Francisco Javier Sáez Maldonado
José Antonio Álvarez Ocete

Ejercicios Programación Lineal
Optimización

March 27, 2022

Exercise 1
Show that if S is an open set, its complement S^c is closed, and vice versa.

Authors: 50% *José Antonio,* %50 *Javier*

$\boxed{\Rightarrow}$

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 $\{x_n\} \longrightarrow x$, 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 $\{x_n\} \longrightarrow x$. 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 $\{x_n\} \longrightarrow x$, 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 sequence 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.

$\boxed{\Leftarrow}$

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:

$$S_1 \cap S_2 = \{x \ : \ x \in S_1 \text{ and } x \in S_2\}$$
$$S_1 + S_2 = \{x + x' \ : \ x \in S_1, x' \in S_2\}$$
$$S_1 - S_2 = \{x - x' \ : \ x \in S_1, x' \in S_2\}$$

Authors: 50% *José Antonio,* %50 *Javier*

- 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 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

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

  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 $f : S \to \mathbb{R}$ 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.

Authors: 50% *José Antonio*, %50 *Javier*

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 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:

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

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 $q(w) = w^T Qw + bw + c$, 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.

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

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

Authors: 50% *José Antonio,* %50 *Javier*

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:

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

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

Authors: 50% *José Antonio*, %50 *Javier*

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

$$\sum_{i=1}^{K} p_i = 1$$
$$p_i \geq 0 \quad \forall i = 1, \ldots, n$$

They will be used later.

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

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

and, hence,

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

where δ_{ij} is the Kroneker delta. 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:

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

Consider the lagrangian of this problem:

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

We can obtain its gradient differentiating with respect to each variable

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

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

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

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:

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

as we wanted to prove. □

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 α, β 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.

Authors: 50% *José Antonio*, %50 *Javier*

Writing the **Lagrangian** in terms of α, β, λ is pretty straightforward:

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

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:

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

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 $(0, 1)$. Any non-negative values of α and β are valid to satisfy both hte KKT conditions and our problem restrictions.

- 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?

Authors: 50% *José Antonio,* %50 *Javier*

Firstly, consider the Lagrangian for this problem

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

Then, we have to compute the gradient of the Lagrangian

$$\nabla_w \mathcal{L}(w,b;\alpha) = w - \sum_p \alpha_p y^p x^p = 0 \implies w = \sum_p \alpha_p y^p x^p$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_p \alpha_p y^p = 0$$

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

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

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

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

At this point we may realize that we have completely removed the dependency b from our problem. The KKT conditions for this problem are:

$$\begin{cases}\nabla_w \mathcal{L} = w - \sum_p \alpha_p y^p x^p & = 0\\ \frac{\partial \mathcal{L}}{\partial b} = -\sum_p \alpha_p y^p & = 0\\ \alpha_p\left(1 - y^p\left(w \cdot x^p + b\right)\right) & = 0\end{cases}$$

Exercise 8
A typical Linear Programming (LP) problem can be stated as the following constrained optimization problem:

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

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

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

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

Authors: 50% *José Antonio*, %50 *Javier*

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

$$\begin{aligned}\mathcal{L}(x,\lambda,\mu) &= c\cdot x - \sum_{i=1}^{d}\lambda_i x_i + \sum_{j=1}^{m}\mu_j(a_j\cdot x - b_j)\\ &= xc - x\lambda + x\left(A^T\mu\right) - b\mu\end{aligned}$$

where $\mu \geq 0$. Hence, the gradient respect to x of the lagrangian is:

$$\nabla_x\mathcal{L} = c - \lambda + A^T\mu \implies c = \lambda - A^T\mu$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = -x = 0 \implies x = 0$$

Hence, the dual problem is:

$$\begin{cases}\max_\mu -b\mu\\ \text{s.t. } c = \lambda - A^T\mu, \quad \mu\geq 0\end{cases}$$

Since λ doesn't appear in the objective function, we may remove it by changing the restriction from $c = \lambda - A^T\mu$ to $-c \geq A^T\mu$. Additionally, we may change the optimization function from $\max_\mu -b\mu$ to $\min_\mu b\mu$, obtaining:

$$\begin{cases}\min_\mu b\mu\\ \text{s.t. } -c \geq A^T\mu, \quad \mu\geq 0\end{cases}$$

which was to be demonstrated.

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 α_p^* and checking that both expressions coincide.

