Содержание

Аннотация

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. Алгоритм обощается на случаи выпуклых и сильно выпуклых, а также гладких и негладких функций. Полученные теоретические оценки на сложность алгоритмов проверяются на вычислительных экспериментах.

Ключевые слова: выпуклая оптимизация, стохастическая оптимизация, тяжелые хвосты.

1 Введение

В данной работе рассматривается задача стохастической оптимизации

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

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

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

Такую сходимость в дальнейшем будем называть сходимостью по математическому ожиданию. Стоимость таких алгоритмов, например, стохастического градиентного спуска (Stochastic Gradient Descent, 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 число $p > 0$ может быть достаточно маленьким. Проще говоря, мы ищем такие решения для которых вероятность того, что невязка меньше желаемой точности ε достаточно большая. Формулу (9) можно переписать в другом виде:

$$\mathbb{P}(f(x_{\varepsilon,p}) - \min f \geq \varepsilon) \leq p, \tag{4}$$

Формулу (9) можно интерпретировать как сходимость «с высокой вероятностью», а формулу (4) как оценку вероятности больших отклонений, что отражено в названии дипломной работы. Из неравенства Маркова ясно, что (9) или (4) можно гарантировать, если найти точку $x_{\varepsilon,p}$ такую, что $\mathbb{E} f(x_{\varepsilon,p}) - \min f \leq p\varepsilon$. 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. Существует 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 [1] 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 (9) 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 [1], 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 Постановка задачи

Пусть $\mathbf{R}^d$ - евклидово пространство со скалярным произведением $\langle\cdot,\cdot\rangle$ и индуцированным им нормой $\|x\|_2 = \langle x, x\rangle, x \in \mathbf{R}^d$. Замкнутый шар с центром в точке x и радиусом ε будем обозначать $B_\varepsilon(x)$.

Будем решать задачу стохастической оптимизации

$$\min_{x\in\mathbf{R}^d} f(x) := \mathbb{E}f(x, \xi). \tag{5}$$

при следующих предположениях на функцию $f(x)$:

Предположение 1. *Исследуемая функция $f : \mathbf{R}^d \to \mathbf{R}$ μ-сильно выпуклая, то есть* *для всех x, y выполнено:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}||y - x||^2 \tag{6}$$

Иначе говоря, $f(x) - \frac{\mu}{2}||x||^2$ - выпуклая

Предположение 2. *Функция $f : \mathbf{R}^d \to \mathbf{R}$ - (L, γ)-гладкая, то есть $\forall x, y \in B_{R,Q}(x^*) = \{x \in Q : \|x - x^*\| \leq R, R = \|x_0 - x^*\|\}$ выполнено:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + \gamma \tag{7}$$

где $\nabla f(x) \in \partial f(x)$ - произвольный субградиент функции f в точке x.

Предположение 3. *Градиент функции $f(x)$ удовлетворяет условию Гёльдера, то есть $\forall x, y \in B_{R,Q}(x^*) = \{x \in Q : \|x - x^*\| \leq R, R = \|x_0 - x^*\|\}$ имеет место неравенство*

$$\|\nabla f(y) - \nabla f(x)\| \leq L_\nu \|y - x\|_2^\nu, \ \nu \in [0, 1], \ L_0 < \infty \tag{8}$$

Заметим, что при $\nu = 1$ предположение (3) является просто условием L_1-гладкости. При $\nu = 0$ же предположение (3) является условием L_0-липшицевости. Далее будем обозначать $L_1 = L$ и $L_0 = M$.

Предположение (2) введено для того, чтобы смотреть на гладкий и негладкий случаи единообразно. Действительно, при $\gamma = 0$ это и есть условие гладкости. Если же функция негладкая, но $M-$липшицева, то есть $\|\nabla f(y) - \nabla f(x)\| \leq M$, то неравенство всё равно будет выполняться при $L = \frac{M^2}{2\gamma}$. Доказательство этого утверждения можно найти в [2].

Лемма 1. *Если для градиента функции $f(x)$ выполнено условие Гёльдера* (8)*, то такая функция (L, γ)-гладкая (см.* (7)*) при*

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

В частности, при $\nu = 0$ $L = \frac{M^2}{2\gamma}$.

