04.12.17 МОМО $\underline{\text{XIII}}$

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w) \to \min_w \quad, N >> 1 \quad, f_i \in C^1$$

$$\nabla F(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w) \quad, w \in \mathbb{R}^D$$

Величина	Сложность вычислений
$f_i(w)$	$O(q)$
$\nabla f_i(w)$	$O(Dq)$ {градиенты}, $O(q)$ {autograd}
$F(w)$	$O(Nq)$
$\nabla F(w)$	$O(Nq)$ {autograd}

SGD

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w) = \mathbb{E}_{i\sim U} f_i(w)$$

$$\nabla F(w) = \nabla_w \mathbb{E}_{i\sim U} f_i(w) = \mathbb{E}_{i\sim U} \nabla_w f_i(w)$$

$$\begin{cases} i_k \sim \mathrm{Unif}(1\ldots N) \\ g_k = \nabla f_{i_k}(w_k) \\ w_{k+1} = w_k - \alpha_k g_k \end{cases}$$

g_k - несмещённая оценка градиента

$$\mathbb{E}_{i_k} g_k = \nabla F(w_k)$$

SGD + mini-batches

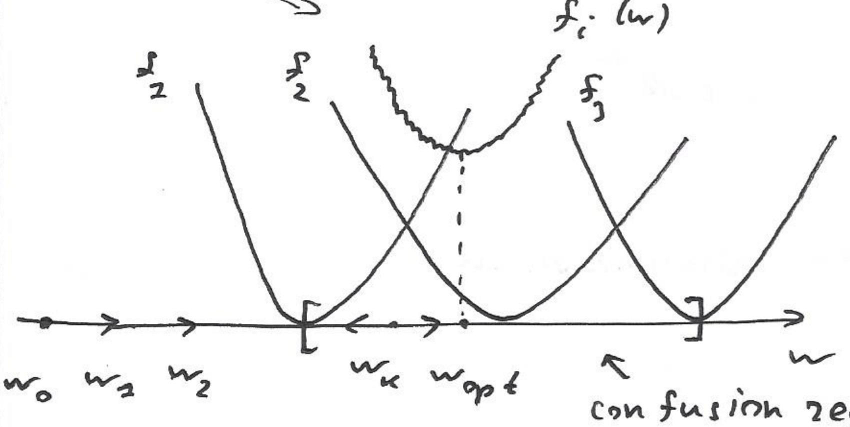
$$\begin{cases} I_k \subseteq \mathrm{Unif}(1\ldots N) \\ g_k = \frac{1}{|I_k|}\sum_{i\in I_k} \nabla f_i(w_k) \\ w_{k+1} = w_k - \alpha_k g_k \end{cases}$$

сходимость оценки g_k к истинному $\nabla F(w)$:

сх-ть дисперсии к нулю: $O(\frac{1}{\sqrt{n}})$

Пример:

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\frac{1}{2}(y_i - w x_i)^2}_{f_i(w)} \to \min_{w \in \mathbb{R}}$$



F - вып. ф-ия и

$\|g_k\| \leq L$

липшицевость градиентов

$$w_{k+1} = w_k - \alpha_k g_k, \quad \mathbb{E}_{i_k} g_k = \nabla F(w_k)$$

$$\|w_{k+1} - w_{opt}\|^2 = \|w_k - \alpha_k g_k - w_{opt}\|^2 = \|w_k - w_{opt}\|^2 - 2\alpha_k \langle g_k, w_k - w_{opt}\rangle + \alpha_k^2 \|g_k\|^2$$

$$\mathbb{E}_{i_k} \|w_{k+1} - w_{opt}\|^2 = \|w_k - w_{opt}\| - 2\alpha_k \langle \nabla F(w_k), w_k - w_{opt}\rangle + \alpha_k^2 \mathbb{E}_{i_k} \|g_k\|^2$$

$$\| \; F(w_{opt}) \geq F(w_k) + \langle \nabla F(w_k), w_{opt} - w_k\rangle$$

$$\alpha_k (F(w_k) - F(w_{opt})) \leq \alpha_k \langle \nabla F(w_k), w_k - w_{opt}\rangle =$$