Authors: 50% *José Antonio*, %50 *Javier*

The primal SVC problem is defined as

$$\min_w \frac{1}{2}\|w\|^2 + C\sum_p \xi_p \text{ s.t. } y^p(wx^p+b)\geq 1-\xi_p, \quad \xi_p\geq 0$$

We are going to see that the minimum of this function, say f^*, is equal to the maximum of the dual function $\mathcal{D}$, say q^*.

Solution. The Lagrangian for this problem is

$$\begin{aligned}\mathcal{L}(w,b,\xi;\alpha) &= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{p=1}^{n}\xi_p^2 - \sum_p\alpha_p[y_p(w^Tx_p+b)-1+\xi_p]\\ &= w^T\left(\frac{1}{2}w - \sum_p\alpha_p y_p x_p\right) + \sum_p\xi_p\left(\frac{C}{2}\xi_p - \alpha_p\right) - b\sum_p\alpha_p y_p + \sum_p\alpha_p,\end{aligned}$$

with $\alpha_p \geq 0$. To get the dual function we solve $\nabla_w \mathcal{L} = 0$, $\frac{\partial \mathcal{L}}{\partial b} = 0$ and $\frac{\partial \mathcal{L}}{\partial \xi_p} = 0$, which yield the KKT stationarity conditions:

$$0 = \nabla_w \mathcal{L} = w - \sum_p \alpha_p y_p x_p \implies w = \sum_p \alpha_p y_p x_p. \tag{2}$$

$$0 = \frac{\partial \mathcal{L}}{\partial b} = \sum_p \alpha_p y_p. \tag{3}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \xi_p} = 2\frac{C}{2}\xi_p - \alpha_p \implies C\xi_p = \alpha_p, \quad 1 \leq p \leq n. \tag{4}$$

Substituting back in the Lagrangian we arrive at the dual function

$$\begin{aligned}\Theta(\alpha) &= -\frac{1}{2}\sum_{p,q} \alpha_p \alpha_q y_p y_q x_p^T x_q - \frac{1}{2C}\sum_p \alpha_p^2 - b\overbrace{\sum_p \alpha_p y_p}^{0} + \sum_p \alpha_p \\ &= -\frac{1}{2}\sum_{p,q} \alpha_p \alpha_q \left(y_p y_q x_p^T x_q + \frac{\delta_{pq}}{C} \right) + \sum_p \alpha_p \\ &= -\frac{1}{2}\alpha^T \left(Q + \frac{I}{C} \right) \alpha + \sum_p \alpha_p,\end{aligned}$$

with constraints $\sum_p \alpha_p y_p = 0$ and $\alpha_p \geq 0$. In the above expression, Q is the symmetric matrix given by $Q_{pq} = y_p y_q x_p^T x_q$, I is the identity matrix, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, and δ_{pq} is Kronecker's delta function, defined to be 1 if $p = q$ and 0 otherwise. We know that strong duality holds in this case, so flipping the sign to get a minimization problem, the dual problem has the following expression:

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{2}\alpha^T \left(Q + \frac{I}{C} \right) \alpha - \sum_p \alpha_p \right\} \quad \text{s.t.} \quad \begin{cases} \sum_p \alpha_p y_p = 0, \\ \alpha_p \geq 0, \\ 1 \leq p \leq n. \end{cases}$$

The KKT complementary slackness conditions in this case are

$$\alpha_p^* [y_p((w^*)^T x_p + b^*) - 1 + \xi_p^*] = 0, \quad 1 \leq p \leq n. \tag{5}$$

To derive the optimal primal solution from the dual one, we resort to the KKT conditions. Firstly, from (2) we have

$$w^* = \sum_{p=1}^n \alpha_p^* y_p x_p.$$

Secondly, from (4) we conclude that

$$\xi_p^* = \frac{\alpha_p^*}{C}, \quad 1 \leq p \leq n.$$

Now, observe that in a non-trivial classification setting where we have at least one positive and one negative example, at least one of the α_p^* is non-zero (otherwise we would have $y_p b^* \geq 1$ for all

p and every example would be of the same class). If we take any $\alpha_p^* > 0$, the conditions in (5) tell us that

$$\begin{aligned} 0 &= y_p((w^*)^T x_p + b^*) - 1 + \xi_p^* \\ &= y_p\left(\sum_{q=1}^n \alpha_q^* y_q x_q^T x_p + b^*\right) - 1 + \frac{\alpha_p^*}{C} \\ &= y_p \sum_{q=1}^n \alpha_q^* y_q \left(x_q^T x_p + \frac{\delta_{pq}}{C}\right) + y_p b^* - 1. \end{aligned}$$

