Об ускоренном спуске по случайному направлению с неевклидовой прокс-структурой

Горбунов Эдуард

Московский Физико-Технический Институт

25 Ноября, 2017

Постановка задачи

Рассматривается задача гладкой выпуклой оптимизации

$$f(x) \to \min_{x \in \mathbb{R}^n}, \tag{1}$$

Постановка задачи

Рассматривается задача гладкой выпуклой оптимизации

$$f(x) \to \min_{x \in \mathbb{R}^n}, \tag{1}$$

где функция $f(x)$, заданная на $\mathbb{R}^n$, имеет градиент, удовлетворяющий условию Гёльдера для некоторого $\nu \in [0, 1]$ с константой L_ν в 2-норме

$$\|\nabla f(y) - \nabla f(x)\|_2 \le L_\nu \|y - x\|_2^\nu, \ \forall x, y \in \mathbb{R}^n.$$

Важные свойства

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1+\nu} ||y - x||_2^{1+\nu}. \tag{2}$$

Важные свойства

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1+\nu} ||y - x||_2^{1+\nu}. \tag{2}$$

Зафиксируем некоторое число $\delta > 0$. Тогда найдётся такая константа L, а именно $L = L_\nu \left[\frac{L_\nu}{2\delta} \cdot \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}}$, что

$$\boxed{f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} ||y - x||_2^2 + \delta} \tag{3}$$

Обозначения

Пусть e — равномерно распределённый случайный вектор на n-мерной евклидовой единичной сфере ($e \in RS_2^n(1)$).

Обозначения

Пусть e — равномерно распределённый случайный вектор на n-мерной евклидовой единичной сфере ($e \in RS_2^n(1)$). Вместо градиента $\nabla f(x)$ метод будет использовать его стохастическую аппроксимацию $n\langle \nabla f(x), e\rangle e$. Можно показать, что $\mathbb{E}_e[n\langle \nabla f(x), e\rangle e] = \nabla f(x)$.

Обозначения

Пусть $d : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая 1-сильно выпуклая по отношению к p-норме (везде далее $1 \leqslant p \leqslant 2$) функция (*прокс-функция*).

Обозначения

Пусть $d : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая 1-сильно выпуклая по отношению к p-норме (везде далее $1 \leqslant p \leqslant 2$) функция (*прокс-функция*). Например, для $p = 1$ можно взять $d(x) = \frac{1}{2(a-1)}||x||_a^2$, где $a = \frac{2\log n}{2\log n - 1}$.

Обозначения

Пусть $d : \mathbb{R}^n \to \mathbb{R}$ — дифференцируемая 1-сильно выпуклая по отношению к p-норме (везде далее $1 \leqslant p \leqslant 2$) функция (*прокс-функция*). Например, для $p = 1$ можно взять $d(x) = \frac{1}{2(a-1)}||x||_a^2$, где $a = \frac{2\log n}{2\log n - 1}$. *Дивергенцией Брегмана* по отношению к прокс-функции d будем называть следующую функцию двух аргументов:

$$V_z(y) \stackrel{\text{def}}{=} d(y) - d(z) - \langle \nabla d(z),\, y - z \rangle. \tag{4}$$

Обозначения

$$\mathrm{Grad}_e(x) \stackrel{\mathrm{def}}{=} x - \frac{1}{L}\langle \nabla f(x), e\rangle e \tag{5}$$

$$\mathrm{Mirr}_e(x, z, \alpha) \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{y \in \mathbb{R}^n} \{\alpha \langle n \langle \nabla f(x), e\rangle e, y - z\rangle + V_z(y)\} \tag{6}$$

Метод

Algorithm 1 ACDS

Вход: f — выпуклая дифференцируемая функция на $\mathbb{R}^n$, удовлетворяющая условию (3); x_0 — некоторая стартовая точка; N — количество итераций.

Выход: точка y_N, удовлетворяющая $\mathbb{E}_{e_1, e_2, \ldots, e_N}[f(y_N)] - f(x^*) \leqslant \frac{4\Theta L C_{n,p}}{N^2} + \frac{2N+3}{16}\delta$.

