TD - SD-TSIA 211

Exercise 1 (Gradient descent)**.**
The exercises 1 to 3 are aimed at proving that the gradient algorithm for minimizing f, where f is convex and differentiable has convergence rate $O(1/k)$ in general (where k is the number of iterations) and $O((\frac{Q-1}{Q})^k)$ when f is strongly convex (where Q is called the *condition number*)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous *i.e.* $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ for all x, y.

1. Prove that for all x, y, $\langle \nabla f(y) - \nabla f(x), y - x\rangle \leq L\|y - x\|^2$.
2. Set $\varphi(t) = f(x + t(y - x))$ for all $t \in [0, 1]$. Prove that
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \varphi(1) - \varphi(0) - \varphi'(0)\,.$$
3. Deduce that
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle dt$$
4. Using the first question, conclude that
$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2\,.$$

We consider the gradient algorithm *i.e.*, the sequence (x_k) defined by $x_{k+1} = x_k - \gamma\nabla f(x_k)$ where $\gamma > 0$ is a constant step size.

5. Show that
$$x_{k+1} = \arg\min_{y\in\mathbb{R}} \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2\gamma}\|x_k - y\|^2.$$
6. Prove that for all $z \in \mathbb{R}^n$,
$$\langle \nabla f(x_k), x_{k+1}-x_k\rangle+\frac{1}{2\gamma}\|x_k-x_{k+1}\|^2 = \langle \nabla f(x_k), z-x_k\rangle+\frac{1}{2\gamma}\|x_k-z\|^2-\frac{1}{2\gamma}\|x_{k+1}-z\|^2. \tag{1}$$
7. Deduce that $f(x_{k+1}) \leq f(x_k) - \frac{1}{\gamma}(1 - \frac{\gamma L}{2})\|x_{k+1} - x_k\|^2$.
8. Provide a condition on γ which ensures that when $x_{k+1} \neq x_k$, $f(x_{k+1}) < f(x_k)$.

From now on, we set $\gamma = \frac{1}{L}$.

9. Using (1), show that for all $z \in \mathbb{R}^n$,
$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), z - x_k\rangle + \frac{L}{2}\|x_k - z\|^2 - \frac{L}{2}\|x_{k+1} - z\|^2\,. \tag{2}$$

We assume from now on that f is convex and admits (at least) one minimizer $x^\star$.

10. Show that
$$f(x_{k+1}) \leq f(x^\star) + \frac{L}{2}\|x_k - x^\star\|^2 - \frac{L}{2}\|x_{k+1} - x^\star\|^2 \,.$$
11. Deduce that for all $k \geq 1$,
$$\sum_{i=1}^{k} f(x_i) \leq k f(x^\star) + \frac{L}{2}\|x_0 - x^\star\|^2 \,.$$
12. Show that
$$f(x_k) - f(x^\star) \leq \frac{L\|x_0 - x^\star\|^2}{2k} \,.$$

Exercise 2 (Gradient descent – strongly convex functions)**.**
We assume from now on that f is μ-strongly convex. Thus, for any x, y,
$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \,.$$

1. Using Eq. (2), prove that
$$f(x_{k+1}) \leq f(x^\star) + \frac{L-\mu}{2}\|x_k - x^\star\|^2 - \frac{L}{2}\|x_{k+1} - x^\star\|^2 \,.$$
2. Define $\Delta_{k+1} = f(x_{k+1}) - f(x^\star) + \frac{L}{2}\|x_{k+1} - x^\star\|^2$. Show that
$$\Delta_{k+1} \leq \left(1 - \frac{\mu}{L}\right)\Delta_k \,.$$
3. Conclude that
$$\begin{aligned} f(x_k) - f(x^\star) &\leq \left(1 - \frac{\mu}{L}\right)^k \Delta_0 \\ \|x_k - x^\star\|^2 &\leq \left(1 - \frac{\mu}{L}\right)^k \frac{2\Delta_0}{L} \,. \end{aligned}$$
4. The ratio $Q = L/\mu$ is called the *condition number* of f. Discuss the influence of Q on the convergence rate.

Exercise 3 (Quadratic case)**.**
From now on, we define $f(x) = \frac{1}{2}x^T H x + c^T x$ where H is positive semidefinite $n \times n$ matrix, and $g(x) = 0$. We denote by λ_{max} and λ_{min} the largest and smallest eigenvalues of H respectively.