$$= \frac{1}{2}\|w_k - w_{opt}\|^2 + \frac{\alpha_k^2}{2}\mathbb{E}_{i_k}\|g_k\|^2 - \frac{1}{2}\mathbb{E}_{i_k}\|w_{k+1} - w_{opt}\|^2$$

$$\sum_{i=0}^{k} \alpha_i (\mathbb{E}_{i_0 \ldots i_{i-1}} F(w_i) - F_{opt}) \leq \frac{1}{2}\|w_0 - w_{opt}\|^2 + \frac{1}{2}\sum_{i=0}^{k}\alpha_i^2 \mathbb{E}\|g_i\|^2 - \underbrace{\frac{1}{2}\mathbb{E}\|w_{k+1} - w_{opt}\|^2}_{\geq 0} \leq \frac{1}{2}\|w_0 - w_{opt}\|^2 + \frac{1}{2}\sum_{i=0}^{k}\alpha_i^2 \mathbb{E}\|g_i\|^2$$

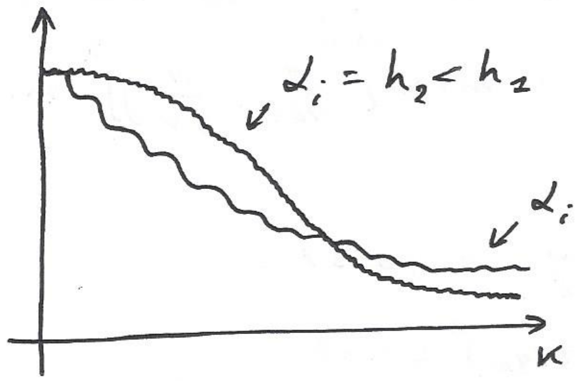
$$\mathbb{E} F\Bigg(\underbrace{\frac{\sum_{i=0}^{k}\alpha_i w_i}{\sum_{i=0}^{k}\alpha_i}}_{\overline{w}_k}\Bigg) - F_{opt} \leq \frac{\sum_{i=0}^{k}\alpha_i(\mathbb{E}F(w_i) - F_{opt})}{\sum_{i=0}^{k}\alpha_i} \leq$$

неравенство Йенсена

$$\leq \frac{\|w_0 - w_{opt}\|^2 + \sum_{i=0}^{k}\alpha_i^2 \mathbb{E}\|g_i\|^2}{2\left(\sum_{i=0}^{k}\alpha_i\right)}$$

① $\alpha_i = h$

$$\mathbb{E}\, F(\bar{w}_k) - F_{opt} \le \frac{R^2 + h^2 (k+1) L^2}{2(k+1)} = \underbrace{\frac{R^2}{2h(k+1)}}_{\to 0} + \frac{hL^2}{2} \to \frac{hL^2}{2}$$



② $\alpha_i = \dfrac{h}{\sqrt{\mathbb{E}\|g_i\|^2}}$

невозможно рассчитывать $\mathbb{E}\|g_i\|^2$ напрямую

③ $\sum\limits_{i=1}^{\infty} \alpha_i = \infty, \quad \sum\limits_{i=1}^{\infty} \alpha_i^2 < \infty; \quad \alpha_i = \dfrac{h}{i^\tau}, \quad \tau \in (\tfrac{1}{2}, 1]$

④ $\sum\limits_{i=1}^{\infty} \alpha_i = \infty, \quad \sum\limits_{i=1}^{\infty} \alpha_i^2 \Big/ \sum\limits_{i=1}^{\infty} \alpha_i \to 0, \quad \sum\limits_{i=1}^{\infty} \alpha_i^2 = \infty, \quad \tau \in (0, \tfrac{1}{2}]$

скорость сх-ти $O(\frac{1}{\sqrt{k}})$

Ф-ия	Метод	Ск-ть сх-ти	Ск-ть стох. вар.
$\in C_L^{1,1}$ и μ-с. вып.	Град. спуск	$O(c^k)$	$O(1/k)$
$\in C^{1,1}$ и вып.	Град. спуск	$O(1/k)$	$O(1/\sqrt{k})$
$\in C^{0,0}$ и вып.	Субгр. спуск	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$

