HOMEWORK 3
Elena Gómez
January 31, 2023

PROBLEM 0: Homework checklist
I worked with Alejandro Mayo, Mario Utiel and Lucía Ostolaza

PROBLEM 1: Convexity and least squares

1. Show that $f(x) = \| b - Ax\|^2$ is a convex function. Feel free to use the result proved on the last homework.

We can rewrite this norm ($f(x) = \| b - Ax\|^2$) as follows:

$$f(x) = \| b - Ax \|^2 = (b - Ax)^T(b - Ax) = b^T b - (Ax)^T b - b^T Ax + (Ax)^T Ax =$$

$$= b^T b - 2b^T Ax + x^T A^T Ax$$

Because $b^T Ax = (Ax)^T b$ (normal equations)

The sum of convex functions is convex, and affine precomposition is convex.
$b^T b - 2b^T Ax$ is convex because is an affine transformation.

So, we only need to prove that $x^T A^T Ax$ is convex.

In the previous homework (hw 2) we proved that $f(x) = x^T Qx$, where Q is an $n \times n$ symmetric positive semi-definite matrix is a convex function

Now we have $x^T A^T Ax$, so we have to prove that $A^T A$ is a **symmetric positive semi-definite** matrix in order to proof that $x^T A^T Ax$ is convex, and as consequently $f(x)$ is convex

$(A^T A)^T = A^T (A^T)^T = A^T A$. So, $A^T A$ **is symmetric**

$x^T A^T Ax = ||Ax||^2 \geq 0$. So $A^T A$ is **positive semi-definite**

As $A^T A$ is a symmetric positive semi-definite matrix, $x^T A^T Ax$ is convex, and, consequently $f(x)$ is convex

2. Show that the null-space of a matrix is a convex set. (A convex set satisfies the condition that, for every pair of points in the set, any point on the line joining those points is also in the set.)

The null space of any matrix A consists of all the vectors y such that $Ay = 0$ and y is not zero.
Formally define:
$null(A) = \{y \in \mathbb{R}^n | Ay = 0\}$

As it is said in the statement, a convex set satisfies the condition that, for every pair of points in the set, any point on the line joining those points is also in the set:

A set S is **CONVEX** if the line between any two points in S lies in S.
$i.e \ \forall y_1, y_2 \in S, \forall \theta \in [0, 1]$
$\theta y_1 + (1 - \theta) y_2 \in S$

In this case, the set is $null(A)$. So, the set $null(A)$ is **CONVEX** if:

$\forall y_1, y_2 \in null(A), \forall \theta \in [0, 1]$
$\theta y_1 + (1 - \theta) y_2 \in null(A)$

$Ay_1 = 0$ and $Ay_2 = 0$. So:

$A(\theta y_1 + (1 - \theta) y_2) = \theta Ay_1 + (1 - \theta) Ay_2 = 0 \in null(A)$

So, $null(A)$ is a convex set

PROBLEM 2: Ridge Regression
The Ridge Regression problem is a variant of least squares:
$minimize||b - Ax||_2^2 + \lambda||x||_2^2$
This is also known as Tikhonov regularization.

1. Show that this problem always has a unique solution, for any A if λ > 0, using the theory discussed in class so far.
This property is one aspect of the reason that Ridge Regression and used. It is also a common regularization method that can help avoid overfitting in a regression problem.

We want to prove that when we minimize $||b - Ax||_2^2 + \lambda||x||_2^2$, we only find one solution. This means that $||b - Ax||_2^2 + \lambda||x||_2^2$ has only one minimizer, so this minimizer is global.

Let's call this function $f(x)$:
$f(x) = g(x) + h(x)$
where $g(x) = ||b - Ax||_2^2$ and $h(x) = \lambda||x||_2^2$

If a function is convex and it has a minimizer, it has only one minimizer and this minimizer is global

So, if we can prove that $f(x)$ is convex and it has a minimizer, we show that the Ridge Regression problem has a unique solution.

Let's show that $f(x)$ is **convex**.