Напомню, что эта работа сосредоточена на эффективном решении задачи оптимизации со следующей мерой качества:

$$\mathbb{P}(f(x_{\varepsilon,p}) - \min f \leq \varepsilon) \geq 1 - p, \tag{9}$$

3 Техники и алгоритмы

3.1 Robust distance estimation

Пусть исследуемая функция $f : \mathbf{R}^d \to \mathbf{R}$ 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 $x, y \in \mathbf{R}^d$ 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 $\min f = f(x^*) =: f^*$.

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) $\mathcal{D}(\varepsilon,m)$

Вход: доступ к оракулу $\mathcal{D}(\varepsilon)$ и число его вызовов m.

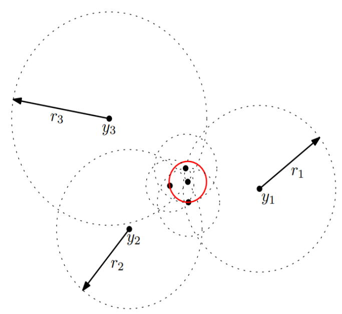
Вызываем оракул $\mathcal{D}(\varepsilon)$ m раз. Обозначим множество его ответов за $X = \{x_1, \ldots, x_m\}$.

В цикле $i = 1, \ldots, m$:

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

Set $i^* = \arg\min_{i \in [1,m]} r_i$

Возвращаем x_{i^*}



Теорема 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 $\mathbb{E}f(x_\varepsilon) - \min f \leq \frac{\varepsilon}{3}$. По неравенству Маркова тогда автоматически следует, что $\mathbb{P}(f(x_i) - f^* \leq \varepsilon) \geq \frac{2}{3}$. Из μ-сильной выпуклости получаем $\mathbb{P}(||x_i - x^*|| < \sqrt{\frac{2\varepsilon}{\mu}} =: \delta) \geq \frac{2}{3}$. 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 [1] был предложена процедура *proxBoost*.

3.2 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, $\bar{x}_0 := \arg\min f$, $\lambda_{-1} := 0$.

Теорема 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,γ)-гладкая, то для всех $j \geq 0$ выполнена оценка:

$$f(x_j) - f^* \leq \frac{L+\lambda_{j-1}}{2}||\bar{x}_j - x_j||^2 + \gamma + \sum_{i=0}^{j-1} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2. \tag{10}$$

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 [1] разработать алгоритм *proxBoost*.

Algorithm 2 *proxBoost*(δ, 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 $1-p$

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

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

  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 $E_j := \{x_i \in B_{\varepsilon_{i-1}}(\bar{x}_i)\ \forall i \in [0,j]\}$

end for

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}

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

Следующая теорема обобщает гарантии для процедуры `proxBoost`.

Теорема 3 (Проксимальный бустинг)**.** *Зафиксируем константу $\delta > 0$, вероятность отказа $p \in (0{,}1)$ и натуральное число $T \in \mathbb{N}$. Тогда с вероятностью не менее $1-(T+2)p$, точка $x_{T+1} =$ `proxBoost`(δ,p,T) удовлетворяет*

$$f(x_{T+1}) - \min f \leq \delta \left(1 + \sum_{i=0}^{T} \frac{\lambda_i}{\mu + \lambda_{i-1}}\right). \tag{11}$$

Доказательство. Сначала докажем по индукции оценку

$$\mathbb{P}[E_t] \geq 1 - (t+1)p \qquad \text{для всех } t = 0, \ldots, T. \tag{12}$$

База индукции $t = 0$ следует непосредственно из определения x_0. Теперь предположим, что (12) выполняется для некоторого индекса $t-1$. Тогда из предположения индукции и определения x_t следует

$$\mathbb{P}[E_t] = \mathbb{P}[E_t \big| E_{t-1}]\mathbb{P}[E_{t-1}] \geq (1-p)(1-tp) \geq 1-(t+1)p,$$

что завершает шаг индукции. Таким образом, неравенства (12) выполняются. Определим событие

$$F = \{f^T(x_{T+1}) - \min f^T \leq \delta\}.$$

Отсюда мы выводим

$$\mathbb{P}[F \cap E_T] = \mathbb{P}[F \mid E_T] \cdot \mathbb{P}[E_T] \geq (1-(T+1)p)(1-p) \geq 1-(T+2)p.$$

