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. Юдин:

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

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

Идея „зеркальности" = шаг град. спуска с сопряж. пр-ве

Опр. Непр. дифф. функция $d : X \to \mathbb{R}$

d — μ-сильно выпукла отн. $\|\cdot\|$ на мн-ве X, если

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

Опр. (дивергенция Брэгмана)

1-сильно выпуклая отн. $\|\cdot\|$ на X, непр. дифф. функция d.

Дивергенция Брэгмана, порожденная d на X, есть функция $V(x, y) : X \times X \to \mathbb{R}$:

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

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 $\Delta_d = \{x \in \mathbb{R}^d \mid \sum x_i = 1,\ x_i \ge 0\}$

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

Свойства

- ассиметричность (см. KL-див.)
- сильная выпуклость: $V(x,y) \ge \frac{1}{2}\|x-y\|^2 \quad \forall x,y \in X$
  (из определения $V(x,y)$)
- неотриц.
- невыпуклость по 2-у аргументу
- Теорема Пифагора $\forall x,y,z \in X$

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

Док-во:

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

$$\min_{x \in X} f(x) \qquad \begin{array}{l} f \text{ - выпуклая} \\ X \text{ - выпуклое} \end{array}$$

- Метод зеркального спуска:

$$\boxed{x^{k+1} = \arg\min_{x \in X} \{ \langle \gamma \nabla f(x^k);\ x \rangle + V(x;x^k) \}}$$

Пример

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

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

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

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

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

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

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

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

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

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

град. спуск с евклидовой проекцией

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

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^d} \underbrace{\left\{\langle \gamma \nabla f(x^k); x\rangle + \underset{d(x) - \cancel{d(x^k)} - \langle \nabla d(x^k); x - \cancel{x^k}\rangle}{\underset{\parallel}{V(x, x^k)}}\right\}}_{\nabla = 0}$$

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

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

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

<u>Опр.</u> Непр. дифф. функцию на X $\quad f: X \to \mathbb{R}$

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

$$\forall x, y \in X \;\Leftrightarrow\; \|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$$

<u>Теорема</u> f – L-гладкая отн. $\|\cdot\|$ на X

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

<u>Док-во:</u>

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

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

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

L-гладкость

$$\leq L\,\tau\,\|y-x\|^2$$

<u>Док-во сходимости:</u>

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

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

$$\langle \nabla^*; x^* - y\rangle \leq 0 \quad \forall y \in X$$

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

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

Св-во див. Брэгмана:

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

$$\langle \nabla d(x^{k+1}) - \nabla d(x^k); x^{k+1}-x^*\rangle \qquad \Big|\; x=x^{k+1} \quad y=x^k \quad z=x^*$$

$$= V(x^*, x^{k+1}) + V(x^{k+1}; x^k) - V(x^*, x^k)$$

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

Небольшие преобр-е

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

L- гладкость

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

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

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

складываем

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

Выпуклость

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

св-во дивергенции Брэгмана: $\frac{1}{2}\|x^k - x^{k+1}\| \le V(x^{k+1}, x^k)$

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

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

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

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

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

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

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

⊙ такая, как у град спуска с почти K

⊕ $L \le L_2$ $\qquad \|\nabla f(x) - \nabla f(y)\|_q \le L\|x - y\|_p$

$\frac{1}{p} + \frac{1}{q} = 1$

$\uparrow$

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

Генерация

град. спуск

$$\frac{L_2 \|x^*, x^0\|_2^2}{2K}$$

$p \in [1;2]$ $\qquad \|\cdot\|_2 \le \|\cdot\|_p \qquad \|\cdot\|_q \le \|\cdot\|_2$

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

Пример ЗС на симплексе с KL-дивергенцией)

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

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

Формально:

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

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

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

Лагранжиан:

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

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

$a_i \qquad b_i$

Минимизируем по x_i и ищем двойств.

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

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

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

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

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

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

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

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

max двойственную $\lambda_i \geq 0 \quad \nu \in \mathbb{R}$

$$\lambda_i^* = 0$$

Вернемся к $L(x, \lambda, \nu)$ с $\lambda_i^* = 0$

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

KKT для $L(x, \nu) \quad \nabla_x L = 0$

$$\nabla_{x_i} L = \left(\gamma [\nabla f(x^k)]_i + \log \frac{x_i^*}{x_i^k} + \nu\right)$$

$$+ \cancel{x_i^*} \cdot \frac{\cancel{x_i^k}}{\cancel{x_i^*}} \cdot \frac{1}{\cancel{x_i^k}}$$

$$= \left(1 + \gamma [\nabla f(x^k)]_i + \log \frac{x_i^*}{x_i^k} + \nu\right) = 0$$

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

ν^* найдем из соображений: $\sum x_i^* = 1$

$\exp(1+\nu^*)$ — нормировка

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

(+) выигрыш у ЗС на симплексе до $\frac{d}{\log d}$ раз