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 обёртка над алгоритмами сходимости по математическому ожиданию, обеспечивающая гарантию сходимости с высокой вероятностью для задач выпуклой оптимизации и седловых задач, причем за эффективную сложность и для функций различной степени гладкости и выпуклости. Полученные гарантии сходимости подтверждаются на вычислительных экспериментах.

Ключевые слова: стохастическая выпуклая оптимизация, стохастические седловые задачи, сходимость с высокой вероятностью, оценка вероятности больших отклонений, проксимальный метод, неравенства концентрации.

1 Введение

В данной работе рассматривается задача стохастической оптимизации

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

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 вызовов стохастического градиентного оракула $\mathcal{O}(\frac{1}{\varepsilon^2})$ в выпуклом случае и $\mathcal{O}(\frac{1}{\varepsilon})$ в сильно выпуклом случае (см. [1], [2], [3], [4]).

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. Здесь под невязкой будет пониматься разность значений функции в точке и минимума этой функции. Формулу (3) можно переписать в другом виде:

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

Формулу (3) 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 (3) 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 $\mathcal{O}(\frac{1}{p\varepsilon})$ вызовов стохастического оракула, 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 ([5], [6], [7]), либо делаются более жесткие ограничения на

шум стохастического градиента ([8], [9], [10], [3], [11], [12]): он предполагается субгауссовским, то есть имеющим «легкие хвосты». Техника клиппирования (см.[13], [14]) хоть и позволяет работать с «тяжелыми хвостами» шума стохастического градиента, но требует исследования теоретических гарантий для каждого нового алгоритма,то есть не является общей оболочкой над алгоритмами.

В работе [15] был разработан общий алгоритм, работающий и в случае «тяжелых хвостов» распределения шума стохастического градиента, при этом требующий не очень большого числа 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). Авторы показали, что для μ-сильно выпуклых L-гладких функций алгоритм, решающий задачу (3) требует $\log(\frac{\log \kappa}{p}) \log \kappa \cdot \mathcal{C}_{\mathcal{M}}(f, \frac{\varepsilon}{\log \kappa})$ вызовов стохастического оракула, где $\mathcal{C}_{\mathcal{M}}(f, \varepsilon)$ - стоимость (сложность) вызова такого оракула. 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.

В данной дипломной работе результаты работы [15] обобщаются на негладкий и не сильно выпуклый случаи. Сложность остается логарифмической по $\frac{1}{p}$, однако ухудшается сложность относительно ε, но лишь логарифмически. Таким образом, здесь существующая обертка над алгоритмами адаптирована для более широких классов минимизируемых функций.

В последней части работы мы также решаем выпукло-вогнутые седловые задачи

$$\min_{x \in X} \max_{y \in Y} \Phi(x, y) := \mathbb{E}\Phi_\xi(x, y) \tag{5}$$

являющиеся актуальными, в частности, в связи с развитием обучения с подкреплением (reinforcement learning; см., например, [16], [17]). Так же, как и в задачах выпуклой оптимизации, в большинстве работ решаются задачи сходимости по математическому ожиданию функций ([8], [18], [19], [20], [21])

$$\mathbb{E}[\Delta_\Phi(\hat{x}, \hat{y})] \leq \varepsilon \qquad \text{или} \qquad \mathbb{E}[\Delta_\Phi^w(\hat{x}, \hat{y})] \leq \varepsilon.$$

. Здесь для любых $(\hat{x},\hat{y}) \in \mathcal{X} \times \mathcal{Y}$ введен зазор двойственности

$$\Delta_\Phi(\hat{x},\hat{y}) := \max_{y\in\mathcal{Y}} \Phi(\hat{x},y) - \min_{x\in\mathcal{X}} \Phi(x,\hat{y}) =: f(\hat{x}) - g(\hat{y}). \tag{6}$$

и его слабая версия

$$\Delta_\Phi^w(\hat{x},\hat{y}) := \Phi(\hat{x},y^*) - \Phi(x^*,\hat{y}), \tag{7}$$

где (x^*, y^*) - решение задачи (5). Целью данного исследования является поиск таких решений, что

$$\mathbb{P}\left[\Delta_\Phi(\bar{x},\bar{y}) \leq \varepsilon\right] \geq 1 - p \tag{8}$$

На базе методов, предложенных в статье [15] для задач выпуклой оптимизации в статье [22] строятся (по аналогии) методы для выпукло-вогнутых седловых задач с обеспечением гарантий сходимости высокой вероятности за небольшую сложность. Так как решение седловых задач можно рассматривать, обобщая результаты для задач минимизации ([23]), начнем с подробного рассмотрения последних.

2 Постановка задачи

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

Будем решать задачу стохастической оптимизации

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E} f(x,\xi). \tag{9}$$

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

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

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}||y - x||^2 \tag{10}$$

Предположение 2. *Функция $f : \mathbb{R}^d \to \mathbb{R}$ - (L,γ)-гладкая, то есть $\forall x, y \in B_R(x^*),\ R =$*

$\|x_0 - x^*\|\}$ *выполнено:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + \gamma \tag{11}$$

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* $\exists \nu \in [0, 1]$ *такое, что* $\forall x, y \in B_R(x^*)$ *имеет место неравенство*

$$\|\nabla f(y) - \nabla f(x)\| \leq L_\nu \|y - x\|_2^\nu, \; L_0 < \infty \tag{12}$$

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 ($\|\nabla f(y) - \nabla f(x)\| \leq M$), то неравенство (11) всё равно будет выполняться при $L = \frac{M^2}{2\gamma}$. Доказательство этого утверждения можно найти в [24].