In the previous problem (problem 1), we showed that $||b - Ax||_2^2$ is convex, so now we just have to show that $\lambda||x||_2^2$ is convex. Because the sum of two convex functions is a convex function. If $g(x)$ is convex and $h(x)$ is convex, then, $f(x) = g(x) + h(x)$ is convex.

Let's show that $h(x)$ is **convex**.

We know that a function h is convex if for all $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$, we have
$h(\theta x + (1-\theta)y) - \theta h(x) - (1-\theta)h(y) \leq 0$

So, let's prove if $h(x)$ fulfills the condition above

$$
\begin{aligned}
&h(\theta x + (1-\theta)y) - \theta h(x) - (1-\theta)h(y) \\
&= \lambda(\theta x + (1-\theta)y)^T(\theta x + (1-\theta)y) - \theta\lambda x^T x - \lambda(1-\theta)y^T y \\
&= \lambda(\theta^2 x^T x + (1-\theta^2)y^T y + 2\theta(1-\theta)x^T y - \theta x^T x - (1-\theta)y^T y) \\
&= \lambda\theta(1-\theta)(2x^T y - x^T x - y^T y) \\
&= -\lambda\theta(1-\theta)||x-y||^2
\end{aligned}
$$

We are asked to show that $f(x)$ has a unique solution if $\lambda > 0$

if $\lambda > 0$, then

$-\lambda \cdot \theta(1-\theta) \cdot ||x-y||^2 \leq 0$

Because
$\lambda > 0$
$\theta \in [0, 1]$, So $\theta(1-\theta) \geq 0$
$||x-y||^2 \geq 0$

So,
$\lambda \cdot \theta(1-\theta) \cdot ||x-y||^2 \geq 0$
And
$-\lambda \cdot \theta(1-\theta) \cdot ||x-y||^2 \leq 0$

So, $h(x)$ is **CONVEX** , and , consequently, $f(x)$ is **CONVEX** because it is the sum of 2 convex functions
So, if $||b - Ax||_2^2 + \lambda||x||_2^2$ has a minimizer, it has only a unique minimizer

Let's see if $f(x)$ has a minimizer

$$\frac{\partial f(\mathbf{x})}{\partial x} = 2A^T(b - Ax) + 2\lambda x$$

$$= 2A^T b - 2A^T Ax + 2\lambda x$$

$$= 2A^T b - 2(A^T A + \lambda)x = 0$$

$$2A^T b = 2(A^T A + \lambda)x$$

$$A^T b = (A^T A + \lambda I)x$$

$(A^T A + \lambda I)$ is invertible and square, so x has a unique solution. And, as f is convex, x is a minimizer which is the global minimizer of f.

2. Use the SVD of A to characterize the solution as a function of λ.

We have to use SVD, $A = U\Sigma V^T$ to characterize the solution as a function os λ

Using the solution of the previous question (question 1), we know that we want to find x such that
$A^T b = (A^T A + \lambda I)x$

In order to solve the unknown x, we have to multiplie both sides of the equality by $(A^T A + \lambda I)^{-1}$ as follows

$(A^T A + \lambda I)^{-1} A^T b = (A^T A + \lambda I)^{-1}(A^T A + \lambda I)x$

Then,
$x = (A^T A + \lambda I)^{-1} A^T b$

By using SVD ($A = U\Sigma V^T$)

$x = ((U\Sigma V^T)^T U\Sigma V^T + \lambda I)^{-1}(U\Sigma V^T)^T b$

$x = (V\Sigma^T U^T U\Sigma V^T + \lambda I)^{-1} V\Sigma^T U^T b$

As $U^T U = I$ because U is orthonormal, and $\Sigma^T = \Sigma$ because Σ is a diagonal matrix :

$x = (V\Sigma\Sigma V^T + \lambda I)^{-1} V\Sigma U^T b$

$x = (V\Sigma^2 V^T + \lambda I)^{-1} V\Sigma U^T b$

We can replace I with VV^T

$x = (V\Sigma^2 V^T + \lambda VV^T)^{-1} V\Sigma U^T b$

$x = (V(\Sigma^2 + \lambda I)V^T)^{-1} V\Sigma U^T b$

If we have 3 matrices A,B and C, $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$. So, we continue solving x as follows

