\|y-x\|^2 + \delta.$$

Имея в распоряжении такого (δ, L, D)-оракула, требуется предложить оптимальный метод. По определению это метод, для которого для данного класса задач в соотношении

$$E\left[f(x_N)\right] - \min_{x\in Q} f(x) \le \varepsilon,$$

$N(\varepsilon)$ – минимально (равномерно по малым ε).

Теорема 1. *Существуют два однопараметрических семейства методов (параметр $p \in [0,1]$), которые дают оценки на требуемое число обращений к (δ, L, D)-оракулу*

$$N_1(\varepsilon) = \max\left\{\mathrm{O}\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p+1}}, \mathrm{O}\left(\frac{DR^2}{\varepsilon^2}\right)\right\}, \delta \le \mathrm{O}\left(\varepsilon\cdot\left(\frac{\varepsilon}{LR^2}\right)^{\frac{p}{p+1}}\right);$$

$$N_2(\varepsilon) = \max\left\{\mathrm{O}\left(\left(\frac{L_2}{\mu_2}\right)^{\frac{1}{p+1}} \ln\left(\frac{\mu_2 R_2^2}{\varepsilon}\right)\right), \mathrm{O}\left(\frac{D_2}{\mu_2\varepsilon}\right)\right\}, \ \delta \le \mathrm{O}\left(\varepsilon\cdot\left(\frac{\mu_2}{L_2}\right)^{\frac{p}{p+1}}\right).$$

При $\delta = 0$ в не много другой форме эти оценки были достаточно давно известны (см., например, [1]). Оценка $N_1(\varepsilon)$ в детерминированном случае ($D = 0$) была получена в работе [4]. В стохастическом случае при $p = 1$ в работе [2]. Оценка $N_2(\varepsilon)$ в детерминированном случае при $p = 1$ была получена в работе [5].

Замечание 1. Здесь и далее вместо $\mathrm{O}(\ )$ можно писать константы ~ 10, однако эти константы улучшаемы, как именно мы не знаем, поэтому ограничимся $\mathrm{O}(\ )$.

Замечание 2. Выписанные оценки достигаются и не улучшаемы (с точностью до логарифмических факторов) [2] – [6]. Здесь и далее мы считаем, что $N(\varepsilon)$ меньше размерности гильбертова пространства, в котором происходит оптимизация (в случае если это конечномерное евклидово пространство).

Замечание 3. Следуя [7], заметим, что за счет допускаемой неточности оракула, можно погрузить задачу с гельдеровым градиентом ($\nu \in [0,1]$)

$$\left\|\nabla f(x) - \nabla f(y)\right\|_* \le L_\nu \|x-y\|^\nu$$

(в том числе и негладкую задачу с ограниченной нормой разности субградиентов при $\nu = 0$) в класс гладких задач с оракулом, характеризующимся точностью δ и

$$L = L_\nu \left[\frac{L_\nu (1-\nu)}{2\delta(1+\nu)} \right]^{\frac{1-\nu}{1+\nu}}.$$

Заметим, в этой связи, что если в предположении 1 считать

$$E_\xi\left[f(y,\xi)\right] - E_\xi\left[F(x,\xi)\right] - \left\langle E_\xi\left[G(x,\xi)\right], y-x\right\rangle \le \frac{L}{2}\|y-x\|^2 + M\|y-x\| + \delta,$$

то вместо D в теореме 1 стоит писать $M^2 + D$.

Следствие 1. *Для детерминированной постановки задачи (1) существуют два однопараметрических семейства методов (параметр $p \in [0,1]$), которые дают оценки на требуемое число итераций*

$$N_1(\varepsilon) = \mathrm{O}\left(\inf_{\nu\in[0,1]} \left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+2p\nu+\nu}} \right), \ \delta \le \mathrm{O}\left(\varepsilon / N_1(\varepsilon)^p\right);$$

$$N_2(\varepsilon) = \mathrm{O}\left(\inf_{\nu\in[0,1]} \frac{1}{\varepsilon^{1-\nu}} \ln\left(\frac{\mu_2 R_2^{1+\nu}}{\varepsilon} \right) \left(\frac{L_{\nu,2}^{\frac{2}{1+\nu}}}{\mu_2} \right)^{\frac{1+\nu}{1+2p\nu+\nu}} \right), \ \delta \le \mathrm{O}\left(\varepsilon / N_2(\varepsilon)^p\right).$$

