Exam 2016-2017

Ex. 1 (Logistic regression with gross outliers)

① $f_i(x,c) = \log\left(1 + e^{-b_i(a_i^T x + c)}\right)$

$\nabla f_i(x,c) = \begin{pmatrix} \nabla_x f_i(x,c) \\ \nabla_c f_i(x,c) \end{pmatrix}$

$\nabla_x f_i(x,c) = -\frac{-b_i a_i e^{-b_i(a_i^T x + c)}}{(1+e^{\cdots})} + \lambda x = -\frac{b_i a_i}{(1+e^{b_i(a_i^T x + c)})} + \lambda x$

$\nabla_c f_i(x,c) = \frac{-b_i}{(1+e^{b_i(a_i^T x + c)})}$

$A^T \in \mathbb{R}^{d \times n}$ $\begin{pmatrix} \frac{-b_1}{1+e^{b_1(a_1^T x + c)}} \\ \frac{-b_n}{1+e^{b_n(a_n^T x + c)}} \end{pmatrix}$

Donc $\nabla_x f(x,c) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x,c) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{b_i a_i}{(1+e^{b_i(a_i^T x + c)})} + \lambda x\right)$

$\nabla_c f(x,c) = -\frac{1}{n}\sum_{i=1}^{n} \frac{b_i}{(1+e^{b_i(a_i^T x + c)})}$

Computing $\nabla f_i(x,c)$ is $O(d)$

Computing $\nabla f(x,c)$ is $O(nd)$ (for a single x)

② For $t = 1, \dots, T$,

$\begin{cases} \text{sample } i \in \{1, \dots, n\} \\ x_{t+1} = x_t - \alpha \nabla f_i(x_t) \end{cases}$ where α = stepsize

③ We rewrite the iterate $x_t = \beta_t z_t$, $\beta_t \in \mathbb{R}$, $z_t \in \mathbb{R}^d$

$x_{t+1} = x_t - \eta\left[-\frac{b_i a_i}{(1+e^{b_i(a_i^T x_t + c)})} + \lambda x_t\right]$

$= x_t(1-\lambda\eta) + \eta a_i \ell(a_i^T x_t)$ avec $\ell(a_i^T x_t) = \frac{b_i}{1+e^{b_i(a_i^T x_t + c)}}$

$\Rightarrow \beta_{t+1} z_{t+1} = (1-\lambda\eta)\beta_t z_t + \eta a_i \ell(a_i^T \beta_t z_t)$ $O(1)$

$= \underbrace{(1-\lambda\eta)\beta_t}_{\beta_{t+1}}\left[z_t - \frac{\eta a_i \ell(a_i^T \beta_t z_t)}{(1-\eta\lambda)\beta_t}\right]$ $O(1)$

$\beta_{t+1} = (1-\lambda\eta)\beta_t$ $\} O(1)$

$z_{t+1} = z_t - \frac{\eta a_i \ell(a_i^T \beta_t z_t)}{(1-\eta\lambda)\beta_t}$ $\} O(1)$

④ Now the vector c of the biases is of length $\mathbb{R}^m$. But we want to shrink the c_i of the outliers?

We use the ℓ_1 norm because the $\|\cdot\|$ is more robust to outliers. The higher γ, the more c_i are set to zero.

⑤ Ici vérifié

⑥ On peut réécrire $f_i(x,c)$ par : $\log(1+\exp(-b_i(a_i^T x + c^T e_i))) + \frac{1}{2}\|x\|$

Donc $\nabla_x f_i(x,c)$ ne change pas et est toujours égal à :

$$\frac{-b_i a_i}{(1+e^{b_i(a_i^T x + c_i)})} + dx.$$

et : $\nabla_c f_i(x,c) = \frac{-b_i e_i}{1+e^{b_i(a_i^T x + c_i)}} + dx$

Donc $\nabla f_i(x,c) \in \mathbb{R}^{d+n}$ peut s'écrire comme :

$$\alpha_i \begin{pmatrix} \frac{b_i}{1+e^{b_i(a_i^T x + c_i)}} \times \left(-a_i + \frac{dx}{\alpha_i}\right) \in \mathbb{R}^d \\ -e_i \in \mathbb{R}^n \end{pmatrix} := z_i \in \mathbb{R}^d$$

et $\nabla f(x,c) \in \mathbb{R}^{d+n} = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x,c) = \frac{1}{m}\sum_{i=1}^{m} \frac{b_i z_i}{1+e^{b_i(a_i^T x + c_i)}}$

⑦ Algo :

$$\begin{cases} x_{t+1} = x_t - \eta\left(\frac{d}{m}\sum_{i=1}^{m} \frac{-b_i a_i}{1+e^{b_i(a_i^T x_t + c_{i,t})}} + dx\right) \\ c^{t+1} = \text{prox}_{\eta\gamma\|\cdot\|_1^1}\left(c_t + \underbrace{\frac{\eta b_i e_i}{1+e^{b_i(a_i^T x_t + c_{i,t})}}}_{-\nabla_c f(x,c)}\right) \end{cases}$$