$x = (V^T)^{-1}(\Sigma^2 + \lambda I)^{-1} V^{-1} V\Sigma U^T b$

$x = V(\Sigma^2 + \lambda I)^{-1}\Sigma U^T b$

As,

$$(\Sigma^2 + \lambda I)^{-1}\Sigma = \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & & \\ & \frac{\sigma_2}{\sigma_2^2+\lambda} & & \\ & & \ddots & \\ & & & \frac{\sigma_p}{\sigma_p^2+\lambda} \end{bmatrix}$$

We can define x:

$$x = \sum v_i(\frac{\sigma_i}{\sigma_i^2+\lambda})u_i^T b_i$$

3. What is the solution when $\lambda \to \infty$?
What is the solution when $\lambda \to 0$.

When $\lambda \to \infty$
As $x = \sum v_i(\frac{\sigma_i}{\sigma_i^2+\lambda})u_i^T b_i$, the denominator goes to ∞, so x goes to 0

When $\lambda \to 0$
If $\lambda \to 0$ is like there was not a regularization term, so the solution is $x = A^{-1}b$.

4. Give the solutions of the least squares problem using the sports teams data when $\lambda = 0$ and $\lambda = \infty$. (Here, you will not need the constraint that the sum of entries is 1.) Any notable differences from the ranking giving in class?

From the previous question we now:
$x = \sum v_i(\frac{\sigma_i}{\sigma_i^2+\lambda})u_i^T b_i$

When $\lambda \to \infty$
As I said above (in question 3) the solution will be 0 , independently of the data we use. So, it is not necessary to code this problem in order to find the solution.

When $\lambda \to 0$
The solution will be computed as there was not a regularization term. But we will not obtain the same result as we obtained in class because in class we computed constrained optimization and now we are talking about unconstrained optimization

5. Suppose that you only want to regularize one component of the solution, say, x_1, so that your optimization problem is
$minimize \parallel b - Ax \parallel_2^2 + \lambda x_1^2.$
Show how to adapt your techniques in this problem to accomplish this goal.

We want λ to regularize only the first component

The solution for the standar regularized ridge regression problem was

$x = V(\Sigma^2 + \lambda I)^{-1}\Sigma U^T b$

Where

$$(\Sigma^2 + \lambda I)^{-1}\Sigma = \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & & \\ & \frac{\sigma_2}{\sigma_2^2+\lambda} & & \\ & & \ddots & \\ & & & \frac{\sigma_p}{\sigma_p^2+\lambda} \end{bmatrix}$$

In this case, instead of multiplying λ by I , we have to multiply λ by a matrix M of all zeros and and a 1 on the position M_{11}. Thus, lambda only affects x_1

$$\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & 0 \end{bmatrix}$$

So now, we can define x as follows:

$x = V(\Sigma^2 + \lambda M^T M)^{-1}\Sigma U^T b$

Where

$$(\Sigma^2 + \lambda M^T M)^{-1}\Sigma = \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & & \\ & \frac{\sigma_2}{\sigma_2^2} & & \\ & & \ddots & \\ & & & \frac{\sigma_p}{\sigma_p^2} \end{bmatrix} = \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & & \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_p} \end{bmatrix}$$

So, now the value of λ only affects x_1

When $\lambda \to \infty$
x_1 goes to 0

When $\lambda \to 0$
If $\lambda \to 0$ is like there was not a regularization term

PROBLEM 3: Thinking about constraint
Find the minimizer of $f(x, y) = x^2 + 2y^2$ where also $y = 5x + 2$ (that is, we have added one linear constraint) and justify, as concisely and correctly as you can, that you have found a solution. Do the terms local and global minimizer apply? If so, use one of them to describe your solution. Furthermore, evaluate the gradient of f at your solution. What is notable about the gradient at your solution in comparison with the solution where x, y have no constraints?

