Сходимость с оценкой вероятностей больших отклонений для задач выпуклой оптимизации и седловых задач в условиях повышенной гладкости

Рубцов Д.Н.
rubtsov.dn@phystech.edu

Abstract

Классические результаты стохастической оптимизации, как правило, формулируются в терминах числа итераций, необходимых для достижения ε-точности по математическому ожиданию функции. В данной работе разрабатывается алгоритм, обеспечивающий гарантию сходимости с высокой вероятностью, причем предположения о "легкости хвостов"распределения шума стохастического градиента здесь не делаются, то есть Минимизируемая функция здесь предполагается обладающей повышеннной гладкостью.

Ключевые слова : выпуклая оптимизация, седловые задачи, стохастическая оптимизация, тяжелые хвосты, повышенная гладкость, безградиентная оптимизация,

1 Введение

В данной работе рассматривается задача стохастической оптимизации

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}f(x,\xi), \tag{1}$$

где случайная величина ξ из фиксированного, но неизвестного распределения $\mathcal{P}$: $\xi \sim \mathcal{P}$. Как правило, результатом стохастических градиентных методов является точка x_ε такая, что

$$\mathbb{E}f(x_\varepsilon) - \min f \leq \varepsilon. \tag{2}$$

Стоимость таких алгоритмов, например, SGD в терминах количества итераций $\mathcal{O}(\frac{1}{\varepsilon^2})$ в выпуклом случае и $\mathcal{O}(\frac{1}{\varepsilon})$ в сильно выпуклом случае.

В данной работе мы рассматриваем алгоритмы, результатом которых являются точки $x_{\varepsilon,p}$, удовлетворяющие условию

$$\mathbb{P}(f(x_{\varepsilon,p}) - \min f \leq \varepsilon) \geq 1-p, \tag{3}$$

где «уровень уверенности» $p > 0$ модет быть достаточно маленьким. Из неравенства Маркова ясно, что (2) можно гарантировать, если найти точку $x_{\varepsilon,p}$ такую, что $\mathbb{E}f(x_\varepsilon) - \min f \leq p\varepsilon$. Однако для этого необходимо $\mathcal{O}(\frac{1}{p^2\varepsilon^2})$ или $\mathcal{O}(\frac{1}{p\varepsilon})$ итераций, то есть сложность существенно возрастает при малых p. Существо несколько статей, в которых сложность относительно p снижается до логарифмической $\log(\frac{1}{p})$, однако либо в то же время ухудшается сложность относительно ε, либо делаются более жесткие ограничения на шум стохастического градиента: он предполагатся субгауссовским, то есть имеющим "легкие хвосты".

В работе Davis et al. [2021] был разработан общий алгоритм, работающий и в случае "тяжелых хвостов"распределения шума стохастического градиента, при этом требующий не очень большого числа итераций (вызовов оракула). В этой работе рассматривается оракул $\mathcal{M}(f,\varepsilon)$, возвращающий точку x_ε такую, что $\mathbb{P}(f(x_\varepsilon) - \min f \leq \varepsilon) \geq \frac{2}{3}$. В частности, такой оракул может быть порожден любым алгоритмом стохастической оптимизации, возвращающим точку x_ε такую, что $\mathbb{E}f(x_\varepsilon) - \min f \leq \frac{\varepsilon}{3}$

(следствие неравенства Маркова). Стоимость вызова такого оракула обозначим за $\mathcal{C}_{\mathcal{M}}(f, \varepsilon)$. Авторы показали, что для μ-сильно выпуклых L-гладких функций алгоритм, решающий задачу 2 требует $\log(\frac{\log \kappa}{p}) \log \kappa \cdot \mathcal{C}_{\mathcal{M}}(f, \frac{\varepsilon}{\log \kappa})$. Таким образом, задача сходимости с высокой вероятностью сложнее (в смысле оракульной сложности) задачи сходимости по матожиданию лишь в логарфимическое по $\frac{1}{p}$ и полилогарифмическое по числу обусловленности $\kappa := \frac{L}{\mu}$ раз.

