

**Exercise 1**

Show that if  $S$  is an open set, its complement  $S^c$  is closed, and viceversa.

$\Rightarrow$

Let  $S$  be an open set. We know that  $S^c$  is a closed set if and only if for any sequence of elements  $\{x_n\} \subset S^c$  such that  $\{x_n\} \rightarrow x$ , then  $x \in S^c$ .

Let us use this characterization to prove that  $S^c$  is closed. Let  $\{x_n\} \subset S^c$  such that  $\{x_n\} \rightarrow x$ . Suppose that  $x$  is not in  $S^c$ . Then  $x$  must be in its complementary,  $S$ . Since  $S$  is an open set we know that  $\exists \epsilon > 0$  such that  $B(x, \epsilon) \subset S$ .

Since  $\{x_n\} \rightarrow x$ , for any  $\delta > 0 \exists n \in \mathbb{N}$  such that  $\|x - x_n\| < \delta$ . In particular, for  $\delta = \epsilon$  there is a element of the succession  $x_n$  in  $B(x, \epsilon) \subset S$ , but  $\{x_n\} \subset S^c$ . This contradiction implies that  $x$  is, in fact, in  $S^c$ .

$\Leftarrow$

Let  $S$  be a set such that its complement  $S^c$  is closed (that is,  $S^c = cl(S^c)$ ). Let us show that  $S$  is open.

Let  $x \in S$ . Then,  $x \notin S^c$ , which implies  $x \notin cl(S^c)$ . This implies that

$$\exists \epsilon > 0 \text{ such that } B(x, \epsilon) \cap S^c = \emptyset \implies B(x, \epsilon) \subset S$$

Thus,  $S$  is open. □

**Exercise 2**

If  $S_1, S_2$  are convex subsets, prove that the following are also convex sets:

$$S_1 \cap S_2 = \{x : x \in S_1 \text{ and } x \in S_2\}$$

$$S_1 + S_2 = \{x + x' : x \in S_1, x' \in S_2\}$$

$$S_1 - S_2 = \{x - x' : x \in S_1, x' \in S_2\}$$

- Consider  $S_1 \cap S_2 = \{x : x \in S_1 \text{ and } x \in S_2\}$ . Let  $x, x' \in S_1 \cap S_2$  and consider the segment

$$z \equiv \lambda x + (1 - \lambda)x', \lambda \in [0, 1]$$

Since  $x$  is in  $S_1$ , which is a convex set,  $z$  will also be in  $S_1$  for any  $\lambda \in [0, 1]$ . Likewise,  $z$  will be in  $S_2$ , and thus in the intersection  $S_1 \cap S_2$ . Since the segment is contained in the intersection,  $S_1 \cap S_2$  is a convex subset.

- Consider  $S_1 + S_2 = \{x + z : x \in S_1, z \in S_2\}$ . Now, let  $x + z \in S_1 + S_2$  and  $x' + z' \in S_1 + S_2$ . Also, consider the segment

$$\begin{aligned}\lambda(x + z) + (1 - \lambda)(x' + z') &= \lambda x + \lambda z + (1 - \lambda)x' + (1 - \lambda)z' \\ &= \underbrace{\lambda x + (1 - \lambda)x'}_{x'' \in S_1} + \underbrace{\lambda z + (1 - \lambda)z'}_{z'' \in S_2}\end{aligned}$$

Where the underbraces are true due to the convexity of  $S_1$  and  $S_2$ , so we have

$$x'' + z'' \in S_1 + S_2$$

so the segment is in the sum set, and thus, the set  $S_1 + S_2$  is convex.

- We use the same process done in the previous set. Consider  $S_1 - S_2 = \{x - z : x \in S_1, z \in S_2\}$ . Now, let  $x - z \in S_1 - S_2$  and  $x' - z' \in S_1 - S_2$ . Also, consider the segment

$$\begin{aligned}\lambda(x - z) + (1 - \lambda)(x' - z') &= \lambda x - \lambda z + (1 - \lambda)x' - (1 - \lambda)z' \\ &= \underbrace{\lambda x + (1 - \lambda)x'}_{x'' \in S_1} - \underbrace{(\lambda z + (1 - \lambda)z')}_{z'' \in S_2}\end{aligned}$$

Where the underbraces are true due to the convexity of  $S_1$  and  $S_2$ , so we have

$$x'' - z'' \in S_1 - S_2$$

so the segment is in the difference set, and thus, the set  $S_1 - S_2$  is convex.

### Exercise 3

If  $f : S \rightarrow \mathbb{R}$  is a convex function on the convex set  $S$ , the set  $S_{\min} = \{x : x \text{ is a minimum of } f\}$  is a convex set.

