11.12.17 момо XIV

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w) + h(w) \to \min_w$$

$f_i \in C^1$ вып., $h \in C$ вып.

$$\begin{cases} (1-\frac{\mu}{L})^k c = \varepsilon \\ k \log(1-\frac{\mu}{L}) + \log c = \varepsilon \\ k \approx O(\frac{L}{\mu}\log\frac{1}{\varepsilon}) \end{cases}$$

Метод	Число итераций
prox - GD	$O(N\frac{L}{\mu}\log\frac{1}{\varepsilon})$
acc. prox - GD	$O(N\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon})$
SAG / SVRG	$O((N+\frac{L}{\mu})\log\frac{1}{\varepsilon})$
?	$O((N+\sqrt{\frac{L}{\mu}})\log\frac{1}{\varepsilon})$

SDCA стохастический двойственный покоординатный подъём

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} \varphi_i(a_i^T w) + \frac{\lambda}{2}\|w\|^2 \to \min_w, \quad \varphi_i : \mathbb{R}\to\mathbb{R}$$

линейная регрессия : $\varphi_i(z) = \frac{1}{2}(y_i - z)^2$, $a_i = x_i$

логистическая регрессия : $\varphi_i(z) = \log(1+\exp(-z))$, $a_i = y_i x_i$

SVM : $\varphi_i(z) = \max(0, 1-z)$, $a_i = y_i x_i$

$$P(w,z) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\varphi_i(z_i) + \frac{\lambda}{2}\|w\|^2 \to \min_{z,w} \\ Aw = z \end{cases}$$

$$L(w,z,\mu) = \frac{1}{N}\sum_{i=1}^{N}\varphi_i(z_i) + \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\mu^T(z - Aw)$$

$q(\mu) = \inf\limits_{w, z} L(w, z, \mu)$

$\nabla_w L = \lambda w - \frac{1}{N} A^T \mu = 0, \quad w = \frac{1}{\lambda N} A^T \mu = \frac{1}{\lambda N} \sum\limits_{i=1}^{N} \mu_i a_i$

отн. z_i : $\varphi_i(z_i) + \mu_i z_i \to \min\limits_{z_i}$

$\varphi_i^*(u) = \max\limits_{z} (uz - \varphi_i(z))$

$q(\mu) = \frac{1}{N} \sum\limits_{i=1}^{N} \min\limits_{z_i} (\varphi_i(z_i) + \mu_i z_i) + \frac{\lambda}{2} \| \frac{1}{\lambda N} A^T \mu \|^2 -$

$- \lambda \left(\frac{1}{\lambda N} A^T \mu \right)^T \frac{1}{\lambda N} A^T \mu = \frac{1}{N} \sum\limits_{i=1}^{N} \left(- \max\limits_{z_i} (-\mu_i z_i - \varphi_i(z_i)) \right) -$

$- \frac{\lambda}{2} \| \frac{1}{\lambda N} A^T \mu \|^2 = \frac{1}{N} \sum\limits_{i=1}^{N} (-\varphi_i^*(-\mu_i)) - \frac{\lambda}{2} \| \frac{1}{\lambda N} A^T \mu \|^2 \to \max\limits_{\mu}$

Схема SDCA

$\mu^{(0)}; \ w^{(0)} = \frac{1}{\lambda N} A^T \mu^{(0)}$

для $k = 0, 1, 2, \dots$

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

Найти $\delta \mu_{i_k}$:

$\frac{1}{N} \left(-\varphi_{i_k}^*(-\mu_{i_k}^{(k)} - \delta \mu_{i_k}) \right) - \frac{\lambda}{2} \| w^{(k)} + \frac{1}{\lambda N} a_{i_k}^T \delta \mu_{i_k} \|^2 \to \max\limits_{\delta \mu_{i_k}}$

$\mu^{(k+1)} = \mu^{(k)} + \delta \mu_{i_k} \cdot e_{i_k}$

$w^{(k+1)} = w^{(k)} + \frac{1}{\lambda N} a_{i_k}^T \delta \mu_{i_k}$

<u>Утв.</u> $\varphi_i \in C_L^{1,1} \Rightarrow$ для $\mathbb{E} F(w_k) - F_{opt} \le \varepsilon$

в SDCA достаточно сделать $O((N + \frac{L}{\mu}) \log \frac{1}{\varepsilon}))$ итер.