1. What is the Hessian matrix of f? Deduce that f is convex.
2. Justify briefly that ∇f is λ_{max}-Lipschitz continuous.
3. Prove that f is λ_{min}-strongly convex.
4. Write the condition number Q of f. What kind of matrix H yields the smallest condition number?
5. Characterize the set of minimizers of f.

Exercise 4 (Proximal gradient descent).
The aim of this exercise is to prove that the proximal gradient algorithm for minimizing $F := f + g$, where :

- f is convex and differentiable;
- g is convex and possibly nondifferentiable;
- there exists at least one minimizer x^*,

has convergence rate $O(1/k)$ in general (where k is the number of iterations) and $O((\frac{Q-1}{Q})^k)$ when f is strongly convex (where Q is called the *condition number*)
Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function whose proximal operator defined by $\mathrm{prox}_g(x) = \arg\min_{y\in\mathbb{R}^n} g(y) + \frac{1}{2}\|y-x\|^2$ is easy to compute. We consider the proximal gradient algorithm *i.e.*, the sequence (x_k) defined by $x_{k+1} = \mathrm{prox}_{\gamma g}\left(x_k - \gamma\nabla f(x_k)\right)$ where $\gamma > 0$ is a constant step size.

1. Show that
$$x_{k+1} = \arg\min_{y\in\mathbb{R}} g(y) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2\gamma}\|x_k - y\|^2. \tag{3}$$
2. Let $h : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a μ-strongly convex function and note $\partial h(x)$ its subdifferential in x. Recall that a function h is said μ-strongly convex if $h - \frac{\mu}{2}\|\,.\,\|^2$ is convex. Prove that for any x, y and any subgradient $v \in \partial h(x)$,
$$h(y) \geq h(x) + \langle v, y - x\rangle + \frac{\mu}{2}\|y - x\|^2\,.$$
3. By remarking that the mapping defined in (3) is $1/\gamma$-strongly convex, prove that for all $z \in \mathbb{R}^n$,
$$\begin{aligned} g(x_{k+1}) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\gamma}\|x_k - x_{k+1}\|^2 \\ \leq g(z) + \langle \nabla f(x_k), z - x_k\rangle + \frac{1}{2\gamma}\|x_k - z\|^2 - \frac{1}{2\gamma}\|x_{k+1} - z\|^2. \end{aligned} \tag{4}$$
4. Deduce that $F(x_{k+1}) \leq F(x_k) - \frac{1}{\gamma}(1 - \frac{\gamma L}{2})\|x_{k+1} - x_k\|^2$.
5. Provide a condition on γ which ensures that when $x_{k+1} \neq x_k$, $F(x_{k+1}) < F(x_k)$.

From now on, we set $\gamma = \frac{1}{L}$.

6. Show that
$$F(x_k) - F(x^\star) \leq \frac{L\|x_0 - x^\star\|^2}{2k}\,.$$
7. Suppose that f is μ-strongly convex. Define $\Delta_{k+1} = f(x_{k+1}) - f(x^\star) + \frac{L}{2}\|x_{k+1} - x^\star\|^2$. Show that
$$F(x_k) - F(x^\star) \leq \left(1 - \frac{\mu}{L}\right)^k \Delta_0 \qquad \text{and} \qquad \|x_k - x^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \frac{2\Delta_0}{L}\,.$$

Exercise 5 (LASSO)**.** We consider the problem

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 \quad \text{s.t. } \|x\|_1 \leq \epsilon\,.$$

1. Show that there exist $\lambda \geq 0$ such that any minimizer is a solution to the LASSO(λ) problem defined by
$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1\,.$$
2. We now focus on the LASSO(λ). Prove that the solution is $\{0\}$ for large λ.
3. For an arbitrary λ, provide the expression of the proximal gradient algorithm, using the step size suggested at Exercise 1.
4. Assume that the initial point is at distance D from a minimizer. How many iterations are needed (at most) to achieve an ε-minimizer ?

Exercise 6 (Gaussian Channel, Water filling)**.** In signal processing, a *Gaussian channel* refers to a transmitter-receiver framework with Gaussian noise : the transmitter sends an information X (real valued), the receiver observes $Y = X + \epsilon$, where ϵ is a noise.

A Channel is defined by the joint distribution of (X, Y). If it is Gaussian, the channel is called *Gaussian.* In other words, if X and ϵ are Gaussian, we have a Gaussian channel.

