SD-TSIA 211
Optimization for Machine Learning
29 January 2021

Paper documents are allowed (lecture notes, exercises, books and summary sheets)
Electronic devices are forbidden

Exercise 1 (Ridge regression by distributed computations)**.**
We would like to solve the following ridge regression problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{2}\sum_{i=1}^{n}(a_i^\top x - b_i)^2 + \frac{\mu}{2}\|x\|_2^2 \tag{1}$$

where for all i, $a_i \in \mathbb{R}^p$, $b_i \in \mathbb{R}$ and $\mu > 0$.

We denote $A = \begin{bmatrix} a_1^\top \\ \vdots \\ a_n^\top \end{bmatrix}$ $F(x) = \frac{1}{2}\|Ax-b\|_2^2 + \frac{\mu}{2}\|x\|_2^2$ and $f_i(x) = \frac{1}{2}(a_i^\top x - b_i)^2 + \frac{\mu}{2n}\|x\|_2^2$.

1. Calculate $\nabla F(x)$.
2. Show that the solution to (1) is also the solution to a linear system for which you will explicit the matrix and the right hand side.

Suppose that the data a_i and b_i is private information belonging to an agent i. This agent would not like to share it directly to all other agents. Yet, we assume that all agents have an incentive in solving the global machine learning problem. The rest of the exercise will show how this can be achieved thanks to Lagrangian duality.
We define the following auxiliary problem

$$\min_{(x_1,\dots x_n)\in(\mathbb{R}^p)^n} \frac{1}{2}\sum_{i=1}^{n}\left((a_i^\top x_i - b_i)^2 + \frac{\mu}{n}\|x_i\|_2^2\right) \tag{2}$$
$$x_{i+1} = x_i, \qquad \forall i \in \{1,\dots,n-1\}$$

3. Show that the value of (2) is equal to the value of (1) and explain how we can reconstruct the solution of (1) from the solution of (2).
4. Write the Lagrangian $L(x,\phi)$ of Problem (2), where $x = (x_1,\dots,x_n) \in \mathbb{R}^{n\times p}$.
5. What is the dimension of the vector of Lagrange multipliers ?
6. Write the Lagrange multiplier method for this problem. We shall use the notation x_i^k for the ith block of coordinates of the kth iterate of the primal vector.
7. What value of the dual step size guarantees the convergence of the algorithm ?
8. Solve the step of minimization in $(x_1,\dots,x_n)$. Show that x_i^{k+1} depends only on a_i, b_i, ϕ_{i-1}^k and ϕ_i^k.

9. Show that ϕ_i^{k+1} depends only on ϕ_i^k, x_i^{k+1} and x_{i+1}^{k+1}.

Hence, using the Lagrange multiplier method, if Agent i manages x_i^k and ϕ_i^k, then together, the n agents can solve the ridge regression problem. They can do it by sharing only the primal and dual optimization variable to their neighbor agents, they never communicate their private data a_i nor b_i.

Exercise 2 (Subgradient method)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We assume that f has a minimizer x^*.
The subgradient method is the following algorithm that starts at $x^0 \in \mathbb{R}^n$ and for all $k \in \mathbb{N}$:

$$\begin{aligned} &\text{Take } g_k \in \partial f(x_k) \\ &x_{k+1} = x_k - \gamma_k g_k \ . \end{aligned}$$

The sequence $(\gamma_k)_k$ is such that $\gamma_l > 0$ for all l, $\sum_{k=0}^{+\infty} \gamma_k = +\infty$ and $\lim_{N\to\infty} \frac{\sum_{k=0}^{N} \gamma_k^2}{\sum_{k=0}^{N} \gamma_k} = 0$.

1. For what kind of objective function is the subgradient method more appropriate than the gradient descent method ?
2. Using the definition of the subgradient, show that for all $x \in \mathbb{R}^n$ and $g \in \partial f(x)$, we have
$$f(x+g) \geq f(x) + \langle g, g \rangle$$
3. Suppose that f is L-Lipschitz ($|f(x) - f(y)| \leq L \|x - y\|$). Show that there exists $M > 0$ such that for all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, we have $\|g\|_2 \leq M$.
4. Conversely, suppose that there exists $M \geq 0$ such that for all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, $\|g\|_2 \leq M$. Show that f is Lipschitz continuous.

For the rest of the exercise, we assume that f is L-Lipschitz.

