- Градиентный спуск:

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

$x \in (E, \|\cdot\|_{\cancel{2}})$

а что тогда про $\nabla f(x)$? $\nabla f(x) \in (E^*, \|\cdot\|_*)$

до этого $\|\cdot\| = \|\cdot\|_2 \Rightarrow (E^*, \|\cdot\|_*) = (E, \|\cdot\|)$

- Модификация (А. Немировский, Д. Юдин)

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

$\varphi: E \to E^*$ $\quad \varphi^{-1}: E^* \to E$

Шаг град. спуска в „зеркальном" пр-ве.

Опр. (Сильная выпуклость отн. произв. нормы)

$d: X \to \mathbb{R}$ является сильно выпуклой с $\mu > 0$, если

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

произв. норма

Опр. (Дивергенция Брэгмана)

Пусть 1-сильно выпуклая отн. нормы $\|\cdot\|$ на X функция d.

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

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

Примеры

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

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

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

$$V(x,y) = \sum_{i=1}^d x_i \log \frac{x_i}{y_i}$$ ← KL-дивергенция (измерение расст. между распределениями)

- $d(X) = tr(X \log X)$ X – матрица

$$V(X,Y) = tr(X \log X - X \log Y - X + Y)$$
(квантовая дивергенция фон Неймана)

Свойства

- асимметричность (см KL-дивергенцию)
- сильная выпуклость (из определения)

$$V(x,y) \geq \frac{1}{2}\|x-y\|^2$$ $\left(KL(x\|y) \geq \frac{1}{2}\|x-y\|_1^2\right.$ нер-во Пинскера

- неотрицательность
- невыпукла по 2 аргументу
- теорема Пифагора для V:

$$\forall x,y,z \in X \hookrightarrow V(z,x) + V(x,y) - V(z,y) = \langle \nabla d(y) - \nabla d(x), z-x \rangle$$

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

$V(z,x) + V(x,y) =$ (по определению)

$$= d(z) - d(x) - \langle \nabla d(x); z-x \rangle$$

$$+ d(x) - d(y) - \langle \nabla d(y); x-y \rangle$$

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

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

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

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

f - выпуклая

X - выпуклое

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

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

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

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

(градиентный спуск)

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

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

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

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

$$= \arg\min_{x \in \mathcal{X}} \Big\{ \|x - y^k\|_2^2 \Big\} \leftarrow \text{градиентный спуск с проекцией}$$

- Условие оптимальности для

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

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

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

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

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

Опр. Непрер. диффер. функция $f: X \to \mathbb{R}$ — L-гладкая относительно нормы $\|\cdot\|$ на X, если $\forall x, y \in X \hookrightarrow \|\nabla f(x) - \nabla f(y)\|_* \le L \|x - y\|$

Теорема $f: X \to \mathbb{R}$ L-гладкая отн. $\|\cdot\|$, тогда

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

Док-во: ф. Ньютона-Лейбница

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + \tau(y - x)); y - x \rangle d\tau$$

$$= \langle \nabla f(x); y - x \rangle + \int_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x); y - x \rangle d\tau$$

$$\left| f(y) - f(x) - \langle \nabla f(x); y - x \rangle \right| = \left| \int_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x); y - x \rangle d\tau \right|$$

$|\Sigma| \le \Sigma |\ |$

$$\le \int_0^1 \left| \langle \nabla f(x + \tau(y-x)) - \nabla f(x); y - x \rangle \right| d\tau$$

КБШ $\langle a; b \rangle \le \|a\|_* \|b\|$

$$\le \int_0^1 \|\nabla f(x + \tau(y-x)) - \nabla f(x)\|_* \|y - x\| d\tau$$

L-гладкость отн. $\|\ \|$

$$\le \int_0^1 L \tau \|y - x\|^2 d\tau$$

$$= \frac{L \|y - x\|^2}{2} \quad \blacksquare$$

Док-во сходимости

Условие оптимальности для

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

$$\forall x \in X \iff \langle \nabla g(\underset{x^{k+1}}{x^*}); \underset{x^{k+1}}{x^*} - x \rangle \le 0$$

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

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

Свойство V

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

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

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

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

$x^{k+1} \to x^k$ в скалярном произв.

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

Гладкость

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

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

выпуклость f:

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

$$\frac{1}{2}\|x^k - x^{k+1}\|^2 \leq V(x^{k+1}, x^k)$$

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

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

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

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

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

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

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

Сходимость зеркального спуска для выпуклых, L-гладких

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

сублинейная через $\frac{1}{K}$, как и град. спуск
в град спуске L_2, $\frac{1}{2}\|x^0-x^*\|_2^2$

⊕ $L \le L_2$

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

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

$$1\le p\le 2 \Rightarrow \|\cdot\|_2 \le \|\cdot\|_p, \quad \|\cdot\|_q \le \|\cdot\|_2$$

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

- Зеркальный спуск на симплексе

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

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

Формальная запись:

$$\min_{x\in\mathbb{R}^d} \langle\gamma\nabla f(x^k);x\rangle + V(x,x^k)$$

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

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

Лагранжиан:

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

по компонентам

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

минимизируем x_i для получения двойственной

$$\left(a + \log\left(\frac{x_i}{b}\right)\right) x_i \to \max$$

$$a + \log\left(\frac{x_i^*}{b}\right) + \frac{b}{b} = 0$$

$$x_i^* = b \exp\left(-a - \cancel{b}\,1\right)$$

$$\inf_x L(x, \lambda, \nu) = \sum_{i=1}^{d} \left(\cancel{a} + -(\cancel{a} + \cancel{b}\,1)\right) b\left(\exp(-a - \cancel{b}\,1)\right) - \nu$$

$$= \sum_{i=1}^{d} -(x_i^k)^{\cancel{1}\,2} \exp\left(+\lambda_i - \gamma \left[\nabla f(x^k)\right]_i - \nu\right) - \nu$$

Двойственная задача

$$\max_{\lambda_i \geq 0,\, \nu \in \mathbb{R}} \left[\sum_{i=1}^{d} -x_i^k \exp\left(-1 + \lambda_i - \gamma \left[\nabla f(x^k)\right]_i - \nu\right) - \nu\right]$$

$$\lambda_i^* = 0, \qquad KKT$$

$$\nabla_x L(x, \lambda, \nu) = 0$$

$$\nabla_{x_i} \left[\langle \gamma \nabla f(x^k), x \rangle + \underset{\sum x_i \log\left(\frac{x_i}{x_i^k}\right)}{\underset{\parallel}{V(x, x^k)}} + \sum_{i=1}^{d} \lambda_i (-x_i) + \nu\left(\sum_{i=1}^{d} x_i - 1\right) \right]_{\lambda_i^* = 0} = 0$$

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

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

?

$\sum x_i^* = 1$ $\exp(1+J^*)$ – нормировка

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

softmax

⊕ для симплекса можно получить улучшение порядка $\frac{d}{\log d}$.