We omit the case where  $S_{\min} = \emptyset$ , since the empty set is convex. Now, let  $y$  be the minimum of  $f(x)$ :  $y = \min_x f(x)$ . Then  $S_{\min} = \{x \in S : f(x) = y\}$ . We need to show that for all  $x, x' \in S_{\min}$  and for all  $\lambda \in [0, 1]$ :

$$z \equiv \lambda x + (1 - \lambda)x' \in S_{\min} \leftrightarrow f(z) = y$$

Since  $S$  is convex,  $z$  is in  $S$ , and since  $f$  is also convex:

$$\begin{aligned}f(z) &= f(\lambda x + (1 - \lambda)x') \\ &\leq \lambda f(x) + (1 - \lambda)f(x') \\ &= \lambda y + (1 - \lambda)y \\ &= y\end{aligned}$$

where we used that  $x, x' \in S_{\min}$ . But since  $y$  is the minimum of  $f$ , the equality holds  $f(z) = y$ . This means that  $z$  is in  $S_{\min}$ , therefore  $S_{\min}$  is a convex set.

### Exercise 4

Given a quadratic form  $q(w) = w^T Q w + b w + c$ , with  $Q$  a symmetric  $d \times d$  matrix,  $w, b$  being  $d \times 1$  vectors and  $c$  a real number, derive its gradient and Hessian.

$$\nabla q(w) = Q w + b, \quad H q(w) = Q$$

*Hint: expand  $q(w)$  and take the partials with respect to  $w_i$  and  $w_i, w_j$ .*

Let us start by unrolling the quadratic form expression:

$$q(w) = \sum_{i,j=1}^d Q_{ij} w_i w_j + \sum_{i=1}^d b_i w_i + c,$$

and compute the partial derivative over the  $k - th$  component:

$$\frac{\partial q}{\partial w_k}(w) = \sum_{i=1}^d Q_{ik} w_i + \sum_{j=1}^d Q_{kj} w_j + b_k$$

where  $k \in \{1, \dots, d\}$ . By using that  $Q$  is symmetric we obtain:

$$\frac{\partial q}{\partial w_k}(w) = 2 \sum_{j=1}^d Q_{kj} w_j + b_k. \quad (1)$$

That is, we are multiplying the  $k - th$  row of the  $Q$  matrix and multiplying it by  $w$ . We can obtain gradient as a product of matrices using the previous expression:

$$\nabla q(w) = \begin{pmatrix} \frac{\partial q}{\partial w_1}(w) \\ \vdots \\ \frac{\partial q}{\partial w_d}(w) \end{pmatrix} = \begin{pmatrix} 2 \sum_{j=1}^d Q_{1j} w_j + b_1 \\ \vdots \\ 2 \sum_{j=1}^d Q_{dj} w_j + b_d \end{pmatrix} = 2 \begin{pmatrix} \sum_{j=1}^d Q_{1j} w_j \\ \vdots \\ \sum_{j=1}^d Q_{dj} w_j \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix} = 2 Qw + b$$

In order to obtain the Hessian, we take partial derivatives over the  $l - th$  component in 1:

$$\begin{aligned} \frac{\partial^2 q}{\partial w_k \partial w_l}(w) &= \frac{\partial q}{\partial w_l} \left( \frac{\partial q}{\partial w_k} \right) (w) \\ &= \frac{\partial q}{\partial w_l} \left( 2 \sum_{j=1}^d Q_{kj} w_j + b_k \right) (w) \\ &= 2 Q_{kl} \end{aligned}$$

Hence, the Hessian matrix of  $q$  will have  $2Q_{kl}$  in position  $(k, l)$ . That is:

$$\text{Hess } q(w) = 2Q$$

#### Exercise 5

If  $(p_1, \dots, p_K)$  is a probability distribution, prove that its entropy  $H(p_1, \dots, p_K) = -\sum_{i=1}^K p_i \log p_i$  is a concave function. Show also that its maximum is  $\log K$ , attained when  $p_i = \frac{1}{K}$  for all  $i$ .

In this problem, since we are dealing with probabilities, two new constraints appear:

$$\sum_{i=1}^K p_i = 1$$

$$p_i \geq 0 \quad \forall i = 1, \dots, K$$

They will be used later.

Let us compute the gradient and Hessian of  $H$  to see that it is concave. Firstly, we have that

$$\frac{\partial H}{\partial p_i} = -\log(p_i) - 1, \quad \forall i = 1, \dots, K$$

and, hence,

$$\frac{\partial^2 H}{\partial p_i \partial p_j} = -\frac{\lambda_{ij}}{p_i}$$

Lastly, since  $p_i \geq 0$ , we have that the Hessian is a negative-definite diagonal matrix, so  $H$  is concave.

