1.5em 0pt

Промежуточные градиентные методы с относительным шумом

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

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра "Интеллектуальные системы"

16 декабря, 2023

Постановка задачи и цели

Цель

Исследовать поведение промежуточных ускоренных алгоритмов в условии относительного шума и оценить влияние шума на ускорение.

Рассмотрим задачу оптимизации

$$\min_{x} f(x),$$

где f выпуклая или сильно выпуклая функция.
Зашумленный оракул первого порядка со значением $\hat{\varepsilon} \in [0, 1)$:

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \hat{\varepsilon} \|\nabla f(x)\|_2, \quad \forall \hat{\varepsilon} \in [0, 1].$$

Мотивация

Проблема

Для теоретического анализа относительного шума существует мало техник, однако такая постановка является довольно распространённой.

Нахождение градиента — решение подзадачи с нужной точностью точности

1. Решение задачи минимизации функционала
2. Решение систем PDE
3. Машинные ошибки

Решение

Решение

Предлагается использовать численный метод доказательств Performance Estimation Problem(PEP) для получения точных оценок сходимости, из которых получается теория.

PEP позволяет задавать любую выпуклую гладкую функцию как набор векторов и условий на них, тем самым переходя к задаче конечномерной выпуклой оптимизации

Algorithm Intermediate Similar Triangle Method (`ISTM`) [1]

Require: Initial point x^0, number of iterations N, smoothness constant $L > 0$, and step size parameter $a \geq 1$, intermediate parameter $p \in [1, 2]$.

Set $A_0 = \alpha_0 = 0, y^0 = z^0 = x^0$.

for $k = 0, 1, \ldots, N-1$ **do**

Set $\alpha_{k+1} = \frac{(k+2)^{p-1}}{2aL}$, $A_{k+1} = \alpha_{k+1} + A_k$.

$x^{k+1} = \frac{1}{A_{k+1}}\left(A_k y^k + \alpha_{k+1} z^k\right)$.

$z^{k+1} = z^k - \alpha_{k+1} \widetilde{\nabla} f(x^{k+1})$.

$y^{k+1} = \frac{1}{A_{k+1}}\left(A_k y^k + \alpha_{k+1} z^{k+1}\right)$.

Ensure: y^N

Мы можем контролировать a, p, чтобы достичь лучшей сходимости. Чем больше $p \in [1, 2]$, тем сильнее влияние шума ($\sim \hat{\varepsilon} N^{p-1}$), но быстрее сходимость $\sim \frac{1}{N^p}$.

PEP

Запускаем алгоритм ISTMstep на N итераций на выпуклой гладкой функции f с градиентом g и зашумленным градиентом $\hat{g}$ и смотрим максимальное отклонение

$$\max_{\substack{n, x^*, f^*, g^* \\ \{x^i, y^i, z^i\}_{i=0}^N \\ \{f^i, g^i, \widetilde{g}^i\}_{i=0}^N}} \quad f^N - f^*$$

$$\begin{aligned}
& f : \mathbb{R}^n \to \mathbb{R} \text{ is } L\text{-smooth and convex}, \\
& f^k = f(x^k), \quad g^k \in \nabla f(x^k), \quad k = *, 0, 1 \ldots, N, \\
& g^* = 0, \\
& \|x^0 - x^*\|_2^2 \le R^2, \\
& \|\widetilde{g}^k - g^k\|_2^2 \le \hat{\varepsilon}^2 \|g^k\|_2^2, \quad k = \overline{0, N-1}, \\
& x^{k+1}, y^{k+1}, z^{k+1} = \text{ISTMstep}(x^k, y^k, z^k, \widetilde{g}^k), \ k = \overline{0, N-1}.
\end{aligned}$$

PEP 2

Theorem ([3])

Для набора $\{x^i, f^i, g^i\}_{i\in I}$ *существует выпуклая и L-гладкая функция* f *такая, что для всех* $i \in I$ *мы имеем* $g^i \in \partial f(x^i)$ *и* $f^i = f(x^i)$ *тогда и только тогда, когда для любой пары индексов* $i \in I$ *и* $j \in I$ *верно следующее неравенство*

$$f^i - f^j - (g^j)^\top (x_i - x_j) \geq \frac{\|g^i - g^j\|^2}{2}.$$

Заменяем оптимизацию на бесконечном домене $f : \mathbb{R}^n \to \mathbb{R}^n$ на набор неравенств для всех точек.

PEP 3

$$\max_{\substack{n, x^*, f^*, g^* \\ \{x^i, y^i, z^i\}_{i=0}^N \\ \{f^i, g^i, \widetilde{g}^i\}_{i=0}^N}} \quad f^N - f^*$$

$$f^i - f^j - (g^j)^\top (x_i - x_j) \geq \frac{\|g^i - g^j\|_2^2}{2}, \;\; i, j = *, 0, \ldots, N,$$
$$\|x^0 - x^*\|_2^2 \leq R^2, g^* = 0,$$
$$\|\widetilde{g}^k - g^k\|_2^2 \leq \hat{\varepsilon}^2 \|g^k\|_2^2, \quad k = \overline{0, N-1},$$
$$x^{k+1}, y^{k+1}, z^{k+1} = \mathrm{ISTMstep}(x^k, y^k, z^k, \widetilde{g}^k), \;\; k = \overline{0, N-1}.$$

Проблема линейна по скалярным произведениям относительно оптимизируемых векторов и скаляров, так что, определив матрицу Грамма $G := V^\top V \in \mathbb{R}^{2(N+2)\times 2(N+2)}$, где $V = \left(x^0, x^*, \{g^i\}_{i\in I}, \{\widetilde{g}^i\}_{i\in I}\right) \in \mathbb{R}^{d\times 2(N+2)}$ и $\mathbf{f} = (f_*, f_0, \ldots, f_N) \in \mathbb{R}^{N+2}$, мы получим задачу SDP.

