$$\min_{x\in\mathbb{R}^d} f(x)$$

f – ~~L-гладкая~~

$$f(x) = |x|$$

$$|f'(0+\delta) - f'(0-\delta)| = 2 \; \cancel{\le L\cdot 2\delta}$$

~~гладкость~~

~~$\delta \to 0$~~

~~$L \to \infty$~~

Предположение (вместо гладкости)

$f: \mathbb{R}^d \to \mathbb{R}$ является M-липшицевой, если

$$\forall x, y \in \mathbb{R}^d \hookrightarrow |f(x) - f(y)| \le M\|x-y\|_2$$

f – выпуклая и M-липшицевая

Субградиент

$f: \mathbb{R}^d \to \mathbb{R}$ выпуклая. Вектор $g \in \mathbb{R}^d$ называется субградиентом ф. f в т. $x \in \mathbb{R}^d$, если

$$\forall y \in \mathbb{R}^d \hookrightarrow f(y) \ge f(x) + \langle \underset{\nabla f(x)}{g}; y - x\rangle$$ (выпуклость)

Субдифференциал $\partial f(x)$ – мн-во субград. ф. f в т. x

Условие оптимальности

x^* – минимум вып. ф. $f \iff 0 \in \partial f(x^*)$

Док-во:

$\Leftarrow \quad 0 \in \partial f(x^*)$

определение субградиента

$$f(\underset{x}{\underset{\parallel}{y}}) \ge f(\underset{x^*}{\underset{\parallel}{x}}) + \langle \underset{\in \partial f(x^*)}{g}; \underset{x}{\underset{\parallel}{y}} - \underset{x^*}{\underset{\parallel}{x}} \rangle$$

$$f(x) \ge f(x^*) + \langle 0, x - x^* \rangle = f(x^*) \qquad \forall x \in \mathbb{R}^d$$

$\leftarrow$ по опред. глоб. миним.

$\Rightarrow \quad f(x) \ge f(x^*) \quad \forall x \in \mathbb{R}^d$, тогда

$$\underbrace{f(x) \ge f(x^*) + \langle 0; x - x^* \rangle \qquad \forall x \in \mathbb{R}^d}_{\text{по опред. субград.}} \qquad 0 \in \partial f(x^*) \qquad \blacksquare$$

Лемма

$f: \mathbb{R}^d \to \mathbb{R}$ выпуклая, тогда

f является M-липшицевой

$$\iff$$

$\forall x \in \mathbb{R}^d$ и $\forall g \in \partial f(x) \hookrightarrow \|g\|_2 \le M$

Док-во:

$\Rightarrow$ f – выпуклая и M-липшицева

$\triangleleft \quad g \in \partial f(x)$

опред. субград.

$$f(y) - f(x) \geq \langle g; y - x \rangle \quad \forall y \in \mathbb{R}^d$$

M- липшицевость

$$\langle g; y - x \rangle \leq f(y) - f(x) \leq |f(y) - f(x)| \leq M\|x - y\|_2$$

$y = x + g$

$$\|g\|_2^{\cancel{2}} \leq M\cancel{\|g\|_2} \Rightarrow \boxed{\|g\|_2 \leq M}$$

$\Leftarrow$ f - выпуклая, $\|g\|_2 \leq M$ $\forall x \in \mathbb{R}^d$ $\forall g \in \partial f(x)$

$\triangle$ $g \in \partial f(x)$

опред. субградиента

$$\langle g; x - y \rangle \geq f(x) - f(y)$$

КБШ

$$f(x) - f(y) \leq \langle g; x - y \rangle \leq \|g\|_2 \cdot \|x - y\|_2$$

$\|g\|_2 \leq M$

$$f(x) - f(y) \leq M \cdot \|x - y\|_2$$

аналогично

$$f(y) - f(x) \leq M\|x - y\|_2$$

$\Rightarrow |f(x) - f(y)| \leq M\|x - y\|_2$

Субградиентный метод

$$\boxed{x^{k+1} = x^k - \gamma g^k \qquad g^k \in \partial f(x^k)}$$

Док-во сходимости:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - \gamma g^k - x^*\|_2^2$$

$$= \|x^k - x^*\|_2^2 - 2\gamma \langle g^k; x^k - x^* \rangle + \gamma^2 \|g^k\|_2^2$$

