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

f – ~~L-гладкая~~

гладкости нет

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

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

~~гладкость~~

$\delta \to 0$

- 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-липшицева

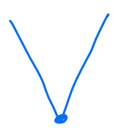
- <u>Субградиент</u>

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

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

(выпукло)

<u>Субдифференциал</u> $\partial f(x)$ – мн-во субгр. ф f в т. x



$$f(y) \ge f(0) + \langle g; y - 0\rangle$$

$$|y| \ge gy$$

$$\text{sign}(y) \ge g \Rightarrow g \in [-1; 1]$$

• Условие оптимальности

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

Док-во: $\Leftarrow$ $0 \in \partial f(x^*)$, то тогда по опр. субград.

$$\forall x \in \mathbb{R}^d \quad f(x) \geq f(x^*) + \langle \underset{0}{\underset{\parallel}{g(x^*)}}; x - x^* \rangle = f(x^*)$$

по опред. глоб. миним.

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

$$\underbrace{f(x) \geq f(x^*) + \underbrace{\langle 0; x - x^* \rangle}_{0} \quad \text{для } \forall x \in \mathbb{R}^d}$$

опред. субград. в т. $x^* \Rightarrow 0 \in \partial f(x^*)$ ∎

Лемма

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

f является M-липшицевой

$\iff$

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

Док-во:

$\Rightarrow$ f - выпуклая и M-липшицева

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

опред. субград. вып. ф.

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

M-липшицевости

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

$y = x + g$

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

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

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

опред. субград. вып. ф.

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

КБШ

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

$\|g\|_2 \le M$

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

аналогично

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

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

- Субградиентный спуск

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

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

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

опред. субградиента для выпуклой функции

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

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

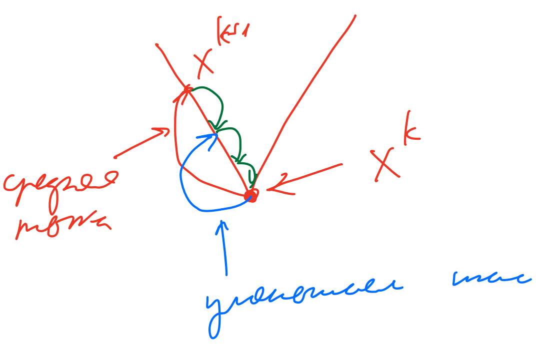
$$\frac{1}{K}\sum_{k=0}^{K-1}$$

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

нер-во Йенсен

$$f\Big(\underbrace{\frac{1}{K}\sum_{k=0}^{K-1} x^k}_{\text{средняя точка}}\Big) - f(x^*) \le \underbrace{\frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}}_{\min\limits_{\gamma}}$$

$$\gamma = \frac{\|x^0 - x^*\|_2}{\sqrt{K}\cdot M} \qquad \gamma_k \sim \frac{1}{\sqrt{k}}$$



Сходимость

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

для гладкого случая $\frac{1}{K}$

Проблема: $\gamma_k = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}}$ зависимость от M и $\|x^0-x^*\|_2$

Решение:

AdaGradNorm

$$\boxed{\gamma_k = \frac{\|x^0-x^*\|_2}{\sqrt{kM^2}} \approx \frac{\overset{\sim \|x^0-x^*\|_2}{D}}{\sqrt{\underbrace{\sum_{t=0}^{k-1}\|g^t\|_2^2}_{M^2}}}}$$

D ← диаметр

AdaGradNorm $\Rightarrow$ AdaGrad (скалярн. подбор шага $\Downarrow$ покоорд. подбор)

AdaGrad

$$\boxed{\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}}}$$

i ← номер коорд.

Док-во сходимости:

$$|x_i^{k+1} - x_i^*|^2 = |x_i^k - \gamma_{k,i} g_i^k - x^*|^2$$

$$= |x_i^k - x_i^*|^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2$$

$$g_i^k(x_i^k - x_i^*) \le \frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i}(g_i^k)^2}{2}$$

$\sum_{i=1}^{d}$ по всем коорд.

$$\langle g^k; x^k - x^* \rangle \le \sum_{i=1}^{d}\left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i}(g_i^k)^2}{2}\right]$$

Выпуклость и субградиент:

$$f(x^k) - f(x^*) \le \sum_{i=1}^{d} \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i}(g_i^k)^2}{2} \right]$$