Метод SAG

FULL GD: $w_{k+1} = w_k - \alpha_k \frac{1}{N} \sum\limits_{i=1}^{N} \nabla f_i(w_k)$

SAG: $w_{k+1} = w_k - \alpha_k \frac{1}{N} \sum\limits_{i=1}^{N} \nabla f_i(v_i^k)$

$i_k \sim Unif(1 \ldots N)$

$v_j^k = v_j^{k-1} \;\; \forall j \ne i_k, \quad v_{i_k}^k = w_k$

$$g_k = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(v_i^k) = g_{k-1} + \frac{1}{N}\nabla f_{i_k}(w_k) - \frac{1}{N}\nabla f_{i_k}(v_{i_k}^{k-1})$$

$w_{k+1} = w_k - \alpha_k g_k$ память $O(ND)\{-\frac{1}{N}\nabla f_{i_k}(v_{i_k}^{k-1})\}$

критерий останова: $\|g_k\|^2 \le \varepsilon$

Утв. $F \in C_L^{1,1}$ и μ с. вып. и $\alpha = \frac{1}{16L}$ в SAG:

тогда $\mathbb{E}\, F(w_k) - F_{opt} \le \left(1 - \min\left(\frac{\mu}{16L}, \frac{1}{8N}\right)\right)^k \cdot const$

Пример:

$N = 700.000$ GD: $1 - \frac{\mu}{L} \approx 0.9999\ldots$

$L = \frac{1}{4}$, $\mu = \frac{1}{N}$ SAG: $(1 - 1/8N)^N \approx 0.88\ldots$

$f_i \in C_{L_i}^{1,1}$: $F(v) = \frac{1}{N}\sum_{i=1}^{N} f_i(v) \le \frac{1}{N}\sum_{i=1}^{N}(f_i(w) +$

$+ \langle \nabla f_i(w_k), v - w\rangle + \frac{L_i}{2}\|v - w\|^2) = F(w) +$

$+ \langle \nabla F(w), v - w\rangle + \frac{\left(\frac{1}{N}\sum_{i=1}^{N} L_i\right)}{2}\|v - w\|^2$

отсюда: $L_F \le \frac{1}{N}\sum_{i=1}^{N} L_i \le \max_i(L_i)$

$L \leftarrow L/2$

повторять

  $i_k \sim \mathrm{Unif}(1\ldots N)$

  $\alpha = 1/L$

  $w_{k+1} = w_k - \alpha g_k \leftarrow$ SAG

  если $f_{i_k}(w_{k+1}) \ge f_{i_k}(w_k) + \langle \nabla f_{i_k}(w_k), w_{k+1} - w_k\rangle +$

  $+ \frac{L}{2}\|w_{k+1} - w_k\|^2$, то $L \leftarrow L \cdot 2$

  иначе, выход

  $L \leftarrow L \cdot 2$

Метод SVRG

$\tilde{w}$, $\tilde{\mu} = \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\tilde{w})$ несмещённая оценка на градиент

$g_k = \nabla f_{i_k}(w_k) - \nabla f_{i_k}(\tilde{w}) + \tilde{\mu}$

$\mathbb{E}_{i_k} g_k = \underbrace{\mathbb{E}_{i_k}\nabla f_{i_k}(w_k)}_{=\nabla F(w_k)} - \underbrace{\mathbb{E}_{i_k}\nabla f_{i_k}(\tilde{w})}_{=\nabla F(\tilde{w})} + \tilde{\mu} = \nabla F(w_k)$

Схема SVRG

$\nabla f_{i_t}(w_t) - \nabla f_{i_t}(\tilde{w}_k) + \tilde{\mu}_k$

$\to \nabla f_{i_t}(w_t) - \nabla f_{i_t}(\tilde{w}_k) + 0$

$\to 0$

$\tilde{w}_0 = w_0$

для $k = 0, 1, 2, \dots$

$\tilde{\mu}_k = \nabla F(\tilde{w}_k)$

$w_0 = \tilde{w}_k$

для $t = 0, 1, 2, \dots$

$i_t \sim Unif(1 \dots N)$

$g_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(\tilde{w}_k) + \tilde{\mu}_k$