In order to find the minimum entropy, we have to solve the following optimization problem:

$$\begin{aligned} & \max_{(p_1, \dots, p_K)} H(p_1, \dots, p_K) \\ & \text{s.t.} \\ & \sum_{i=1}^K p_i - 1 = 0 \\ & p_i \geq 0 \quad \forall i = 1, \dots, K \end{aligned}$$

Consider the lagrangian of this problem:

$$L(\{p_i\}_{i=1}^K, \lambda) = -\sum_{i=1}^K p_i \log(p_i) + \lambda \left( \sum_{i=1}^K p_i - 1 \right).$$

We can obtain its gradient derivating with respect to each variable

$$\frac{\partial L}{\partial p_i} = -\log p_i - 1 + \lambda, \quad \frac{\partial L}{\partial \lambda} = \sum_{i=1}^K p_i - 1.$$

Using this derivatives, we have to equate them to zero, which is solving the following equations system:

$$\begin{cases} \log p_i = \lambda - 1, & i = 1, \dots, K \\ \sum_{i=1}^K p_i = 1 \end{cases}$$

Looking at the first equation, the  $p_i$  have no interdependencies, they are constant and have the same value. Also, since they have to add 1, the only possible solution is that each  $p_i = \frac{1}{K}$ . Lastly, we can compute the maximum value of the Entropy:

$$H(\{p_i\}_{i=1}^K) = -\sum_{i=1}^K \frac{1}{K} \log\left(\frac{1}{K}\right) = -\frac{1}{K} \left( \sum_{i=1}^K \log 1 - \log K \right) = \frac{1}{K} \cdot K \log K = \log K,$$

as we wanted to see. □

**Exercise 6**

We want to solve the following constrained restriction problem:

$$\begin{aligned} \min \quad & x^2 + 2y^2 + 4xy \\ \text{s.t} \quad & x + y = 1 \\ & x, y \geq 0. \end{aligned}$$

1. Write its Lagrangian with  $\alpha, \beta$  the multipliers of the inequality constraints.
2. Write the KKT conditions.
3. Use them to solve the problem. For this consider separately the  $(\alpha = \beta = 0)$ ,  $(\alpha > 0, \beta = 0)$ ,  $(\alpha = 0, \beta > 0)$ ,  $(\alpha > 0, \beta > 0)$  cases.

Writing the **Lagrangian** in terms of  $\alpha, \beta, \lambda$  is pretty straightforward:

$$L(x, y, \alpha, \beta, \lambda) = x^2 + 2y^2 + 4xy + \lambda(x + y - 1) + \alpha x + \beta y.$$

Now, to write the KKT conditions. As a very brief summarization, the KKT conditions are: the gradient of the Lagrangian equals to zero and the inequality restrictions (multiplied by its corresponding constant) also equal to zero. In our case, the **KKT conditions** are:

$$\begin{aligned} \frac{\partial L}{\partial x} &= 2x + 4y + \lambda + \alpha = 0 \\ \frac{\partial L}{\partial y} &= 4y + 4x + \lambda + \beta = 0 \\ \alpha x &= 0 \\ \beta y &= 0 \end{aligned}$$

Now, we want to solve this equations system to see if we can find a minimum of our problem. We have the following cases:

- Case  $\alpha = \beta = 0$ .  
In this case, the system is

$$\begin{aligned} 2x + 4y + \lambda &= 0 \\ 4y + 4x + \lambda &= 0 \end{aligned}$$

If we substitute  $4y$  from the first equation into the second one, we obtain

$$-2x - \lambda + 4x + \lambda = 0 \implies 2x = 0 \implies x = 0,$$

and, since  $x + y = 1$ , we obtain that our *KKT point* is  $(0, 1)$ .

- Case  $\alpha, \beta > 0$ .  
In this case, we obtain from the KKT conditions that  $x = y = 0$ , which does not match our initial conditions  $x + y = 1$ , so no *KKT points* are obtained.
- Case  $\alpha > 0, \beta = 0$ .  
Looking at our KKT conditions, since  $\alpha > 0$ , we have that  $x = 0$ , resulting in  $y = 1$  and a *KKT point*  $(0, 1)$ , which is the same that we obtained in the first case.

- Case  $\alpha = 0, \beta > 0$ .

Using the same reasoning, we obtain  $(1, 0)$  as a new KKT point.