Лемма 1. *Если для градиента функции* $f(x)$ *выполнено условие Гёльдера* (12)*, то такая функция* (L, γ)*-гладкая (см.* (11)*) 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 (3):

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

При дальнейшем изложении нам будет важно, сохраняется ли «слабая гладкость» (предположение 2) при добавлении регуляризационного слагаемого $\frac{\lambda}{2}\|x - z\|^2$, где z - некоторая фиксированная точка. На этот вопрос отвечает следующая лемма.

Лемма 2. *Пусть функция* $f(x)$ *-* (L, γ)*-гладкая. Тогда функция* $h(x) = f(x) + \frac{\lambda}{2}\|x - z\|^2$ *является* $(L + \lambda, \gamma)$*-гладкой.*

Доказательство. Действительно, так как функция $f(x)$ - (L, γ)-гладкая, то

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

Это эквивалентно

$$h(y) - \frac{\lambda}{2}\|y - z\|^2 \leq h(x) - \frac{\lambda}{2}\|x - z\|^2 + \langle \nabla h(x) - \lambda(x - z), y - x \rangle + \frac{L}{2}\|y - x\|^2 + \gamma$$

или

$$h(y) \leq h(x) + \langle \nabla h(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + \frac{\lambda}{2}\|y - z\|^2 - \frac{\lambda}{2}\|x - z\|^2 - \lambda\langle x - z, y - x \rangle + \gamma$$

В силу того, что $\frac{\lambda}{2}\|y - z\|^2 - \frac{\lambda}{2}\|x - z\|^2 - \lambda\langle x - z, y - x \rangle = \frac{\lambda}{2}\|y - x\|^2$ получаем $(L + \lambda, \gamma)$-гладкость функции $h(x) = f(x) + \frac{\lambda}{2}\|x - z\|^2$ по определению:

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

□

3 Техники и алгоритмы

Далее описываемая оболочка над алгоритмами сходимости по математическому ожиданию для обеспечения сходимости с высокой вероятностью зиждется на двух методах: RDE (Robust Distance Estimation, робастная оценка расстояний) и проксимальный метод.

3.1 Robust distance estimation

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 $\nabla f(x^*) = 0$):

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

Отмечу, что в дальнейшем будет полезно более общее неравенство для (L, γ)-гладких функций:

$$f(x) - f(x^*) \leq \frac{L}{2}||x - x^*||^2 + \gamma \tag{13}$$

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

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

Возвращаем x_{i^*}

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

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

Доказательство теоремы основано на неравенствах концентрации, его можно найти в [1] или в [25].

Опишем, как алгоритм RDE 1 может обеспечивать сходимость с высокой вероятностью. Пусть точки x_i $(i = 1, ..., m)$ таковы, что $\mathbb{E}f(x_\varepsilon) - \min f \leq \frac{\varepsilon}{3}$, то есть могут быть сгенерированы алгоритмом сходимости по мат. ожиданию. 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 (1), 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, RDE обеспечивает 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 [15] был предложена процедура *proxBoost*, которая будет описана далее. На процедуру RDE можно посмотреть и под другим углом. Желаемое качество сходимости (3) обеспечивается, если $m \sim \ln \frac{1}{p}$ раз вызвать оракул $\mathcal{M}(f, \frac{\varepsilon}{\kappa})$, что требует итоговой оракульной сложности $\mathcal{O}\left(\log \frac{1}{p} \cdot \mathcal{C}_{\mathcal{M}}(f, \frac{\varepsilon}{\kappa})\right)$, в которой содержится нежелаемый множитель κ. Предлагаемый в [15] подход уменьшает сложность по числу обусловленности κ до логарифмического.

3.2 Проксимальный метод

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. *(Неточный проксимальный метод) 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:

$$f(x_j) - f^* \leq \frac{L + \lambda_{j-1}}{2}||\bar{x}_j - x_j||^2 + \gamma + \sum_{i=0}^{j-1} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2. \tag{14}$$

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 $\sum_{i=0}^{T} \frac{\lambda_j}{2}||\bar{x}_i - x_i||^2$. Доказательство можно найти в [15]. Последний пункт теоремы следует непосредственно из леммы 2. Для достаточно больших T можно гарантировать то, что функция f^T хорошо обусловлена. Использование преимуществ результатов теорем (1) и (2) навело авторов [15] разработку алгоритм *proxBoost*, который здесь представлен в обобщенном виде для (L, γ)-гладких функций.

Algorithm 2 `proxBoost`(δ, p, T)

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

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

Получить точку x_0, удовлетворяющую $\|x_0 - \bar{x}_0\| \leq \varepsilon_{-1}$ с вероятностью $1 - p$.

В цикле $j = 0, \ldots, T-1$

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

Получить точку x_{j+1}, удовлетворяющую

$$\mathbb{P}\left[\|x_{j+1} - \bar{x}_{j+1}\| \leq \varepsilon_j \mid E_j\right] \geq 1 - p, \tag{15}$$

где E_j обозначает событие $E_j := \left\{x_i \in B_{\varepsilon_{i-1}}(\bar{x}_i) \text{ for all } i \in [0,j]\right\}$.