Теперь предположим, что событие $F \cap E_T$ наступает. Тогда, используя оценку (**??**), мы заключаем

$$f(x_{T+1}) - \min f \leq (f^T(x_{T+1}) - f^T(\bar{x}_{T+1})) + \sum_{i=0}^{T} \frac{\lambda_i}{2}\|\bar{x}_i - x_i\|^2 \leq \delta + \sum_{i=0}^{T} \frac{\delta\lambda_i}{\mu + \lambda_{i-1}},$$

где последнее неравенство использует определения x_{T+1} и ε_j. Это завершает доказательство. $\square$

Глядя на оценку (11), мы видим, что итоговая ошибка $f(x_{T+1}) - \min f$ контролируется суммой $\sum_{i=0}^{T} \frac{\lambda_i}{\mu+\lambda_{i-1}}$. Недолгое размышление приводит к привлекательному выбору проксимальных параметров $\lambda_i = \mu 2^i$. Действительно, в этом случае каждый член суммы $\frac{\lambda_i}{\mu+\lambda_{i-1}}$ ограничен сверху двойкой. Более того, если f является L-гладкой, то число обусловленности $\frac{L+\lambda_T}{\mu+\lambda_T}$ для функции f^T оказывается ограничено двойкой уже после $T = \lceil \log(L/\mu) \rceil$ итераций.

4 Основной алгоритм

Как правило, сложность стохастических градиентных методов, то есть количество итераций, необходимых для достижения желаемой точности $\mathbb{E}[f(x_i)] - f^* \leq \delta$ зависит от начальной невязки $f(x_0) - f^*$. Так что мы должны иметь доступ к верхней оценке этой невязки $\Delta : \Delta \geq f(x_0) - f^*$. В предложенном далее алгоритме мы будем динамически обновлять соответствующие верхние оценки.

Предположение 4. Введем вспомогательную проксимальную задачу

$$\min_y \ \varphi_x(y) := f(y) + \frac{\lambda}{2}\|y - x\|^2,$$

Пусть $\Delta > 0$ такое, что $\varphi_x(x) - \min \varphi_x \leq \Delta$. Будем обозначать $\texttt{Alg}(\delta,\lambda,\Delta, x)$ процедуру, которая возвращает точку y такую, что

$$\mathbb{P}[\varphi_x(y) - \min \varphi_x \leq \delta] \geq \frac{2}{3}.$$

Так как функция φ_x is $(\mu + \lambda)$-сильно выпукла, она имеет единственную точку минимума $\bar{y}_x$, и выполнено неравенство

$$\frac{\mu + \lambda}{2}\|y - \bar{y}_x\|^2 \leq \varphi_x(y) - \min \varphi_x.$$

Таким образом, $\texttt{Alg}(\delta,\lambda,\Delta, x)$ возвращает точку y в которой не просто значение функции близко к минимальному, но и сама точка близка к точке минимума функции $\mathbb{P}(\|y - \bar{y}_x\| \leq \varepsilon) \geq \frac{2}{3}$, где $\varepsilon = \sqrt{\frac{2\delta}{\mu+\lambda}}$. Технику Robust Distance Estimation (1) мы можем снабдить предложенным оракулом. Приведем этот алгоритм отдельно.

Algorithm 3 $\texttt{Alg-R}(\delta,\lambda,\Delta,x,m)$

Вход: функциональная точность $\delta > 0$, коэффициент $\lambda > 0$, верхняя оценка $\Delta > 0$, центральная точка $x \in \mathbf{R}^d$,

число вызовов оракула $m \in \mathbb{N}$.

Вызываем $\texttt{Alg}(\delta,\lambda,\Delta,x)$ m раз. Его ответы обозначим за $Y = \{y_1, \ldots, y_m\}$.

В цикле $j = 1,\ldots,m$:

Вычисляем $r_i = \min\{r \geq 0 : |B_r(y_i) \cap Y| > \frac{m}{2}\}$.

Возьмем $i^* = \arg\min_{i \in [1,m]} r_i$

Возвращаем y_{i^*}

Теперь объединим идеи `proxBoost` с только что предложенным робастным оценщиком расстояния `Alg-R`. Оформим это в виде отдельного алгоритма 4.