Thus, solving for b^* we get the optimal primal bias term in closed form[1]:

$$b^* = \frac{1}{y_p} - \sum_{q=1}^n \alpha_q^* y_q \left(x_q^T x_p + \frac{\delta_{pq}}{C}\right) = y_p - \sum_{q=1}^n \alpha_q^* y_q \left(x_q^T x_p + \frac{\delta_{pq}}{C}\right).$$

Exercise 10

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

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

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

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

Authors: 50% *José Antonio*, %50 *Javier*

Let us write the Lagrangian for this problem:

$$\mathcal{L}(w,;\lambda) = \frac{1}{2}\mathrm{mse}(w) + \lambda\left(\| \ w \ \|^2 - \rho\right)$$

Then, we compute the gradient of the Lagrangian:

$$0 = \nabla_w \mathcal{L}(w;\lambda) = \frac{1}{n} x^T(xw - y) + \lambda w$$

The complete KKT conditions are:

$$\begin{cases} \nabla_w \mathcal{L} = \frac{1}{n} x^T(xw - y) + \lambda w & = 0 \\ \lambda\left(\| \ w \ \|^2 - \rho\right) & = 0 \end{cases}$$

[1] Although in theory the formula works for any support vector, to increase the stability of the method in practice and compensate for numerical errors it is advised to average the values of b^* obtained for each $\alpha_p^* > 0$.

At the optimal point (w^*, λ^*) we obtain:

$$\begin{cases} \frac{1}{n} x^T(xw^* - y) + \lambda^* w^* & = 0 \\ \lambda^* \left(\| \; w^* \; \|^2 - \rho\right) & = 0 \end{cases} \tag{6}$$

Let us distinguish cases:

- Case $\lambda^\star > 0$: Using the second KKT condition we obtain $\|\boldsymbol{w}^\star\|^2 = \rho > 0$, so $\boldsymbol{w} \neq 0$. Using the following equality
$$\frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda^\star \boldsymbol{I})\boldsymbol{w}^\star - \boldsymbol{X}^T\boldsymbol{t} = 0,$$
it is clear that $\boldsymbol{w}^\star$ verifies
$$\begin{aligned} \boldsymbol{w}^\star &= \left[\frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda^\star \boldsymbol{I})\right]^{-1} \boldsymbol{X}^T\boldsymbol{t} \\ &= \arg\min_{\boldsymbol{w}} \frac{1}{2n}\|\boldsymbol{X}^T\boldsymbol{w} - \boldsymbol{t}\| + \lambda^\star\|\boldsymbol{w}\|^2. \end{aligned}$$
As a result, $\boldsymbol{w}^\star$ is the solution of the Ridge regression problem with regularizer $\lambda^\star$.

- Case $\lambda^\star = 0$: The KKT conditions simplify as
$$\frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^\star\right) - \boldsymbol{X}^T\boldsymbol{t} = 0,$$
with
$$\boldsymbol{w}^\star = \left[\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1} \boldsymbol{X}^T\boldsymbol{t} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}.$$
Using the problem restriction,
$$\|\boldsymbol{w}^\star\|_2 \leq \rho \implies \|n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}\|_2 \leq \rho.$$

Exercise 11
If Q is a symmetric, positive definite $d \times d$ matrix, show that $f(x) = x^TQx$, $x \in \mathbb{R}^d$, is a convex function.

Authors: 50% *José Antonio*, %50 *Javier*

If a function f is twice differentiable, then it is convex if and only if its Hessian matrix is definite positive. In our case, Hess $f = Q$, which is symmetric and positive definite by hypothesis, proving that f is convex.

Exercise 12
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function and assume that $epi(f) \subset \mathbb{R}^d \times \mathbb{R}$ is convex. Prove that then f is convex.

Authors: 50% *José Antonio*, %50 *Javier*

Consider the set

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

It is clear that, for each $x \in \mathbb{R}^d$, $(x, f(x)) \in \text{epi}(f)$. Now, epi$(f)$ is convex by hypothesis, so if we consider $x, x' \in \mathbb{R}^d$, we have:

$$\lambda(x, f(x)) + (1-\lambda)(x', f(x')) = \big(\lambda x + (1-\lambda)x', \lambda f(x) + (1-\lambda)f(x')\big) \in \text{epi}(f) \quad \forall \lambda \in [0,1].$$

Since for each $\lambda \in [0,1]$ the obtained point is in epi(f), we have

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda)f(x'), \quad \forall \lambda \in [0,1].$$

So f is convex. □

Exercise 13
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Prove that epi(f) is a closed set and that $(x, f(x)) \in \partial\,\text{epi}(f)$.