Получить точку x_{T+1}, удовлетворяющую

$$\mathbb{P}\left[f^T(x_{T+1}) - \min f^T \leq \delta \mid E_T\right] \geq 1 - p. \tag{16}$$

Выход: x_{T+1}

Теорема 3 (О proxBoost)**.** *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*

$$f(x_{T+1}) - \min f \leq \delta \left(1 + \sum_{i=0}^{T} \frac{\lambda_i}{\mu + \lambda_{i-1}}\right). \tag{17}$$

Доказательство. Сначала докажем по индукции оценку

$$\mathbb{P}[E_t] \geq 1 - (t+1)p \qquad \text{для всех } t = 0, \ldots, T. \tag{18}$$

База индукции $t = 0$ следует непосредственно из определения x_0. Теперь предположим, что (18) 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 (18) выполняются. Определим событие

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

Тогда

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

Теперь предположим, что выполнено событие $F \cap E_T$. Тогда

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

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

Отметим, что данная теорема не использует свойства гладкости или негладкости функции. Глядя на оценку (17), мы видим, что итоговая ошибка $f(x_{T+1}) - \min f$ контролируется суммой $\sum_{i=0}^{T} \frac{\lambda_i}{\mu+\lambda_{i-1}}$. Если выбрать проксимальные параметры геометрически возрастающими $\lambda_i = \mu 2^i$, то 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

Часто сложность стохастических градиентных методов, то есть количество оракульных вызовов, 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 (оракул), 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 $(\mu + \lambda)$-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 $\varepsilon = \sqrt{\frac{2\delta}{\mu+\lambda}}$. Если же функция f - (L,γ)-гладкая, то φ_x - $(L+\lambda,\gamma)$ (как было показано в лемме 2), следовательно выполняется двойное неравенство:

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

Технику Robust Distance Estimation (1) мы можем снабдить предложенным оракулом $\texttt{Alg}(\cdot)$. 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 $x \in \mathbb{R}^d$,

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`(δ,$\Delta_{\text{in}}, x_{\text{in}}$,$T$,$m$)

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

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

В цикле $j = 0, \ldots, T$:

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

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

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

Докажем работоспособность и эффективность этого алгоритма.

Теорема 4 (Эффективность `BoostAlg`)**.** *Пусть $x_{\text{in}} \in \mathbb{R}^d$ - 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`(δ,$\Delta_{\text{in}}, x_{\text{in}}$,$T$,$m$) удовлетворяет*

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

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

$$\mathbb{P}[\|x_{j+1} - \bar{x}_{j+1}\| \leq \varepsilon_j | E_j] \geq 1 - p \tag{19}$$

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

$$\mathbb{P}[f^T(x_{j+1}) - \min f^T \leq \delta + \gamma | E_T] \geq 1 - p \tag{20}$$

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 (19) 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 (14), 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:

$$f^{j-1}(x_{j-1}) - \min f^{j-1} \le f(x_{j-1}) - f^* \le \Delta_{j-1}. \tag{21}$$

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 (19) выполнено для x_j, что и требовалось.

Теперь предположим, что выполнено событие E_T. Аналогично (21) 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 - $(L+\lambda_T,\gamma)$-гладкая, получим

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

тем самым установив (20). Осталось лишь применить теорему (3) □

Следующая теорема собирает всё воедино.

Теорема 5 (Эффективность `BoostAlg` с геометрически возрастающими проксимальными параметрами)**.** *Зафиксируем стартовую точку $x_{\text{in}} \in \mathbb{R}^d$. Пусть Δ_{in} - некоторая верхняя оценка начальной невязки $\Delta_{\text{in}} \ge f(x_{\text{in}}) - \min f$. Зафиксируем желаемую функциональную*

точность $\varepsilon > 0$ *и вероятность* $p \in (0,1)$. *При параметрах алгоритма*

$$T = \lceil \log_2(\kappa) \rceil, \; m = \left\lceil 18 \ln\left(\frac{2+T}{p}\right)\right\rceil, \; \lambda_i = \mu 2^i, \; \left(\delta = \gamma = \frac{\varepsilon}{2(2+2T)} \text{ или } \delta = \frac{\varepsilon}{2+2T}, \gamma = 0\right)$$

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

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

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

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

при этом максимальная начальная невязка

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

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 $G\colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ — 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. Процедура, представленная в этой работе, вызывает оракул минимизации многократно, чтобы обеспечить сходимость с высокой вероятностью. 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 $\Delta_{\text{in}} > 0$ удовлетворяет $\Delta_{\text{in}} \geq f(x_0) - f^*$. Хорошо известно, что в сильно выпуклом гладком случае стохастический градиентный метод может сгенерировать точку x, удовлетворяющую $\mathbb{E}f(x) - f^* \leq \varepsilon$ за

$$\mathcal{O}\left(\kappa\log\left(\frac{\Delta_{\text{in}}}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}\right). \tag{22}$$

вызовов оракула. Ускоренные стохастические градиентные методы имеют меньшую сложность ([3])

$$\mathcal{O}\left(\sqrt{\kappa}\log\left(\frac{\Delta_{\text{in}}}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}\right). \tag{23}$$

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

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

за следующую оракульную сложность:

$$\mathcal{O}\left(\ln(\kappa)\ln\left(\frac{\ln\kappa}{p}\right)\cdot\left(\kappa\ln\left(\frac{\Delta_{\text{in}}\ln(\kappa)}{\varepsilon}\vee\kappa\right)+\frac{\sigma^2\ln(\kappa)}{\mu\varepsilon}\right)\right), \tag{24}$$

и

$$\mathcal{O}\left(\ln(\kappa)\ln\left(\frac{\ln\kappa}{p}\right)\cdot\left(\sqrt{\kappa}\ln\left(\frac{\Delta_{\text{in}}\ln(\kappa)}{\varepsilon}\vee\kappa\right)+\frac{\sigma^2\ln(\kappa)}{\mu\varepsilon}\right)\right), \tag{25}$$

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

Пусть исследуемая функция f не является гладкой, но является M-липшицевой. Тогда по лемме 1 она является $(\frac{M^2}{2\gamma}, \gamma)$ - гладкой для любого γ. Так как эффективность основного алгоритма 4 был доказан для случая (L, γ) - гладких функций, то достаточно взять $\gamma \sim \frac{\varepsilon}{\ln \kappa}$ из теоремы для обеспечения нужной сходимости 3. Так что $\kappa = \frac{L}{\mu} \approx \frac{M^2 \ln \kappa}{\mu \varepsilon}$ или $\frac{\kappa}{\ln \kappa} \approx \frac{M^2}{\mu \varepsilon}$. Для нахождения из этого соотношения κ понадобится следующая лемма:

Лемма 3. *Пусть* $\frac{x}{\ln x} = a \in \mathbb{R}_{++}$, *причем* $x > e$. *Тогда* $a \ln a < x < 2a \ln a$

Доказательство. Пусть $f(x) = \frac{x}{\ln(x)}$. Найдём её производную:

$$f'(x) = \frac{1 \cdot \ln(x) - x \cdot (1/x)}{(\ln(x))^2} = \frac{\ln(x) - 1}{(\ln(x))^2}$$

Производная положительна, когда $\ln(x) - 1 > 0$, то есть при $x > e$. Это означает, что для $x > e$ функция $f(x)$ строго возрастает.