Нет необходимости подбирать оптимальную длину шага α_k. Все рассчёты на итерации производятся аналитически.

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} \varphi_i(a_i^T w) + h(w) \to \min_w$$

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N} \varphi_i(z_i) + h(w) \to \min_w \\ Aw = z \end{cases}$$

$$L(z, w, \mu) = \frac{1}{N}\sum_{i=1}^{N} \varphi_i(z_i) + h(w) + \frac{1}{N}\mu^T(z - Aw)$$

$$q(\mu) = \frac{1}{N}\sum_{i=1}^{N} \left(-\varphi_i^*(-\mu_i)\right) + \min_w \left(h(w) + \frac{1}{N}\mu^T A w\right) =$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left(-\varphi_i^*(-\mu_i)\right) - \underbrace{\max_w \left(-w^T \frac{A^T\mu}{N} - h(w)\right)}_{h^*(-\frac{1}{N}A^T\mu)} \to \max_\mu$$

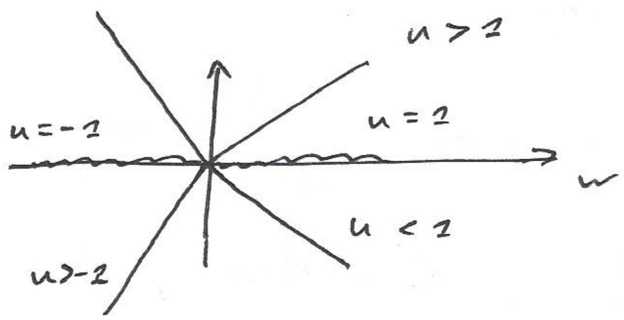
$$h(w) = \|w\|_1$$

$$h^*(u) = \max_w \left(w^T u - \|w\|_1\right) = \left\{\max_{w_i} \left(w_i u_i - |w_i|\right)\right\}$$

$$wu - |w| \to \max$$

$w \geq 0$, $(u-1)w$

$w \leq 0$, $(u+1)w$



$\max_w \{wu - |w|\} = 0$ при $|u| \leq 1$, ∞ иначе, $\quad h^*(u) = [|u| \leq 1] = \delta_{|u| \leq 1}(u)$

$$F_k(w) = F(w) + \frac{\varkappa}{2}\|w - y_{k-1}\|^2 \to \min_w$$

$L_{F_k} = L + \varkappa$, $\quad \operatorname{cond} F = \frac{L}{\mu} \searrow$

$\mu_{F_k} = \mu + \varkappa$, $\quad \operatorname{cond} F_k = \frac{L + \varkappa}{\mu + \varkappa}$

Схема Acc. SDCA

$$R^2 = \max_i \|x_i\|^2$$

$$\varkappa = \frac{R^2 L}{N} - \mu \; ; \; \eta = \sqrt{\frac{\mu}{\mu + \varkappa}} \; ; \; \beta = \frac{1-\eta}{1+\eta}$$

$$y_1 = w_1 = 0 \; , \; \alpha_1 = 0, \; \xi_1 = (1 + \eta^{-2})(P(0) - q(0))$$

для $k = 2, 3, \dots$

$$F_k(w) = F(w) + \frac{\varkappa}{2}\|w - y_{k-1}\|^2$$

$$(w_k, \alpha_k) = \text{prox-SDCA}\Big(F_k, \alpha_{k-1}, \underbrace{\frac{\eta}{2(1+\eta^{-2})}}_{=\varepsilon_k} \xi_{k-1}\Big)$$

$$y_k = w_k + \beta(w_k - w_{k-1})$$

$$\xi_k = (1 - \eta/2)\xi_{k-1}$$

$$\tilde{O}\left(\left(N + \sqrt{\frac{N R^2 L}{\mu}}\right)\log(1/\varepsilon)\right)$$