Algorithm 4 `BoostAlg`$(\delta,\Delta_{\text{in}}, x_{\text{in}},T,m)$

Вход: функциональная точность $\delta > 0$, верхняя оценка $\Delta_{\text{in}} > 0$, центральная точка $x_{\text{in}} \in \mathbf{R}^d$, и числа $m,T \in \mathbb{N}$

Установим $\lambda_{-1} = 0$, $\Delta_{-1} = \Delta_{\text{in}}$, $x_{-1} = x_{\text{in}}$

В цикле $j = 0, \ldots, T$:

$x_j =$ `Alg-R`$(\delta/9,\lambda_{j-1},\Delta_{j-1},x_{j-1},m)$

$\Delta_j = \delta\left(\frac{L+\lambda_{j-1}}{\mu+\lambda_{j-1}} + \sum_{i=0}^{j-1}\frac{\lambda_i}{\mu+\lambda_{i-1}}\right) + \gamma$

Возвращаем $x_{T+1} =$ `Alg-R`$(\frac{\mu+\lambda_T}{L+\lambda_T}\cdot\frac{\delta}{9},\lambda_T,\Delta_T,x_T,m)$

Докажем работоспособность и эффективность этого алгоритма.

Теорема 4 (Эффективность `BoostAlg`)**.** *Пусть $x_{\text{in}} \in \mathbf{R}^d$ - фиксированная стартовая точка, а Δ_{in} - некоторая верхняя оценка на невязку $\Delta_{\text{in}} \geq f(x_{\text{in}}) - \min f$. Зафиксируем числа $T,m \in \mathbb{N}$. Тогда с вероятностью не меньше $1-(T+2)\exp\left(-\frac{m}{18}\right)$ точка $x_{T+1} =$ `BoostAlg`$(\delta,\Delta_{\text{in}}, x_{\text{in}},T,m)$ удовлетворяет*

$$f(x_{T+1}) - \min f \leq \delta\left(1+\sum_{i=0}^{T}\frac{\lambda_i}{\mu+\lambda_{i-1}}\right).$$

Доказательство. Обозначим $p = \exp(-\frac{m}{18})$ и $E_j := \{x_i \in B_{\varepsilon_{i-1}}(\bar{x}_i)\ \forall i \in [0,j]\}$ Покажем, что с таким выбором p точки x_j удовлетворяют

$$\mathbb{P}[\|x_{j+1} - \bar{x}_{j+1}\| \leq \varepsilon_j | E_j] \geq 1-p \tag{13}$$

для каждого $j = 0, \ldots, T$ и x_{T+1} удовлетворяет

$$\mathbb{P}[f^T(x_{j+1}) - \min f^T \leq \delta + \gamma | E_T] \geq 1-p \tag{14}$$

Для $j = 0$ лемма RDE гарантирует, что с вероятностью не менее $1-p$ точка x_0, порожденная алгоритмом `Alg-R` удовлетворяет

$$\|x_0 - \bar{x}_0\| \leq 3\sqrt{\frac{2\cdot\delta/9}{\mu}} = \varepsilon_{-1}.$$

На шаге индукции предположим, что (13) выполняется для $x_0, \ldots, x_{j-1}$ для некоторого $j \geq 1$. Докажем для x_j. Для этого предположим, что выполнено событие E_{j-1}. Тогда, используя (10), получаем

$$f(x_{j-1})-f^* \le \frac{L+\lambda_{j-2}}{2}\|\bar{x}_{j-1}-x_{j-1}\|^2+\gamma+\sum_{i=0}^{j-2}\frac{\lambda_i}{2}\|\bar{x}_i-x_i\|^2 \le \frac{\delta(L+\lambda_{j-2})}{\mu+\lambda_{j-2}}+\gamma+\sum_{i=0}^{j-2}\frac{\delta\lambda_i}{\mu+\lambda_{i-1}} = \Delta_{j-1}.$$

Второе неравенство следует из $x_i \in B_{\varepsilon_{i-1}}(\bar{x}_i)$ с $\varepsilon_{i-1} = \sqrt{\frac{2\delta}{\mu+\lambda_{i-1}}}$ для всех $i \in [0,j-1]$. По определению f^{j-1} имеем $f^{j-1}(x_{j-1}) = f(x_{j-1})$ и $\min f^{j-1} \ge \min f = f^*$. Тогда имеем следующее неравенство:

$$f^{j-1}(x_{j-1}) - \min f^{j-1} \le f(x_{j-1}) - f^* \le \Delta_{j-1}. \qquad (15)$$