M - липшицева $\Rightarrow$ $\|g\|_2 \le M$

$$\le \|x^k - x^*\|_2^2 - 2\gamma \langle g^k; x^k - x^* \rangle + \gamma^2 M^2$$

опред. субград. для вып. ф.

$$\underset{\substack{\| \\ x^*}}{f(y)} \ge \underset{\substack{\| \\ x^k}}{f(x)} - \langle \underset{\in \partial f(x^k)}{g}; \underset{x^k}{x} - \underset{x^*}{y} \rangle$$

$$\le \|x^k - x^*\|_2^2 + 2\gamma \left(f(x^*) - f(x^k)\right) + \gamma^2 M^2$$

$$2\gamma\left(f(x^k) - f(x^*)\right) \le \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

$$\frac{1}{K} \sum_{k=0}^{K-1} \qquad r^i - \cancel{r^{i+1}} + \cancel{r^{i+1}} - r^{i+2}$$

$$2\gamma \cdot \frac{1}{K} \sum_{k=0}^{K-1} \left(f(x^k) - f(x^*)\right) \le \frac{\|x^0 - x^*\|_2^2 - \cancel{\|x^K - x^*\|_2^2}}{K} + \gamma^2 M^2$$

нер-во Йенсена

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

$$2\gamma \left(f(\bar{x}^K) - f(x^*) \right) \le \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$

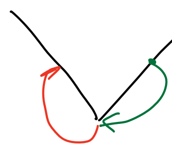
$$f(\bar{x}^K) - f(x^*) \le \underbrace{\frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}}_{\min\limits_{\gamma}}$$

$$\gamma = \frac{\|x^0 - x^*\|_2}{\sqrt{K} \cdot M}$$

<u>Сходимость</u>

$$\boxed{f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{M \|x^0 - x^*\|_2}{\sqrt{K}}}$$

⊙ сходимость по средней точке



⊖ сходимость хуже, чем в гладком случае ($\frac{1}{K}$ для град. спуска)

⊕ (суб) град. спуск является оптимальным

<u>Проблема</u>: $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$ зависимость от M, $\|x^0 - x^*\|_2$, K

<u>Решение</u>:

K: $\gamma \to \gamma_k = \frac{\|x^0 - x^*\|_2}{M\sqrt{k+1}}$

текущий номер

M: 1) $$\frac{\|x^0-x^*\|_2}{\sqrt{(k+1)M^2}} \approx \frac{D}{\sqrt{\underbrace{\sum_{t=0}^{k}\|g^t\|_2^2}_{M^2}}}$$

$D \sim \|x^0-x^*\|_2$ — диаметр

M — оцен. $\|\ \|$ градиента

AdaGrad Norm

---

2) подбор инд. шага => покоорд. оценка M_i

$$\gamma^k \to \gamma^{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}, \quad D_i \approx D$$

i — номер коорд.

AdaGrad

⊕ сублинейная сход. $\frac{1}{\sqrt{k}}$ — как в субград. спуске

---

3) $$\gamma^{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k} g_i^t}} \to \frac{D_i}{h_i^{k+1}}, \quad D_i \approx D$$

старые градиенты имеют такой же вес, что и новые

$$(h_i^{k+1})^2 = \beta_2 (h_i^k)^2 + (1-\beta_2)(g_i^k)^2$$

$\beta_2 \in (0;1)$

RMSProp

---

4) RMSProp + тяжелого шарика

= Adam

$(\beta_2, \beta_1 \to \beta_1^k, \beta_2^k)$ $(D_i \to D)$

$\|x^0 - x^*\|_2$: если метод сходится, то $x^t \to x^*$

$$\|x^0-x^*\|_2 \to \|x^0-x^k\|_2 \to \underbrace{d_k = \max_{t\in[0,k]} \{\|x^0-x^t\|_2\}}_{\text{эмпирический диаметр}}$$

по коорд.

$$|x_i^0 - x_i^*| \to d_{k,i} = \max_t \{|x_i^0 - x_i^t|\}$$

$$\text{AdaGrad} + \underset{\text{вместо } D_i}{d_{k,i}} = \text{DoG}$$

(Distance over Gradients)

Другой взгляд

AdaGrad, Adam, RMSProp, DoG можно описать в виде

$$x^{k+1} = x^k - \gamma H_k g^k$$

матрица $\mathbb{R}^{d\times d}$ (диагональная)