Avec $\text{prox}_{\eta\|\cdot\|_1} = (S_{\eta\gamma}(c_1), \ldots, S_{\eta\gamma}(c_n))$

Exo.2: Element of proof of the SAGA algorithm

① $T^k = \frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(x^*) - \frac{1}{n}\sum_{i=1}^n \langle \nabla f_i(x^*), \phi_i^k - x^* \rangle + c\|x^k - x^*\|^2$

$= \frac{1}{n}\sum_{i=1}^n \left(f_i(\phi_i^k) - f(x^*) - \langle \nabla f_i(x^*), \phi_i^k - x^* \rangle \right) + c\|x^k - x^*\|^2$

somme de fonctions convexes.

Donc $T^k \geq c\|x^k - x^*\|^2$

② According to the previous question, $T^k \geq c\|x^k - x^*\|^2$

Therefore, $c\,\mathbb{E}[\|x^k - x^*\|^2] \leq \mathbb{E}[T^k]$

$= \mathbb{E}\left(\mathbb{E}[T^k \mid \mathcal{F}_{k-1}]\right)$

$\leq \mathbb{E}\left[\left(\frac{1}{2} - \frac{1}{4k}\right) T^{k-1}\right]$ d'après l'énoncé

$\leq \left(1 - \frac{1}{k}\right)\mathbb{E}(T^{k-1})$

$= \left(1 - \frac{1}{k}\right)\mathbb{E}\left[\mathbb{E}\left[T^{k-1} \mid \mathcal{F}_{k-2}\right]\right]$

$\leq \left(1 - \frac{1}{k}\right)^2 \mathbb{E}[T^{k-2}]$

$\leq \left(1 - \frac{1}{k}\right)^k \mathbb{E}[T_0]$

Mais T_0 est déterministe, donc : $\mathbb{E}(T_0) = T_0$ et :

$$c\,\mathbb{E}[\|x^k - x^*\|^2] \leq \left(1 - \frac{1}{k}\right)^k T_0$$

~~Donc $c\,\mathbb{E}[\|x^k - x^*\|^2]$~~

Lets now prove that (1) holds :

$T_0 = \frac{1}{n}\sum_{i=1}^n f_i(\phi_i^0) - f(x^*) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*)^T(\phi_i^0 - x^*) + c\|x_0 - x^*\|^2$ ($\phi_i^0 = x_0$)

$= f(x_0) - f(x^*) - \nabla f(x^*)^T(x^0 - x^*) + c\|x_0 - x^*\|^2$

Donc $\mathbb{E}[\|x^k - x^*\|^2] \leq \left[\frac{1}{c}\left(f(x^0) - \nabla f(x^*)^T(x_0 - x^*) - f(x^*) + \|x_0 - x^*\|^2\right)\right]\left(1 - \frac{1}{k}\right)^k$

D'après l'énoncé, ~~$\alpha = \frac{\mu}{2\eta n}$~~

$\frac{1}{\alpha} = \gamma\mu = \frac{\mu}{2(\eta\mu + L)} \Rightarrow 1 - \frac{1}{k} = 1 - \frac{\mu}{2(\eta\mu + L)}$

et $\frac{1}{c} = 2\eta n(1 - \gamma\mu) = \frac{\eta n}{\eta(n\mu + L)}\left(1 - \frac{\mu}{2(\mu n + L)}\right)$

$\leq \frac{n}{n\mu + L}$ ≤ 1

Par convexité, $f(x_0^0) - \langle \nabla f(x^*), x^0 - x^* \rangle \geq 0$.
(car $f(x^*)=0$)

Donc : $\frac{1}{C}(\quad) \leq \frac{n}{\mu n + L}(\quad)$.

$\rightsquigarrow$ donc on a prouvé (1).

(3) $\mathbb{E}[T^{k+1} \mid \mathcal{F}_k] = \mathbb{E}_{I_k}[T^{k+1}]$

$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{I_k}[f_i(\phi_i^{k+1})] - f(x^*) - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{I_k}(\nabla f_i(x^*)^T(\phi_i^k - x^*))}_{B} + C\,\mathbb{E}_{I_k}[\|x^{k+1}-x^k\|^2]$

$\mathbb{E}_{I_k}[f_i(\phi_{i_\alpha}^{k+1})] = \frac{1}{n} f_i(x^k) + (1-\frac{1}{n}) f_i(\phi_i^k)$

$\alpha \sim \frac{1}{n}$ ↪ pour chaque I_k

- Donc $A = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{I_k}[f_i(\phi_i^{k+1})] = \frac{1}{n}\left(\sum_{i=1}^{n}\left(\frac{1}{n} f_i(x^k) + (1-\frac{1}{n}) f_i(\phi_i^k)\right)\right)$

