

Сходимость с оценкой вероятностей больших отклонений для задач выпуклой оптимизации

Денис Николаевич Рубцов

Московский физико-технический институт

67-я Всероссийская научная конференция МФТИ

Секция проблем интеллектуального анализа данных, распознавания и
прогнозирования

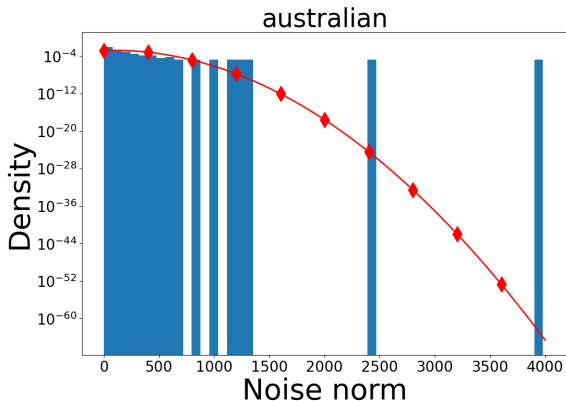
Научный руководитель: д.ф.-м.н. А. В. Гасников

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

Тяжелые хвосты распределения шума стохастического градиента

Гистограмма распределения шума стох.градиента для датасета "australian" из библиотеки LIBSVM. Красная линия аппроксимирует гистограмму функцией плотности нормального распределения.



Постановка задачи

Задача стохастической оптимизации

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

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)

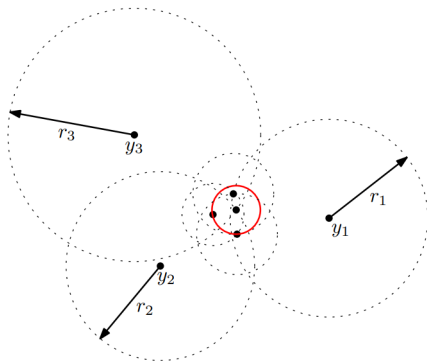
Пусть имеется m точек x_1, \dots, x_m , для которых $\mathbb{E}x_i - x^* \leq \frac{\epsilon}{3}$, т.е. $P(\|x - x^*\| \leq \epsilon) \geq \frac{2}{3}$. Тогда среди этих точек можно выбрать такую x_{i^*} , вокруг которой кластеризуются остальные точки.

Рис.: Идея метода RDE

Theorem

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

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



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 ($i = 1, \dots, m$) 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

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

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

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 $\lambda_0, \dots, \lambda_T$ и последовательность точек x_0, \dots, x_T . На каждой итерации $i = 0, \dots, T$ 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 $P\{f(x_{\varepsilon,p}) - \min f \leq \varepsilon\} \geq 1 - p$:

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

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

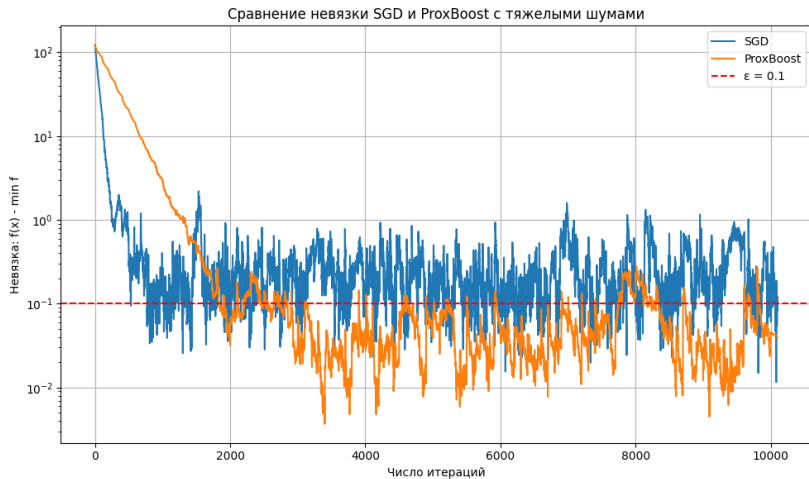
Сложность обобщенного алгоритма proxBoost для **выпуклых** L -гладких функций, решающего задачу

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

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

Вычислительный эксперимент

SGD vs ProxBoost



Выносятся на защиту

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