В случае $p=1$ и $\delta = 0$ первая оценка была получена в работе [7]. Она не улучшаемая [1], [7], [8]. В остальном эти оценки являются новыми и также неулучшаемыми. Важно отметить, что эти методы могут ничего априорно не знать о свойствах гладкости задачи (то есть не знать L_ν, $\nu \in [0,1]$). Они сами оптимально настраиваются на разных участках итерационного процесса на соответствующую этим участкам гладкость функционала. Если отказаться от возможности самонастраивания, то следствие 1 можно обобщить.

Следствие 2. *Существуют два однопараметрических семейства методов (параметр $p \in [0,1]$), которые дают оценки на требуемое число обращений к (δ, L, D)-оракулу*

$$N_1(\varepsilon) = \max\left\{ \underbrace{\mathrm{O}\left(\left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+2p\nu+\nu}} \right)}_{\bar{N}_1(\varepsilon)}, \mathrm{O}\left(\frac{DR^2}{\varepsilon^2} \right) \right\}, \ \delta \le \mathrm{O}\left(\varepsilon / \bar{N}_1(\varepsilon)^p\right);$$

$$N_2(\varepsilon) = \max\left\{ \underbrace{\mathrm{O}\left(\frac{1}{\varepsilon^{1-\nu}} \ln\left(\frac{L_{\nu,2} R_2^{1+\nu}}{\mu_2^{1-\nu}\varepsilon} \right) \left(\frac{L_{\nu,2}^{\frac{2}{1+\nu}}}{\mu_2} \right)^{\frac{1+\nu}{1+2p\nu+\nu}} \right)}_{\bar{N}_2(\varepsilon)}, \mathrm{O}\left(\frac{D_2}{\mu_2 \varepsilon} \right) \right\}, \ \delta \le \mathrm{O}\left(\varepsilon / \bar{N}_2(\varepsilon)^p \right).$$

Приведенные в этом следствие оценки также нелучшаемые.

Предположим теперь, что у нас оракул может выдавать только реализацию значения функции при этом с шумом не только случайной природы. Далее в этом пункте будем считать, что гильбертово пространство конечномерное $\mathbb{R}^n$.

Предположение 2. *δ-оракул выдает (на запрос, в котором указывается только одна точка x) $f(x,\xi)+\delta(x,\xi)$, где с.в. ξ независимо разыгрывается из одного и того же распределения, фигурирующего в постановке (1), случайная величина $\delta(x,\xi)=\tilde{\delta}(x)+\bar{\delta}(\xi)$, где $\bar{\delta}(\xi)$ – независимая от x случайная величина (случайность которой может быть обусловлена не только зависимостью от ξ), ограниченная по модулю δ, $\tilde{\delta}(x)/(R\delta)$ – 1-липшицева функция в норме $\|\ \|$.*

Считаем

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|, \ E_\xi\left[\left\| \nabla f(x,\xi) - E_\xi\left[\nabla f(x,\xi)\right] \right\|_*^2 \right] \le D.$$

Для того чтобы понять соответствие между требованиями к уровню шума в предположениях 1, 2, полезно привести в простейшем случае аналог стохастического градиента, который используется в прямых методах

$$g_{\tau,\delta}(x,s,\xi) = \frac{n}{\tau}\left(f(x+\tau s,\xi)+\delta(x+\tau s,\xi) - \left(f(x,\xi)+\delta(x,\xi) \right) \right) s,$$

где s – случайный вектор (независимый от ξ), равномерно распределенный на единичной сфере в 2-норме в пространстве $\mathbb{R}^n$. Из этого представления сразу видно, что липшицева составляющая шума δ_2 из предположения 2 и уровень шума δ_1 из предположения 1 связаны соотношением $\delta_2 \sim \delta_1 / n$.

