Градиентный спуск

$$x^{k+1} = x^k - \gamma \nabla f(x^k)$$

$$x \in (E, \|\cdot\|) \Rightarrow \nabla f(x) \in ? \quad \notin (E, \|\cdot\|)$$

$$\in (E_*, \|\cdot\|_*)$$

Пример $\|\cdot\| = \|\cdot\|_2 \Rightarrow (E_*, \|\cdot\|_*) = (E, \|\cdot\|)$

- А. Немировский и Д. Юдина:

$$\varphi(x^{k+1}) = \varphi(x^k) - \gamma \nabla f(x^k)$$

подобрать $\varphi: E \to E_*$, $\varphi^{-1}: E_* \to E$

Шаг град. спуска в сопряж. пр-ве = „двойственность"

Опр. Непр. дифф. функция $d: X \to \mathbb{R}$ d – μ-сильно выпуклая отн $\|\cdot\|$ на мн-ве X, если

$$\forall x, y \in X \hookrightarrow d(x) \geq d(y) + \langle \nabla d(y); x - y \rangle + \frac{\mu}{2}\|x-y\|^2$$

Опр. (дивергенция Брэгмана) 1 – сильно выпуклая отн $\|\cdot\|$ на X, непр. дифф. функция d. Дивергенция Брэгмана, порожденная d на X есть функция $V(x,y): X \times X \to \mathbb{R}$:

$$\forall x, y \in X \qquad V(x,y) = d(x) - d(y) - \langle \nabla d(y); x - y \rangle$$

Примеры

- $d(x) = \frac{1}{2}\|x\|_2^2$ на $\mathbb{R}^d$

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

- $d(x) = \sum_{i=1}^{d} x_i \log x_i$ (энтропия) на $\Delta_d = \{x \in \mathbb{R}^d \mid \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$

$$V(x,y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i} \quad \text{(KL-дивергенция)}$$

Свойства

- асимметричность (см. KL-дивергенцию)
- сильная выпуклость $V(x,y) \geq \frac{1}{2}\|x-y\|^2 \quad \forall x,y \in X$ (из определения)
- неотриц.
- невыпуклость по 2-ому аргументу
- Теорема Пифагора $\forall x,y,z \in X \hookrightarrow$

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

Док-во: Из опред. V

$$V(z,x) + V(x,y) = d(z) - d(x) - \langle \nabla d(x); z-x \rangle$$
$$+ d(x) - d(y) - \langle \nabla d(y); x-y \rangle$$
$$= d(z) - d(y) - \langle \nabla d(y); z-y \rangle = V(z,y)$$
$$\langle \nabla d(y) - \nabla d(x); z-x \rangle \quad \blacksquare$$

$$\min_{x \in X} f(x) \qquad f \text{ - выпукла}, \quad X \text{ - выпуклое}$$

- Метод зеркального спуска:

$$\boxed{x^{k+1} = \operatorname*{argmin}_{x \in X} \{\langle \gamma \nabla f(x^k); x \rangle + V(x, x^k)\}}$$

argmin отыскивается явно

Примеры

- $d(x) = \frac{1}{2}\|x\|_2^2 \qquad X = \mathbb{R}^d$

$$x^{k+1} = \arg\min_{x \in X} \Big\{ \underbrace{\langle \gamma \nabla f(x^k); x \rangle + \frac{1}{2}\|x - x^k\|_2^2}_{\nabla = 0} \Big\}$$

$$\gamma \nabla f(x^k) + \underset{\substack{\uparrow \\ x^{k+1}}}{x} - x^k = 0$$

$$x^{k+1} = x^k - \gamma \nabla f(x^k) \quad \leftarrow \text{градиентный спуск}$$

- $d(x) = \frac{1}{2}\|x\|_2^2 \qquad X$ – выпуклое

$$x^{k+1} = \arg\min_{x \in X} \Big\{ \gamma \langle \nabla f(x^k); x \rangle + \frac{1}{2}\|x - x^k\|_2^2 \Big\}$$

$$+ \frac{\gamma^2}{2}\|\nabla f(x^k)\|_2^2 - \gamma \langle \nabla f(x^k); x^k \rangle$$

$$= \arg\min_{x \in X} \Big\{ \frac{1}{2}\Big(\|x - x^k\|_2^2 + 2\gamma \langle \nabla f(x^k); x - x^k \rangle + \gamma^2 \|\nabla f(x^k)\|_2^2 \Big) \Big\}$$

$$= \arg\min_{x \in X} \Big\{ \frac{1}{2}\| x - x^k + \gamma \nabla f(x^k)\|_2^2 \Big\}$$

$$= \arg\min_{x \in X} \Big\{ \|x - y^k\|_2^2 \Big\} \qquad y^k = x^k - \gamma \nabla f(x^k)$$