- H_k – матрица шкалирования
- H_k – не квазиньютоновская
- H_k – „рабоче-крестьянский" способ выч. ∇^2

Опр. (проксимальный оператор)

$r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Проксимальный оператор, порожденный ф. r есть

$$\mathrm{prox}_r(x) = \underset{\hat{x} \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ r(\hat{x}) + \frac{1}{2} \|x - \hat{x}\|_2^2 \right\}$$

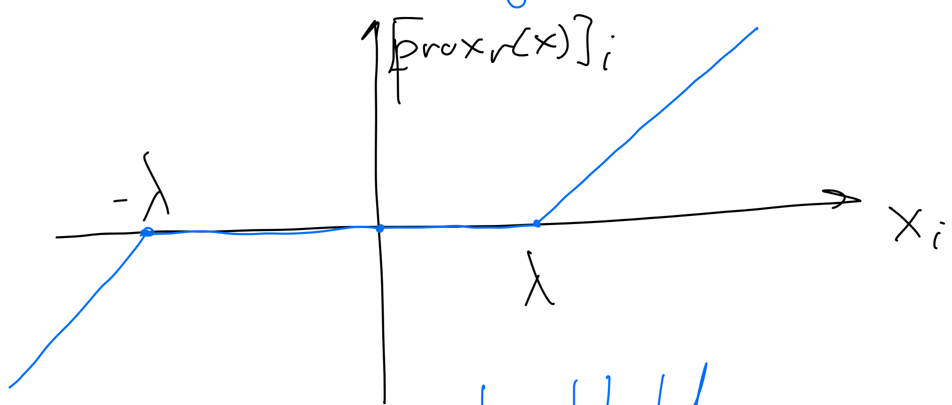
Сущ. решения:

Пусть $\exists\, \hat{x} \in \mathbb{R}^d \hookrightarrow r(\hat{x}) < +\infty$ (конечное знач.), тогда если r выпукла, то проксимальный оператор однозн. опред.

Пример:

- $r(x) = \lambda \|x\|_1$, где $\lambda > 0$

$$[\mathrm{prox}_r(x)]_i = \underbrace{\left[|x_i| - \lambda \right]_+}_{y_+ = \max\{0; y\}} \mathrm{sign}(x_i)$$



trash hold

- $r(x) = \frac{\lambda}{2} \|x\|_2^2$, где $\lambda > 0$

$$\mathrm{prox}_r(x) = \frac{x}{1+\lambda}$$

• $r(x) = \mathbb{I}_{\bar{X}}(x)$, где $\bar{X}$ – выпуклое мн-во

$$\mathbb{I}_{\bar{X}}(x) = \begin{cases} 0 & x \in \bar{X} \\ +\infty & \text{иначе} \end{cases}$$

$$\operatorname{prox}_r(x) = \arg\min_{\hat{x} \in \mathbb{R}^d} \left\{ \mathbb{I}_{\bar{X}}(\hat{x}) + \frac{1}{2}\|\hat{x} - x\|_2^2 \right\}$$

$$= \begin{cases} \arg\min_{\hat{x} \in \bar{X}} \left\{ \frac{1}{2}\|\hat{x} - x\|_2^2 \right\} \\ \text{нет решений} & x \notin \bar{X} \end{cases}$$

$$= \operatorname{proj}_{\bar{X}}(x)$$

<u>Свойство 1</u> (техническое)

$r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпукла. prox_r определен.

Тогда $\forall x, y \in \mathbb{R}^d$ следующие условия эквивал.

1) $\operatorname{prox}_r(x) = y$

2) $x - y \in \partial r(y)$

3) $\langle x - y, z - y \rangle \leq r(z) - r(y) \quad \forall z \in \mathbb{R}^d$

<u>Док-во:</u>

Первое условие:

$$y = \arg\min_{\hat{x} \in \mathbb{R}^d} \left\{ r(\hat{x}) + \frac{1}{2}\|\hat{x} - x\|_2^2 \right\}$$

Условие оптимальности

$$0 \in \partial\left(\hat{r}(\hat{x}) + \frac{1}{2}\|\hat{x} - x\|_2^2\right)\Big|_{\hat{x} = y}$$

$$= \partial r(y) + y - x$$

$$x - y \in \partial r(y)$$

доказано $1 \Leftrightarrow 2$

Определение субградиента

$$\langle g; z-y \rangle \le r(z) - r(y) \quad \forall z \in \mathbb{R}^d$$

$\in \partial r(y) \Leftarrow x-y \in$

$$\langle x-y; z-y \rangle \le r(z) - r(y) \quad \forall z \in \mathbb{R}^d$$

доказано $2 \Leftrightarrow 3$

Свойство 2:

$r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпукла и prox_r определен.

