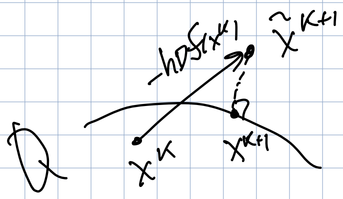
Лекция 3. Методы проекции градиента

$$\min_{x\in\mathbb{R}^n} f(x) \to \text{вып. ф-ция}$$

GD: $$x^{k+1} = x^k - h\nabla f(x^k) =$$
$$= \operatorname{argmin}_{x\in\mathbb{R}^n}\left\{ f(x^k) + \langle \nabla f(x^k), x-x^k\rangle + \frac{1}{2h}\|x-x^k\|_2^2 \right\}$$

$$\min_{x\in Q \to \text{вып. мн-во}} f(x) \to \text{вып. ф-ция}$$

PGD: $$x^{k+1} = \pi_Q(\underbrace{x^k - h\nabla f(x^k)}_{\tilde{x}^{k+1}}) =$$
$$= \operatorname{argmin}_{x\in Q}\left\{ f(x^k) + \langle \nabla f(x^k), x-x^k\rangle + \frac{1}{2h}\|x-x^k\|_2^2 \right\}$$



Q – простая структура

$$\min_{x\in Q} f(x) \to \text{вып.}, \to \text{лип.}$$

$$|f(y)-f(x)| \le M_p\|y-x\|_p,\quad p\in[1,2]$$

$\|\cdot\|_p$. $\quad d(x)$ – прокс. ф-ция:

1-сильно-выпуклая отн. p-нормы

$$V(y,x) = d(y) - d(x) - \langle \nabla d(x), y-x \rangle$$

дивергенция

$$d(x) = \frac{1}{2}\|x\|_2^2$$

1-сильно вып. отн. 2-нормы

$$V(y,x) = \frac{1}{2}\|y-x\|_2^2$$

$$x^{k+1} = \operatorname{argmin}_{x\in Q} \left\{ f(x^k) + \langle \nabla f(x^k), x-x^k \rangle + \cancel{\frac{1}{2h}\|x-x^k\|_2^2} \right\}$$

$$\frac{1}{h}V(x,x^k)$$

$$\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$$

§2 Гасников МЦНМО

$$f(\bar{x}^N) - f(x_*) \le \frac{M_p R_p}{\sqrt{N}}$$

$$\sqrt{2V(x_*,x^0)}$$

$p=2$

$d(x) = \frac{1}{2}\|x\|_2^2$

$V(y,x) = \frac{1}{2}\|y-x\|_2^2$

$M_p \to |f(y)-f(x)| \le$

$\le M\|y-x\|_2$

$R_p = \|x^0 - x_*\|_2$

Пример (оптимизация на симплексе)

$$Q = S_n(1) = \left\{ x \in \mathbb{R}^n_+ : \sum_{i=1}^{n} x_i = 1 \right\}$$

плотная

$$A = n\underbrace{\left[\begin{matrix} \vdots \end{matrix}\right]}_{n}$$

$$\min_{x\in S_n(1)} \left\{ \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle \right\}$$

$$d(x) = \sum_{i=1}^{n} x_i \ln x_i, \quad x^0 = (1/n, \ldots, 1/n)$$

$$V(y,x) = \sum_{i=1}^{n} y_i \ln(y_i/x_i)$$

$$x^{k+1} = \underset{x\in S_n(1)}{\operatorname{argmin}} \{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle +$$
$$+ \frac{1}{h}\sum_{i=1}^{n} x_i \ln(x_i/x_i^k) \}$$

$$h = \frac{\varepsilon}{M_p^2}$$

$$x_i^{k+1} = \frac{x_i^k \exp(-h\, \partial f(x^k)/\partial x_i)}{\sum_{j=1}^{n} x_j^k \exp(-h\, \partial f(x^k)/\partial x_j)}$$

$\nabla f(x^k) \longrightarrow O(n^2)$

стоимость "проектирования"

$\downarrow$

$O(n)$

$$M_p = \max_{x\in S_n(1)} \|\nabla f(x)\|_q \qquad \frac{1}{p}+\frac{1}{q}=1$$

$\downarrow$

$p=1$

$q=\infty$

$$\max_{x\in S_n(1)} \|Ax-b\|_\infty \leq \max_{i,j} |A_{ij}| + \max_i |b_i|$$

$$R_p = \sqrt{2V(x_0, x^0)} \leq \max_{x\in S_n(1)} \sqrt{2\sum_{i=1}^{n} x_i \ln(x_i/(1/n))} \leq$$
$$\leq \sqrt{2\ln n}$$