Say the transmitter wants to send a word of size p to the receiver. He does so by encoding each possible word w of size p by a a certain vector of size n, $\boldsymbol{x}_n^w = (x_1^w, \ldots, x_n^w)$. To stick with the Gaussian channel setting, we assume that the x_i^w's are chosen as i.i.d. replicates of a Gaussian, centered random variable, with variance x.
The receiver knows the code (the dictionary of all 2^p possible $\boldsymbol{x}_n^w$'s) and he observes $\boldsymbol{y}_n = \boldsymbol{x}_n^w + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n))$. We want to recover w.
The *capacity* of the channel, in information theory, is (roughly speaking) the maximum ratio $C = n/p$, such that it is possible (when n and p tend to ∞ while $n/p \equiv C$), to recover a word w of size p using a code $\boldsymbol{x}_n^w$ of length n.

For a Gaussian Channel , $C = \log(1+x/\sigma^2)$. (x/σ^2 is the ratio signal/noise). For n Gaussian channels in parallel, with $\alpha_i = 1/\sigma_i^2$, then

$$C = \sum_{i=1}^{n} \log(1 + \alpha_i x_i).$$

The variance x_i represents a *power* affected to channel i. The aim of the transmitter is to maximize C under a *total power constraint* : $\sum_{i=1}^n x_i \leq P$. In other words, the problem is

$$\max_{x\in\mathbb{R}^n} \sum_{i=1}^{n} \log(1+\alpha_i x_i) \qquad \text{under constraints} : \forall i, x_i \geq 0, \quad \sum_{i=1}^{n} x_i \leq P. \tag{5}$$

1. Write problem (5) as a minimization problem under constraint $g(x) \preceq 0$. Show that this is a convex problem (objective and constraints both convex).
2. Show that the constraints are qualified. (hint : Slater).
3. Write the Lagrangian function
4. Using the KKT theorem, show that a primal optimal x^* exists and satisfies :
   — $\exists K > 0$ such that $x_i = \max(0, K - 1/\alpha_i)$.
   — K is given by
   $$\sum_{i=1}^{n} \max(K - 1/\alpha_i, 0) = P$$
5. Justify the expression *water filling*

Exercise 7 (Distance to an hyperplane)**.** Set $\mathcal{X} = \mathbb{R}^d$. Define the hyperplane

$$\mathcal{H}_{w,b} = \{x \in \mathcal{X} : \langle w, x \rangle + b = 0\}$$

for some fixed $w \in \mathcal{X}$ ($w \neq 0$) and $b \in \mathbb{R}$. For a fixed $z \in \mathcal{X}$, consider the problem

$$\min_{x \in \mathcal{H}_{w,b}} \frac{1}{2} \|x - z\|^2 .$$

1. Write the Lagragian function $L(x; \nu)$ associated with this problem.
2. Solve the KKT conditions and characterize the solution.
3. Prove that the distance of a point z to $\mathcal{H}$ is equal to

$$d(z, \mathcal{H}_{w,b}) = \frac{|\langle w, z \rangle + b|}{\|w\|} .$$

Exercise 8 (SVM - linearly separable case)**.** Consider a training set formed by couples (x_i, y_i) for $i \in \{1, \dots, n\}$ where x_i is a feature vector in $\mathcal{X}$ and $y_i \in \{-1, +1\}$ for all i. The hyperplane $\mathcal{H}_{w,b}$ is called *separating* if

$$\forall i, \quad y_i(\langle w, x_i \rangle + b) > 0 .$$

In the sequel, we assume that a separating hyperplane exists. Among all separating hyperplanes, we seek to find the one which maximizes the minimum distance

$$f(w, b) = \min_{i=1,\dots,n} d(x_i, \mathcal{H}_{w,b}) .$$

1. Show that if (w, b) defines a separating hyperplane, then $f(w, b) = c(w, b)/\|w\|$ where $c(w, b) = \min_i y_i(\langle w, x_i \rangle + b)$.

