EXAM 2016 - 2017

Within exercises the questions are not independent, but, at each step, you may use the results of previous questions, even if you have not succeeded in proving them.

Documents are allowed, but not electronic devices
3 Hours

Exercise 1 (Logistic regression with gross outliers)

Consider the logistic regression problem with ridge regularization

$$\underset{x\in\mathbb{R}^d, c\in\mathbb{R}}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-b_i(a_i^\top x + c))) + \frac{\lambda}{2}\|x\|_2^2 \right\}$$

where $a_i \in \mathbb{R}^d$ are the features, $b_i \in \{-1, 1\}$ are the labels, and $x \in \mathbb{R}^d$ are the weights and $c \in \mathbb{R}$ is the intercept of the model. Put

$$f_i(x,c) = \log(1+\exp(-b_i(a_i^\top x + c))) + \frac{\lambda}{2}\|x\|_2^2 \quad \text{and} \quad f(x,c) = \frac{1}{n}\sum_{i=1}^{n} f_i(x,c).$$

1. Compute $\nabla f_i(x,c)$ and $\nabla f(x,c)$. What is the complexity of computing ∇f_i and ∇f for a single i ?
2. Write the iteration of the stochastic gradient algorithm to minimize f.
3. Assuming that features vectors a_i are sparse, namely $s_i = \{j : a_i^j \neq 0\}$ is much smaller than d for most i, explain how to make the numerical complexity of one iteration of SGD equal to $O(s_i)$.
4. Consider now that a small subset of sample points (a_i, b_i) from the dataset contains gross outliers: we replace the population intercept $c \in \mathbb{R}$ by individual intercepts $c_i \in \mathbb{R}$ for $i = 1, \ldots, n$, leading to the problem

$$\underset{x\in\mathbb{R}^d, c\in\mathbb{R}^n}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-b_i(a_i^\top x + c_i))) + \frac{\lambda}{2}\|x\|_2^2 + \gamma\|c\|_1 \right\}.$$

   where $\gamma > 0$ is an extra regularization parameter, and where $\|c\|_1$ stands for the ℓ_1-norm of $c \in \mathbb{R}^n$. Explain with words why ℓ_1-penalization on c is used for, and what is the effect of $\gamma > 0$.
5. Decompose this objective into a sum $\frac{1}{n}\sum_{i=1}^{n} f_i(x,c) + g(x,c)$ where the f_i are smooth and g is non-smooth.
6. Compute again $\nabla f_i(x,c)$ and $\nabla f(x,c)$ for this problem.
7. Write explicitly the iterations of the proximal gradient algorithm for this problem.

Exercise 2 (Elements of proof for the SAGA algorithm)

We want to minimize a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

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

where all functions f_i are L-smooth and μ-strongly convex. For this purpose, we consider the SAGA algorithm. Consider a starting point $x^0 \in \mathbb{R}^d$ and initial gradients $\nabla f_i(\phi_i^0)$ where $\phi_i^0 = x^0$ for all i. At iteration k, given x^k and each $\nabla f_i(\phi_i^k)$, the next iterate x^{k+1} is obtained as follows: pick uniformly at random I_k in $\{1, \ldots, n\}$ and take $\phi_{I_k}^{k+1} = x^k$ and $\phi_i^{k+1} = \phi_i^k$ for $i \neq I_k$. Then, compute $\nabla f_{I_k}(\phi_{I_k}^{k+1})$ and take the next iterate as

$$x^{k+1} \leftarrow x^k - \gamma\Big(\nabla f_{I_k}(\phi_{I_k}^{k+1}) - \nabla f_{I_k}(\phi_{I_k}^k) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\phi_i^k)\Big),$$

where $\gamma = 1/(2(n\mu + L))$. The aim of this exercise is to give some elements of proof for the following inequality

$$\mathbb{E}\|x^k - x^*\|^2 \leq \Big(1 - \frac{\mu}{2(\mu n + L)}\Big)^k \Big(\|x^0 - x^*\|^2 + \frac{n}{\mu n + L}\big(f(x^0) - \langle \nabla f(x^*), x^0 - x^*\rangle - f(x^*)\big)\Big), \quad (1)$$

where $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$, which proves linear convergence of the algorithm. In what follows $\mathbb{E}[\cdot|\mathcal{F}_k]$ stands for the conditional expectation with respect to the random draws $I_1, \ldots, I_{k-1}$, and $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ for the Euclidean inner product and norm.