$x_0 = (1, 0, \dots, 0)$

$$M_1 R_1 = \underbrace{\left(\max_{i,j} |A_{ij}| + \max_i |b_i|\right)}_{O(1)} \sqrt{2\ln n}$$

$p=2$

$$M_2 = \max_{x\in S_n(1)} \|Ax-b\|_\infty \simeq \max_{k=1,n} \|A^{\langle k\rangle}\|_2 + \|b\|_2$$
$$\gtrsim \Omega(\sqrt{n})$$

$$R_2 = \max_{x\in S_n(1)} \|x^0 - x\|_2 \simeq O(1)$$

$M_2 R_2 \sim \sqrt{n}$ vs $M_1 R_1 \sim \sqrt{\ln n}$

Глава 1 книги по стох. оптим

Стох. оптимизация

$$\min_{x \in Q} f(x)$$

$|f(y) - f(x)| \le$
$\le M \|y - x\|$

$$x^{k+1} = \arg\min_{x \in Q} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{h} V(x, x^k) \right\}$$

$h = \varepsilon / M^2$

$d(x)$
1-сильно вып.
отн. $\|\cdot\|$

$\nabla f(x)$ — не доступен

Например, $Ax - b \to$ град.
A — огром. разм.

$\nabla f(x, \xi)$ — стох. град.

1) $\mathbb{E}_{\xi} \nabla f(x, \xi) \equiv \nabla f(x)$, для всех $x \in Q$

2) $\mathbb{E}\left[\|\nabla f(x, \xi)\|_*^2 \right] \le M^2$

$\|\cdot\|_*$ - сопряж. норма к $\|\cdot\|$, т.е.

$$\|y\|_* = \sup_{\|x\| \le 1} \langle y, x \rangle$$

стохастическ. зеркальный спуск

$$x^{k+1} = \arg\min_{x \in Q} \left\{ f(x^k) + \langle \nabla f(x^k, \xi^k), x - x^k \rangle + \frac{1}{h} V(x, x^k) \right\}$$

Пример $Q = \mathbb{R}^n$: $x^{k+1} = x^k - h \nabla f(x^k, \xi^k)$, ξ^k i.i.d.
SGD

$$\mathbb{E}[f(\bar{x}^N)] - f(x_*) \le \frac{MR}{\sqrt{N}}$$

Немировский и др. 2008

$$R = \sqrt{2V(x_*, x^0)}$$

Пример (оптимизация на симплексе)

$$Q = S_n(1) = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$$

плотная

$$A = n\underbrace{\begin{bmatrix} \cdots \\ \cdots \\ \cdots \end{bmatrix}}_{n} \qquad p = 1 \qquad \min_{x \in S_n(1)} \{\underbrace{\tfrac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle}_{f(x)}\}$$

$$d(x) = \sum_{i=1}^n x_i \ln x_i, \qquad x^0 = (1/n, \dots, 1/n)$$

$$x_i^{k+1} = \frac{x_i^k \exp(-h\partial f(x^k, \xi^k)/\partial x_i)}{\sum_{j=1}^n x_j^k \exp(-h\partial f(x^k, \xi^k)/\partial x_j)}$$

если $\nabla f(x^k, \xi^k)$ известен, то проекц. стоит $O(n)$

$$\nabla f(x, \xi) = A^{\langle \xi \rangle} - b, \text{ где}$$

вычисление стоит $O(n)$

$A^{\langle \xi \rangle}$ – это ξ-ый столбец матрицы A

$$\mathbb{P}(\xi = i) = x_i, \quad i = \overline{1, n}. \quad // \; x \in S_n(1)$$

1) $\mathbb{E}_\xi[\nabla f(x, \xi)] \overset{?}{\equiv} \nabla f(x) = Ax - b$

$$\mathbb{E}_\xi[\nabla f(x,\xi)] = \sum_{i=1}^n (A^{\langle i \rangle} - b)\mathbb{P}(\xi = i) = \sum_{i=1}^n A^{\langle i \rangle} x_i - b\sum_{i=1}^n x_i = Ax - b$$

// !