Покажем, что для $y = 2a \ln(a)$ $f(y) > f(x)$, а значит и $y > x$ из строгого возрастания.

$$f(y) = \frac{2a \ln a}{\ln(2a \ln a)} = \frac{2a \ln(a)}{\ln(2) + \ln(a) + \ln(\ln(a))} >^? a = f(x)$$

$$2 \ln(a) >^? \ln(2) + \ln(a) + \ln(\ln(a))$$

$$\ln(a) - \ln(\ln(a)) >^? \ln(2)$$

Это неравенство верно для всех $a > e$. Действительно, функция $g(z) = z - \ln(z)$ при $z > 1$ возрастает, и её минимальное значение равно 1 (при $z = 1$). Поскольку $\ln(2) \approx 0.693$, неравенство выполняется. Верхняя оценка доказана.

Для доказательства нижней оценки покажем, что $f(a \ln a) < a$:

$$\frac{a \ln a}{\ln(a \ln a)} = \frac{a \ln a}{\ln a + \ln(\ln a)} <^? a$$

Последнее неравенство выполнено при $\ln \ln a > 0 \Rightarrow \ln a > 1 \Rightarrow a > e$. Нижняя оценка доказана. Таким образом, $x = \Theta(a \ln a)$

□

Таким образом,

$$\kappa \approx \frac{M^2}{\mu\varepsilon}\log\left(\frac{M^2}{\mu\varepsilon}\right)$$

Теорема 6 (Рубцов, 2025)**.** *Итоговая оракульная сложность задачи сходимости с высокой вероятностью* (3) *в случае μ-сильно выпуклых негладких, но M-липшицевых функций порядка*

$$\log\left(\frac{\log\left(\frac{M^2}{\mu\varepsilon}\log\left(\frac{M^2}{\mu\varepsilon}\right)\right)}{p}\right)\log\left(\frac{M^2}{\mu\varepsilon}\log\left(\frac{M^2}{\mu\varepsilon}\right)\right)\cdot\mathcal{C}_{\mathcal{M}}\left(f,\frac{\varepsilon}{\log\left(\frac{M^2}{\mu\varepsilon}\log\left(\frac{M^2}{\mu\varepsilon}\right)\right)}\right).$$

Пренебрегая «вложенными логарифмами», выражение можно упростить до

$$\log\left(\frac{1}{p}\right)\log\left(\frac{M^2}{\mu\varepsilon}\right)\cdot\mathcal{C}_{\mathcal{M}}\left(f,\frac{\varepsilon}{\log\left(\frac{M^2}{\mu\varepsilon}\right)}\right).$$

Типичная оракульная сложность решения задачи по мат.ожиданию (см., например, [26]) $\mathcal{C}_{\mathcal{M}}(f,\varepsilon)=\mathcal{O}\left(\frac{\max\{M^2,\sigma^2\}}{\mu\varepsilon}\right)$. Тогда оракульная сложность решения задачи сходимости по математическому ожиданию

$$\mathcal{O}\left(\log\left(\frac{1}{p}\right)\log^2\left(\frac{M^2}{\mu\varepsilon}\right)\cdot\frac{\max\{M^2,\sigma^2\}}{\mu\varepsilon}\right). \tag{26}$$

5.4 Выпуклый гладкий случай

От сильно выпуклому случая к выпуклому можно перейти с помощью метода регуляризации (см. [23]).

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 оракульная сложность задачи сходимости с высокой вероятностью в случае выпуклых гладких функций порядка*

$$\log\left(\frac{\log(\frac{LR^2}{\varepsilon})}{p}\right)\log(\frac{LR^2}{\varepsilon})\cdot\mathcal{C}_{\mathcal{M}}\left(f,\frac{\varepsilon}{\log(\frac{LR^2}{\varepsilon})}\right).$$

Пренебрегая «вложенными логарифмами», выражение можно упростить до

$$\log\left(\frac{1}{p}\right)\log(\frac{LR^2}{\varepsilon})\cdot\mathcal{C}_{\mathcal{M}}\left(f,\frac{\varepsilon}{\log(\frac{LR^2}{\varepsilon})}\right).$$

Типичная оракульная сложность решения задачи по мат.ожиданию $\mathcal{C}_{\mathcal{M}}(f,\varepsilon)=\mathcal{O}(\max\{\frac{LR_0^2}{\varepsilon};\frac{\sigma^2R_0^2}{\varepsilon^2}\})$. Тогда оракульная сложность решения задачи сходимости по математическому ожиданию

$$\mathcal{O}(\max\{\frac{LR^2}{\varepsilon};\frac{\sigma^2R^2\log(\frac{LR^2}{\varepsilon})}{\varepsilon^2}\}\cdot\log^2(\frac{LR^2}{\varepsilon})\log\{\frac{1}{p}\}) \tag{27}$$

5.5 Выпуклый негладкий случай

Обобщая предыдущие результаты, получим следующую теорему.