Основываясь на техниках, предложенных в статье Davis et al. [2021], мы разрабатываем алгоритм для μ-сильно выпуклых β-гёльдеревых функций, учитывающей повышенную гладкость минимизируемых функций и тем самым, уменьшая полную стоимость алгоритма.

В последней части работы мы решаем седловые задачи

$$\min_{x \in X} \max_{x \in Y} \Phi(x, y) := \mathbb{E}\Phi_{\xi}(x, y), \tag{4}$$

являющиеся актуальными в связи с развитием обучения с подкреплением (reinforcement learning). Разрабатываются алгоритмы поиска приближенного решения с высокой вероятностью в условиях повышенной гладкости.

2 Техники и алгоритмы

Пусть $\mathbb{R}^d$ - евклидаово пространство со скалярным произведением $\langle \cdot, \cdot \rangle$ и индуцированным им нормой $||x|| = \langle x, x \rangle, x \in \mathbb{R}^d$. Замкнутый шар с центром в точке x и радиусом ε будем обозначать $B_{\varepsilon}(x)$. Пусть исследуемая функция $f : \mathbb{R}^d \to \mathbb{R}$ μ-сильно выпуклая (т.е. $f(x) - \frac{\mu}{2}||x||^2$ - выпуклая) и L-гладкая (т.е. дифференцируемая с L-липшицевым градиентом). Для такой функции для всех точек $x, y \in \mathbb{R}^d$ справедливо:

$$\langle \nabla f(x), y - x \rangle + \frac{\mu}{2}||y - x||^2 \leq f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}||y - x||^2.$$

Для точки x^*, в которой достигается минимум функции f тогда справедливо (с учетом необходимого условия $\nabla f(x^*) = 0$):

$$\frac{\mu}{2}||x - x^*||^2 \leq f(x) - f(x^*) \leq \frac{L}{2}||x - x^*||^2$$

Далее $\min f = f(x^*) =: f^*$.

2.1 Robust distance estimation

Обозначим за $\mathcal{D}(\varepsilon)$ - оракул, возвращающий точку $\mathbb{P}[||x - x^*|| \leq \varepsilon] \geq \frac{2}{3}$. Можно сделать m вызовов этого оракула $x_1, ..., x_m$ и выбрать среди полученных точек такую x_{i^*}, вокруг которой класстеризуются остальные точки.

Algorithm 1 Robust Distance Estimation (RDE) $\mathcal{D}(\varepsilon, m)$

Input: оракул $\mathcal{D}(\varepsilon)$ и число его вызовов m

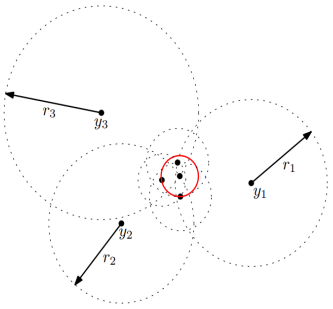
0: for $i \in 1, ..., m$ do

0: $\quad r_i = \min\{r \geq 0 : |B_r(x_i) \cap X| > \frac{m}{2}\} \leftarrow Compute$

