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

f – ~~L-гладкая~~ $\Rightarrow$ гладкости нет

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

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

$\cancel{\delta\to 0}$

- **Новое предположение**

$f:\mathbb{R}^d\to\mathbb{R}$ f – M-липшицева, т.е. $\forall x,y\in\mathbb{R}^d \hookrightarrow$

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

μ-сильно выпуклая растет как парабола
M-липшицева – как линейная (не быстрее чем) $\Big\}$ на $\mathbb{R}^d$ не сочет.

выпуклые, M-липшицевы

- **Субградиент**

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

$$f(y)\geq f(x) + \langle g; y-x\rangle \quad \text{(выпуклость)}$$

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

- **Условие оптимальности**

x^* – минимум выпуклой функции $f \Leftrightarrow 0\in\partial f(x^*)$

- **Лемма**

$f:\mathbb{R}^d\to\mathbb{R}$ выпуклая, тогда f – M-липшицева $\Leftrightarrow$
$\forall x\in\mathbb{R}^d$ и $\forall g\in\partial f(x) \hookrightarrow \|g\|_2\leq M$

Док-во:

$\Rightarrow$ f - выпуклая и M- липшицева

$\exists\, g \in \partial f(x)$

выпуклость и опред. субградиента

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

M- липшицевость

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

$y = g+x$

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

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

$\exists\, g \in \partial f(x)$

выпуклость и субград.

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

КБШ

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

$\|g\|_2 \leq M$

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

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

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

выпуклость и опр. субградиента

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

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

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

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

для гладких $f(x^k) \le f(x^{k-1})$ — здесь так нельзя

Йенсен

$$2\gamma\, f\Big(\underbrace{\frac{1}{K}\sum_{k=0}^{K} x^k}_{\text{по средней точке}}\Big) - f(x^*) \le \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$

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

$$M^2\gamma^2 = \frac{\|x^0 - x^*\|_2^2}{K}$$

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

На практике $\gamma_k \sim \frac{1}{\sqrt{k}}$

Сходимость

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

в гладком случае GD давал $\frac{1}{K}$

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

Решение:

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

$D \sim \|x^0-x^*\|_2$

← AdaGradNorm

D - гиперпар (можно оценить или подобрать)

AdaGradNorm $\Rightarrow$ AdaGrad (индивид. шаги)

$\Downarrow$

покоординатный подбор

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

← номер коорд.

Док-во сходимости:

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

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

Разделим на $2\gamma_{k,i}$

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

$\sum_{i=1}^{d}$ по всем коорд.

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

выпуклость

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

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

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

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

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

группируем $|x_i^k - x_i^*|^2$

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

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

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

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

$|x_i^k - x_i^*| \le D_i$ — диаметр

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

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

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

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

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

$\{a_k\} \geq 0$ $\qquad \displaystyle\sum_{k=0}^{K-1} \frac{(a_k)^2}{\sum_{t=0}^{k} (a_t)^2} \leq 2 \sqrt{\sum_{k=0}^{K-1} (a_k)^2}$

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

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

$$= \frac{3}{2K} \sum_{i=1}^{d} \underbrace{\sqrt{\sum_{k=0}^{K-1} (g_i^k)^2}}_{\leq K M^2}\, D_i$$

$$= \frac{3M}{2\sqrt{K}} \sum_{i=1}^{d} D_i = \tilde{D}$$

$$= \frac{3M\tilde{D}}{2\sqrt{K}}$$

$\frac{1}{\sqrt{K}}$ — сублинейно, как и у субград. спуска

AdaGrad $\Rightarrow$ RMSProp

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k} (g_i^t)^2}} \Rightarrow \boxed{\begin{gathered}\gamma_{k,i} = \frac{\gamma}{\sqrt{\beta_2 (h_i^k)^2 + (1-\beta_2)(g_i^k)^2}} \\ (h_i^{k+1})^2 = \beta_2 (h_i^k)^2 + (1-\beta_2)(g_i^k)^2 \\ \beta_2 \in (0;1) \quad \beta_2 \approx 0{,}99\end{gathered}}$$

Старые градиенты имеют тот же вес, что и текущие

RMSProp $\Rightarrow$ Adam $\quad \boxed{\beta_2 = \beta_{2,k}}$

Связь с методом Ньютона

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

$$H_{k+1}^2 = \beta_2 H_k^2 + (1-\beta_2)\, \mathrm{diag}\begin{pmatrix} g_{k,1}^2 & & 0 \\ & \ddots & \\ 0 & & g_{k,d}^2 \end{pmatrix}$$

OASIS: $\quad H_{k+1} = \beta_2 H_k + (1-\beta_2)\, \mathrm{diag}\left(u_k \odot \nabla^2 f(x^k) u_k\right)$

1) $\nabla f(x^k)$

2) $\langle \nabla f(x^k); u^k \rangle$

$\{-1;1\}$

u_i^k – случ.

$\mathbb{E} u_i^k = 0$

$(u_i^k)^2 = 1$

3) $\nabla \langle \nabla f(x^k); u^k \rangle$

4) $\nabla^2 f(x^k) u^k$

$\mathrm{diag}(\nabla^2 f(x^k)) = \xrightarrow{\mathbb{E}}$ 5) $u^k \odot \nabla^2 f(x^k) u^k$

$$(\nabla^2 f(x^k))_{ii} \underbrace{(u_i^k)^2}_{1}$$

$$\mathbb{E}\, (\nabla^2 f(x^k))_{ij}\, u_i^k u_j^k = 0$$