Теорема 9 (Рубцов, 2025). *Итоговая оракульная сложность задачи сходимости с высокой вероятностью* (3) *в случае выпуклых негладких, но M-липшицевых функций порядка (пренебрегая «вложенными логарифмами»)*

$$\log\left(\frac{1}{p}\right)\log\left(\frac{MR}{\varepsilon}\right)\cdot\mathcal{C}_{\mathcal{M}}\left(f,\frac{\varepsilon}{\log\left(\frac{MR}{\varepsilon}\right)}\right).$$

Типичная оракульная сложность решения задачи по мат.ожиданию $\mathcal{C}_{\mathcal{M}}(f,\varepsilon)=\mathcal{O}\left(\frac{\max\{M^2,\sigma^2\}R^2}{\varepsilon^2}\right)$. Тогда оракульная сложность решения задачи сходимости по математическому ожиданию

$$\mathcal{O}\left(\log\left(\frac{1}{p}\right)\log^3\left(\frac{MR}{\varepsilon}\right)\cdot\frac{\max\{M^2,\sigma^2\}R^2}{\varepsilon^2}\right). \tag{28}$$

5.6 Замечания

Особо стоит обратить внимание, что результатами теорем 5, 6, 8, 9 являются вычисленные сложности алгоритмов сходимости с высокой вероятностью, выраженные через сложность алгоритмов сходимости по матожиданию $\mathcal{C}_{\mathcal{M}}(f,\varepsilon)$. Важно, что последние могут быть решены с помощью применения различных оракулов: стохастического гради-

ентного оракула (про него шла речь ранее), оракулов нулевого порядка и прочих, а также с помощью различных методов (SGD, SSTM и пр.). Таким образом, разработанный подход является не просто очередным алгоритмом оптимизации, но оберткой над целым семейством алгоритмов, которая позволяет от сходимости по математическому ожиданию перейти к сходимости с высокой вероятностью за довольно скромную стоимость, логарифмическую по критическим параметрам $\frac{1}{\varepsilon}$, $\frac{1}{p}$ и другим.

6 Обзор применения техник в случае седловых задач

6.1 Постановка задачи

Рассмотрим стохастическую седловую задачу (Stochastic Saddle Point Problem, SSP):

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x,y) := \mathbb{E}[\Phi_\xi(x,y)] \tag{29}$$

где $\mathcal{X}$ и $\mathcal{Y}$ — замкнутые выпуклые множества, а ξ — случайная величина из неизвестного распределения $\mathcal{P}$.

Для любого допустимого решения $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ определяется **зазор двойственности** (duality gap):

$$\Delta_\Phi(\hat{x},\hat{y}) := \max_{y \in \mathcal{Y}} \Phi(\hat{x},y) - \min_{x \in \mathcal{X}} \Phi(x,\hat{y}) \tag{30}$$

Также вводится более слабый вариант разности двойственности:

$$\Delta_\Phi^w(\hat{x},\hat{y}) := \Phi(\hat{x},y^*) - \Phi(x^*,\hat{y}) \tag{31}$$

где (x^*, y^*) — оптимальное решение задачи (29).

В работе [22] разработан подход, при котором имея произвольный оракул, который возвращает решение с малым ожидаемым зазором двойственности, строится решение $(\bar{x}, \bar{y})$ такое, что

$$\mathbb{P}[\Delta_\Phi(\bar{x},\bar{y}) \leq \varepsilon] \geq 1 - p \tag{32}$$

с использованием лишь небольшого числа вызовов этого оракула.

6.2 Основные предположения

Предположение 5 (Сильная выпуклость-вогнутость)**.** *Существуют $\mu_x, \mu_y > 0$ такие, что для почти всех $\xi \sim \mathcal{P}$ функция $\Phi_\xi(\cdot,y)$ является μ_x-сильно выпуклой по x, а $\Phi_\xi(x,\cdot)$*

является μ_y-сильно вогнутой по y.

$$\Phi_\xi(x_2,y) \geq \Phi_\xi(x_1,y) + \langle \nabla_x \Phi_\xi(x_1,y), x_2 - x_1 \rangle + \frac{\mu_x}{2}\|x_1 - x_2\|^2, \quad \forall x_1,x_2 \in \mathcal{X}, y \in \mathcal{Y},$$
$$\Phi_\xi(x,y_2) \leq \Phi_\xi(x,y_1) + \langle \nabla_y \Phi_\xi(x,y_1), y_2 - y_1 \rangle - \frac{\mu_y}{2}\|y_1 - y_2\|^2, \quad \forall y_1,y_2 \in \mathcal{Y}, x \in \mathcal{X}.$$

Будем говорить, что Φ - (μ_x, μ_y)*-сильно выпукла-сильно вогнута.*

Предположение 6 (Гладкость)**.** *Существуют $L_x, L_y, L_{xy} > 0$ такие, что градиенты $\nabla_x \Phi$ и $\nabla_y \Phi$ являются липшицевыми с соответствующими константами.*

$$\|\nabla_x \Phi(x_1,y_1) - \nabla_x \Phi(x_2,y_1)\| \leq L_x \|x_1 - x_2\|, \quad \|\nabla_y \Phi(x_1,y_1) - \nabla_y \Phi(x_1,y_2)\| \leq L_y \|y_1 - y_2\|,$$
$$\|\nabla_x \Phi(x_1,y_1) - \nabla_x \Phi(x_1,y_2)\| \leq L_{xy} \|y_1 - y_2\|, \quad \|\nabla_y \Phi(x_1,y_1) - \nabla_y \Phi(x_2,y_1)\| \leq L_{xy} \|x_1 - x_2\|.$$

Предположение 7 (Липшицевость функции)**.** *Для ограниченных областей существуют $\ell_x, \ell_y > 0$ такие, что функция Φ_ξ липшицева по каждой переменной.*

$$|\Phi_\xi(x_2,y) - \Phi_\xi(x_1,y)| \leq \ell_x \|x_1 - x_2\| \quad and \quad |\Phi_\xi(x,y_1) - \Phi_\xi(x,y_2)| \leq \ell_y \|y_1 - y_2\|.$$

Обозначения: $\mu := \min\{\mu_x, \mu_y\}$, $L := \max\{L_x, L_y, L_{xy}\}$, $\ell := \max\{\ell_x, \ell_y\}$, число обусловленности $\kappa := L/\mu$.