$$Ax = [A^{\langle 1\rangle}, \ldots, A^{\langle n\rangle}] \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} =$$

$$= A^{\langle 1\rangle} x_1 + \ldots + A^{\langle n\rangle} x_n$$

2) $\mathbb{E}_\xi \| \nabla f(x, \xi) \|_\infty^2 \le M^2$ $\quad // \; p = 1, \; q = \infty, \; \frac{1}{p} + \frac{1}{q} = 1$

$$\downarrow$$

$$?$$

$$\max_{\xi = 1, \ldots, n} \| A^{\langle \xi \rangle} - b \|_\infty$$

$$\wedge\!\!|$$

$$\underbrace{\max_{i,j} |A_{ij}| + \max_i |b_i|}_{O(1)}$$

$$M = O(1)$$

$$R = O(\sqrt{\ln n})$$

$$\mathbb{E} f(\bar{x}^N) - f(x_*) \le \frac{MR}{\sqrt{N}} = O\left(\frac{\sqrt{\ln n}}{\sqrt{N}}\right)$$

$$\# \nabla f(x^k, \xi^k) \to O(n)$$

$$\mathbb{E} f(\bar{x}^N) - f(x_*) \le \varepsilon$$

$$N \sim \frac{\ln n}{\varepsilon^2}$$

$$T = N \cdot [\underbrace{\# \nabla f(x^k, \xi^k) + \text{проекция}}_{O(n)}] =$$

$$= O\left(\frac{n \ln n}{\varepsilon^2}\right).$$

§2.6 Воронцова и др. Выпукл. оптим.

Метод условного градиента

(Франк-Вульф)

$$x^{k+1} = \arg\min_{x\in Q} \{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \cancel{\tfrac{1}{h}\|x-x^k\|_2^2} \}$$

$\cancel{\tfrac{1}{h} V(x, x^k)}$

$$x^{k+1} - x^k = \gamma_k (y^k - x^k)$$

$$\begin{cases} x^{k+1} = (1-\gamma_k) x^k + \gamma_k y^k, \quad \gamma_k = \dfrac{2}{k+1} \\ y^k = \arg\min_{y\in Q} \{ f(x^k) + \langle \nabla f(x^k), y - x^k \rangle \} = \\ \qquad = \arg\min_{y\in Q} \langle \nabla f(x^k), y \rangle \qquad (*) \end{cases}$$

Теорема. Пусть 1) $\|\nabla f(y) - \nabla f(x)\|_\infty \le L \|y-x\|_1$ $\quad q \quad \frac{1}{p}+\frac{1}{q}=1 \quad p$

2) $R = \sup_{x,y\in Q} \|y-x\|_1$. Тогда

$$f(x^N) - f(x_*) \le \frac{2LR^2}{N+2}$$

для ускор. метода Нестерова $\sim \frac{1}{N^2}$

Д-во. Из 1) $\forall x, y \in Q$

$$0 \le f(y) - f(x) - \langle \nabla f(x), y-x \rangle \le \frac{L}{2}\|y-x\|_1^2$$

вып. f $\qquad$ 1)

$$f(x^{k+1}) - f(x^k) \le f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|_1^2 =$$

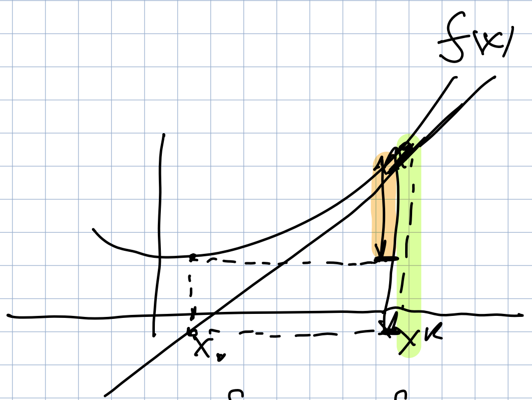
$$= \gamma_k \langle \nabla f(x^k), y^k - x^k \rangle + \frac{L}{2}\gamma_k^2 \underbrace{\|y^k - x^k\|_1^2}_{\le R^2} \le$$

$$\leq \gamma_k \langle \nabla f(x^k), x_* - x^k \rangle + \gamma_k^2 \frac{LR^2}{2}$$

