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

- на пред. лекциях: f – гладкой
- далеко не все функции гладкие: $f(x) = |x|$

$$|f'(0+\delta) - f'(0-\delta)| =$$
$$= |1-(-1)| = 2 \cancel{\leq} L\delta$$
$$\delta \to 0$$

нет, L-Липшиц. градиента

Новое предположение

$f: \mathbb{R}^d \to \mathbb{R}$ $\quad$ f – M-Липшицева, если $\forall x, y \in \mathbb{R}^d$ выполнено

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

растет не быстрее линейной, μ-сильно выпуклые $\Rightarrow$ квадратичные

не уживаются на $\mathbb{R}^d$

выпуклые, M-Липшицева f

Субградиент

Опр. $f: \mathbb{R}^d \to \mathbb{R}$ выпуклая. Вектор g будем называть субградиентом f в $x \in \mathbb{R}^d$, если $\forall y \in \mathbb{R}^d$ выполнено:

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

Субдифференциал $\partial f(x)$ – мн-во всех субградиентов f в точке x

- f дифференцируема $\partial f(x) = \{\nabla f(x)\}$. Определение выпуклости
-  $f(x) = |x|$ $\quad$ $\partial f(0)$

$$|y| \geq \langle g; y \rangle \quad g \in [-1; 1]$$

Условие оптимальности

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

Док-во:

$\Leftarrow$ $0 \in \partial f(x^*)$, тогда $\forall x \in \mathbb{R}^d$

$$f(x) \geq f(x^*) + \langle 0; x - x^* \rangle = f(x^*) \leftarrow \text{минимум по опр.}$$

опред. субградиента

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

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

субградиент в 0 (по определению)

Лемма

$f: \mathbb{R}^d \to \mathbb{R}$ – выпуклая, тогда f M-Липшицева

$$\iff$$

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

Субградиентный метод

$$x^{k+1} = x^k - \gamma \underset{\in \partial f(x^k)}{g^k} \leftarrow \text{субградиент в точке } x^k$$

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 $\Rightarrow$ $\|g^k\|_2 \leq M$

$$= \|x^k - x^*\|_2^2 - 2\gamma \underbrace{\langle g^k, x^k - x^* \rangle} + \gamma^2 M^2$$

выпуклость и субградиент

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

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

$\frac{1}{K}\sum$

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

Нер-во Йенсена

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

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

$$\underbrace{f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right)}_{\text{эргодическая}} - f(x^*) \leq \frac{M\|x^0 - x^*\|_2}{\sqrt{K}}$$

для гладкого случая $\frac{1}{K}$, здесь $\frac{1}{\sqrt{K}}$

подбор шага зависит $\|x^0 - x^*\|$, M, $\gamma_k = \frac{1}{\sqrt{K}}$

Теорема о сходимости

f - выпуклая и M-липшицева, то для субград. спуска справедлива эргодическая сходимость

$$\underbrace{f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*)}_{\sim\varepsilon} \leq \frac{M\|x^0 - x^*\|_2}{\sqrt{K}}$$

$$\boxed{\text{Для достижения } \varepsilon\text{-точности:} \quad K = \frac{M^2\|x^0-x^*\|_2^2}{\varepsilon^2}}$$

<u>Проблема:</u> шаг $\gamma_k = \dfrac{\|x^0-x^*\|_2}{M\sqrt{k}}$ зависит M, $\|x^0-x^*\|_2$

<u>Как решить?</u>

AdaGradNorm $\rightarrow$
$$\boxed{\gamma_k = \frac{\|x^0-x^*\|_2}{\sqrt{kM^2}} \approx \frac{D}{\sqrt{\sum_{t=0}^{k}\|g^t\|_2^2}}}$$

$\|x^0-x^*\|_2 \le D$ (можем оценить или сразу ∓ ограничим задачу)

AdaGradNorm $\Rightarrow$ AdaGrad
(индивидуальность шага)

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

i – номер координаты

<u>Док-во сходимости:</u>

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

разделим на $2\gamma_{k,i}$

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

$\sum_{i=1}^{d}$ по всем координатам

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

выпуклость

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

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

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

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

сгруппировать $|x_i^k - x_i^*|^2$

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

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

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

$$|x_i^k - x_i^*|^2 \leq D_i^2$$

$$\leq \frac{1}{k}\sum_{i=1}^{d}\sum_{k=0}^{K-2}\left[\left(\frac{1}{2\gamma_{k,i}} - \frac{1}{2\gamma_{k-1,i}}\right)D_i^2\right.$$

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

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

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

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

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

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

$$\{a_k\} \geq 0$$

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

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

$$= \frac{3}{2K}\sum_{i=1}^{d}D_i\sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}$$

M-липшицевость $\longrightarrow$ $\le \sqrt{KM^2}$

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

$\frac{1}{\sqrt{K}}$ — как у субградиентного спуск.

AdaGrad $\Rightarrow$ RMS Prop

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}} \Rightarrow \boxed{\begin{aligned}\gamma_{k,i} &= \frac{\gamma}{\sqrt{\beta_2(h_i^k)^2+(1-\beta_2)(g_i^k)^2}}\\ (h_i^{k+1})^2 &= \beta_2(h_i^k)^2+(1-\beta_2)(g_i^k)^2\end{aligned}}$$

$$\boxed{\begin{aligned}&\beta\in(0;1)\\ &\beta\approx 0{,}99\end{aligned}}$$

RMS Prop $\Rightarrow$ Adam

$$\boxed{\beta_2 \Rightarrow \beta_{2,k}}$$

Связь с методом Ньютона

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

RMS Prop:

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

OASIS:

$$H_{k+1} = \beta H_k + (1-\beta)\,\mathrm{diag}\big(\underbrace{u_k \odot \nabla^2 f(x^k)\, u_k}\big)$$

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

2) $\langle \nabla f(x^k); u_k \rangle$ $\qquad u_{k,i} = \{1; -1\}$, $p = 1/2$

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

4) $\nabla^2 f(x^k)\, u_k$

5) $u_k \odot \nabla^2 f(x^k)\, u_k$

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