$= \frac{1}{n} f(x^k) + (1-\frac{1}{n}) \frac{1}{n}\sum_{i=2}^{n} f_i(\phi_i^k)$ ($f(x^k)$)

De même, $\mathbb{E}_{I_k}[\nabla f_i(x^*)^T(\phi_i^{k+1} - x^*)]$

$= \frac{1}{n}\nabla f_i(x^*)^T(x^k - x^*) + (1-\frac{1}{n})\nabla f_i(x^*)(\phi_i^k - x^*)$

Donc $B = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{I_k}[\nabla f_i(x^*)^T(\phi_i^{k+1} - x^*)]$

$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n}\nabla f_i(x^*)^T(x^k - x^*) + (1-\frac{1}{n})\nabla f_i(x^*)^T(\phi_i^k - x^*)$

$\nabla f(x^*)^T(x^k - x^*)$

Donc $-\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{I_k}[\nabla f_i(x^*)^T(\phi^{k+1} - x^*)]$

$= \frac{1}{n}\nabla f(x^*)^T(x^k - x^*) - \frac{1}{n}(1-\frac{1}{n})\sum_{i=1}^{n}\nabla f_i(x^*)^T(\phi_i^k - x^*)$

En recollant les morceaux, on a bien l'identité recherchée.

④ D'après l'énoncé, $\phi_{I_k}^{k+1} = x^k$.

$$\mathbb{E}_{I_k}\left[\nabla f_{I_k}(\phi_{I_k}^{k+1})\right] = \mathbb{E}_{I_k}\left[\nabla f_{I_k}(x^k)\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \nabla_i f(x^k) = \nabla f(x^k).$$

Pour prouver l'égalité, il reste à montrer que :

$$\mathbb{E}\left[-\nabla I_k(\phi_{I_k}^k) + \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\phi_i^k)\right] = 0.$$

On a : $\mathbb{E}(\quad) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\phi_i^k) + \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\phi_i^k) = 0$ ✓

⑤ $\|x^{k+1} - x^*\| = \|x^{k+1} - x^k + x^k - x^*\|^2$

$$= \|x^k - x^*\|^2 + 2(x^{k+1} - x^k)^T(x^k - x^*) + \|x^{k+1} - x^k\|^2$$

mais $x^{k+1} - x^k = -\gamma\left(\nabla f_{I_k}(x^k) - \nabla f_{I_k}(\phi_{I_k}^k) + \frac{1}{m}\sum_{i=1}^{m} f_i(\phi_i^k)\right)$

$$\mathbb{E}_{I_k}\left[\|x^{k+1} - x^*\|^2\right] = \mathbb{E}_{I_k}\left[\|x^k - x^*\|^2\right] + 2(x^k - x^*)^T \mathbb{E}_{I_k}(x^{k+1} - x^k)$$

(déterministe)

$$+ \mathbb{E}_{I_k}\left(\|x^{k+1} - x^k\|^2\right).$$

$$= \|x^k - x^*\|^2 - 2\gamma(x^k - x^*)^T \nabla f(x^k) + \gamma^2 \mathbb{E}_{I_k}(\quad)$$

Comme f est convexe, et on minimise sur $\mathbb{R}^d$ alors $\nabla f(x^*) = 0$

Doi le result.

⑥

⑦ Prouvons que g est $L-\mu$ smooth.

$$g(y)-g(x) = f(y)-f(x) - \frac{\mu}{2}\|y\|^2 + \frac{\mu}{2}\|x\|^2$$
$$\leq \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 - \frac{\mu}{2}\|y\|^2 + \frac{\mu}{2}\|x\|^2$$

(f L-smooth)

Mais $\nabla f(x)^T = (\nabla g(x) + \mu x)^T$

Donc: $g(y)-g(x) \leq \nabla g(x)^T(y-x) + \mu x^T(y-x) - \frac{\mu}{2}\|y\|^2 + \frac{\mu}{2}$
$+ \frac{L}{2}\|y$

Mais $\|y\|^2 = \|y-x+x\|^2 = \|y-x\|^2 + \|x\|^2 + 2(y-x)^T x$

$$\Rightarrow -\frac{\mu}{2}\|y\|^2 = -\frac{\mu}{2}\|y-x\|^2 - \frac{\mu}{2}\|x\|^2 - \mu x^T y$$

Donc $g(y)-g(x) \leq \nabla g(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 - \frac{\mu}{2}\|y-x\|^2$
$= \nabla g(x)^T(y-x) + \frac{L-\mu}{2}\|y-x\|^2$

Donc g est $(L-\mu)$ smooth

Con: $\Rightarrow \langle \nabla g(y) - \nabla g(x), y-x \rangle \geq \frac{1}{(L-\mu)}\|\nabla g(y) - \nabla g(x)$