1: $y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$
2: **for** $k = 0, \ldots, N-1$
3: $\quad \alpha_{k+1} \leftarrow \frac{k+2}{2LC_{n,p}}$, $\tau_k \leftarrow \frac{1}{\alpha_{k+1} L C_{n,p}} = \frac{2}{k+2}$
4: $\quad$ Сгенерировать $e_{k+1} \in RS_2^n(1)$ независимо от предыдущих итераций
5: $\quad x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$
6: $\quad y_{k+1} \leftarrow \mathrm{Grad}_{e_{k+1}}(x_{k+1})$
7: $\quad z_{k+1} \leftarrow \mathrm{Mirr}_{e_{k+1}}(x_{k+1}, z_k, \alpha_{k+1})$
8: **end for**
9: **return** y_N

Сходимость по функции

Теорема

Пусть $f(x)$ — выпуклая дифференцируемая функция на $Q = \mathbb{R}^n$ с градиентом, удовлетворяющим условию Гёльдера для некоторого $\nu \in [0,1]$ с константой L_ν в 2-норме, $d(x)$ — 1-сильно выпуклая в p-норме функция на Q, N – число итераций метода. Тогда ACDS на выходе даст точку y_N, удовлетворяющую неравенству

$$\mathbb{E}_{e_1,e_2,\dots,e_N}[f(y_N)] - f(x^*) \leqslant \frac{4\Theta L C_{n,p}}{N^2} + \frac{2N+3}{16}\delta,$$

где $\Theta = V_{x_0}(x^*)$, $C_{n,p} \leqslant \frac{4}{3}\min\{q-1,\, 4\ln n\} \cdot n^{\frac{2}{q}+1}$.

Замечание

Оказывается, что $C_{n,2} = n^2$ *и* $C_{n,1} \sim n \ln n$.

Параллельные траектории

Предположим, что мы хотим найти такую точку y, что $f(y) - f(x^*) \leqslant 2\varepsilon$. В таком случае можно выбрать $N = \lceil \sqrt{\frac{4\Theta L C_{n,p}}{\varepsilon}} \rceil$, чтобы обеспечить $\mathbb{E}_{e_1,e_2,\dots,e_N}[f(y_N)] - f(x^*) \leqslant \varepsilon \Leftrightarrow \mathbb{E}_{e_1,e_2,\dots,e_N}[f(y_N) - f(x^*)] \leqslant \varepsilon$. По неравенству Маркова

$$\mathbb{P}\{f(y_N) - f(x^*) \geqslant 2\varepsilon\} \leqslant \frac{\varepsilon}{2\varepsilon} = \frac{1}{2}. \tag{7}$$

Параллельные траектории

Предположим, что мы хотим найти такую точку y, что $f(y) - f(x^*) \leqslant 2\varepsilon$. В таком случае можно выбрать $N = \left\lceil \sqrt{\frac{4\Theta L C_{n,p}}{\varepsilon}} \right\rceil$, чтобы обеспечить $\mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N)] - f(x^*) \leqslant \varepsilon \;\Leftrightarrow\; \mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N) - f(x^*)] \leqslant \varepsilon$. По неравенству Маркова

$$\mathbb{P}\{f(y_N) - f(x^*) \geqslant 2\varepsilon\} \leqslant \frac{\varepsilon}{2\varepsilon} = \frac{1}{2}. \tag{7}$$

Это означает, что если запустить $m = \lceil \log_2(\sigma^{-1}) \rceil$ независимых траектории метода ACDS мы получим точки $y_N^1, y_N^2, \ldots, y_N^m$, для которых

$$\mathbb{P}\{\min_{i=\overline{1,m}} f(y_N^i) - f(x^*) \geqslant 2\varepsilon\} \leqslant \left(\frac{1}{2}\right)^m \leqslant \sigma. \tag{8}$$

Обсуждение

В 2014 Z. Allen-Zhu и L. Orrechia предложили ускоренный метод, основанный на идее комбинирования градиентного и зеркального спусков, который после N итераций выдавал точку y_N, удовлетворяющую

$$f(y_N) - f(x^*) \leqslant \frac{4\Theta L}{N^2}. \tag{9}$$

Обсуждение

В 2014 Z. Allen-Zhu и L. Orrechia предложили ускоренный метод, основанный на идее комбинирования градиентного и зеркального спусков, который после N итераций выдавал точку y_N, удовлетворяющую

$$f(y_N) - f(x^*) \leqslant \frac{4\Theta L}{N^2}. \tag{9}$$

В случае $p = 1$ алгоритму ACDS нужно примерно в $\sim \frac{n}{\ln n}$ раз меньше арифметических операций в предположении, что $f(x)$ задаётся моделью чёрного ящика и её градиент восстанавливается по $n + 1$ значению функции $f(x)$.