Authors: 50% *José Antonio*, %50 *Javier*

We know that a set S is closed if $cl(S) = S$. Let us prove this when $S = \text{epi}(f)$, using a double inclusion argument. Firstly, it is clear that epi$(f) \subseteq cl(\text{epi}(f))$, by the definition of $cl(S)$. To check the second inclusion, we will see that if $(x,t) \in cl(\text{epi}(f))$, then $(x,t) \in \text{epi}(f)$.

Let $(x,t) \in cl(\text{epi}(f))$. Then, for any $\delta > 0$,

$$B((x,t), \delta) \cap \text{epi}(f) \neq \emptyset. \tag{7}$$

Suppose that $(x,t) \notin \text{epi}(f)$. Since f is convex, epi(f) is also convex. Using the Projection Theorem, we can find a minimum distance $d > 0$ from (x,t) to epi(f). Thus,

$$\text{epi}(f) \cap B((x,t), \frac{d}{2}) = \emptyset,$$

which is a contradiction with (7), so $(x,t) \in \text{epi}(f)$, and $cl(\text{epi}(f)) = \text{epi}(f)$.

Let us now see that $(x, f(x)) \in \partial\,\text{epi}(f)$. We can write epi$(f) = cl(\text{epi}(f)) = int(\text{epi}(f)) \cup \partial\,\text{epi}(f)$, so

$$\text{epi}(f) = cl(\text{epi}(f)) = int(\text{epi}(f)) \cup \partial\,\text{epi}(f) \implies \partial\,\text{epi}(f) = cl(\text{epi}(f)) \backslash int(\text{epi}(f))$$

If $(x, f(x)) \in int(\text{epi}(f))$, it would exist $\delta > 0$ such that:

$$B((x, f(x)), \delta) \subset \text{epi}(f).$$

If this happened, we could say that moving the value of $(x, f(x) - \frac{\delta}{2}) \in \text{epi}(f)$. However, this is not true, since this could mean that $f(x) \leq f(x) - \frac{\delta}{2}$, which can not happen since $\delta > 0$. This proves that $(x, f(x)) \notin int(\text{epi}(f))$, but $(x, f(x))$ is clearly in epi(f), so this point has to be in $\partial\,\text{epi}(f)$, as we wanted to see.

Exercise 14
Prove that if f is strictly convex, it has a unique global minimum.

Authors: 50% *José Antonio*, %50 *Javier*

This result is incomplete: with the given hypothesis, it is simply not true. For instance, the function $x \mapsto e^x$ is strictly convex and doesn't have a unique global minimum.

Let us add an additional hypothesis to create a new result to prove.

- Case 1: If we suppose f is born on a compact set, then using Weierstrass theorem we will have at least 1 minimum.
- Case 2: We may simply suppose that we have at least 1 minimum, without imposing any restrictions on the dominion of f.

In either one of those cases, the additionally hypothesis is summarized in having at least 1 minimum. However, this is still not enough. For instance, the function $f : \{-1, 1\} \to \mathbb{R}, f(x) = x^2$ is strictly convex and has two minimums (in the two points of its dominion). The result we will prove is:

Proposition. Let f be a strictly convex function. Then it has at most one global minimum in each connected component.

Suppose $x \neq z$ are both global minimums of f in the same connected component and let $\lambda \in (0, 1)$. Then:

$$f(\lambda x + (1 - \lambda)z) < \lambda f(x) + (1 - \lambda) f(z) = f^*$$

We have find an element $\lambda x + (1 - \lambda)z$ (include in the dominion of f because x and z are in the same connected component) that has a lower value of f than the minimum, which is impossible. Hence, $x = z$. □

Exercise 15
Let $f, g : S \subset \mathbb{R}^d \to \mathbb{R}$ 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 $x \in S$.

Authors: 50% *José Antonio*, %50 *Javier*

We already know that $\xi \in \partial f(x)$ implies that $f(x') > f(x) + \xi(x - x')$ for all $x' \in S$. Let us apply this definition to obtain the result.
Consider $\xi_1 \in \partial f(x)$ and $\xi_2 \in \partial g(x)$. Then, $\xi_1 + \xi_2 \in \partial f(x) + \partial g(x)$. Now, using the definition for each of the ξ_i with $i = 1, 2$, we obtain:

$$f(x') > f(x) + \xi_1(x - x'), \quad g(x') > g(x) + \xi_2(x - x')$$

And, if we add both inequalities:

$$\begin{gathered} f(x') + g(x') > f(x) + g(x) + (\xi_1 + \xi_2)(x - x') \\ (f + g)(x') > (f + g)(x) + (\xi_1 + \xi_2)(x - x') \end{gathered}$$

which means that $\xi_1 + \xi_2 \in \partial (f + g)(x)$, as we wanted to see. □

Exercise 16
Compute the proximal of $f(x) = 0$ and of $g(x) = \frac{1}{2}\|x\|^2$.