Так что Δ_{j-1} является верхней оценкой для невязки $f^{j-1}(x_{j-1}) - \min f^{j-1}$ для всех j в случае выполнения E_{j-1}. Более того, теорема (1) обеспечивает, что при выполнении события E_{j-1}, с вероятностью не менее $1-p$ выполняется следующее неравенство:

$$\|x_j - \bar{x}_j\| \le 3\sqrt{\frac{2\cdot\delta/9}{\mu+\lambda_{j-1}}} = \varepsilon_{j-1}.$$

Так что (13) выполнено для x_j, что и требовалось.

Теперь предположим, что выполнено событие E_T. Аналогично (15) получим $f^T(x_T) - \min f^T \le \Delta_T$. Теперь теорема (1) гарантирует, что с вероятностью не менее $1-p$ при выполнении события E_T имеем

$$\|x_{T+1} - \bar{x}_{T+1}\| \le 3\sqrt{\frac{2}{\mu+\lambda_T}\cdot\frac{\delta}{9}\cdot\frac{\mu+\lambda_T}{L+\lambda_T}} = \sqrt{\frac{2\delta}{L+\lambda_T}}.$$

Используя тот факт, что f^T is $(L+\lambda_T,\gamma)$-гладкая, получим

$$\mathbb{P}[f^T(x_{T+1}) - \min f^T \le \delta+\gamma \mid E_T] \ge 1-p,$$

тем самым установив (14). Осталось лишь применить теорему. □

следующая теорема собирает всё воедино.

Теорема 5 (Efficiency of `BoostAlg` with geometric decay)**.** *Зафиксируем стартовую точку* $x_{\text{in}} \in \mathbf{R}^d$*. Пусть* Δ_{in} *- некоторая верхняя оценка начальной невязки* $\Delta_{\text{in}} \ge f(x_{\text{in}}) - \min f$*. Зафиксируем желаемую функцинальную точность* $\varepsilon > 0$ *и вероятность* $p \in (0,1)$*. При* *параметрах алгоритма*

$$T = \lceil \log_2(\kappa) \rceil, \qquad m = \left\lceil 18 \ln \left(\frac{2+T}{p} \right) \right\rceil, \qquad \delta = \frac{\varepsilon}{2(2+2T)}, \qquad \gamma = \frac{\varepsilon}{2}, \qquad \lambda_i = \mu 2^i.$$

точка $x_{T+1} = \texttt{BoostAlg}(\delta, \Delta_{\mathrm{in}}, x_{\mathrm{in}}, T, m)$ *удовлетворяет*

$$\mathbb{P}(f(x_{T+1}) - \min f \leq \varepsilon) \geq 1 - p.$$

Общее число обращений к $\texttt{Alg}(\cdot)$

$$\left\lceil 18 \ln \left(\frac{\lceil 2 + \log_2(\kappa) \rceil}{p} \right) \right\rceil \lceil 2 + \log_2(\kappa) \rceil,$$

при этом максимальная начальная невязка

$$\max_{i=0,\ldots,T+1} \Delta_i \leq \frac{\kappa + 1 + 2\lceil \log_2(\kappa) \rceil}{2 + 2\lceil \log_2(\kappa) \rceil} \varepsilon + \frac{\varepsilon}{2}$$

5 Обсуждение результатов

5.1 Стохастический градиентный оракул

Предположим, что мы имеем доступ к функции f через стохастический градиентный оракул. А именно, зафиксируем вероятностное пространство $(\Omega, \mathcal{F}, \mathcal{P})$ и пусть $G\colon \mathbf{R}^d \times \Omega \to \mathbf{R}$ — измеримое отображение, удовлетворяющее

$$\mathbb{E}_z G(x,z) = \nabla f(x) \qquad \text{и} \qquad \mathbb{E}_z \| G(x,z) - \nabla f(x) \|^2 \leq \sigma^2.$$

Мы предполагаем, что для любой точки x мы можем сэмплировать $z \in \Omega$ и вычислить вектор $G(x,z)$, который служит несмещенной оценкой градиента $\nabla f(x)$. Сложность стандартных численных методов в рамках этой модели вычислений оценивается по количеству вызовов стохастического градиента $G(x,z)$ с $z \sim \mathcal{P}$, требуемых алгоритмом для получения приближенного решения задачи.