5. Using the formula $x_{k+1} = x_k - \gamma_k g_k$, find $\beta(k)$ such that for all $k \in \mathbb{N}$,
$$\frac{1}{2}\|x_{k+1} - x^*\|_2^2 = \frac{1}{2}\|x_k - x^*\|_2^2 + \gamma_k \langle g_k, x_* - x_k \rangle + \beta(k)\|g_k\|_2^2 \ .$$
6. Show that for all $k \in \mathbb{N}$, $f(x^*) \geq f(x_k) + \langle g_k, x_* - x_k \rangle$
7. Deduce from this inequality that
$$\gamma_k\big(f(x_k) - f(x_*)\big) \leq \frac{1}{2}\|x_k - x^*\|_2^2 - \frac{1}{2}\|x_{k+1} - x^*\|_2^2 + \beta(k)M^2$$
8. Show that
$$\sum_{l=0}^{k} \gamma_l\big(f(x_l) - f(x_*)\big) \leq \frac{1}{2}\|x_0 - x_*\|_2^2 + \sum_{l=0}^{k} \beta(l)M^2$$

9. Denote $\bar{x}_k = \dfrac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0}^{k} \gamma_j x_j$. Using the fact that $\bar{x}_k$ is a convex combination of the previous iterates, find a bound on $f(\bar{x}_k) - f(x_*)$.
10. Using the properties of the sequence (γ_k), show that $f(\bar{x}_k)$ converges to $f(x_*)$.

Application. We consider the square-root lasso problem

$$\min_{x \in \mathbb{R}^n} \lambda \|x\|_1 + \|Ax - b\|_2$$

where A is a $m \times n$ matrix, $b \in \mathbb{R}^m$ and $\lambda > 0$. We denote $f_1(x) = \lambda \|x\|_1$ and $f_2(x) = \|Ax - b\|_2$. We admit that $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$.

11. Let x be such that $Ax - b \neq 0$. Give a subgradient of f_2 at x. Same question when $Ax - b = 0$.
12. For all $x \in \mathbb{R}^n$, propose $g \in \partial f_1(x)$.
13. Write the subgradient method for the resolution of the square-root lasso problem.

Ex 1.

1. $\nabla F(x) = A^T(Ax - b) + \mu x$

2. $\lim_{\|x\| \to +\infty} F(x) = +\infty$, which means $F(x)$ is coercive.

   Thus, $\exists x^*$ is a solution to (1) $\Leftrightarrow \nabla F(x^*) = 0$

   $A^T(Ax^* - b) + \mu x^* = 0 \Leftrightarrow x^* = (A^TA - \mu I_d)^{-1} A^T b$

3. $x_{i+1} = x_i \quad \forall\, i \in \{1, \cdots, n-1\} \Leftrightarrow x_n = x_{n-1} = \cdots = x_2 = x_1 = x$

   Auxiliary problem $\Leftrightarrow \min \frac{1}{2} \sum_{i=1}^{n} \left((a_i x^T - b_i)^2 + \frac{\mu}{n} \|x\|^2 \right)$

   $\Leftrightarrow \min \frac{1}{2} \sum_{i=1}^{n} (a_i x^T - b_i)^2 + \frac{\mu}{2} \|x\|^2$

   Thus, (2) $\Leftrightarrow$ (1) under the condition of $x_{i+1} = x_i$.

   which indicates that we could reconstruct the solution of (1)

   by letting $x_{i+1} = x_i$ in the solution of (2).

4. $L(x, \phi) = \min \frac{1}{2} \sum_{i=1}^{n} \left[(a_i x_i^T - b_i)^2 + \frac{\mu}{n} \|x_i\|^2 \right] + \sum_{i=1}^{n-1} \phi_i (x_{i+1} - x_i)$

5. $\phi \in \mathbb{R}^{n-1}$

6. $x_i^{(k+1)} = x_i^{(k)} - \gamma^{(k)} \nabla f_i(x_i^{(k)})$ gradient descent ?

7.

8.

9.

Ex 2.

1. The object function is convex but not differentiable.

2. $g = \{ g : \forall y,\ f(y) - f(x) \geq \langle g, y-x \rangle \}$

   let $y = x+g$. then: $f(x+g) \geq f(x) + \langle g, g \rangle$.

3. $f(x+g) - f(x) \geq \langle g, g \rangle = \|g\|^2$ (L-Lipschitz)

   $f(x+g) - f(x) \leq |f(x+g) - f(x)| \leq L \|g\|$

   then $\|g\|^2 \leq L\|g\| \iff \|g\| \leq L = M$.

4. $\|g\| \leq M \iff \|g\|^2 \leq M\|g\|$.

   g: $f(y) - f(x) \geq \langle g, y-x \rangle \Rightarrow f(x) - f(y) \leq \langle g, x-y \rangle$.