Численные эксперименты

Считаем $N_{\text{pep}}(a, \hat{\varepsilon}, p)$ на котором $\{\tau_i\}_{i=0}^{N_{\max}}$ перестаёт уменьшаться. Ориентируясь на $N_{\text{pep}}(a, \hat{\varepsilon}, p)$, мы выводим функцию $a = C^2 N^p \hat{\varepsilon}^2$ или $N_{\text{theory}}(a, \hat{\varepsilon}, p) = \left(\frac{C^2 a}{\hat{\varepsilon}^2}\right)^{\frac{1}{p}}$.

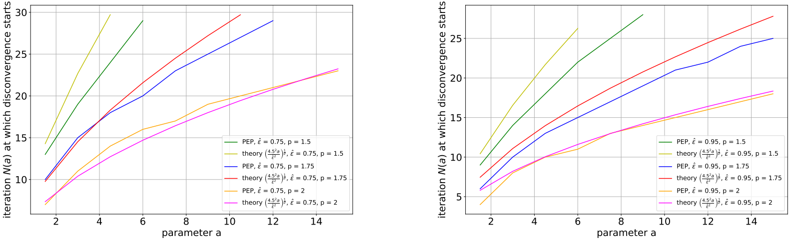

Рис.: Графики $N_{\text{pep}}(a, \hat{\varepsilon}, p)$ для различных p с $L = 1, R = 1$ и **Слева:** $\hat{\varepsilon} = 0.75$, **Справа:** $\hat{\varepsilon} = 0.75$.

Theorem ([2])

Пусть f выпуклая и L-гладкая функция с относительным шумом $\hat{\varepsilon} \in [0,1]$. Тогда после $N \geq 1$ итераций ISTM с параметром $p \in [1,2]$ и

$$a = O\left(\max\left\{1, N^{\frac{p}{4}}\sqrt{\hat{\varepsilon}}, N^{\frac{p}{2}}\hat{\varepsilon}, N^{p}\hat{\varepsilon}^2\right\}\right), \tag{1}$$

верна следующая оценка

$$f(y^N) - f(x^*) \leq \frac{16aLR_0^2}{(N+1)^p}, \quad R_0 = \|x^0 - x^*\|_2. \tag{2}$$

Учитывая a из (1), мы имеем

$$f(y^N) - f(x^*) \leq O\left(\max\left\{\frac{LR_0^2}{N^p}, \frac{\sqrt{\hat{\varepsilon}}LR_0^2}{N^{\frac{3p}{4}}}, \frac{\hat{\varepsilon}LR_0^2}{N^{\frac{p}{2}}}, \hat{\varepsilon}^2 LR_0^2\right\}\right). \tag{3}$$

Сильно выпуклый случай

В случае μ-сильно выпуклой функции мы применяем технику рестартов (алгоритм `RISTM`)

1. Запустить `ISTM` с теоретическими параметрами (x^i, N^i, L^i, a^i, p) и получить y^i
2. Задать ответ как новую начальную точку $x^{i+1} = y^i$

Теоретические результаты RISTM

Theorem ([2])

Пусть f L-гладкая и μ-сильно выпуклая функция с относительным шумом $\hat{\varepsilon}$. Если $\hat{\varepsilon}$ достаточно мал, а именно

$$\hat{\varepsilon} \lesssim \sqrt{\frac{\mu}{4L}},$$

то для достижения $f(x) - f(x^*) \leq \varepsilon$, `RISTM` с параметром $p \in [1, 2]$ необходимо

$$\begin{aligned} K &= \left\lceil \log_2\left(\frac{\mu R_0^2}{\varepsilon}\right) + 1 \right\rceil \text{ рестартов,} \\ N_{total} &= \left\lceil \left(\frac{L}{\mu}\right)^{\frac{1}{p}} \log_2\left(\frac{\mu R_0^2}{\varepsilon}\right) \right\rceil \text{ оракульных вызовов.} \end{aligned}$$

Обсуждение результатов

- В выпуклом случае алгоритм сходится к значению $\hat{\varepsilon}^2 L R_0^2$, промежуточность p никак не влияет.
- В сильно выпуклом случае сходимость остаётся такой же, как и без шума, при $\hat{\varepsilon} \lesssim \sqrt{\frac{\mu}{4L}}$. Промежуточность вновь не никак влияет.
- Исследован лишь один конкретный алгоритм, хотя оценки в силу PEP являются точными.

Публикации

1. Результаты представлены как часть статьи
Kornilov, N., Gorbunov, E., Alkousa, M., Stonyakin, F., Dvurechensky, P., Gasnikov, A. (2023). Intermediate Gradient Methods with Relative Inexactness. arXiv preprint arXiv:2310.00506.

Литература

Olivier Devolder, François Glineur, Yurii Nesterov, et al.
Intermediate gradient methods for smooth convex problems with inexact oracle.
Technical report, Technical report, CORE-2013017, 2013.

Nikita Kornilov, Eduard Gorbunov, Mohammad Alkousa, Fedor Stonyakin, Pavel Dvurechensky, and Alexander Gasnikov.
Intermediate gradient methods with relative inexactness.
arXiv preprint arXiv:2310.00506, 2023.

Adrien B Taylor, Julien M Hendrickx, and François Glineur.
Smooth strongly convex interpolation and exact worst-case performance of first-order methods.
Mathematical Programming, 161:307–345, 2017.