Thus, we are interested in solving the problem

$$\max_{w,b} \frac{c(w,b)}{\|w\|} \text{ such that } \forall i,\ y_i(\langle w, x_i\rangle + b) \geq 0\,.$$

Let $(w^\star, b^\star)$ be a solution and define

$$v^\star = \frac{w^\star}{c(w^\star, b^\star)} \text{ and } a^\star = \frac{b^\star}{c(w^\star, b^\star)}$$

2. Justify that $(w^\star, b^\star)$ and $(v^\star, a^\star)$ define the same separating hyperplane.
3. Prove that $(v^\star, a^\star)$ solves the optimization problem

$$\max_{v,a} \frac{1}{\|v\|} \text{ such that } \forall i,\ y_i(\langle v, x_i\rangle + a) \geq 1\,.$$

4. Deduce that $(v^\star, a^\star)$ solves the optimization problem

$$\min_{v,a} \frac{\|v\|^2}{2} \text{ such that } \forall i,\ 1 - y_i(\langle v, x_i\rangle + a) \leq 0\,. \tag{6}$$

5. Write the Lagrangian $L(v, a; \phi)$.
6. Write the KKT conditions.
7. Let $(v, a; \phi)$ be a saddle point of the Lagrangian. Show that ϕ_i is non-zero only if $y_i(\langle v, x_i\rangle + a) = 1$.

The training points (x_i, y_i) satisfying the above property are the closest to the hyperplane $\mathcal{H}_{v,a}$. The corresponding x_i's are often called *support vectors*.

8. If one is given a dual solution $\phi^\star$, how to recover a primal solution $(v^\star, a^\star)$ from $\phi^\star$?

Define the $n \times n$ matrices $K = (\langle x_i, x_j\rangle)_{i,j=1\ldots n}$, $D = \mathrm{diag}(y_1 \ldots y_n)$ and $\mathbf{1}^T = (1, \ldots, 1)$.

9. Prove that the dual problem reduces to

$$\min_{\substack{\phi \geq 0 \\ y^T\phi = 0}} \frac{1}{2}\phi^T DKD\phi - \mathbf{1}^T\phi\,.$$

10. Assume that this algorithm has identified a dual solution $\phi^\star$. Write explicitely the classifier as a function of $\phi^\star$.
11. What part of the training data do you need in order to implement the above classifier?

Exercise 9 (SVM - non separable case)**.**
Consider the case when a separable hyperplane might not exist. The constraints $1 - y_i(\langle v, x_i\rangle + a) \leq 0$ in Problem (6) may not be jointly feasible. For a fixed $c > 0$, we consider the relaxed problem

$$\min_{v,a} \frac{\|v\|^2}{2} + c\sum_i \xi_i \text{ such that } \forall i,\ 1 - y_i(\langle v, x_i\rangle + a) \leq \xi_i \text{ and } \xi_i \geq 0\,. \tag{7}$$

1. How many constraints has this problem ?
2. Write the Lagrangian function.
3. Show that the dual problem reduces to

$$\min_{\substack{c \geq \phi \geq 0 \\ y^T \phi = 0}} \frac{1}{2} \phi^T DKD\phi - \mathbf{1}^T \phi\,.$$

Exercise 10 (Dual of the Lasso problem)**.**
We consider the Lasso problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1\,,$$

where A is a $m \times n$ matrix and b is a vector of $\mathbb{R}^m$. The goal of this exercise is to give a dual Lasso problem.

1. Show that the objective function of this problem is convex.
2. By considering an auxiliary variable z and the constraint $z = Ax - b$, write an equivalent Lasso problem with a separable objective, which means that it can be written as $f_1(x) + f_2(z)$.
   Two optimization problems are said to be equivalent if there exists a bijection between their set of optimal solutions and their optimal value is equal.
3. Write the Lagrangian of this new problem.
4. Compute the dual problem.

Exercise 11 (Total-Variation-regularized least squares regression)**.**
Let $x \in \mathbb{R}^n$ be a vector. We consider the following problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \alpha \sum_{i=1}^{n-1} |x_{i+1} - x_i|\,,$$

where A is a $m \times n$ matrix, b is a vector of $\mathbb{R}^m$. The second term is called the total-variation (TV) regularization term.

1. Can you guess what type of solution is promoted by the TV regularization ?
2. Show that the problem writes as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \alpha \|Dx\|_1\,, \tag{8}$$

   where matrix D should be explicited.
3. Show that the problem (8) is convex.
4. By considering an auxiliary variable z and the constraint $z = Dx$, write an equivalent problem with an objective that can be written as $f_1(Ax) + f_2(z)$.
5. Write the Lagrangian of this new problem.
6. Write the Lagrange multipliers method for this problem.