1.5em 0pt

Сходимость с оценкой вероятностей больших отклонений для задач выпуклой оптимизации

Денис Николаевич Рубцов

Московский физико-технический институт

Научный руководитель: д.ф.-м.н., чл.-корр. РАН А. В. Гасников

2025

Цель исследования

Цель

- разработать быстрые алгоритмы для решения задач выпуклой стохастической оптимизации, обеспечивающие сходимость с высокой вероятностью
- исследовать эти алгоритмы с помощью теоретического анализа и вычислительных экспериментов

Постановка задачи

Задача стохастической оптимизации

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E} f(x, \xi), \; \xi \sim \mathcal{P}$$

Как правило, результатом стохастических градиентных методов является точка x_ε такая, что

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

Мы рассматриваем алгоритмы, результатом которых являются точки $x_{\varepsilon,p}$, удовлетворяющие условию

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

где «уровень уверенности» $1 - p$ может быть достаточно большим

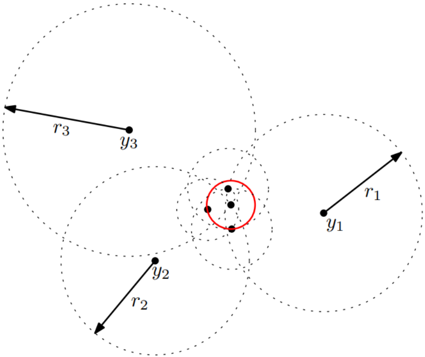
Постановка задачи

Если решить задачу $\mathbb{E}f(x_\varepsilon) - \min f \leq p\varepsilon$, то желаемое неравенство $P\{f(x_{\varepsilon,p}) - \min f \leq \varepsilon\} \geq 1 - p$ следует автоматически по неравенству Маркова.
Сложность решения задачи сходимости по матожиданию, обычно, порядка $\mathcal{O}(\frac{1}{\varepsilon})$. Тогда сложность наивного решения задачи сходимости с высокой вероятностью $\mathcal{O}(\frac{1}{p\varepsilon})$. Хочется уменьшить множитель $\frac{1}{p}$ до $\ln(\frac{1}{p})$

Robust distance estimation (RDE)

Пусть имеется m точек $x_1, ..., x_m$, для которых $\mathbb{E}x_i - x^* \leq \frac{\varepsilon}{3}$, т.е. $P[||x - x^*|| \leq \varepsilon] \geq \frac{2}{3}$. Тогда среди этих точек можно выбрать такую x_{i^*}, вокруг которой кластеризуются остальные точки.

Рис.: Идея метода RDE



Theorem

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

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

Применение RDE для обеспечения сходимости с высокой вероятностью: описание подхода

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

Пусть мы имеем точки x_i $(i = 1, ..., m)$ такие, что

$$\mathbb{E}f(x_i) - \min f \leq \frac{\varepsilon}{3} \xRightarrow{\text{неравенство Маркова}}$$

$$P(f(x_i) - f^* \leq \varepsilon) \geq \frac{2}{3} \xRightarrow{\text{сильная выпуклость}}$$

$$P(||x_i - x^*|| < \sqrt{\frac{2\varepsilon}{\mu}} =: \delta) \geq \frac{2}{3} \xRightarrow{\text{RDE}}$$

$$P(||x_{i^*} - x^*|| < 3\delta) \geq 1 - e^{-\frac{m}{18}} \xRightarrow{\text{гладкость}}$$

$$P(f(x_{i^*}) - f^* \leq 9\frac{L}{\mu}\varepsilon) \geq 1 - e^{-\frac{m}{18}}$$

Применение RDE для обеспечения сходимости с высокой вероятностью: проблема

$$\mathbb{E}f(x_i) - \min f \leq \frac{\varepsilon}{3} \overset{\dots}{\Longrightarrow}$$

$$P(f(x_{i^*}) - f^* \leq 9\frac{L}{\mu}\varepsilon) \geq 1 - e^{-\frac{m}{18}}$$

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.