град. спуск с проекцией

- $d(x)$ – произвольно $\qquad X = \mathbb{R}^d$

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \Big\{ \underbrace{\langle \gamma \nabla f(x^k); x \rangle + d(x) - d(x^k) - \langle \nabla d(x^k); x - x^k \rangle}_{\nabla = 0} \Big\}$$

$$\gamma \nabla f(x^k) + \underset{\substack{\uparrow \varphi \\ x^{k+1}}}{\nabla d(x)} - \nabla d(x^k) = 0$$

$$\nabla d(x^{k+1}) = \nabla d(x^k) - \gamma \nabla f(x^k)$$

$$\varphi(x^{k+1}) = \varphi(x^k) - \gamma \nabla f(x^k)$$

- $d(x) = \sum_{i=1}^{d} x_i \log x_i$ на Δ_d

$$x^{k+1} = \operatorname*{argmin}_{x \in \Delta_d} \{ \langle \gamma \nabla f(x^k); x \rangle + V(x, x^k) \}$$

$$V(x,y) = \sum_{i=1}^{d} x_i \log\left(\frac{x_i}{y_i}\right)$$

Формальная запись задачи:

$$\min_{x \in \mathbb{R}^d} \langle \gamma \nabla f(x^k); x \rangle + \sum_{i=1}^{d} x_i \log\left(\frac{x_i}{x_i^k}\right)$$

$$\text{s.t.} \quad -x_i \le 0$$

$$\sum_{i=1}^{d} x_i - 1 = 0$$

Лагранжиан:

$$L(x, \lambda, \nu) = \langle \gamma \nabla f(x^k), x \rangle + \sum_{i=1}^{d} x_i \log \frac{x_i}{x_i^k} + \sum_{i=1}^{d} \lambda_i (-x_i) + \nu\left(\sum_{i=1}^{d} x_i - 1\right)$$

$$= \left[\sum_{i=1}^{d} \Big(\underbrace{\gamma [\nabla f(x^k)]_i + \log \frac{x_i}{x_i^k} + \nu - \lambda_i}_{q_i} \Big) x_i \right] - \nu$$

Минимизируем по x_i (ищем двойств.)

$$\inf_x L(x, \lambda, \nu)$$

$$\left(a_i + \log\frac{x_i}{x_i^k}\right)x_i \longrightarrow \inf_{x_i} \quad \Rightarrow \quad \nabla_{x_i} = 0$$

$$\left(a_i + \log\frac{x_i^*}{x_i^k}\right) + \cancel{x_i^*}\cdot\frac{\cancel{x_i^k}}{\cancel{x_i^*}}\cdot\frac{1}{\cancel{x_i^k}} = 0$$

$$a_i + \log\frac{x_i^*}{x_i^k} + 1 = 0 \quad \Rightarrow \quad x_i^* = x_i^k \exp(-1-a_i)$$

$$\inf_x L(x,\lambda,\nu) = \left[\sum_{i=1}^d \left(a_i + \log\frac{x_i^*}{x_i^k}\right)x_i^*\right] - \nu$$

$$= \left[\sum_{i=1}^d - x_i^*\right] - \nu$$

$$= \left[\sum_{i=1}^d - x_i^k \exp(-1-a_i)\right] - \nu$$

$$= \left[\sum_{i=1}^d - x_i^k \exp\left(-1+\lambda_i-\nu-\gamma[\nabla f(x^k)]_i\right)\right] - \nu$$

Максимизируем по $\lambda_i \geq 0$ $\quad \nu \in R$

$$\lambda_i^* = 0$$

$$x_i^* = x_i^k \exp(-1-a_i)$$

$$= x_i^k \exp\left(-1-\nu-\gamma[\nabla f(x^k)]_i\right)$$

$$= x_i^k \exp\left(-\gamma[\nabla f(x^k)]_i\right)\cdot \exp(-1-\nu)$$

ν^*

как подобрать?

ν^* найдём из соображен. $\sum x_i^* = 1$

$\exp(-1-\nu)$ – нормировка

$$\boxed{x_i^* = \frac{x_i^k \exp\left(-\gamma [\nabla f(x^k)]_i\right)}{\sum\limits_{j=1}^{d} x_j^k \exp\left(-\gamma [\nabla f(x^k)]_j\right)}}$$

Опр. Непр. дифферен. функция на X, $f: X \to \mathbb{R}$

f является L-гладкой отн. $\|\cdot\|$ на X, если

$$\forall x, y \in X \hookrightarrow \|\nabla f(x) - \nabla f(y)\|_* \le L \|x - y\|$$

Теорема f − L-гладкая отн. $\|\cdot\|$ на X

$$\forall x, y \in X \hookrightarrow |f(y) - f(x) - \langle \nabla f(x); y - x \rangle| \le \frac{L}{2} \|y - x\|^2$$