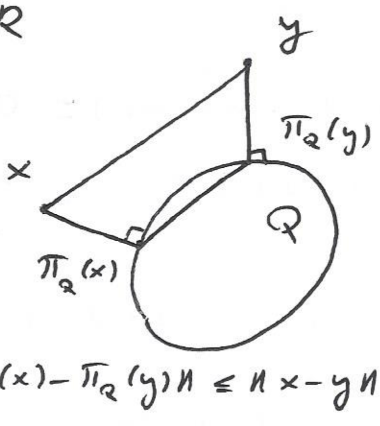
число итераций

семинар

$$\frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

$$f(x) = \mathbb{E}\, F(x, z) = \int_{\Omega} F(x, w)\, dP(w)$$

$$\min_{x \in Q} f(x) \quad , \quad Q - \text{вып.}, \quad f : Q \to \mathbb{R}$$



$$\text{SVM}: \frac{1}{N}\sum_{i=1}^{N} \max\{0; 1 - \langle a_i, x\rangle\}$$

robust regression : $\displaystyle \frac{1}{N}\sum_{i=1}^{N} |\langle a_i, x\rangle - b_i|$

$$\|\pi_Q(x) - \pi_Q(y)\| \le \|x - y\|$$

$$\min_x f(x) \quad , \quad \mathbb{E}(g_k | x_k) \in \partial f(x_k)$$

$$\{x_k\} \subseteq Q \quad , \quad f(x) - f(x_k) \geq \langle G_k, x_k - x_k \rangle \quad \forall x \in Q$$

$$G_k \in \partial f(x_k)$$

$$x_{k+1} = \pi_Q(x_k - \alpha_k g_k)$$

$$\|x_{k+1} - x^*\|^2 \le \| x_k - x^* - \alpha_k g_k \|^2 = \| x_k - x^*\|^2 - 2\alpha_k \langle g_k, x_k - x^*\rangle +$$

$$+ \alpha_k^2 \|g_k\|^2 \quad ; \quad \alpha_k \langle g_k, x_k - x^*\rangle \le \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \frac{\alpha_k^2}{2}\|g_k\|^2$$

телескопирующая сумма: $\sum_{k=1}^{T} (a_k - a_{k+1}) = a_1 - a_{T+1}$

$$\sum_{k=1}^{T} \alpha_k \underbrace{\langle g_k, x_k - x^*\rangle}_{\ge f(x_k) - f^* \ ?} \le \frac{1}{2}\underbrace{\|x_1 - x^*\|^2}_{R^2} + \sum_{k=1}^{T} \frac{\alpha_k^2}{2}\|g_k\|^2$$

$$\sum_{k=1}^{T} \mathbb{E}(\alpha_k \langle g_k, x_k - x^*\rangle) \le \frac{R^2}{2} + \sum_{k=1}^{T} \mathbb{E}\, \frac{\alpha_k^2}{2}\|g_k\|^2$$

необходимо: α_k – детерминированные, тогда

$$\sum_{k=1}^{T} \alpha_k \, \mathbb{E}\langle g_k, x_k - x^*\rangle \le \frac{R^2}{2} + \sum_{k=1}^{T} \frac{\alpha_k^2}{2} \mathbb{E}\|g_k\|^2$$

$$\mathbb{E}\langle g_k, x_k - x^*\rangle = \mathbb{E}\,\mathbb{E}(\langle g_k, x_k - x^*\rangle \mid x_k) = \mathbb{E}(\langle \mathbb{E} g_k, x_k - x^*\rangle \mid x_k) \ge$$

$$\ge \mathbb{E}(\mathbb{E}(f(x_k) - f^* \mid x_k)) = \mathbb{E}(f(x_k) - f^*)$$

$$\frac{\sum_{k=1}^{T} \alpha_k \, \mathbb{E}(f(x_k) - f^*)}{\sum_{k=1}^{T} \alpha_k} \ge \mathbb{E} f\Big(\underbrace{\sum_{k=1}^{T} \alpha_k x_k}_{\bar{x}}\Big) - f^* = \mathbb{E} f(\bar{x}) - f^*$$