$w_{t+1} = w_t - \alpha_t g_t$

$\tilde{w}_{k+1} = w_t$

семинар

① линейное программирование (LP)

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} \langle c, x \rangle \\ \text{s.t. } Ax \le b \\ Gx = h \end{cases}$$

$\langle a_i, x \rangle \le b_i, \ i = 1 \dots m$

барьерная ф-ия

$$F(x) = -\sum_{i=1}^{m} \ln (b_i - \langle a_i, x \rangle)$$

$$\varphi_t(x) = t \langle c, x \rangle + F(x)$$

② кв. прогр. с кв. огр. (QP)

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} \langle c, x \rangle \\ \text{s.t. } \frac{1}{2} \langle Q_i x, x \rangle + \langle p_i, x \rangle \le r_i, \ i = 1 \dots m \end{cases}$$

$$F(x) = -\sum_{i=1}^{m} \ln \left(r_i - \frac{1}{2} \langle Q_i x, x \rangle - \langle p_i, x \rangle\right)$$

③ конические програм. 2-го порядка (SOCP)

$$\begin{cases} \min \langle c, x \rangle \\ \text{s.t. } \| A_i x + b_i \|_2 \le \langle c_i, x \rangle + d_i \end{cases}$$

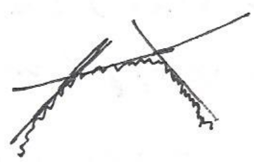
$$F(x) = -\sum_{i=1}^{m} \ln \left((\langle c_i, x \rangle + d_i)^2 - \| A_i x + b_i \|_2^2 \right)$$

④ полуопределённые програм. SDP

$$\min_{x \in \mathbb{R}^n} \langle c, x \rangle$$
$$\text{s.t. } x_1 A_1^{(i)} + \ldots + x_n A_n^{(i)} \le A_0^{(i)} \quad (LMI)$$

$$F(x) = -\sum_{i=1}^{m} \ln \det \left(A_0^{(i)} - \ldots - x_n A_n^{(i)} \right)$$

$$X = \sum_{i,j} x_{ij} E_{ij} \succcurlyeq 0 \in S_+^n$$



Техника сглаживания, Nesterov 2005

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_1 = \sum_{i=1}^{m} |\langle a_i, x \rangle - b_i| , \quad \underset{m \times n}{A} = \begin{bmatrix} - a_1 - \\ \vdots \\ - a_m - \end{bmatrix}$$

субградиенты: $O\left(\frac{1}{\sqrt{k}}\right) \; : \; O\left(\frac{1}{\varepsilon^2}\right) \to O\left(\frac{1}{\varepsilon}\right)$

$$\|g_i\| \le M , \quad f(x_k) - f^* \le \frac{MR}{\sqrt{k}} = \varepsilon, \quad T(\varepsilon) = \frac{M^2 R^2}{\varepsilon^2}$$

$$g(x) = \sum_{i=1}^{m} \operatorname{sgn}(\langle a_i, x \rangle - b_i) a_i = A^T \operatorname{sgn}(Ax - b)$$

$$\|g(x)\| = \|A^T \operatorname{sgn}(Ax - b)\| \le \|A\|_{op} \|\operatorname{sgn}(Ax - b)\| \le \|A\|_{op} \sqrt{m}$$

субградиентный метод: $T(\varepsilon) = \dfrac{m \|A\|_{op}^2 R^2}{\varepsilon^2}$

Преобразование Фенхеля

$$f^*(s) = \max_{x \in Q} \left(\langle s, x \rangle - f(x) \right), \quad f: Q \to \mathbb{R}$$
непр. $Q \leftarrow$ вып. компакт

утв. Пусть f сильно вып. с $\mu > 0$. Тогда f^* дифф-ма,
$$\nabla f^*(s) = x^*(s) = \operatorname*{argmax}_{x \in Q} \{\langle s, x \rangle - f(x)\} \qquad \{\Leftrightarrow\}$$

причём ∇f^* явл-ся липшицевым с константой $\frac{1}{\mu}$

$\square$ $\|\underbrace{\nabla f^*(s_1)}_{x_1^*} - \underbrace{\nabla f^*(s_2)}_{x_2^*}\|^* \le \frac{1}{\mu} \|s_1 - s_2\|$?