Authors: 50% *José Antonio*, %50 *Javier*

Let us directly compute the proximal of f directly:

$$\text{prox}_f(x) = \arg\min_z 0 + \frac{1}{2} \parallel x - z \parallel^2 = x.$$

We will assume g is born in $\mathbb{R}$ for more generality. For its proximal we have:

$$\text{prox}_g(x) = \arg\min_z \underbrace{\frac{1}{2}z^2 + \frac{1}{2} \parallel x - z \parallel^2}_{\equiv h(z)}$$

We equalize the gradient of h to 0 to find the minimum:

$$0 = \nabla h(z) = z + z - x \implies z = \frac{1}{2}x$$

Hence

$$\text{prox}_g(x) = \frac{1}{2}x$$

Exercise 17
Assume that f is convex. Prove that for any $\lambda > 0, \partial(\lambda f)(x) = \lambda \partial f(x)$ as subsets.

Authors: 50% *José Antonio*, %50 *Javier*

We will prove this result with a double inclusion. Let $A \equiv \partial(\lambda f)(x)$ and $B \equiv \lambda \partial f(x)$

- Case $A \subseteq B$: Let $xi \in A$, then for all z:

  $$\lambda f(z) \geq \lambda f(x) + \xi(z - x) \implies f(z) \geq f(x) + \frac{\xi}{\lambda}(z - x)$$

  where we used that $\lambda > 0$. This implies that $\frac{\xi}{\lambda} \in \partial f(x)$. Defining $\mu \equiv \frac{\xi}{\lambda} \in \partial f(x)$ we obtain $\xi = \lambda\mu \in \lambda \cdot \partial f(x) = B$.

- Case $B \subseteq A$: Let $\xi \in B$, then $\xi = \lambda\mu$ with $\mu \in \partial f(x)$. Hence, for all z:

  $$\begin{aligned} f(z) \geq f(x) + \mu(z - x) &\implies \lambda f(z) \geq \lambda f(x) + \lambda\mu(z - x) \\ &\implies (\lambda f)(z) \geq (\lambda f)(x) + (\lambda\mu)(z - x) \\ &\implies \xi = \lambda\mu \in \partial(\lambda f)(x) \end{aligned}$$

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 $x \in \mathbb{R}$.

Authors: 50% *José Antonio*, %50 *Javier*

Let us see the convexity of $\ell_\epsilon(z)$. We will use the following proposition a couple times:

Proposition 1 *Let S be an nEC open set and $f : S \to \mathbb{R}$. f is convex if, and only if, epi(f) is convex.*

Now, we will show that $\operatorname{epi}(\ell_\epsilon)$ is convex. Consider the function $f_1(x) = |x| - \epsilon$. This function is clearly convex.

$$\begin{aligned} f_1(\lambda z + (1-\lambda z')) &= |\lambda z + (1-\lambda z')| - \epsilon \\ &\leq \lambda|x| + (1-\lambda)|x'| - \epsilon \\ &= \lambda|x| - (\lambda\epsilon) + (1-\lambda)|x'| - (1-\lambda)\epsilon \\ &= \lambda f_1(x) + (1-\lambda) f_1(x') \end{aligned}$$

Hence, f_1 is convex, and using the proposition, $\operatorname{epi}(f_1)$ is convex. Now, consider the *upper half plane* (which is a convex not empty set)

$$\mathcal{H} = \{(x,y) \in \mathbb{R}^2 \;:\; y > 0\}.$$

Then, it is clear that

$$\operatorname{epi}(\ell_\epsilon(z)) = \operatorname{epi}(f_1) \cap \mathcal{H}.$$

Lastly, using that the finite intersection of convex sets is convex (this is pretty straightforward to see, since each pair of points x, x' in the intersection is in both convex sets, in particular it is in one of them, and so does the segment that aligns the points, so the segment is contained in the intersection, proving that the intersection is convex), we obtain that $\operatorname{epi}(\ell_\epsilon(z))$ is convex and the proposition tells us that $\ell_\epsilon(z)$ is also convex.