(*)

$$\leq \gamma_k (f(x_*) - f(x^k)) + \gamma_k^2 \frac{LR^2}{2}$$

выпукл. f



$$\langle \nabla f(x^k), x_* - x^k \rangle \leq f(x_*) - f(x^k)$$

$$f(x^k) - f(x_*) \leq \langle \nabla f(x^k), x^k - x_* \rangle$$

$$f(x^{k+1}) - f(x^k) \leq \gamma_k (f(x_*) - f(x^k)) + \gamma_k^2 \frac{LR^2}{2}$$

$$\delta_k = \frac{f(x^k) - f(x_*)}{LR^2} \leq \frac{2}{k+2}$$

$$\delta_{k+1} \leq (1 - \gamma_k)\delta_k + \frac{\gamma_k^2}{2} = \left(1 - \frac{2}{k+1}\right)\delta_k + \frac{2}{(k+1)^2}$$

$k = 1$ $\;\; = 0$

$$\delta_2 \leq \frac{2}{2^2} = \frac{1}{2}$$

По индукции док-ть, что $\delta_k \leq \frac{2}{k+2}$

Снова рассмотрим пример

Гасников МЗНМО упр. 1.6

$$\min_{x \in S_n(1)} \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$$

$$\nabla f(x) = Ax - b$$

(*) $$\min_{y \in S_n(1)} \langle \nabla f(x^k), y\rangle \Rightarrow$$

$$\min_{\substack{\sum y_i = 1 \\ y_i \geq 0}} \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^k)\, y_i$$

$$i_0(k) = \arg\min_{i=\overline{1,n}} \frac{\partial f}{\partial x_i}(x^k)$$

$O(n)$

$$y^k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{array}{l} \\ \\ \\ \to i_0(k) \\ \\ \\ \\ \end{array}$$

$$x^{k+1} = (1-\gamma_k)x^k + \gamma_k y_k$$

$$x^k = (1-\gamma_1)(1-\gamma_2)\ldots(1-\gamma_{k-1})\, z^k$$

$$\underbrace{(1-\gamma_1)\ldots(1-\gamma_k)\, z^{k+1}} = \underbrace{(1-\gamma_1)\ldots(1-\gamma_{k-1})\cdot(1-\gamma_k)}\, z^k + \gamma_k y_k$$

$$z^{k+1} = z^k + \underbrace{\frac{\gamma_k}{\prod_{i=1}^{k}(1-\gamma_i)}}_{\beta_k \to \alpha_k} \cdot y_k = z^{k+1} + \alpha_k y_k.$$

$$\beta_{k+1} = \beta_k (1 - \gamma_{k+1})$$

$$\alpha_{k+1} = \frac{\gamma_{k+1} \cdot \alpha_k}{\gamma_k (1 - \gamma_{k+1})}$$

$$\nabla f(x^k) = A x^k - b = A \beta_k z^k - b =$$

$$= \beta_k \underbrace{A z^k} - b$$

$$A z^{k+1} = A (z^k + \alpha_k y^k) =$$

$$= A z^k + \alpha_k A y^k = \underbrace{A z^k + \alpha_k A^{\langle i_k \rangle}}_{O(n)}$$

Стоимость итерации $O(n)$!

$$f(x^N) - f(x_*) \leq \varepsilon \to N \simeq \frac{2LR^2}{\varepsilon} = O\left(\frac{1}{\varepsilon}\right)$$

$$L = \max_{\substack{\|h\|_1 \leq 1 \\ x \in S_n(1)}} \langle h, \underbrace{\nabla^2 f(x)}_{A} h \rangle = \max_{i,j} |A_{ij}| = O(1)$$

$$R^2 = \max_{x, y \in S_n(1)} \|y - x\|_1 \leq 2$$

Стоимость итерации $O(n)$ 1-я итерация $\uparrow$

Число итераций $O\left(\frac{1}{\varepsilon}\right)$

Общие трудозатраты $= O\left(\frac{n}{\varepsilon}\right) + O(n^2)$

А если использ. оптим. метод Нестеров

Стоимость итерации $O(n^2)$

$$\nabla f(x) = Ax - b$$

число итераций $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ //

$$f(x^N) - f(x_*) \le \frac{4LR^2}{N^2}$$

Общие трудозатраты $= O\left(\frac{n^2}{\sqrt{\varepsilon}}\right)$

ускор. метод Нестерова

$O\left(\frac{n^2}{\sqrt{\varepsilon}}\right)$

vs

Франк-Вульф

$O\left(n^2 + \frac{n}{\varepsilon}\right)$

SGD

$O\left(\frac{n \ln n}{\varepsilon^2}\right)$

Лекция 10 Задачи с аффинными ограничениями

§4 пособия МФТИ Гасников

$$\min_{Ax=b} f(x) = \sum_{i=1}^{n} f_i(x_i)$$

$x \in Q$?

$x \in \mathbb{R}^n$