6.3 Алгоритм PB-SSP для неограниченных задач

Для неограниченных задач ($\mathcal{X} = \mathbb{R}^{d_x}$, $\mathcal{Y} = \mathbb{R}^{d_y}$) предлагается алгоритм **PB-SSP** (Proximal Boost for Stochastic Saddle Point problems).

Основная идея, как и в proxBoost, использовать неточный проксимальный метод для последовательного решения возмущенных подзадач:

$$f^i(x) = f(x) + \frac{\lambda_x^i}{2}\|x - x_i^c\|^2 \tag{33}$$
$$g^i(y) = g(y) - \frac{\lambda_y^i}{2}\|y - y_i^c\|^2 \tag{34}$$

где $f(x) = \max_{y \in \mathcal{Y}} \Phi(x,y)$ и $g(y) = \min_{x \in \mathcal{X}} \Phi(x,y)$.

Algorithm 5 PB-SSP(δ, p, T)

Вход: Точность $\delta > 0$, вероятность $p \in (0,1)$, число итераций T

Установить $\lambda_x^{-1} = \lambda_y^{-1} = 0$, $x_{-1}^c = y_{-1}^c = 0$

for $i = 0, \ldots, T$ **do**

Установить $\varepsilon_x^i = \sqrt{\frac{2\delta}{\mu_x + \lambda_x^{i-1}}}$, $\varepsilon_y^i = \sqrt{\frac{2\delta}{\mu_y + \lambda_y^{i-1}}}$

Найти точку (x_i^c, y_i^c) такую, что

$$\mathbb{P}[\|x_i^c - x_i^*\| \le \varepsilon_x^i] \ge 1 - \frac{p}{2T+4} \quad \text{и} \quad \mathbb{P}[\|y_i^c - y_i^*\| \le \varepsilon_y^i] \ge 1 - \frac{p}{2T+4}$$

Найти точку (x_{T+1}^c, y_{T+1}^c) такую, что

$$\mathbb{P}[f^T(x_{T+1}^c) - f^T(x_{T+1}^*) \le \delta] \ge 1 - \frac{p}{2T+4} \quad \text{и} \quad \mathbb{P}[g^T(y_{T+1}^*) - g^T(y_{T+1}^c) \le \delta] \ge 1 - \frac{p}{2T+4}.$$

return (x_{T+1}^c, y_{T+1}^c)

Теорема 10 (Эффективность PB-SSP)**.** *С вероятностью не менее* $1-p$ *точка* $(x_{T+1}^c, y_{T+1}^c) =$ *PB-SSP*(δ, p, T) *удовлетворяет*

$$\Delta_\Phi(x_{T+1}^c, y_{T+1}^c) \le \delta \left(2 + \sum_{i=0}^{T} \frac{\lambda_x^i}{\mu_x + \lambda_x^{i-1}} + \frac{\lambda_y^i}{\mu_y + \lambda_y^{i-1}}\right) \tag{35}$$

Теорема 11 (Геометрическое убывание параметров)**.** *При выборе* $\lambda_x^i = \mu_x \cdot 2^i$, $\lambda_y^i = \mu_y \cdot 2^i$, $T = \lceil \log_2(\kappa) \rceil$ *и* $\delta = \frac{\varepsilon}{4+4T}$ *получаем решение с* $\mathbb{P}[\Delta_\Phi(\bar{x}, \bar{y}) \le \varepsilon] \ge 1 - p$ *и сложностью*

$$\mathcal{O}\left(\ln\left(\frac{\ln(\kappa)}{p}\right) \ln(\kappa) \cdot C_{\mathcal{M}}^w\left(\Phi, \frac{\varepsilon}{\ln(\kappa)}\right)\right) \tag{36}$$

6.4 Связь с proxBoost

Данный подход является естественным обобщением алгоритма ProxBoost из статьи [15] на случай седловых задач. Оба подхода используют неточный проксимальный метод с геометрически возрастающими параметрами регуляризации для улучшения числа обусловленности, а также метод RDE для получения высоковероятностных гарантий. В обоих случаях достигается полилогарифмическая зависимость от числа обусловленности κ и логарифмическая по $\frac{1}{p}$. Важно, что оба подхода являются мета-алгоритмическими, то есть работают с произвольными оракулами. Аналогично задачам выпуклой оптимизации,

данный подход может быть расширен на случай невыпуклых и/или не сильно выпуклых функций, что является планами на будущую работу.

7 Вычислительный эксперимент

Целью эксперимента является практическая демонстрация надежности мета-алгоритма `BoostAlg`, описанного в работе [15], в сравнении со стандартным стохастическим градиентным спуском (`SGD`).

7.1 Постановка задачи

Возьмем функцию $f(x, \xi) = \frac{Lx_1^2}{2} + \frac{\mu x_2^2}{2} + \langle \xi, x \rangle$, где $x = (x_1, x_2) \in \mathbb{R}^2$ и $L \geq \mu > 0$. Эта функция является стохастической версией квадратичной функции $f(x) := \mathbf{E}_\xi[f(x, \xi)] = \frac{Lx_1^2}{2} + \frac{\mu x_2^2}{2}$, где предполагается, что $\mathbf{E}[\xi] = 0$. Стохастический градиент по x имеет вид $\nabla_x f(x, \xi) = [Lx_1, \mu x_2]^T + \xi = \nabla f(x) + \xi$.