This function is also differentiable in $R\backslash\{-\epsilon, \epsilon\}$. We can give its subgradient dividing it in parts. Let us rewrite the function:

$$\ell_\epsilon(z) = \begin{cases} -x-\epsilon & \text{if } x < -\epsilon \\ 0 & \text{if } -\epsilon < x < \epsilon \\ x - \epsilon & \text{if } x > \epsilon \end{cases}$$

Lastly, its subgradient is:

$$\partial\ell_\epsilon(z) = \begin{cases} -1 & \text{if } x < -\epsilon \\ 0 & \text{if } -\epsilon < x < \epsilon \\ 1 & \text{if } x > \epsilon \end{cases}$$

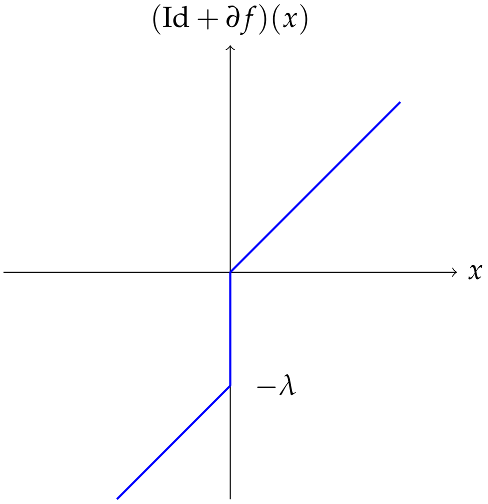
Exercise 19
Compute the proximals of the hinge $f(x) = \max\{0, -x\}$ and the ϵ-insensitive $g(x) = \max\{0, |x| - \epsilon\}$ loss functions.

Authors: 50% *José Antonio*, %50 *Javier*

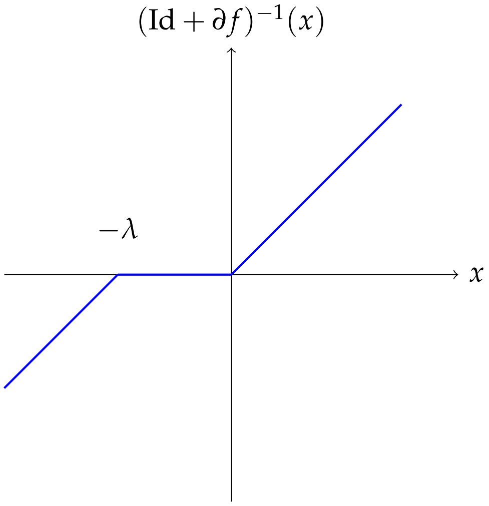
We start by computing $(\mathrm{Id} + \partial f)(x)$:

$$\partial f(x) = \begin{cases} -1 & x < 0 \\ [-1, 0] & x = 0 \\ 0 & x > 0 \end{cases} \quad \lambda\partial f(x) = \begin{cases} -\lambda & x < 0 \\ [-\lambda, 0] & x = 0 \\ 0 & x > 0 \end{cases} \quad (\mathrm{Id} + \lambda\partial f)(x) = \begin{cases} -\lambda + x & x < 0 \\ [-\lambda, 0] & x = 0 \\ x & x > 0 \end{cases}$$

Let us plot the previous function:



To obtain the proximal we simply compute the inverse by rotating 90 degrees around the origin and flipping around the vertical axis.