йенсен

$$\{ \mathbb{E}(f(x_k) - f^* \mid x_k) \ge (\langle \mathbb{E} g_k, x_k - x^*\rangle \mid x_k) \}$$

$$\frac{R^2}{\sum_{k=1}^{T} \alpha_k} + \frac{\sum_{k=1}^{T} \alpha_k^2 \mathbb{E}\|g_k\|^2}{\sum_{k=1}^{T} \alpha_k}$$

$$\alpha_k = \alpha : \quad \frac{R^2}{\alpha T} + \frac{\alpha \sum_{k=1}^{T} \mathbb{E}\|g_k\|^2}{T} \to \min_{\alpha} \quad \underline{(*)}$$

$$\frac{R^2}{\alpha} = \alpha \sum_{k=1}^{T} \mathbb{E}\|g_k\|^2, \quad \alpha = \frac{R}{\left(\sum_{k=1}^{T} \mathbb{E}\|g_k\|^2\right)^{\frac{1}{2}}}$$

$$\frac{R}{\sqrt{T}} \left(\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\|g_k\|^2 \right)^{\frac{1}{2}}, \quad \alpha = \frac{R}{\sqrt{T}\, M}$$

① стох. субгр. метод

$$\alpha_k = \alpha = \frac{R}{M\sqrt{T}}, \quad (*) = \frac{MR}{\sqrt{T}}$$

пример: $f(x) = \frac{1}{N} \sum_{i=1}^{N} |\langle a_i, x\rangle - b_i|$

i_k равномерное из $\{1 \dots N\}$

$$g_k = \operatorname{sgn}(\langle a_{i_k}, x_k\rangle - b_{i_k})\, a_{i_k}$$

$$\mathbb{E}(g_k | x_k) \in \partial f(x_k)$$

$$\|g_k\|^2 \le \|a_{i_k}\|^2, \quad \mathbb{E}\|g_k\|^2 \le \frac{1}{N}\sum_{i=1}^{N} \|a_i\|^2 = M$$

$$(*) = \frac{\left(\frac{1}{N}\sum_{i=1}^{N} \|a_i\|^2\right)^{\frac{1}{2}} R}{\sqrt{T}} \le \max_{1\le i\le N} \|a_i\|^2 R / \sqrt{T}$$

② стох. субгр. метод с адаптивными длинами шагов

α_k - случайные, зависят от $(g_1 \dots g_k)$

...

$$\langle g_k, x_k - x^*\rangle \le \frac{1}{2\alpha_k}\|x_k - x^*\|^2 - \frac{1}{2\alpha_k}\|x_{k+1} - x^*\|^2 + \frac{\alpha_k \|g_k\|^2}{2}$$

$$\sum_{k=1}^{T} \frac{1}{2\alpha_k}\|x_k - x^*\|^2 - \sum_{k=1}^{T}\frac{1}{2\alpha_k}\|x_{k+1} - x^*\|^2 =$$

$$= \sum_{k=1}^{T} \frac{1}{2\alpha_k}\|x_k - x^*\|^2 - \sum_{k=2}^{T+1} \frac{1}{2\alpha_{k-1}}\|x_k - x^*\|^2 =$$

$$= \frac{1}{2\alpha_1}\|x_1 - x^*\|^2 - \frac{1}{2\alpha_T}\|x_{T+1} - x^*\|^2 + \sum_{k=2}^{T}\left(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}}\right)\underbrace{\|x_k - x^*\|^2}_{\le R^2} \le$$

пусть мн-во Q-ограничено: $\|x_k - x^*\| \le R$

и $\{\alpha_k\}$ монотонно убывают

$$\le \frac{R^2}{2\alpha_1} + \frac{R^2}{2}\left(\frac{1}{\alpha_T} - \frac{1}{\alpha_1}\right) = \frac{R^2}{2\alpha_T}$$

$$\sum_{k=1}^{T} \langle g_k, x_k - x^*\rangle \le \frac{R^2}{2\alpha_T} + \sum_{k=1}^{T}\frac{\alpha_k}{2}\|g_k\|^2$$

$$\overline{x}_T = \frac{1}{T} \sum_{k=1}^{T} x_k$$

$$\mathbb{E} f(\overline{x}_T) - f^* \le \mathbb{E}\left(\frac{R^2}{2T\alpha_T} + \frac{1}{2T} \sum_{k=1}^{T} \alpha_k \|g_k\|^2 \right) \quad \underline{(*)}$$

$$\alpha_k \equiv \alpha : \quad \frac{R^2}{\alpha} + \alpha \sum_{k=1}^{T} \|g_k\|^2 \to \min_{\alpha}$$

$$\alpha_k = \frac{R}{\left(\sum_{k=1}^{T} \|g_k\|^2\right)^{\frac{1}{2}}}$$

выбираем $\alpha_k = \dfrac{R}{\left(\sum_{s=1}^{k-1} \|g_k\|^2\right)^{\frac{1}{2}}}$

$$(*) = \frac{R}{T} \left(\sum_{s=1}^{T-1} \|g_s\|^2 \right)^{\frac{1}{2}} + \frac{R}{T} \sum_{k=1}^{T} \frac{\|g_k\|^2}{\left(\sum_{s=1}^{k-1} \|g_s\|^2\right)^{\frac{1}{2}}}$$