условие оптимальности: $0 \in s_1 - \partial f(x_1^*)$

$$\Leftrightarrow \; s_1 \in \partial f(x_1^*), \quad s_2 \in \partial f(x_2^*)$$

$$\begin{cases} s_1 \in \partial f(x_1^*) \\ s_2 \in \partial f(x_2^*) \end{cases} \iff \begin{cases} f(x) \geq f(x_1^*) + \langle s_1, x - x_1^* \rangle + \frac{\mu}{2}\|x - x_1^*\|^2 \\ f(x) \geq f(x_2^*) + \langle s_2, x - x_2^* \rangle + \frac{\mu}{2}\|x - x_2^*\|^2 \end{cases} \quad \forall x$$

$$\iff \begin{cases} f(x_2^*) \geq f(x_1^*) + \langle s_1, x_2^* - x_1^* \rangle + \frac{\mu}{2}\|x_2^* - x_1^*\|^2 \\ f(x_1^*) \geq f(x_2^*) + \langle s_2, x_1^* - x_2^* \rangle + \frac{\mu}{2}\|x_1^* - x_2^*\|^2 \end{cases}$$

$$\iff \langle s_1 - s_2, x_1^* - x_2^* \rangle \geq \mu \|x_1^* - x_2^*\|^2$$

$$\|s_1 - s_2\| \|x_1^* - x_2^*\| \geq \langle s_1 - s_2, x_1^* - x_2^* \rangle$$

$$\iff \underbrace{\|x_1^* - x_2^*\| \leq \frac{1}{\mu}\|s_1 - s_2\|}$$ ∎

$f(x) = \max\limits_{s \in Q} \{\langle s, x\rangle - f^*(s)\}$ обратное преобразование

$\min\limits_{x \in \mathbb{R}^n} f(x)$, $f^*: Q \to \mathbb{R}$ непр. вып. ф-ия

$$f_\mu(x) = \max_{s \in Q} \{\langle s, x\rangle - f^*(s) - \frac{\mu}{2}\underbrace{\|s - s_0\|^2}_{\leq D}\}$$

<u>утв.</u> $f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D \quad \forall x$

$D = \max\limits_{s \in Q} \frac{1}{2}\|s - s_0\|^2$, равномерная аппроксимация

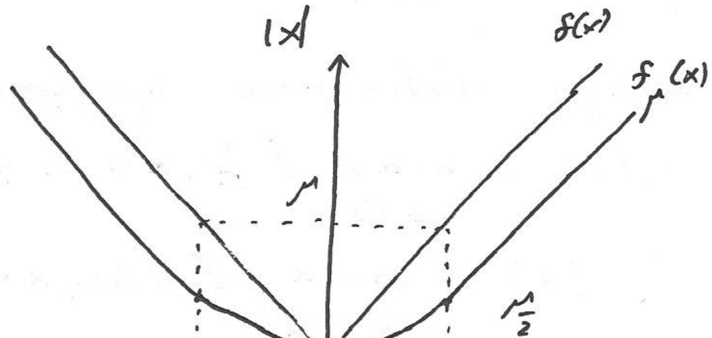
<u>Пример</u> $f: \mathbb{R} \to \mathbb{R}$, $f(x) = |x| \Rightarrow f(x) = \max\limits_{|s| \leq 1} s \cdot x$

$Q = [-1, 1]$, $D = \frac{1}{2}$

$f_\mu(x) = \max\limits_{|s| \leq 1} \{sx - \frac{\mu}{2}s^2\}$, $x = \mu s$, $s = \frac{x}{\mu}$