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

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{d} \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i}(g_i^k)^2}{2} \right]$$

$$\sum_k \sum_i = \sum_i \sum_k$$

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1} \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i}(g_i^k)^2}{2} \right]$$

Группируем $|x_i^k - x_i^*|^2$

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \le \frac{1}{K}\sum_{i=1}^{d}\sum_{k=1}^{K-1} \left(\frac{1}{2\gamma_{k,i}} - \frac{1}{2\gamma_{k-1,i}}\right)|x_i^k - x_i^*|^2$$

$$+ \frac{1}{K}\sum_{i=1}^{d}\left(\frac{1}{2\gamma_{0,i}}|x_i^0 - x_i^*|^2 - \cancel{\frac{1}{2\gamma_{K-1,i}}|x_i^K - x_i^*|^2}\right)$$

≤ 0

$$+ \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1} \frac{\gamma_{k,i}}{2}(g_i^k)^2$$

$$\gamma_{-1,i} = +\infty$$

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \le \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1} \left(\frac{1}{2\gamma_{k,i}} - \frac{1}{2\gamma_{k-1,i}}\right)|x_i^k - x_i^*|^2$$

$$+ \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1} \frac{\gamma_{k,i}}{2}(g_i^k)^2$$

$|x_i^k - x_i^*| \le D_i$ — диаметр по i коорд.

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \le \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left(\frac{1}{2\gamma_{k,i}} - \frac{1}{2\gamma_{k-1,i}}\right) D_i^2$$

$$+ \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\frac{\gamma_{k,i}}{2}(g_i^k)^2$$

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}}$$

телескоп. сумма

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \le \frac{1}{K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left(\frac{\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}}{\cancel{D_i}} - \frac{\sqrt{\sum_{t=0}^{k-2}(g_i^t)^2}}{\cancel{D_i}}\right) D_i^{\cancel{2}}$$

$$+ \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\frac{D_i}{\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}}(g_i^k)^2$$

$$\le \frac{1}{K}\sum_{i=1}^{d} D_i \sqrt{\sum_{t=0}^{K-2}(g_i^t)^2}$$

$$+ \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\frac{D_i}{\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}}(g_i^k)^2$$

$\{a_k\} \ge 0$ $$\sum_{k=0}^{K-1}\frac{(a_k)^2}{\sqrt{\sum_{t=0}^{k-1}(a_t)^2}} \le 2\sqrt{\sum_{k=0}^{K-1}(a_k)^2}$$

$$\le \frac{1}{K}\sum_{i=1}^{d} D_i \sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}$$

$$+ \frac{1}{K}\sum_{i=1}^{d} D_i \sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}$$

$$= \frac{2}{K} \sum_{i=1}^{d} D_i \underbrace{\sqrt{\sum_{t=0}^{K-1} (g_i^t)^2}}_{K \cdot M^2}$$

$$= \frac{2M}{\sqrt{K}} \underbrace{\sum_{i=1}^{d} D_i}_{\tilde{D}}$$

$$= \frac{2M\tilde{D}}{\sqrt{K}}$$

добились сход $\frac{1}{\sqrt{K}}$ — как в субград. методе

AdaGrad $\Rightarrow$ RMSProp

$$\gamma_{k,i} = \frac{D_i}{\underbrace{\sqrt{\sum_{t=0}^{k-1} (g_i^t)^2}}} \Rightarrow \gamma_{k,i} = \frac{\gamma}{h_i^k}, \text{ где}$$

старые град.
имеют
тот же
вес, что и новые

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

$$\beta_2 \in (0;1) \qquad \beta_2 = 0,99$$
$$= 0,999$$

RMSProp $\Rightarrow$ Adam

$$\beta_2 = \text{const} \Rightarrow \beta_{2,k}$$

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

$$(h_i^0) = 0 \qquad (h_i^1)^2 = \underbrace{(1-\beta_2)}_{\neq 1}(g_i^0)^2$$

h^{k-1} — сумма старых град с весами

$$g_i^k \qquad g_i^{k-1} \qquad \dots \qquad g_i^0$$

$$1-\beta_2 \qquad (1-\beta_2)\beta_2 \qquad (1-\beta_2)\beta_2^k \leftarrow \sum \neq 1$$

Adam $\sum = 1$ $\quad \beta_2 \Rightarrow \beta_{2,k}$ с доп. нормировкой

$$\sum_{i=0}^{k} (1-\beta_2)\beta_2^i = \frac{(1-\beta_2)^2}{(1-\beta_2)^k} \leftarrow \text{нормировка}$$