нер-во: $\displaystyle \sum_{k=1}^{n} \frac{a_i}{\left(\sum_{j=1}^{k} a_j\right)^{\frac{1}{2}}} \le 2 \left(\sum_{i=1}^{n} a_i \right)^{\frac{1}{2}}$

$$\mathbb{E}(*) = \frac{R}{\sqrt{T}} \mathbb{E} \left(\frac{1}{T} \sum_{k=1}^{T} \|g_k\|^2 \right)^{\frac{1}{2}} \le \frac{R}{\sqrt{T}} \left(\frac{1}{T} \sum_{k=1}^{T} \underbrace{\mathbb{E}\|g_k\|^2}_{\le M^2} \right)^{\frac{1}{2}} \le \frac{MR}{\sqrt{T}}$$

3 AdaGrad

$$x_{k+1} = \pi_Q^{B_k} (x_k - B_k^{-1} g_k) = \operatorname*{argmin}_{x \in Q} \{ f(x_k) + \langle g_k, x - x_k \rangle +$$
$$+ \frac{1}{2} \langle B_k (x - x_k), x - x_k \rangle \} \qquad \text{Раньше: } B_k = \frac{1}{\alpha_k} I$$

теперь: $B_k = \operatorname{diag}\{b_k\}$

$$\|x_{k+1} - x^*\|^2_{B_k} = \|x_k - x^*\|_{B_k} - \dots$$

$$\sum_{k=1}^{T} \|x_k - x^*\|^2_{B_k} - \sum_{k=1}^{T} \|x_{k+1} - x^*\|_{B_k}$$

$$\langle Ds, s \rangle = \sum_j d_j s_j^2 \le \|s\|^2_\infty \operatorname{tr}(D)$$

$$\langle \underbrace{(B_k - B_{k-1})}_{D} \underbrace{(x_k - x^*)}_{S}, \underbrace{x_k - x^*}_{S} \rangle$$

$$\bar{x}_T = \frac{1}{T} \sum_{k=1}^{T} x_k$$

$$\mathbb{E} f(\bar{x}_T) - f^* \le \frac{R_\infty^2 \, t_2(B_T)}{2T} + \frac{1}{2T} \sum_{k=1}^{T} \langle B_k^{-1} g_k, g_k \rangle \ \underline{(*)}$$

$$B_k = \operatorname{diag} \{ \sum_{s=1}^{k-1} g_s g_s^T \}^{\frac{1}{2}}$$

$$(*) = \frac{R_\infty}{\sqrt{T}} \sum_{i=1}^{n} \mathbb{E} \left(\frac{1}{T} \sum_{k=1}^{T} g_{ki}^2 \right)$$

$$\|S\|_\infty \le \|S\|_2, \quad R_2 = \sqrt{n} \, R_\infty, \quad R_2 \le \sqrt{n} \, R_\infty$$

Ada Grad vs Ada Step Size

$$\frac{1}{n} \sum_{j=1}^{n} \sqrt{a_j} \le \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{n} a_j \right)^{\frac{1}{2}}, \quad \frac{1}{n} \sum_{j=1}^{n} \sqrt{a_j} \le \left(\frac{1}{n} \sum_{j=1}^{n} a_j \right)^{\frac{1}{2}}$$

Вогнутость корня

$$\sum_{j=1}^{n} \sqrt{a_j} \le \sqrt{n} \left(\sum_{j=1}^{n} a_j \right)^{\frac{1}{2}}$$

ada grad $\le$ ada step size $\le$ sgd

$$R_\infty = \|x - x_*\|_\infty = \max_j \{ |x - x_*|_j \}$$
$$R_2 = \|x - x_*\|_2 = \left(\sum_j (x - x_*)_j^2 \right)^{\frac{1}{2}}$$