   Let $x = x+g$, $y = x$. $f(x+g) - f(x) \leq \langle g, g \rangle \leq M\|g\|$

   thus, $|f(x+g) - f(x)| \leq M\|g\|$ which indicate that:

   $|f(x) - f(y)| \leq M\|x-y\|$ with $L = M$

   Therefore, f is L-Lipschitz.

5. $\frac{1}{2}\|x_{k+1} - x^*\|^2 = \frac{1}{2}\|x_k - \gamma_k g_k - x^*\|^2$

   $= \frac{1}{2}\|x_k - x^*\|^2 + \frac{\gamma_k^2}{2}\|g_k\|^2 - \gamma_k \langle x_k - x^*, g_k \rangle$

   thus, $\beta(k) = \frac{\gamma_k^2}{2}$

6. $g_k \in \partial f(x_k) \quad \forall k \in \mathbb{N}$.

$f(x^*) - f(x_k) \geq \langle g, x^* - x_k \rangle \implies f(x^*) \geq f(x_k) + \langle g_k, x^* - x_k \rangle$

7. $\gamma_k \langle g_k, x_k - x^* \rangle = \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \beta(k)\|g_k\|^2$.

$\gamma_k \langle g_k, x_k - x^* \rangle \geq \gamma_k \left(f(x_k) - f(x^*) \right)$

$\beta(k)\|g_k\|^2 \leq \beta(k) M^2$

therefore: $\gamma_k \left(f(x_k) - f(x^*) \right) \leq \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \beta(k) M^2$

8. $\sum_{l=0}^{k} \gamma_l \left(f(x_l) - f(x^*) \right) \leq \sum_{l=0}^{k} \left[\frac{1}{2}\|x_l - x^*\|^2 - \frac{1}{2}\|x_{l+1} - x^*\|^2 + \beta(l) M^2 \right]$

$\sum_{l=0}^{k} \gamma_l \left(f(x_l) - f(x^*) \right) \leq \frac{1}{2}\|x_0 - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \sum_{l=0}^{k} \beta(l) M^2$

$\leq \frac{1}{2}\|x_0 - x^*\|^2 + \sum_{l=0}^{k} \beta(l) M^2$.

9. $f(\overline{x_k}) = f\left(\frac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0}^{k} \gamma_j x_j \right) \leq \frac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0}^{k} \gamma_j f(x_j)$

$f(\overline{x_k}) - f(x^*) \leq \frac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0}^{k} \gamma_j f(x_j) - \frac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0}^{k} \gamma_j f(x^*)$

$= \frac{1}{\sum_{l=0}^{k} \gamma_l} \left[\sum_{j=0}^{k} \gamma_j \left(f(x_j) - f(x^*) \right) \right]$

$\leq \frac{\frac{1}{2}\|x_0 - x^*\|^2 + \sum_{l=0}^{k} \frac{\gamma_l^2}{2} M^2}{\sum_{l=0}^{k} \gamma_l}$

10. props for $(\gamma_k)_k$: $\sum_{k=0}^{+\infty} \gamma_k = +\infty$, $\lim_{N\to\infty} \frac{\sum_{k=0}^{N} \gamma_k^2}{\sum_{k=0}^{N} \gamma_k} = 0$

$$f(\overline{x}_k) - f(x^*) \le \frac{\frac{1}{2}\|x_0 - x_*\|^2}{\sum_{\ell=0}^{k} \gamma_\ell} + \frac{\frac{1}{2}\sum_{\ell=0}^{k} \gamma_\ell^2 M^2}{\sum_{\ell=0}^{k} \gamma_\ell} \xrightarrow{k\to\infty} 0$$

therefore, $f(\overline{x}_k) \to f(x^*)$.

11. if $Ax - b \neq 0$.

$$\partial f_2(x) = \partial\left(\|Ax-b\|^2\right)^{1/2} = \frac{1}{2}\left(\|Ax-b\|^2\right)^{-1/2} 2A^T(Ax-b)$$

$$= \frac{A^T(Ax-b)}{\|Ax-b\|}$$

if $Ax - b = 0$, f_2 is indifferentiable.

$$\partial f_2 = \left\{ g : f(x) \ge f(y) + \langle g, x-y \rangle \right\}.$$

$$\|Ax-b\| - \|Ay-b\| \ge \|Ax\| - \|Ay\| \ge \|A\|\|x-y\|$$

$$g = A^T$$

12. $f_1(x) = \lambda\|x\|$. $g \in \partial f_1(x) = \begin{cases} -\lambda, & x < 0 \\ \lambda[-1, 1], & x = 0 \\ \lambda, & x > 0 \end{cases}$

13. let $g^k \in \partial f(x^k)$

$$x^{k+1} = x^k - \gamma_k g^k$$