

TD - SD-TSIA 211

Exercise 1 (Picard's fixed point theorem).

Prove the following theorem:

If $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfies

$$\exists 0 < \rho < 1, \forall x \in \mathbb{R}^d, \forall y \in \mathbb{R}^d, \|T(x) - T(y)\| \leq \rho \|x - y\|$$

then T has a unique fixed point x^* such that $x^* = T(x^*)$.

Moreover, every sequence of the form $x_{k+1} = T(x_k)$ converges to x^* with a linear convergence rate given by $\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$.

Exercise 2 (Gradient calculus).

- Calculate the gradient of the following functions. A , M and Q are fixed matrices, b is a fixed vector. f_1 is useful for least squares and regression problems, f_2 is useful for logistic regression and binary classification, f_3 is useful for nonnegative matrix factorization.

$$f_1 : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$x \mapsto \frac{1}{2} \|Ax - b\|^2 = \frac{1}{2} \sum_{i=1}^m \left(\sum_{j=1}^n A_{ij} x_j - b_i \right)^2$$

$$f_2 : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$z \mapsto \sum_{i=1}^n \log(1 + \exp(z_i))$$

$$f_3 : \mathbb{R}^{m \times p} \rightarrow \mathbb{R}$$

$$P \mapsto \frac{1}{2} \|M - PQ\|_F^2 = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n \left(M_{ij} - \sum_{k=1}^p P_{ik} Q_{kj} \right)^2$$

- Let g_1, g_2, g_3 be functions such that $g_1 : \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$, $g_2 : \mathbb{R}^{n_2} \rightarrow \mathbb{R}^{n_3}$, $g_3 : \mathbb{R}^{n_3} \rightarrow \mathbb{R}$ and let

$$f_4 = g_3 \circ g_2 \circ g_1 .$$

Compute the gradient of f_4 using the Jacobian matrices of g_i for $i \in \{1, 2, 3\}$.

Suppose that computing one element of the Jacobian matrices costs C_J and that multiplying two numbers costs C_M . How much does it cost to compute $\nabla_4 f(x)$?

Exercise 3 (Convergence of gradient descent for strongly convex C^2 functions). Consider a C^2 function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $0 \prec \mu I \preceq \nabla^2 f(x) \preceq LI$.

1. Show that the fixed point operator $T : x \mapsto x - \gamma \nabla f(x)$ is contractant for any $0 < \gamma < \frac{2}{L}$.
2. Show that the gradient method converges linearly.
3. How many iterations are necessary to ensure that $\|x_k - x^*\| \leq \epsilon$?

Exercise 4 (Proximal operator).

Let $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex lower-semicontinuous function such that $\text{dom } f \neq \emptyset$.

1. Recall the definition of the domain of a convex function.
2. It is possible to prove (but we do not ask you to do it) that $\exists x_0 \in \text{dom } f$ such that $\exists q_0 \in \partial f(x_0)$. Using this information, show that there exists $\alpha \in \mathbb{R}$ and $w \in \mathbb{R}^n$ such that for all x , $f(x) \geq \alpha + \langle w, x \rangle$.
3. Let us fix $x \in \mathbb{R}^n$. Let us define $g : y \mapsto f(y) + \frac{1}{2}\|x - y\|^2$. Show that g is strongly convex.
4. Show that $\lim_{\|y\| \rightarrow +\infty} g(y) = +\infty$.
5. Show that g has a minimizer and that it is unique.

We will denote this minimizer as $\text{prox}_f(x)$. The function $\text{prox}_f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called the proximal operator of f .

Exercise 5. Let us denote

$$\text{prox}_g(y) = \arg \min_{x \in \mathcal{X}} g(x) + \frac{1}{2} \|x - y\|^2$$

the proximal operator of g at y .

Fix $\gamma > 0$. Show that if f and g are convex, then the fixed points of the nonlinear equation

$$x = \text{prox}_{\gamma g}(x - \gamma \nabla f(x))$$

are the minimizers of the function $F = f + g$.

Exercise 6 (Taylor-Lagrange inequality). The goal of this exercise is to prove Taylor-Lagrange inequality. This is a fundamental inequality for the study of gradient descent and related methods.