5.2 Сильно выпуклый гладкий случай

В случае, когда f является μ-сильно выпуклой и L-гладкой (дифференцируемой с L-липшицевым градиентом) стоимость $\mathcal{C}_{\mathcal{M}}(f, \varepsilon)$ обычно зависит от числа обусловленности

$\kappa := L/\mu \gg 1$, а также от начальной невязки,дисперсии градиентов и т.д. Процедуры, представленные в этой статье, выполняют оракул минимизации многократно, чтобы повысить его надежность, при этом общая стоимость составляет порядка

$$\log\left(\frac{\log(\kappa)}{p}\right)\log(\kappa)\cdot\mathcal{C}_{\mathcal{M}}\left(f,\tfrac{\varepsilon}{\log(\kappa)}\right).$$

Таким образом, гарантии высокой вероятности достигаются с небольшим увеличением стоимости, которое зависит лишь логарифмически от $1/p$ и полилогарифмически от числа обусловленности κ.

Зафиксируем начальную точку x_{in} и пусть $\Delta_{\mathrm{in}} > 0$ удовлетворяет $\Delta_{\mathrm{in}} \geq f(x_0) - f^*$. Хорошо известно, что в сильно выпуклом гладком случае стохастический градиентный метод может сгенерировать точку x, удовлетворяющую $\mathbb{E}f(x) - f^* \leq \varepsilon$ за

$$\mathcal{O}\left(\kappa\log\left(\frac{\Delta_{\mathrm{in}}}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}\right). \tag{16}$$

вызовов оракула. Ускоренные стохастические градиентные методы имеют меньшую сложность

$$\mathcal{O}\left(\sqrt{\kappa}\log\left(\frac{\Delta_{\mathrm{in}}}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}\right). \tag{17}$$

Очевидно, мы можем использовать любую из этих двух процедур в качестве `Alg`$(\cdot)$ в рамках `proxBoost`. Действительно, используя Теорему 5, мы получаем точку x, удовлетворяющую

$$\mathbb{P}[f(x) - f^* \leq \varepsilon] \geq 1 - p$$

за такую оракульную сложность:

$$\mathcal{O}\left(\ln(\kappa)\ln\left(\frac{\ln\kappa}{p}\right)\cdot\left(\kappa\ln\left(\frac{\Delta_{\mathrm{in}}\ln(\kappa)}{\varepsilon}\vee\kappa\right)+\frac{\sigma^2\ln(\kappa)}{\mu\varepsilon}\right)\right), \tag{18}$$

и

$$\mathcal{O}\left(\ln(\kappa)\ln\left(\frac{\ln\kappa}{p}\right)\cdot\left(\sqrt{\kappa}\ln\left(\frac{\Delta_{\mathrm{in}}\ln(\kappa)}{\varepsilon}\vee\kappa\right)+\frac{\sigma^2\ln(\kappa)}{\mu\varepsilon}\right)\right), \tag{19}$$

для неускоренного и ускоренного методов соответственно. Таким образом, `proxBoost` наделяет стохастический градиентный метод и его ускоренный вариант гарантиями высокой вероятности с дополнительными множителями, которые являются лишь полилогарифмическими по κ и логарифмическими по $1/p$.

5.3 Сильно выпуклый негладкий случай

Так как основную теорему мы доказывали для общего случая, не составляет труда найти итоговую сложность в сильно выпуклом негладком случае.

Для этого положим $L = \mathcal{O}\left(\frac{M^2}{\varepsilon}\right)$. Сложность сходимости по матожиданию, как известно, $\mathcal{C}_{\mathcal{M}}(f, \varepsilon) = \mathcal{O}\left(\frac{M^2+\sigma^2}{\mu\varepsilon}\right)$.

Теорема 6 (Рубцов, 2025)**.** *Итоговая сложность задачи сходимости с высокой вероятностью в случае сильно выпуклых негладких функций порядка*

$$\ln\left(\frac{\ln \frac{M^2}{\mu\varepsilon}}{p}\right) \ln^2 \frac{M^2}{\mu\varepsilon} \cdot \frac{M^2 + \sigma^2}{\mu\varepsilon} \tag{20}$$

5.4 Выпуклый гладкий случай

От сильно выпуклому случая к выпуклому можно перейти с помощью метода регуляризации.