Далее предполагается, что на каждом шаге можно обращаться только к δ-оракулу, причем не более $2 \le k \le 2n$ раз с одной реализацией ξ (на одной итерации).

Теорема 2. *Существуют два однопараметрических семейства методов (параметр $p \in [0,1]$), которые дают оценки на требуемое число обращений к δ-оракул*

$$N_1(\varepsilon) = n \max\left\{ O\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p+1}}, O\left(\frac{DR^2}{\varepsilon^2}\right)\right\}, \delta \le \frac{1}{n} O\left(\varepsilon \cdot \left(\frac{\varepsilon}{LR^2}\right)^{\frac{p}{p+1}}\right);$$

$$N_2(\varepsilon) = n \max\left\{ O\left(\left(\frac{L_2}{\mu_2}\right)^{\frac{1}{p+1}} \ln\left(\frac{L_2 R_2^2}{\varepsilon}\right)\right), O\left(\frac{D_2}{\mu_2 \varepsilon}\right)\right\}, \ \delta \le \frac{1}{n} O\left(\varepsilon \cdot \left(\frac{\mu_2}{L_2}\right)^{\frac{p}{p+1}}\right).$$

При этом число итераций будет в k раз меньше.

Эти оценки в детерминированном случае при $p=1$ и с $\delta \equiv 0$ были получены в работе [9]. В стохастическом случае при $p=1$ и с $\delta = 0$ похожие оценки были приведены в [10]. Выписанные оценки не улучшаемые. Однако здесь имеются лишь частичные результаты [10].

Следствие 1, 2 естественным образом переносятся и на теорему 2. Мы опускаем здесь соответствующие переформулировки.

Отметим, что для всех описанных методов (за исключением детерминированного сильно выпуклого случая) можно получить аналоги соответствующих результатов для онлайн постановок задач (в стохастическом сильно выпуклом случае в онлайн контексте приобретается дополнительный логарифмический фактор). При этом в таких постановках регрет формируется, вообще говоря, уже не с одинаковыми весами, а весами специального вида, например,

$$\frac{1+p}{N^{1+p}}\left[\sum_{k=1}^{N} k^p f_k(x_k) - \min_{x \in Q} \sum_{k=1}^{N} k^p f_k(x)\right].$$

Некоторые постановки (например, содержащиеся в результатах следствия 1 и теоремы 2) требуют обращение на одной итерации (шаге) к оракулу несколько раз. Причем в следствии 1 требуется обращаться как за значением соответствующей функции, так и за ее градиентом. Не много проясняет сказанное выше работа [11].

Литература

1. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2013. http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf
2. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. CORE UCL, PhD thesis, March 2013. http://www.ecore.be/DPs/dp_1327057920.pdf
3. *Devolder O., Glineur F., Nesterov Yu.* First order methods of smooth convex optimization with inexact oracle // Math. Progr. Ser. A. Accepted. 2013.
4. *Devolder O., Glineur F., Nesterov Yu.* Intermediate gradient methods for smooth convex problems with inexact oracle // CORE Discussion Paper 2013/17. 2013.

5. *Devolder O., Glineur F., Nesterov Yu.* First order methods with inexact oracle: the smooth strongly convex case. CORE Discussion Paper 2013/16. 2013.
6. *Agarwal A., Bartlett P.L., Ravikumar P., Wainwright M.J.* Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization // e-print, 2011. arXiv:1009.0571
7. *Nesterov Yu.* Universal gradient methods for convex optimization problems. CORE Discussion Paper 2013/63. 2013.
8. *Guzman C., Nemirovski A.* On lower complexity bounds for large-scale smooth convex optimization // Journal of Complexity. 2015. arXiv:1307.5001
9. *Nesterov Yu.* Random gradient-free minimization of convex functions // CORE Discussion Paper 2011/1. 2011.
10. *Duchi J.C., Jordan M.I., Wainwright M.J., Wibisono A.* Optimal rates for zero-order convex optimization: the power of two function evaluations // e-print, 2014. arXiv:1312.2139
11. *Shi Z., Liu R.* Online universal gradient method // e-print, 2013. arXiv:1311.3832v2