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

Exercise 7 (Gradient descent).

This exercise is 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). The next one shows that the rate is $O((\frac{Q-1}{Q})^k)$ when f is strongly convex (where Q is called the *condition number*)

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function whose gradient is L -Lipschitz continuous.

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.

1. Show that

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2\gamma} \|x_k - y\|^2.$$

2. Prove that for all $z \in \mathbb{R}^n$,

$$\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\gamma} \|x_k - x_{k+1}\|^2 = \langle \nabla f(x_k), z - x_k \rangle + \frac{1}{2\gamma} \|x_k - z\|^2 - \frac{1}{2\gamma} \|x_{k+1} - z\|^2. \quad (1)$$

3. Deduce that $f(x_{k+1}) \leq f(x_k) - \frac{1}{\gamma} (1 - \frac{\gamma L}{2}) \|x_{k+1} - x_k\|^2$.

4. 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 $\gamma = \frac{1}{L}$.

5. Using (1), show that for all $z \in \mathbb{R}^n$,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), z - x_k \rangle + \frac{L}{2} \|x_k - z\|^2 - \frac{L}{2} \|x_{k+1} - z\|^2. \quad (2)$$

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

6. Show that

$$f(x_{k+1}) \leq f(x^*) + \frac{L}{2}\|x_k - x^*\|^2 - \frac{L}{2}\|x_{k+1} - x^*\|^2.$$

7. Deduce that for all $k \geq 1$,

$$\sum_{i=1}^k f(x_i) \leq kf(x^*) + \frac{L}{2}\|x_0 - x^*\|^2.$$

8. Show that

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

Exercise 8 (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

$$f(x_{k+1}) \leq f(x^*) + \frac{L - \mu}{2}\|x_k - x^*\|^2 - \frac{L}{2}\|x_{k+1} - x^*\|^2.$$

2. Define $\Delta_{k+1} = f(x_{k+1}) - f(x^*) + \frac{L}{2}\|x_{k+1} - x^*\|^2$. Show that

$$\Delta_{k+1} \leq \left(1 - \frac{\mu}{L}\right) \Delta_k.$$

3. Conclude that

$$\begin{aligned} f(x_k) - f(x^*) &\leq \left(1 - \frac{\mu}{L}\right)^k \Delta_0 \\ \|x_k - x^*\|^2 &\leq \left(1 - \frac{\mu}{L}\right)^k \frac{2\Delta_0}{L}. \end{aligned}$$

4. The ratio $Q = L/\mu$ is called the *condition number* of f . Discuss the influence of Q on the convergence rate.

Exercise 9 (Quadratic case).

From now on, we define $f(x) = \frac{1}{2}x^T Hx + c^T x$ 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 10 (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 $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex function whose proximal operator defined by $\text{prox}_g(x) = \arg \min_{y \in \mathbb{R}^n} g(y) + \frac{1}{2} \|y - x\|^2$ is easy to compute. We consider the proximal gradient algorithm *i.e.*, the sequence (x_k) defined by $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$ where $\gamma > 0$ is a constant step size.

1. Show that

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} g(y) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2\gamma} \|x_k - y\|^2. \quad (3)$$

2. Let us define

$$h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$$

$$y \mapsto g(y) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2\gamma} \|x_k - y\|^2 - \frac{1}{2\gamma} \|x_{k+1} - y\|^2.$$

Show that h is convex and that $0 \in \partial h(x_{k+1})$.

3. Prove that for all $z \in \mathbb{R}^n$,

$$g(x_{k+1}) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\gamma} \|x_k - x_{k+1}\|^2$$

$$\leq g(z) + \langle \nabla f(x_k), z - x_k \rangle + \frac{1}{2\gamma} \|x_k - z\|^2 - \frac{1}{2\gamma} \|x_{k+1} - z\|^2. \quad (4)$$

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

$$F(x_k) - F(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

7. Suppose that f is μ -strongly convex. Define $\Delta_{k+1} = f(x_{k+1}) - f(x^*) + \frac{L}{2}\|x_{k+1} - x^*\|^2$. Show that

$$F(x_k) - F(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k \Delta_0 \quad \text{and} \quad \|x_k - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \frac{2\Delta_0}{L}.$$

Exercise 11 (Proximal operator of the absolute value).