Теорема 7 (Метод регуляризации)**.** *Пусть функция $f(x)$ выпукла. Будем решать задачу минимизации функции*

$$f^{\mu}(x) = f(x) + \frac{\mu}{2}||x - x_0||^2,$$

где $\mu \sim \frac{\varepsilon}{R^2}, R = ||x^* - x_0||$.

Пусть мы нашли точку x такую, что

$$f^{\mu}(x) - \min f^{\mu} < \frac{\varepsilon}{2}$$

Тогда

$$f(x) - \min f < \varepsilon$$

Теорема 8 (Рубцов, 2025)**.** *Итоговая сложность задачи сходимости с высокой вероятностью в случае выпуклых гладких функций порядка*

$$\mathcal{O}(\max\{\sqrt{\frac{LR_0^2}{\varepsilon}}; \frac{\sigma^2 R_0^2}{\varepsilon^2}\} \cdot \ln^2(\frac{LR_0^2}{\varepsilon}) \ln\{\frac{\ln(\frac{LR_0^2}{\varepsilon})}{\beta}\}) \tag{21}$$

6 Вычислительные эксперименты

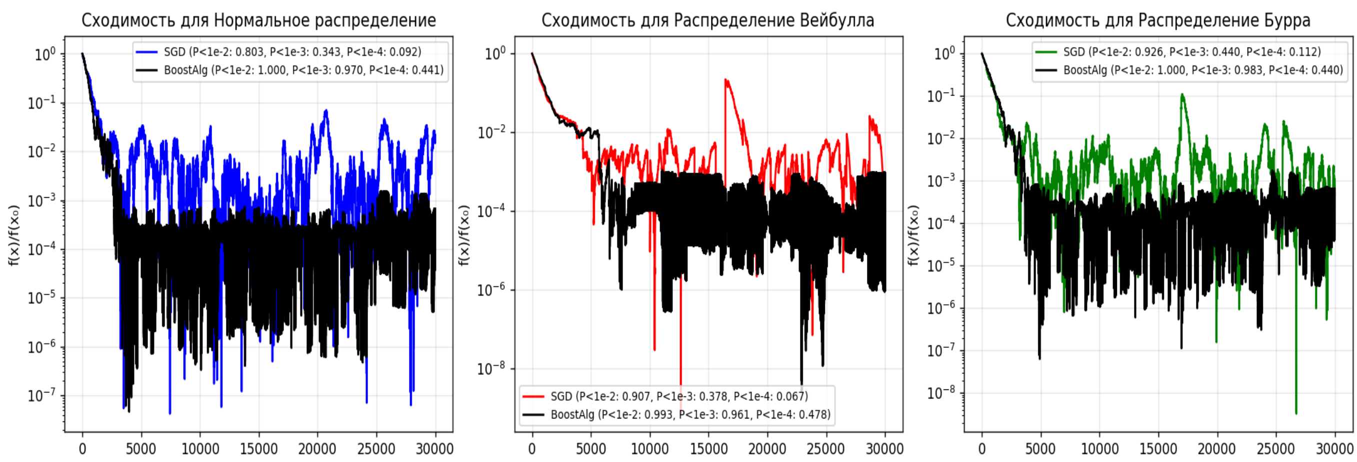
Возьмем функцию $f(x,\xi) = \frac{Lx_1^2}{2} + \frac{\mu x_2^2}{2} + \langle \xi, x \rangle$, где $x = (x_1, x_2) \in R^2, L \geq \mu$. Эта функция является зашумленной версией функции $f(x) := \mathbf{E}_{\xi \sim \mathcal{P}}[f(x,\xi)] = \frac{Lx_1^2}{2} + \frac{\mu x_2^2}{2}$. Стохастический градиент $\nabla f(x,\xi) = [L,\mu]^T + \xi$, 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 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

1. From low probability to high confidence in stochastic convex optimization / D. Davis [и др.] // Journal of machine learning research. — 2021. — Т. 22, № 49. — С. 1—38.

2. *Nesterov Y.* Universal gradient methods for convex optimization problems // Mathematical Programming. — 2015. — Т. 152, № 1. — С. 381—404.