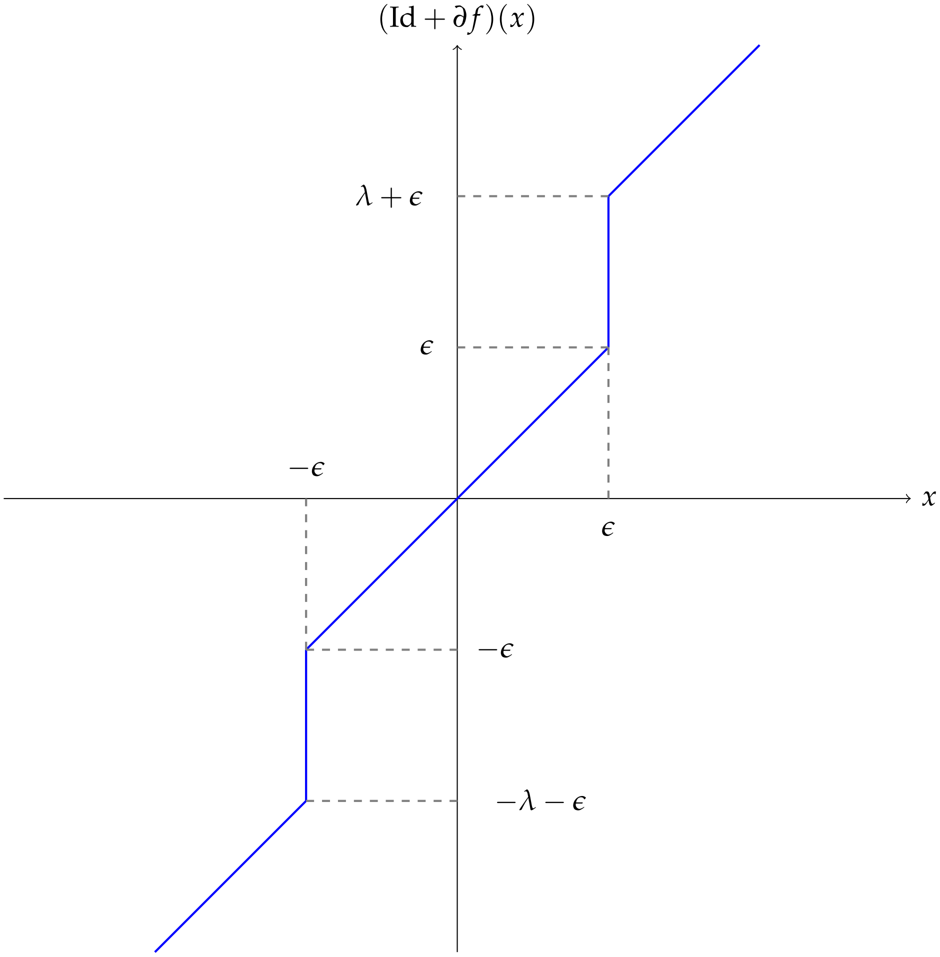
Thus, the requested proximal is:

$$\mathrm{prox}_f(x) \begin{cases} x + \lambda & x \leq -\lambda \\ 0 & -\lambda \leq x \leq 0 \\ x & x \geq 0 \end{cases}$$

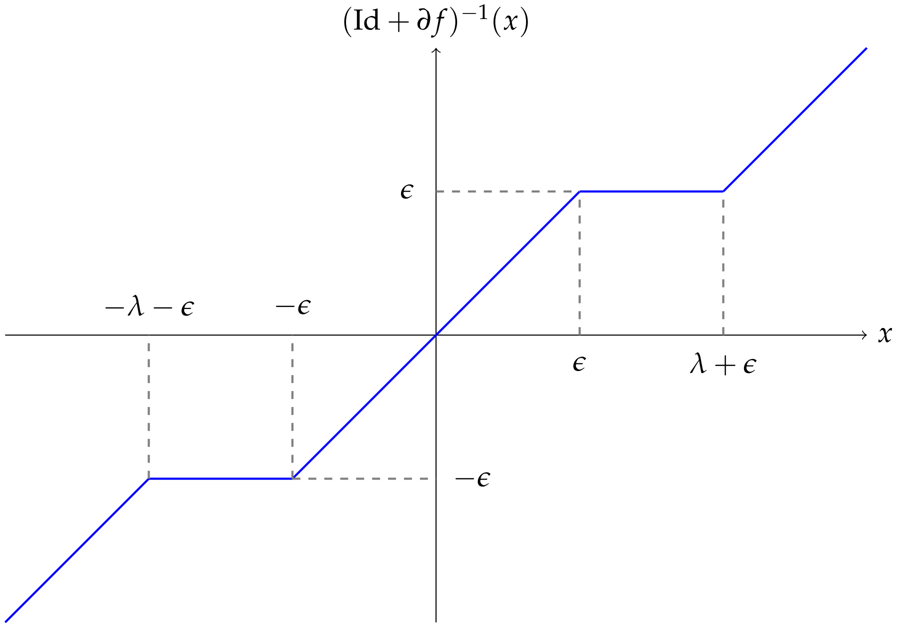
We repeat this process for the ϵ-insensitive $g(x) = \max\{0, |x| - \epsilon\}$ loss function:

$$\partial g(x) = \begin{cases} -1 & x < -\epsilon \\ [-1, 0] & x = -\epsilon \\ 0 & |x| < \epsilon \\ [0, 1] & x = \epsilon \\ 1 & x > \epsilon \end{cases} \quad \lambda \partial g(x) = \begin{cases} -\lambda & x < -\epsilon \\ [-\lambda, 0] & x = -\epsilon \\ 0 & |x| < \epsilon \\ [0, \lambda] & x = \epsilon \\ \lambda & x > \epsilon \end{cases} \quad (\mathrm{Id} + \lambda \partial g)(x) = \begin{cases} -\lambda + x & x < -\epsilon \\ [-\lambda - \epsilon, -\epsilon] & x = -\epsilon \\ x & |x| < \epsilon \\ [\epsilon, \lambda + \epsilon] & x = \epsilon \\ \lambda + x & x > \epsilon \end{cases}$$