Let f be the absolute value, that is $f(x) = |x|$ for all $x \in \mathbb{R}$. We recall that the proximal operator of f at x is given by

$$\text{prox}_f(x) = \arg \min_{y \in \mathbb{R}} f(y) + \frac{1}{2}|y - x|^2$$

1. Show that f is convex.
2. What is the subdifferential of f ?

We are now interested in $p = \text{prox}_f(x)$.

3. Show that $p = \text{prox}_f(x)$ if and only if

$$\begin{cases} x - p = -1 & \text{if } p < 0 \\ x - p \in [-1, 1] & \text{if } p = 0 \\ x - p = 1 & \text{if } p > 0 \end{cases} \quad (5)$$

In order to get a formula, we need p as a function of x . We define the soft-thresholding operator as

$$S(x) = \text{sign}(x) \max(0, |x| - 1).$$

4. Show that

$$\begin{aligned} x - S(x) &\in [-1, 1] \\ \text{if } S(x) > 0, &\text{ then } x - S(x) = 1 \\ \text{if } S(x) < 0, &\text{ then } x - S(x) = -1 \end{aligned}$$

5. Conclude

Exercise 12 (Proximal operator of the 1-norm).

We say that a function $\phi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is separable if there exists n functions $\phi_i : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ such that for all $x \in \mathbb{R}^n$,

$$\phi(x) = \sum_{i=1}^n \phi_i(x_i).$$

1. Let ϕ be a separable function. Show that

$$\partial\phi(x) = \partial\phi_1(x_1) \times \dots \times \partial\phi_n(x_n)$$

where \times denotes the cartesian product.

2. Show that

$$\inf_{x \in \mathbb{R}^n} \sum_{i=1}^n \phi_i(x_i) = \sum_{i=1}^n \inf_{x \in \mathbb{R}} \phi_i(x)$$

and

$$\arg \min_{x \in \mathbb{R}^n} \sum_{i=1}^n \phi_i(x_i) = \arg \min_{x \in \mathbb{R}} \phi_1(x) \times \dots \times \arg \min_{x \in \mathbb{R}} \phi_n(x) .$$

3. Let ϕ be a separable function. Show that

$$\text{prox}_{\phi}(x) = (\text{prox}_{\phi_1}(x_1), \dots, \text{prox}_{\phi_n}(x_n))$$

4. Let F be the 1-norm, that is $F(x) = \sum_{i=1}^n |x_i|$.

Show that F is convex and separable.

5. Recall the proximal operator of the absolute value and give the formula for the proximal operator of the 1-norm.

Exercise 13 (LASSO). Let $\lambda > 0$, A be some $m \times n$ matrix and b be a vector of size m . We consider the problem

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

1. Prove that the solution is $\{0\}$ for large λ .
2. For an arbitrary λ , provide the expression of the proximal gradient algorithm, using the step size suggested in Exercise 4.
3. 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 14 (Link between two LASSO formulations). We consider the problem

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

Show that there exist $\lambda \geq 0$ such that any minimizer is a solution to the LASSO(λ) problem defined by

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

Exercise 15 (Proximal gradient for logistic regression).

We consider a classification problem defined by observations $(x_i, y_i)_{1 \leq i \leq n}$ where for all i , $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. We propose the following linear model for the generation of the data. Each observation is supposed to be independent and there exists a vector $w \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$ such that for all i , (y_i, x_i) is a realization of the random variable (Y, X) whose law satisfies

$$\mathbb{P}_{w, w_0}(Y = 1|X) = \frac{\exp(X^\top w + w_0)}{1 + \exp(X^\top w + w_0)}.$$

$$1. \text{ Show that } \forall i \in \{1, \dots, n\}, \mathbb{P}(Y_i = y_i|x_i) = \frac{1}{1 + \exp(-y_i(x_i^\top w + w_0))}.$$

2. Show that the maximum likelihood estimator is

$$(\hat{w}, \hat{w}_0) = \arg \min_{w, w_0} \sum_{i=1}^n \log(1 + \exp(-y_i(x_i^\top w + w_0)))$$