1. The main trick is to introduce
$$T^k = \frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - f(x^*) - \frac{1}{n}\sum_{i=1}^n \langle \nabla f_i(x^*), \phi_i^k - x^*\rangle + c\|x^k - x^*\|^2$$
with $c = \frac{1}{2\gamma(1-\gamma\mu)n}$. Prove that $T^k \geq c\|x^k - x^*\|^2$.

2. Assume that
$$\mathbb{E}(T^{k+1}|\,\mathcal{F}_k) \leq \Big(1 - \frac{1}{\kappa}\Big)T^k \quad (2)$$
for all $k \geq 1$ and $\kappa = 1/(\gamma\mu)$. Prove that
$$c\mathbb{E}\|x^k - x^*\|^2 \leq \Big(1 - \frac{1}{\kappa}\Big)^k T^0,$$
and deduce from it that (1) holds.

3. Now, we give some elements for the proof of (2). Prove that
$$\mathbb{E}(T^{k+1}|\mathcal{F}_k) = \frac{1}{n}f(x^k) + \Big(1 - \frac{1}{n}\Big)\frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - \frac{1}{n}\langle \nabla f(x^*), x^k - x^*\rangle - f(x^*)$$
$$- \Big(1 - \frac{1}{n}\Big)\frac{1}{n}\sum_{i=1}^n \langle \nabla f_i(x^*), \phi_i^k - x^*\rangle + c\mathbb{E}\big[\|x^{k+1} - x^*\|^2|\mathcal{F}_k\big].$$

4. Prove that
$$\mathbb{E}\Big[\nabla f_{I_k}(\phi_{I_k}^{k+1}) - \nabla f_{I_k}(\phi_{I_k}^k) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\phi_i^k))\Big|\mathcal{F}_k\Big] = \nabla f(x^k).$$

5. Deduce that
$$\mathbb{E}[\|x^{k+1} - x^*\|^2|\mathcal{F}_k] = \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle$$
$$+ \gamma^2\mathbb{E}\left[\Big\|\nabla f_{I_k}(\phi_{I_k}^{k+1}) - \nabla f_{I_k}(\phi_{I_k}^k) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\phi_i^k) - \nabla f(x^*)\Big\|^2\Big|\mathcal{F}_k\right]$$

6. Using the fact that for any random vector $X \in \mathbb{R}^d$ we have $\mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2$ and the fact that $\|x + y\|^2 \leq (1 + \frac{1}{\beta})\|x\|^2 + (1 + \beta)\|y\|^2$ for any $\beta > 0$ and $x, y \in \mathbb{R}^d$, prove that
$$\mathbb{E}\left[\Big\|\nabla f_{I_k}(\phi_{I_k}^{k+1}) - \nabla f_{I_k}(\phi_{I_k}^k) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\phi_i^k) - \nabla f(x^*)\Big\|^2\Big|\mathcal{F}_k\right]$$
$$\leq \Big(1 + \frac{1}{\beta}\Big)\mathbb{E}\big[\|\nabla f_{I_k}(\phi_{I_k}^k) - \nabla f_{I_k}(x^*)\|^2|\mathcal{F}_k\big] + (1 + \beta)\mathbb{E}\big[\|\nabla f_{I_k}(x^k) - \nabla f_{I_k}(x^*)\|^2|\mathcal{F}_k\big]$$
$$- \beta\mathbb{E}\|\nabla f(x^k) - \nabla f(x^*)\|^2.$$

7. The rest of the proof uses some extra technicalities. An important result is the fact that for a function f which is μ-strongly convex and L-smooth with have
$$f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle + \frac{1}{2(L-\mu)}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{2(L-\mu)}\|x - y\|^2$$
$$+ \frac{\mu}{L-\mu}\langle \nabla f(x) - \nabla f(y), y - x\rangle$$
for any $x, y \in \mathbb{R}^d$. Prove this result using the following hints: put $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$, and remark that g is $L - \mu$-smooth, and use the coercivity property of the gradient with g.