Тогда $\forall x, y \in \mathbb{R}^d \hookrightarrow$

$$\|\mathrm{prox}_r(x) - \mathrm{prox}_r(y)\|_2 \le \|x-y\|_2$$

Док-во: $u = \mathrm{prox}_r(x), \quad v = \mathrm{prox}_r(y)$

из пред. св-ва $(1 \Rightarrow 3)$

$$\langle x-u, z_1-u \rangle \le r(z_1) - r(u) \quad \forall z_1 \in \mathbb{R}^d$$

$$\langle y-v; z_2-v \rangle \le r(z_2) - r(v) \quad \forall z_2 \in \mathbb{R}^d$$

$z_1 = v, \quad z_2 = u$

$$+\left(\begin{array}{l} \langle x-u, v-u \rangle \le \cancel{r(v)} - \cancel{r(u)} \\ \langle y-v; u-v \rangle \le \cancel{r(u)} - \cancel{r(v)} \end{array}\right.$$

$$\langle x - y - u + v;\ v - u \rangle \le 0$$

$$\langle v - u;\ v - u \rangle \le \langle x - y;\ u - v \rangle$$

$$\|v - u\|_2^2 \le \underbrace{\langle x - y;\ u - v \rangle}_{\text{КБШ}}$$

$$\|v - u\|_2^{\cancel{2}} \le \|x - y\|_2 \cdot \cancel{\|u - v\|_2}$$

$$\|\mathrm{prox}_r(x) - \mathrm{prox}_r(y)\|_2 \le \|x - y\|_2 \quad \blacksquare$$

$$\min_{x \in \mathbb{R}^d} \left[f(x) + r(x) \right] \longleftarrow \text{композитная задача}$$

- f — μ-с. в., L-гладкая
- r — выпуклая, проксимально дружественная
  (prox_r считается аналитически)

Проксимальный градиентный метод

$$\boxed{x^{k+1} = \mathrm{prox}_{\gamma r}\left(x^k - \gamma \nabla f(x^k)\right)}$$

Физика

- если r дифференцируема

$$\arg\min_{\hat{x}} \Big\{ \underbrace{\gamma r(\hat{x}) + \frac{1}{2} \|\hat{x} - x^k + \gamma \nabla f(x^k)\|_2^2}_{\nabla = 0} \Big\}$$

$$\gamma \nabla r(x) + x - x^k + \gamma \nabla f(x^k) = 0$$

x^{k+1}

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

неявная схема

<u>Свойство</u> (неподвижн. точка)

$$\forall \gamma > 0 \qquad x^* = \operatorname{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*))$$

$\operatorname{argmin} f(x) + r(x)$

Док-во: Условие оптимальности

$$0 \in \nabla f(x^*) + \partial r(x^*) \quad \Big| \cdot \gamma$$

$$0 \in \gamma(\nabla f(x^*) + \partial r(x^*))$$

$$-\gamma \nabla f(x^*) \in \gamma \partial r(x^*) \quad \Big| \pm x^*$$

$$x^* - \gamma \nabla f(x^*) - x^* \in \gamma \partial r(x^*)$$

Св-во prox $(1 \Leftrightarrow 2)$

$$x^* = \operatorname{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)) \quad \blacksquare$$

Док-во сходимости проксим. град. спуска:

$$\|x^{k+1} - x^*\|_2^2 = \|\operatorname{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k)) - x^*\|_2^2$$

$$x^* = \operatorname{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*))$$

$$= \|\operatorname{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k)) - \operatorname{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*))\|_2^2$$

$$\|\operatorname{prox}(x) - \operatorname{prox}(y)\|_2 \le \|x - y\|_2$$

$$\le \|x^k - \gamma \nabla f(x^k) - x^* + \gamma \nabla f(x^*)\|_2^2$$

Дальше, как в методе с проекцией

⊕ сходимость, как у GD на гладких задач

⊕ можно ускорять

⊖ для „простых" функций r