Проксимальный метод *proxBoost*

Зафиксируем возрастающую последовательность $\lambda_0, ..., \lambda_T$ и последовательность точек $x_0, ..., x_T$. На каждой итерации $i = 0, ..., T$ будем решать задачу минимизации не функции f, а функции f^i

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

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

Число обусловленности новых функций можно сделать значительно меньше $\kappa_i = \frac{L+\lambda_i}{\mu+\lambda_i} = (\lambda_i = \mu \cdot 2^i) = \frac{L+\mu\cdot 2^i}{\mu+\mu\cdot 2^i} = \mathcal{O}(1)$ при $i > \log \frac{L}{\mu}$

При этом решение новых задач будет приближенным решением основных задач

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

Сложность алгоритма proxBoost

Theorem

Пусть имеется оракул $\mathcal{M}(f, \varepsilon)$, возвращающий точку x_ε такую, что $P(f(x_\varepsilon) - \min f \leq \varepsilon) \geq \frac{2}{3}$. Стоимость вызова такого оракула обозначим за $\mathcal{C}_\mathcal{M}(f, \varepsilon)$. Тогда для μ-сильно выпуклых L-гладких функций сложность алгоритма, решающего задачу $P\{f(x_{\varepsilon,p}) - \min f \leq \varepsilon\} \geq 1 - p$:

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

Метод регуляризации для решения не сильно выпуклых задач

Theorem

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

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

Сложность алгоритма proxBoost в выпуклом случае

Theorem (Рубцов, 2024)

Сложность обобщенного алгоритма proxBoost для ***выпуклых*** *L-гладких функций, решающего задачу* $P\{f(x_{\varepsilon,p}) - \min f \leq \varepsilon\} \geq 1 - p$*:*

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

Более конкретно,

$$\mathcal{O}\left(\max\{\sqrt{\frac{LR_0^2}{\varepsilon}}; \frac{\sigma^2 R_0^2}{\varepsilon^2}\}\cdot\ln^2(\frac{LR_0^2}{\varepsilon})\ln\{\frac{\ln(\frac{LR_0^2}{\varepsilon})}{\beta}\}\right) \qquad (1)$$

Метод сглаживания

Функция f - (L, γ)-гладкая, если

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

Градиент функции $f(x)$ удовлетворяет условию Гёльдера, если

$$\|\nabla f(y) - \nabla f(x)\| \leq L_\nu \|y - x\|_2^\nu, \ \nu \in [0, 1], \ L_0 < \infty \qquad (3)$$

При $\nu = 1$ условие Гёльдера является условием L_1-гладкости. При $\nu = 0$ оно является условием L_0-липшицевости. Далее $L_1 = L$ и $L_0 = M$.
Предположение «слабой гладкости» введено для того, чтобы смотреть на гладкий и негладкий случаи единообразно.Если функция негладкая, но $M-$липшицева, то есть $\|\nabla f(y) - \nabla f(x)\| \leq M$, то (2) выполняется при $L = \frac{M^2}{2\gamma}$.

Сложность алгоритма в негладком случае

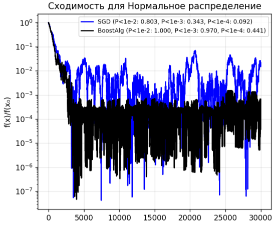
Theorem (Рубцов, 2025)

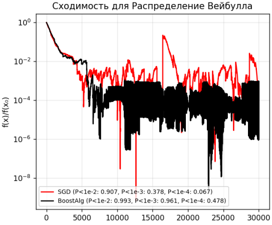
Сложность обобщенного алгоритма proxBoost для ***негладких*** *μ-сильно выпуклых функций, решающего задачу $P\{f(x_{\varepsilon,p}) - \min f \leq \varepsilon\} \geq 1 - p$ порядка*

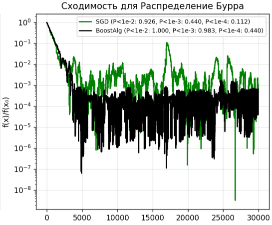
$$\ln\left(\frac{\ln \frac{M^2}{\mu\varepsilon}}{p}\right) \ln^2 \frac{M^2}{\mu\varepsilon} \cdot \frac{M^2 + \sigma^2}{\mu\varepsilon} \tag{4}$$

Вычислительный эксперимент

SGD vs ProxBoost







Заключение

Выносится на защиту

- обобщение алгоритма proxBoost для решения задач выпуклой и негладкой стохастической оптимизации
- теорема сходимости метода
- вычислительные эксперименты, демонстрирующие сходимость метода