Until now, we have two candidates to be the optimal one:  $\{(0, 1), (1, 0)\}$ . Now, we make use of the following theorem:

**Theorem 1** *If in a minimization problem with restrictions  $g_i(x), h_j(x) \in C^1$ , if we assume  $f$  to be convex and  $h_j$  to be affine, then a KKT point  $x^*$  is an optimum of this problem. (Slide 18)*

So, we can evaluate the function on our KKT points to find the minimum. We obtain that  $f(1, 0) = 1, f(0, 1) = 2$ , so the minimum is reached in  $(1, 0)$  with optimal value 1.

#### Exercise 7

We have worked out the dual problem for the soft SVC problem. Do the same for the simpler **hard** SVC problem

$$\min_{w, b} \frac{1}{2} \|w\|^2$$

subject to  $y^p (w \cdot x^p + b) \geq 1$ . What are here the KKT conditions?

Firstly, consider the Lagrangian for this problem

$$\begin{aligned} L(w, b, \alpha) &= \frac{1}{2} \|w\|^2 - \sum_p \alpha_p [y^p (w \cdot x^p + b) - 1] \\ &= \frac{1}{2} w \cdot w - w \sum_p \alpha_p y^p x^p - b \sum_p \alpha_p y^p + \sum_p \alpha_p \\ &= w \left( \frac{1}{2} w - \sum_p \alpha_p y^p x^p \right) - b \sum_p \alpha_p y^p + \sum_p \alpha_p \end{aligned}$$

Then, we have to compute the gradient of the Lagrangian

$$\begin{aligned} \nabla_w L(w, b, \alpha) &= w - \sum_p \alpha_p y^p x^p = 0 \implies w = \sum_p \alpha_p y^p x^p \\ \frac{\partial L}{\partial b} &= - \sum_p \alpha_p y^p = 0 \end{aligned}$$

Lastly, we have to use these equalities in the expression of the Lagrangian:

$$\begin{aligned} L(w, b, \alpha) &= \sum_p \alpha_p - \frac{1}{2} \left( \sum_p \alpha_p y^p x^p \right) \left( \sum_q \alpha_q y^q x^q \right) \\ &= \sum_p \alpha_p - \frac{1}{2} \sum_{p, q} \alpha_p \alpha_q y^p y^q x^p x^q \end{aligned}$$

By defining the matrix  $Q$  with value  $y^p y^q x^p x^q$  in position  $(p, q)$ , our dual optimization problem gets simplified into

$$\begin{cases} \max_{\alpha} \sum_p \alpha^p - \frac{1}{2} \alpha^T Q \alpha \\ \text{s.t. } \alpha^p, \sum_p \alpha_p y^p = 0 \end{cases}$$

At this point we may realize that we have completely removed the dependency  $b$  from our problem. Let us state the KKT conditions for this problem. There are:

$$\begin{aligned} \lambda_w f(x^*) + \sum_i \lambda_i \lambda g_i(x^*) &= 0 \\ \lambda_i g_i(x^*) &= 0 \end{aligned}$$

For our particular problem we obtain:

$$\begin{aligned} w + \sum_p \alpha_p y^p x^p &= 0 \\ \lambda_i g_i(x^*) &= 0 \end{aligned}$$

#### Exercise 8

A typical Linear Programming (LP) problem can be stated as the following constrained optimization problem:

$$\min_x c \cdot x \quad \text{s.t.} \quad x \geq 0, Ax \leq b$$

with  $x \in \mathbb{R}^d$ ,  $A$  an  $m \times d$  matrix and  $b \in \mathbb{R}^m$ . A tool often used in LP is to study the so called dual problem, which in this case is

$$\min_z b \cdot z \quad \text{s.t.} \quad z \geq 0, A^t z \leq -c$$

with now  $z \in \mathbb{R}^m$ . Apply our Lagrangian dual construction technique to show that this is indeed the dual formulation of the initial LP problem

Firstly, we have to write the Lagrangian for this problem:

$$L(x, \lambda, \mu) = c \cdot x - \sum_{i=1}^d \lambda_i x_i + \sum_{j=1}^m \mu_j (a_j \cdot x - b_j)$$

Hence, the gradient respect to  $x$  of the lagrangian is:

$$\nabla_x L = c - \lambda + A^t \mu$$

CHECK!!!

(this Lagrangian must have dimension  $d$ ).

#### Exercise 9

We know that, theoretically, the minimum SVC primal  $f^*$  and the maximum SVC dual  $q^*$  are equal. Check this in this case by writing  $q^*$  and  $f^*$  in terms of the  $\alpha_p^*$  and checking that both

expressions coincide.