Док-во:

$$|\langle \nabla f(x + \tau(y - x)) - \nabla f(x); y - x \rangle|$$

нер-во Гельдера $\langle a; b \rangle \le \|a\|_* \|b\|$

$$\le \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_* \|y - x\|$$

L-гладкость

$$\le L\tau \|y - x\|^2 \quad \blacksquare$$

Док-во сходимости:

$$x^{k+1} = \underset{x \in X}{\operatorname{argmin}} \left\{ \langle \gamma \nabla f(x^k); x \rangle + V(x, x^k) \right\}$$

$$V(x, x^k) = d(x) - d(x^k) - \langle \nabla d(x^k); x - x^k \rangle$$

$\forall y \in X \quad \langle \nabla^*, x^* - y \rangle \le 0$ (условие оптим.)

$$\langle \gamma \nabla f(x^k) + \nabla d(\underset{x^{k+1}}{x}) - \nabla d(x^k); \underset{x^{k+1}}{x} - \underset{x^*}{y} \rangle \le 0$$

$$\langle \gamma \nabla f(x^k) + \nabla d(x^{k+1}) - \nabla d(x^k); x^{k+1} - x^* \rangle \le 0$$

св-ва див Брэгмана:

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

$x = x^{k+1}$ $\quad y = x^k$ $\quad z = x^*$

$$\langle \gamma \nabla f(x^k); x^{k+1} - x^* \rangle + V(x^*; x^{k+1}) - V(x^*, x^k) + V(x^{k+1}; x^k) \le 0$$

$$\gamma \langle \nabla f(x^k); x^{k+1} - x^* \rangle \le V(x^*; x^k) - V(x^*, x^{k+1}) - V(x^{k+1}, x^k)$$

Небольшие преобр.

$$\gamma \langle \nabla f(x^k); x^k - x^* \rangle \le V(x^*; x^k) - V(x^*, x^{k+1}) - V(x^{k+1}, x^k) + \gamma \langle \nabla f(x^k); x^k - x^{k+1} \rangle$$

L – гладкость

$$f(y) - f(x) \le \frac{L}{2}\|x - y\|^2 + \langle \nabla f(x); y - x \rangle$$

$x = x^k$ $\quad y = x^{k+1}$

$$\gamma \cdot \Big| \; f(x^{k+1}) - f(x^k) \le \frac{L}{2}\|x^k - x^{k+1}\|^2 + \langle \nabla f(x^k); x^{k+1} - x^k \rangle$$

Складываем

$$\gamma\left(f(x^{k+1}) - f(x^k)\right) + \gamma \langle \underline{\nabla f(x^k); x^k - x^*} \rangle \le V(x^*; x^k) - V(x^*, x^{k+1}) - V(x^{k+1}, x^k) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2$$

Выпуклость:

$$\gamma \cancel{f(x^k)} - \gamma f(x^*) + \gamma f(x^{k+1}) - \cancel{\gamma f(x^k)} \le V(x^*; x^k) - V(x^*, x^{k+1}) - V(x^{k+1}, x^k) + \frac{\gamma L}{2}\|x^k - x^{k+1}\|^2$$

Свойство дивергенции: $\frac{1}{2}\|x^k - x^{k+1}\|^2 \le V(x^{k+1}; x^k)$

$$\gamma\left(f(x^{k+1}) - f(x^*)\right) \le V(x^*; x^k) - V(x^*, x^{k+1}) - (1-\gamma L)\underbrace{V(x^{k+1}, x^k)}_{\ge 0}$$

$\gamma \le \frac{1}{L}$

$$\gamma\left(f(x^{k+1}) - f(x^*)\right) \le V(x^*; x^k) - V(x^*, x^{k+1})$$

$\frac{1}{K}\sum_{k=0}^{K-1}$ и Йенсен:

$$\gamma\left(f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^{k+1}\right) - f(x^*)\right) \le \frac{V(x^*; x^0) - \cancel{V(x^*; x^K)}}{K}$$

$\gamma = \frac{1}{L}$

$$\boxed{f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^{k+1}\right) - f(x^*) \le \frac{L\, V(x^*, x^0)}{K}} \quad \blacksquare$$

Град. спуска: $O\left(\frac{L_2 \|x^* - x^0\|_2^2}{K}\right)$

$\odot$ как и град спуск с точки зрения K

$\oplus$ $L \le L_2$ $\quad \|\nabla f(x) - \nabla f(y)\|_q \le L \|x-y\|_p \quad \frac{1}{p} + \frac{1}{q} = 1$

$\uparrow$

$p \in [1;2] \rightarrow q \in [2;+\infty]$

$\downarrow$

$\|\cdot\|_2 \le \|\cdot\|_p \qquad \|\cdot\|_q \le \|\cdot\|_2$

$\ominus$ $p \in [1;2] \quad V(x,y) \ge \frac{1}{2}\|x-y\|_p^2 \ge \frac{1}{2}\|x-y\|_2^2$

$\oplus$ на симплексе может выиграть до $\frac{d}{\log d}$ раз.