Для создания плохо обусловленной задачи были выбраны следующие параметры:

- Параметр гладкости $L = 100$.
- Параметр сильной выпуклости $\mu = 0.0001$.
- Число обусловленности $\kappa = L/\mu = 10^6$.
- Шум моделируется как гауссовский: $\xi \sim \mathcal{N}(0, \sigma^2 I_2)$ со ст. отклонением $\sigma = 0.5$.
- Начальная точка $x_{init} = [1.0, 1.0]^T$.

7.2 Сравнение алгоритмов

Сравниваются два алгоритма:

1. `BoostAlg` (`proxBoost`): Мета-алгоритм, который итеративно «усиливает» надежность решения, последовательно решая регуляризованные подзадачи. Теоретически гарантирует достижение точности ε с вероятностью не менее $1 - p$.
2. `SGD`: Классический стохастический градиентный спуск с постоянным шагом $\eta = 1/L$.

Методология. Бюджет вызовов стохастического оракула был зафиксирован. Сначала мы запускаем `BoostAlg` для достижения целевой точности $\varepsilon = 0.001$ с вероятностью ошибки не более $p = 0.05$. Общее число вызовов градиента, которое потребовалось `BoostAlg` (около 400000), используется как бюджет для `SGD`. Для статистической оценки надежности оба алгоритма были запущены по 20 раз. Параметры алгоритмов брались из теорем с возможным небольшим изменением (не по порядку величины).

7.3 Результаты

Результаты 20 независимых запусков сведены в Таблицу 1. «Неудачей» считался запуск, если итоговая ошибка $f(x_{final})$ превышала целевую $\varepsilon = 0.001$.

Алгоритм	**Бюджет вызовов ∇f**	**Кол-во неудач**	**Эмп. вер-ть неудачи**
`BoostAlg`	≈ 400000	0 из 20	0%
`SGD`		7 из 20	35%

Таблица 1: Сравнение надежности алгоритмов.

На Рис. 1 показаны траектории сходимости. Для `BoostAlg` (синяя линия) видна "ступенчатая"сходимость, при этом и медиана, и квантили в конце оказываются значительно ниже целевой ошибки.`SGD` (красная линия) сходится быстрее вначале, но его медианная траектория останавливается около целевой ошибки, а большой разброс результатов указывает на низкую надежность.

7.4 Выводы

Эксперимент подтверждает теоретические преимущества `BoostAlg`:

- **Высокая надежность**: `BoostAlg` достиг целевой точности во всех запусках, что соответствует теоретической гарантии ($p < 0.05$).
- **Ненадежность SGD**: При том же бюджете `SGD` потерпел неудачу почти в половине случаев (35%).
- **Цена надежности**: Более медленная начальная сходимость `BoostAlg` является платой за внутренние процедуры, которые гарантируют достижение результата.

Таким образом, `BoostAlg` является эффективным инструментом для задач стохастической оптимизации, где требуется высокая и предсказуемая вероятность успеха.

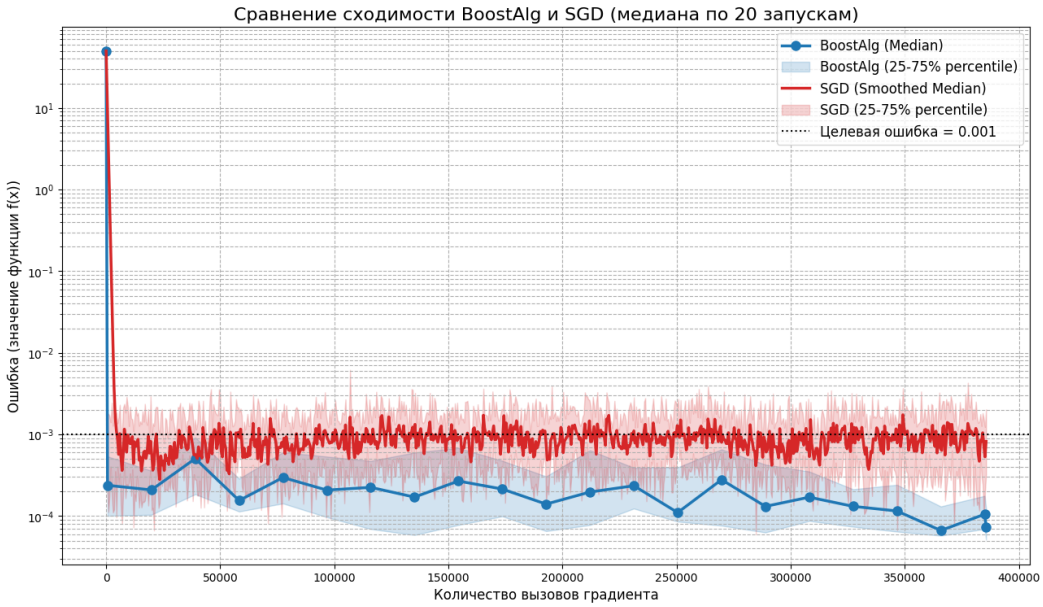

Рис. 1: График сходимости: ошибка $f(x) - f^*$ от количества вызовов оракула. Сплошная линия — медиана, закрашенная область — разброс между 0.25 и 0.75 квантилями.

8 Заключение

В настоящей дипломной работе были исследованы и разработаны методы для решения задач стохастической выпуклой оптимизации и седловых задач, обеспечивающие сходимость с высокой вероятностью. В отличие от классических подходов, гарантирующих сходимость лишь по математическому ожиданию, предложенные алгоритмы позволяют получить решение заданной точности ε с вероятностью не менее $1 - p$ при малых p.