Безградиентный ускоренный метод

Предположим, что значения функции известны с некоторым шумом:

$$\begin{gathered} \tilde{f}(x) = f(x) + \delta(x), \\ \forall x \in \mathbb{R}^n \hookrightarrow |\delta(x)| \leqslant \delta. \end{gathered} \tag{10}$$

Безградиентный ускоренный метод

Предположим, что значения функции известны с некоторым шумом:

$$\begin{gathered}\tilde{f}(x) = f(x) + \delta(x),\\ \forall x \in \mathbb{R}^n \hookrightarrow |\delta(x)| \leqslant \delta.\end{gathered} \quad (10)$$

Кроме того, теперь будем использовать аппроксимацию производной по направлению:

$$\langle \nabla f(x), e \rangle e \rightsquigarrow \frac{\tilde{f}(x+te) - \tilde{f}(x)}{t} e. \quad (11)$$

Безградиентный ускоренный метод

Теорема

Пусть $f(x)$ — выпуклая дифференцируемая функция на $Q = \mathbb{R}^n$ с константой Липшица для градиента, равной L, $d(x)$ — 1-сильно выпуклая в p-норме функция на Q, $N \in \mathbb{N}$. Тогда ACDS на выходе даст точку y_N, удовлетворяющую неравенству

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{16\Theta L C_{n,p}}{N^2} + \frac{7(2N+3)\delta}{4} + \frac{16\sqrt{2\Theta nL\delta}}{N^2} + \frac{8nN^2\delta}{C_{n,p}}.$$

где $\Theta = V_{x_0}(x^*)$, $C_{n,p} = \frac{4}{3}\min\{q-1,\, 4\ln n\} \cdot n^{\frac{2}{q}+1}$.

Безградиентный ускоренный метод

Теорема

Пусть $f(x)$ — выпуклая дифференцируемая функция на $Q = \mathbb{R}^n$ с константой Липшица для градиента, равной L, $d(x)$ — 1-сильно выпуклая в p-норме функция на Q, $N \in \mathbb{N}$. Тогда ACDS на выходе даст точку y_N, удовлетворяющую неравенству

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{16\Theta L C_{n,p}}{N^2} + \frac{7(2N+3)\delta}{4} + \frac{16\sqrt{2\Theta n L\delta}}{N^2} + \frac{8nN^2\delta}{C_{n,p}}.$$

где $\Theta = V_{x_0}(x^*)$, $C_{n,p} = \frac{4}{3}\min\{q-1,\, 4\ln n\} \cdot n^{\frac{2}{q}+1}$.

Замечание

Оказывается, что для получение решения с точностью ε нужно обеспечивать $\delta \sim \frac{\varepsilon^2}{n}$.

Безградиентный неускоренный метод

Рассмотрим метод

$$x_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} \{\alpha \langle n \langle \widetilde{\nabla} f(x_k), e_{k+1} \rangle e_{k+1}, y - x_k \rangle + V_{x_k}(y)\}.$$

Безградиентный неускоренный метод

Теорема

При сделанных предположениях о функции f неускоренная версия ACDS через N итераций выдаёт точку $\bar{x}$, удовлетворяющую неравенству

$$\mathbb{E}[f(\bar{x})] - f(x^*) \leqslant \frac{16\Theta L C_{n,p}}{nN} + \frac{8\sqrt{2n\Theta L\delta}}{N} + 3n\delta + \frac{8n^2 N\delta}{C_{n,p}} \qquad (12)$$

Безградиентный неускоренный метод

Теорема

При сделанных предположениях о функции f неускоренная версия ACDS через N итераций выдаёт точку $\bar{x}$, удовлетворяющую неравенству

$$\mathbb{E}[f(\bar{x})] - f(x^*) \leqslant \frac{16\Theta L C_{n,p}}{nN} + \frac{8\sqrt{2n\Theta L\delta}}{N} + 3n\delta + \frac{8n^2 N\delta}{C_{n,p}} \tag{12}$$

Замечание

В неускоренном случае тоже приходится обеспечивать $\delta \sim \frac{\varepsilon^2}{n}$.