Again, we plot the previous function:



Again, we use the graphical method to compute the inverse:



Finally, the requested proximal is:

$$\operatorname{prox}_g(x)\begin{cases} x+\lambda & x\le -\epsilon-\lambda \\ -\epsilon & -\epsilon-\lambda\le x\le -\epsilon \\ x & |x|<\epsilon \\ \epsilon & \epsilon\le x\le \epsilon+\lambda \\ x-\lambda & x\ge \epsilon+\lambda \end{cases}$$

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.

Authors: 50% *José Antonio,* %50 *Javier*

Consider that we have a data matrix $\mathbf{X}$ and a target vector $\mathbf{y}$. Then, the constrained Ridge problem is

$$\min_{w,b}\frac{1}{2}\operatorname{mse}(w,b)=\frac{1}{2n}\sum_{p=1}^{n}(y_p-w^Tx_p-b)^2 \quad \text{s.t.} \quad \|w\|_2\le\rho.$$

This can be matricially formulated as

$$\min_{\mathbf{w}\in S} f(\mathbf{w})=\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2,$$

where $S=\mathbb{R}\times\bar{B}(0,\rho)$, which is a nEC set. This problem can be reformulated as follows:

$$\min_{\mathbf{w}\in\mathbb{R}^{d+1}} f(\mathbf{w})+i_S\mathbf{w}), \quad \text{where} \quad i_S(\mathbf{w})=\begin{cases}0 & \text{if } \mathbf{w}\in S, \\ +\infty & \text{if } \mathbf{w}\notin S.\end{cases}$$

Recall that $\mathbf{w} \in \mathbb{R}^{d+1}$ because we are adding the *bias b* to the $\mathbf{w}$ vector. We will have to take this into account later. In the Projected Gradient algorithm, we need to compute $\text{prox}_{\lambda i_S}(\mathbf{w})$. However, in this case it can be proved that it is the same as computing $P_S(\mathbf{w})$, where P_S is the (Euclidean) projection operator

$$P_S(x) = \arg\min_{y \in S} \|x - y\|_2.$$

The Projected algorithm is the following: We want to apply the Projected Gradient algorithm to a concrete problem. Recall that this algorithm can be described as:

Algorithm 1: Projected Gradient.

```
Data: $\epsilon, x_0$
k=0
for $k = 1, 2, \ldots$ do
    Choose a step size $\lambda_k$
    $x_{k+1} = P_S(x_k - \lambda_k \nabla f(x_k))$
    if $\|x_{k+1} - x_k\| \leq \epsilon$ then
        return $x_{k+1}$
    end
end
```

As we can see, this algorithm is almost the classic *Gradient Descent* algorithm. The difference between this version and the classic one, is that in this case we are projecting back to the S set to be sure that our solution satisfies the constraints.

In our case, the projection P_S is easy to compute, since it is a projection of **only the second component** to the euclidean ball:

$$P_S(\mathbf{w}) = P_S(b, w) = \left(b, \frac{\rho w}{\max\{\rho, \|w\|\}}\right).$$

The only remaining thing to do is to compute the gradient $\nabla f(\mathbf{w})$:

$$\nabla f(\mathbf{w})^T = \left(\frac{\partial f(b,w)}{\partial b}, \nabla_w f(b,w)^T\right) = -\frac{1}{n}\left(\mathbf{1}^T(y - \mathbf{X}\mathbf{w}),\, X^T(y - \mathbf{X}\mathbf{w})\right),$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones. All in all, the gradient update step for a point $\mathbf{w}_k = (b_k, w_k^T)^T$ amounts to computing

$$\mathbf{w}_{k+1} = P_S(\mathbf{w}_k - \lambda_k \nabla f(\mathbf{w}_k)) = \left(b_k + \frac{\lambda_k}{n}\mathbf{1}^T(y - \mathbf{X}\mathbf{w}),\, \rho\frac{w_k + \frac{\lambda_k}{n}X^T(y - \mathbf{X}\mathbf{w})}{\max\left\{\rho, \|w_k + \frac{\lambda_k}{n}X^T(y - \mathbf{X}\mathbf{w})\|\right\}}\right)^T.$$

And these would be all the needed components to apply this algorithm to the constrained Ridge Regression problem.