На защиту выносятся следующие основные результаты:

1. **Обобщенный мета-алгоритм для сходимости с высокой вероятностью.** Разработана и теоретически обоснована модификация алгоритма `proxBoost` для широкого класса задач выпуклой оптимизации, включая негладкие и не сильно выпуклые случаи. Данный подход представляет собой универсальную «обертку», позволяющую преобразовать любой алгоритм со сходимостью по математическому ожиданию в алгоритм со сходимостью с высокой вероятностью. При этом оракульная сложность возрастает лишь на полилогарифмический множитель по параметрам задачи и логарифмический по $1/p$.

2. **Новые оценки оракульной сложности.** Получены детальные оценки сложности для разработанного мета-алгоритма в следующих классах задач:

- сильно выпуклые негладкие функции;
- выпуклые гладкие функции;
- выпуклые негладкие функции.

Показано, что итоговая сложность имеет слабую (логарифмическую) зависимость от вероятности отказа p и близкую к оптимальной зависимость от точности ε.

3. **Применимость подхода к седловым задачам.** Продемонстрирована универсальность лежащих в основе метода идей (неточный проксимальный метод и робастная оценка расстояний) путем их применения для решения стохастических выпукло-вогнутых седловых задач, что подтверждает общность и фундаментальность подхода.

4. **Экспериментальное подтверждение надежности.** Результаты вычислительного эксперимента наглядно демонстрируют теоретические преимущества предложенного подхода. В условиях плохо обусловленной задачи алгоритм `BoostAlg` обеспечивает гарантированную сходимость при том же бюджете вызовов оракула, при котором стандартный `SGD` показывает крайне низкую надежность.

Таким образом, в работе представлен комплексный подход к построению надежных и эффективных алгоритмов стохастической оптимизации, подкрепленный теоретическими оценками и практическими результатами.

Список литературы

1. *Nemirovskij A. S.*, *Yudin D. B.* Problem complexity and method efficiency in optimization. — 1983.
2. *Polyak B. T.*, *Juditsky A. B.* Acceleration of stochastic approximation by averaging // SIAM journal on control and optimization. — 1992. — Т. 30, № 4. — С. 838—855.
3. *Ghadimi S.*, *Lan G.* Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms // SIAM Journal on Optimization. — 2013. — Т. 23, № 4. — С. 2061—2089.
4. *Hazan E.*, *Kale S.* Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization // The Journal of Machine Learning Research. — 2014. — Т. 15, № 1. — С. 2489—2512.

5. *Bousquet O.*, *Elisseeff A.* Stability and generalization // Journal of machine learning research. — 2002. — Т. 2, Mar. — С. 499—526.

6. *Nesterov Y.*, *Vial J.-P.* Confidence level solutions for stochastic programming // Automatica. — 2008. — Т. 44, № 6. — С. 1559—1568.

7. Stochastic Convex Optimization. / S. Shalev-Shwartz [и др.] // COLT. Т. 2. — 2009. — С. 5.

8. Robust stochastic approximation approach to stochastic programming / A. Nemirovski [и др.] // SIAM Journal on optimization. — 2009. — Т. 19, № 4. — С. 1574—1609.

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

10. *Ghadimi S.*, *Lan G.* Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework // SIAM Journal on Optimization. — 2012. — Т. 22, № 4. — С. 1469—1492.

11. *Harvey N. J.*, *Liaw C.*, *Randhawa S.* Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent // arXiv preprint arXiv:1909.00843. — 2019.

12. Tight analyses for non-smooth stochastic gradient descent / N. J. Harvey [и др.] // Conference on Learning Theory. — PMLR. 2019. — С. 1579—1613.

13. *Gorbunov E.*, *Danilova M.*, *Gasnikov A.* Stochastic optimization with heavy-tailed noise via accelerated gradient clipping // Advances in Neural Information Processing Systems. — 2020. — Т. 33. — С. 15042—15053.

14. High-Probability Complexity Bounds for Non-smooth Stochastic Convex Optimization with Heavy-Tailed Noise / E. Gorbunov [и др.] // Journal of Optimization Theory and Applications. — 2024. — С. 1—60.

15. 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.

16. *Puterman M. L.* Markov decision processes: discrete stochastic dynamic programming. — John Wiley & Sons, 2014.

17. *Wang M.* Primal-Dual π Learning: Sample Complexity and Sublinear Run Time for Ergodic Markov Decision Problems // arXiv preprint arXiv:1710.06100. — 2017.

18. *Shalev-Shwartz S.*, *Zhang T.* Stochastic dual coordinate ascent methods for regularized loss // The Journal of Machine Learning Research. — 2013. — Т. 14, № 1. — С. 567—599.

19. *Zhang Y.*, *Xiao L.* Stochastic primal-dual coordinate method for regularized empirical risk minimization // Journal of Machine Learning Research. — 2017. — Т. 18, № 84. — С. 1—42.

20. Stochastic Primal-Dual Algorithms with Faster Convergence than $O(1/\sqrt{T})$ $for Problems without Bilinear Structure$ / Y. Yan [и др.] // arXiv preprint arXiv:1904.10112. — 2019.

21. Generalization bounds for stochastic saddle point problems / J. Zhang [и др.] // International Conference on Artificial Intelligence and Statistics. — PMLR. 2021. — С. 568—576.

22. *Li D.*, *Li H.*, *Zhang J.* General procedure to provide high-probability guarantees for stochastic saddle point problems // Journal of Scientific Computing. — 2024. — Т. 100, № 1. — С. 13.

23. Выпуклая оптимизация / Е. Воронцова [и др.] // М.: МФТИ. — 2021.

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

25. *Hsu D.*, *Sabato S.* Loss minimization and parameter estimation with heavy tails // Journal of Machine Learning Research. — 2016. — Т. 17, № 18. — С. 1—40.

26. *Гасников А. В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. — 2018.