### Exercise 10

We want to apply our Lagrangian theory to solve the homogeneous constrained Ridge problem (i.e., with a model  $w \cdot x$

$$\arg \min_w \text{mse}(w) = \frac{1}{n} \sum_{p=1}^n (t^p - w \cdot x^p)^2, \quad \text{s.t.} \quad \|w\|_2^2 \leq \rho^2.$$

Write its Lagrangian and, using the lecture slides, the detailed formulation of the KKT conditions at an optimal  $w^*$  and multiplier  $\lambda^*$ .

Assuming that  $\lambda^* > 0$ , use the gradient KKT condition to show that  $w^*$  also solves a standard Ridge regression problem for the optimal value  $\lambda^*$  of the regularization parameter.

Assuming now that  $\lambda^* = 0$ , use again the slides to write down the solution in this case and use this solution to get a lower bound for  $\rho$ .

## HECHO EN CLASE

### Exercise 11

If  $Q$  is a symmetric, positive definite  $d \times d$  matrix, show that  $f(x) = x^T Q x$ ,  $x \in \mathbb{R}^d$ , is a convex function.

## HECHO EN CLASE

### Exercise 12

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a function and assume that  $\text{epi}(f) \subset \mathbb{R}^d \times \mathbb{R}$  is convex. Prove that then  $f$  is convex.

Consider the set

$$\text{epi}(f) = \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : t \geq f(x)\}.$$

This set is, by hypothesis, convex. That is, for any  $(x, t), (x', t') \in \text{epi}(f)$ , we have

$$\lambda(x, t) + (1 - \lambda)(x', t') = (\lambda x + (1 - \lambda)x', \lambda t + (1 - \lambda)t') \in \text{epi}(f) \quad \forall \lambda \in [0, 1].$$

This implies that

$$f(\lambda x + (1 - \lambda)x') \leq \lambda t + (1 - \lambda)t', \quad \forall \lambda \in [0, 1]. \quad (2)$$

Also, since each of the points belongs to  $\text{epi}(f)$ , we have that:

$$\lambda f(x) + (1 - \lambda)f(x') \leq \lambda t + (1 - \lambda)t', \quad \forall \lambda \in [0, 1] \quad (3)$$

Lastly, if we subtract Equation (3) from Equation (2) we obtain:

$$\begin{aligned} f(\lambda x + (1 - \lambda)x') - (\lambda f(x) + (1 - \lambda)f(x')) &\leq 0, & \forall \lambda \in [0, 1] \\ \implies f(\lambda x + (1 - \lambda)x') &\leq \lambda f(x) + (1 - \lambda)f(x'), & \forall \lambda \in [0, 1]. \end{aligned}$$

Lastly, recalling that  $(x, f(x)) \in \text{epi}(f)$  for all  $x \in S$ , we obtain that  $f$  is convex.  $\square$

### Exercise 13

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a convex function. Prove that  $\text{epi}(f)$  is a closed set and that  $(x, f(x)) \in \partial \text{epi}(f)$ .



**Exercise 14**

Prove that if  $f$  is strictly convex, it has a unique global minimum.

HECHO EN CLASE

**Exercise 15**

Let  $f, g : S \subset \mathbb{R}^d \rightarrow \mathbb{R}$  be two convex functions on the convex set  $S$ . Prove that, as subsets,  $\partial f(x) + \partial g(x) \subset \partial(f + g)(x)$  for any  $x \in S$ .

JAVI LO TIENE HECHO ES TRIVIAL

**Exercise 16**

Compute the proximal of  $f(x) = 0$  and of  $g(x) = \frac{1}{2}\|x\|^2$ .

HECHO EN CLASE

**Exercise 17**

Assume that  $f$  is convex. Prove that for any  $\lambda > 0$ ,  $\partial(\lambda f)(x) = \lambda \partial f(x)$  as subsets.

HECHO EN CLASE

**Exercise 18**

Prove that the  $\epsilon$ -insensitive loss function  $\ell_\epsilon(z) = \max\{0, |z - \epsilon|\}$  is convex. Give also its subgradient  $\partial \ell_\epsilon(x)$  at any  $x \in \mathbb{R}$

**Exercise 19**

Compute the proximals of the hinge  $f(x) = \max\{0, -x\}$  and the  $\epsilon$ -insensitive  $g(x) = \max\{0, |x| - \epsilon\}$  loss functions.

HECHO EN CLASE

**Exercise 20**

We have seen that we can solve the constrained Ridge problem by a Projected Gradient algorithm. Using the lecture slides, write down in as much detail as you can the computations needed at each iteration of the algorithm.