0: end for

0: $i^* = \arg\min_{i \in 1,...,m} r_i \leftarrow Set$ =0

Output: x_{i^*}



Theorem 1 Точка x_{i^*}, возвращаемая алгоритмом RDE удовлетворяет условию

$$\mathbb{P}(||x_{i^*} - x^*|| \leq 3\varepsilon) \geq 1 - e^{-\frac{m}{18}}$$

Пусть точки x_i $(i = 1, ..., m)$ таковы, что $\mathbb{P}(f(x_i) - f^* \leq \varepsilon) \geq \frac{2}{3}$. Из μ-сильной выпуклости следует, что $\mathbb{P}(||x_i - x^*|| < \sqrt{\frac{2\varepsilon}{\mu}} =: \delta) \geq \frac{2}{3}$. Применив к этим точкам алгоритм RDE, получим точку x_{i^*}, удовлетворяющую неравенству $\mathbb{P}(||x_{i^*} - x^*|| < 3\delta) \geq 1 - e^{-\frac{m}{18}}$. Из L-гладкости функции f тогда следует, что $\mathbb{P}(f(x_{i^*}) - f^* \leq \frac{L}{2}(3\delta)^2 = 9\frac{L}{\mu}\varepsilon) \geq 1 - e^{-\frac{m}{18}}$. Таким образом, генерируя точки алгоритмом, дающим гарантии сходимости с точностью ε по матожиданию, но не с высокой вероятностью, мы предъявили алгоритм, дающий гарантию сходимости с высокой вероятностью, но лишь с $\kappa\varepsilon$-точностью, где число обусловленности $\kappa = \frac{L}{\mu} \gg 1$ может быть достаточно большим. Для нивелирования этой проблемы в статье Davis et al. [2021] был предложена процедура *proxBoost*.

2.2 proxBoost

Зафиксируем возрастающую последовательность $\lambda_0, ..., \lambda_T$ и последовательность точек $x_0, ..., x_T$. Для каждого $i = 0, ..., T$ введем функцию

$$f^i(x) := f(x) + \frac{\lambda_i}{2}||x - x_i||^2$$

$$\bar{x}_{i+1} := \arg\min_x f^i(x)$$

В качестве x_i можно брать $x_i = \bar{x}_i$ для $i \geq 1$. Так как точное вычисление точки минимума чаще всего невозможно, будем следить лишь за $||\bar{x}_i - x_i||$. Для простоты, $\bar{x}_0 := \arg\min f$, $\lambda_{-1} := 0$.

Theorem 2 (Inexact proximal point method) Для всех $j \geq 0$ выполняется следующее неравенство:

$$f^j(\bar{x}_{j+1}) - f^* \leq \sum_{i=0}^{j} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2.$$

Слндовательно, имеем декомпозицию функциональной ошибки:

$$f(x_{j+1}) - f^* \leq (f^j(x_{j+1}) - f^j(\bar{x}_{j+1})) + \sum_{i=0}^{j} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2.$$

Если функция f еще и L-гладкая, то для всех $j \geq 0$ выполнена оценка:

$$f(x_j) - f^* \leq \frac{L + \lambda_{j-1}}{2}||\bar{x}_j - x_j||^2 + \sum_{i=0}^{j-1} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2.$$

Основным результатом Теоремы 2 является декомпозиция функциональной ошибки на ошибку на последнем шаге $(f^T(x_{j+1}) - f^T(\bar{x}_{j+1}))$ и накопленную ошибку $\sum_{i=0}^{T} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2$. Для достаточно больших T можно гарантировать то, что функция f^T хорошо обусловлена. Использование результатов теорем 1 и 2 позволило авторам Davis et al. [2021] разработать алгоритм *proxBoost*.

Алгоритм *proxBoost* состоит из 3 шагов. На первом шаге ищется точка, довольно близкая к точке минимума функции f с большой вероятностью. Эта задача может быть решена с помощью техники RDE. На втором шаге в цикле точно также можно решить аналогичные задачи для функциий f^j. На последнем шаге

3 Вычислительные эксперименты

Возьмем функцию $f(x, \xi) = \frac{Lx_1^2}{2} + \frac{\mu x_2^2}{2} + \langle \xi, x \rangle$, где $x = (x_1, x_2) \in R^2, L \geq \mu$. Эта функция является зашумленной версией функции $f(x) := \mathbf{E}_{\xi \sim \mathcal{P}}[f(x, \xi)] = \frac{Lx_1^2}{2} + \frac{\mu x_2^2}{2}$. Стохастический градиент $\nabla f(x, \xi) = [L, \mu]^T + \xi$, где $\mathbf{E}\xi = 0, \mathbf{D}\xi < \sigma^2$. Для демонстрации работоспособности алгоритма не только в случае легких хвостов распределеня, но и в случае тяжелых хвостов, случайную величину будем брать из трех