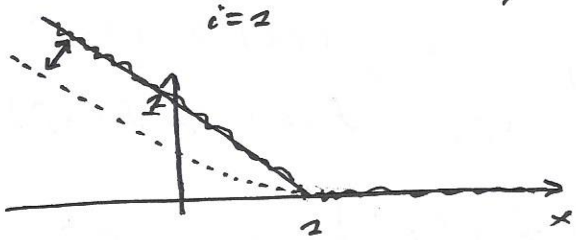
$$s_\mu^*(x) = \begin{cases} \frac{x}{\mu}, & |x| \leq \mu \\ \operatorname{sgn} x, & |x| > \mu \end{cases}$$

$$f_\mu(x) = \begin{cases} \frac{x^2}{2\mu}, & |x| \leq \mu \\ |x| - \frac{\mu}{2}, & |x| > \mu \end{cases}$$



SVM: $\min\limits_{x \in \mathbb{R}^n} \left\{ \frac{1}{2}\|x\|^2 + \sum\limits_{i=1}^{m} \max\{0, 1 - \langle a_i, x\rangle\} \right\}$

hinge loss



$\max\{0, 1-x\}$

$= \max\limits_{s \in [0,1]} s(1-x)$

f_μ липшиц. град. с $L_\mu = \frac{1}{\mu}$

град. спуск: $\frac{LR^2}{k}$ —

ускоренный: $\frac{LR^2}{k^2}$ +

$f_\mu(\bar{x}) - f^* \le ?$

$\underbrace{f_\mu(\bar{x})}_{\ge f(x) - \mu D} - f_\mu^* \le \frac{L_\mu R^2}{k^2}$, $f_\mu(x) \le f(x) \le f_\mu(x) + \mu D$

$\Rightarrow f(\bar{x}) - f^* \le \mu D + f_\mu(\bar{x}) - f_\mu^* \le \frac{L_\mu R^2}{k^2} + \mu D$

$\frac{L_\mu R^2}{k^2} + \mu D = \frac{R^2}{\mu k^2} + \mu D \to \min\limits_\mu$

$-\frac{R^2}{\mu^2 k^2} + D = 0$, $\frac{R^2}{\mu^2 k^2} = D$, $\mu = \frac{R}{k\sqrt{D}}$

" $f(\bar{x}) - f^* \le \frac{R\sqrt{D}}{k}$ "

$f(x) = \|Ax - b\|_2 = \max\limits_{s \in Q} \{\langle s, x\rangle - f^*(s)\}$

$f^*(s) = \max\limits_{x \in \mathbb{R}^n} \{\langle s, x\rangle - \|Ax - b\|_2\}$, $s = 0$ исх. зад.?

не работает!

тогда обобщённое представление Фенхеля:

$f(x) = \max\limits_{u \in Q} \{\langle Au, x\rangle - \varphi(u)\}$

$f_\mu(x) = \max\limits_{u \in Q} \{\langle Au, x\rangle - \varphi(u) - \frac{\mu}{2}\|u - u_0\|^2\}$

f_μ имеет липшицевый градиент с $\frac{\|A\|_{op}^2}{\mu}$

$$\mu_{opt} = \frac{R\|A\|_{op}}{k\sqrt{D}}$$

итоговая сложность:

$$(f(\bar{x}) - f^*) \leq \frac{R\|A\|_{op}\sqrt{D}}{k}$$

задача: $\|Ax - b\|_1 = \max\limits_{\|s\|_\infty \leq 1} \langle Ax - b, s \rangle$

$$f_\mu(x) = \max_{\|s\|_\infty \leq 1} \left\{ \langle Ax - b, s \rangle - \frac{\mu}{2}\|s\|^2 \right\} =$$

$|s_i| \leq 1$

$$= \sum_{i=1}^{m} \psi_\mu(\langle a_i, x \rangle - b_i)$$

$D = \frac{m}{2}$

гладкая аппроксимация модуля

ф-ия Хуберта

$$f(\bar{x}) - f^* \leq \frac{R\|A_{op}\|\sqrt{m}}{k}$$