3. Denote $f(w, w_0) = \sum_{i=1}^n \log(1 + \exp(-y_i(x_i^\top w + w_0)))$. Compute $\nabla f(w, w_0)$.

4. Compute the proximal operator of $(x \mapsto \frac{\lambda}{2} \|x\|^2)$ for $\lambda > 0$.

5. Write the proximal gradient method for the logistic regression problem with ridge regularizer

$$(\hat{w}^{(\lambda)}, \hat{w}_0^{(\lambda)}) = \arg \min_{w, w_0} \sum_{i=1}^n \log(1 + \exp(-y_i(x_i^\top w + w_0))) + \frac{\lambda}{2} \|w\|^2.$$

6. Compute the Hessian matrix and write Newton's method for the same problem.

Exercise 16 (Optimisation with explicit constraints).

We consider the following optimization problem

$$\min_{x \in C} f(x) \tag{6}$$

where $C \subset \mathbb{R}^d$ is a convex set and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable.

1. We define the convex indicator function of the set C as

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$$

Show that (6) is equivalent to

$$\min_{x \in \mathbb{R}^d} f(x) + \iota_C(x) \tag{7}$$

2. Show that for all $x \in C$, $\partial\iota_C(x) = \{q \in \mathbb{R}^n : \forall y \in C, \langle q, y - x \rangle \leq 0\}$ and that $\partial\iota_C(x)$ is a cone (it is called the normal cone to C at x). Show that for all $x \notin C$, $\partial\iota_C(x) = \emptyset$.
3. Show that x^* is a solution to (7) if and only if $-\nabla f(x^*) \in \partial\iota_C(x^*)$.
4. Denote $\mathcal{H}_{w,b} = \{x \in \mathcal{X} : \langle w, x \rangle + b = 0\}$. Compute $\partial\iota_{\mathcal{H}_{w,b}}(x)$ for all $x \in \mathbb{R}^d$.
5. Prove that the distance of a point z to \mathcal{H} is equal to

$$d(z, \mathcal{H}_{w,b}) = \min_{x \in \mathcal{H}_{w,b}} \|x - z\|_2 = \frac{|\langle w, z \rangle + b|}{\|w\|_2}.$$

Exercise 17 (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 Lagrangian 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 18 (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^*, b^*) be a solution and define

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

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

$$\min_{v,a} \frac{\|v\|^2}{2} \text{ such that } \forall i, 1 - y_i(\langle v, x_i \rangle + a) \leq 0. \quad (8)$$

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 ϕ^* , how to recover a primal solution (v^*, a^*) from ϕ^* ?

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

9. Prove that the dual problem reduces to

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

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

Exercise 19 (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 (8) may not be jointly feasible. For a fixed $c > 0$, we consider the relaxed problem

$$\min_{v,a,\xi} \frac{\|v\|^2}{2} + c \sum_i \xi_i \quad \text{such that } \forall i, 1 - y_i(\langle v, x_i \rangle + a) \leq \xi_i \text{ and } \xi_i \geq 0. \quad (9)$$

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

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

Exercise 20 (Projected stochastic gradient).

We consider the following optimization problem

$$\min_{x \in C} \sum_{i=1}^n f_i(x) \quad (10)$$

where $C = [0, 1]^d$ and for all i , $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable.

1. We define the convex indicator function of the set C as $\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$.

Show that (10) is equivalent to $\min_{x \in \mathbb{R}^d} \sum_{i=1}^n f_i(x) + \iota_C(x)$.

2. Compute the proximal operator of ι_C
3. Write the proximal stochastic gradient method for the resolution of (10).

Exercise 21 (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 certain vector of size n , $\mathbf{x}_n^w = (x_1^w, \dots, x_n^w)$. 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 \mathbf{x}_n^w 's) and he observes $\mathbf{y}_n = \mathbf{x}_n^w + \boldsymbol{\epsilon}$, 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 \mathbf{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

$$\max_{x \in \mathbb{R}^n} \sum_{i=1}^n \log(1 + \alpha_i x_i) \quad \text{under constraints: } \forall i, x_i \geq 0, \quad \sum_{i=1}^n x_i \leq P. \quad (11)$$

1. Write problem (11) 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 22 (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 and $\lambda > 0$. 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.