Algorithm 2 $proxBoost(\delta, p, T)$

Input: $\delta \geq 0$, $p \in (0,1)$, $T \in \mathbb{N}$

Set $\lambda_{-1} = 0$, $\varepsilon_{-1} = \sqrt{\frac{2\delta}{\mu}}$

Найти точку x_0 такую, что $||x_0 - \bar{x}_0|| \leq \varepsilon_{-1}$ с вероятностью $1-p$

0: for $j = 0, ..., T-1$ do

0: Set $\varepsilon_j = \sqrt{\frac{2\delta}{\mu+\lambda_j}}$

0: Найти точку x_{j+1} такую, что $\mathbb{P}(||x_{j+1} - \bar{x}_{j+1}|| \leq \varepsilon_j | E_j) \geq 1-p$, где событие $E_j := \{x_i \in B_{\varepsilon_{i-1}}(\bar{x}_i)\ \forall i \in [0, j]\}$

0: end for

0: Найти точку x_{T+1} такую, что $\mathbb{P}(f^T(x_{T+1}) - \min f^T \leq \delta | E_j) \geq 1-p$ =0

Output: x_{T+1}

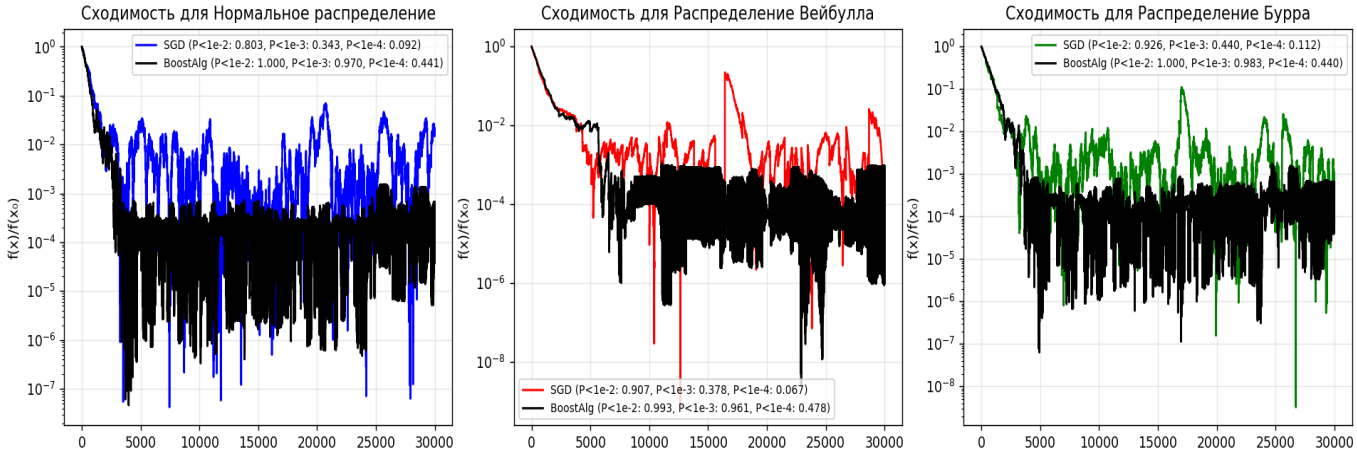
разных распределений: нормального, Вейбулла и Бурра с соответствующими функциями распределения $\mathcal{F}_{\mathcal{N}}, \mathcal{F}_W, \mathcal{F}_B$.

$$\mathcal{F}_{\mathcal{N}}(x) = \int_{-\infty}^{x} f_{\mathcal{N}}(x')dx', \ f_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathcal{F}_W(x) = (1 - e^{-(\frac{x}{\alpha})^c})\mathbb{I}(x \geq 0)$$

$$\mathcal{F}_B(x) = (1 - (1+x^c)^{-d})\mathbb{I}(x > 0)$$

Далее приведены результаты для $L = 100, \mu = 1$, то есть $\kappa = \frac{L}{\mu} = 100 \gg 1$. Цветное - SGD, черное - proxboost.



Список литературы

Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. Journal of machine learning research, 22(49):1–38, 2021.