Let's subtitute y in $f(x, y)$

$f(x) = x^2 + 2(5x+2)^2 = x^2 + 2(25x^2 + 4 + 20x) = 51x^2 + 40x + 8$

Now, let's compute x such that $\frac{\partial f(x)}{\partial x} = 0$

$\frac{\partial f(x)}{\partial x} = 102x + 40 = 0$
$102x = -40$
$x = -\frac{40}{102} = -\frac{20}{51}$

Now let's compute $\frac{\partial^2 f(x)}{\partial x^2}$ for $x = -\frac{20}{51}$

$\frac{\partial^2 f(x)}{\partial x^2} = 102 > 0$

Then, let's compute y.
$y = 5(-\frac{20}{51}) + 2 = -\frac{100}{51} + \frac{102}{51} = \frac{2}{51}$

So, $(x, y) = (-\frac{20}{51}, \frac{2}{51})$

So, $(x, y) = (-\frac{20}{51}, \frac{2}{51})$ is a global minimizer of $f(x, y)$ when it has the constraint $y = 5x + 2$. Because the 2 sufficient conditions are fulfilled. But it is not a minimizer of the function $f(x, y)$ (when we comput belowe the gradient of $f(x, y)$ at the solution , it is different to 0). Again, I clarify that $(x, y) = (-\frac{20}{51}, \frac{2}{51})$ is a global minimizer of $f(x, y)$ **where** $y = 5x + 2$, but not a minimizer of $f(x, y)$.

Now, let's take the gradient of $f(x, y)$ at my solution

$\nabla f(x, y) = \frac{\partial f}{\partial x}\imath + \frac{\partial f}{\partial y}\jmath = 2x\imath + 4y\jmath$

$\nabla f(-\frac{20}{51}, \frac{2}{51}) = 2(-\frac{20}{51})\imath + 4(\frac{2}{51})\jmath = -\frac{40}{51}\imath + \frac{8}{51}\jmath$

As we can see the gradient at my solution is not equal to 0

Let's compute the minimum of x,y when they have no contraints

$\nabla f(x, y) = \frac{\partial f}{\partial x}\imath + \frac{\partial f}{\partial y}\jmath = 2x\imath + 4y\jmath$

if $\nabla f(x, y) = 0$, then:
$x = 0 \; y = 0$

$$\mathbb{Hess}(\mathbb{f}) = \begin{bmatrix} 2 & 0 \\ \\ 0 & 4 \end{bmatrix}$$

$\mathbb{Hess}(\mathbb{f}) > 0$

So $x = 0, y = 0$ is a global minimizer of $f(x, y)$ when there are not constrains since it fulfills the 2 sufficient conditions:

So, the notable thing about the gradient in my solution in comparison with the soltuion where x, y have no contrains is that it is not zero, and when x,y have no contraints it is equal to zero

PROBLEM 4: Alternate formulations of Least Squares
Consider the constrained least squares problem:

$$\begin{array}{ll} \underset{r,y}{\text{minimize}} & ||r||_2 \\ \text{subject to} & r = b - Cy. \end{array}$$

where $C \in R^{m \times n}$,$n \leq m$ and rank n.

1. Convert this problem into the standard constrained least squares form.

The standard constrained least squares form is:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & ||b - Ax||^2 \\ \text{subject to} & Dx = d \end{array}$$

So, let's convert the constrained least square problem given

$r = b - Cy$
$Cy + r = b$

We want the constraint to have the form $Cx = d$

If we define $M = [C \ I_m]$ and $x = [y, r]^T$,
$Cy + r = b \rightarrow Mx = b$
and $Mx = b$ has the same form as $Dx = d$

We want the term $||r||^2$ has the form $||b - Ax||^2$

If we define $A = [0_n \ I_m]$
$r \rightarrow r = Ax$
and Ax has the same form as $||b - Ax||^2$

So, we obtain:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & ||Ax||^2 \\ \text{subject to} & Mx = b \end{array}$$

We add a term in order to minimize

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \frac{1}{2}||Ax||^2 \\ \text{subject to} & Mx = b \end{array}$$

2. Form the augmented system from the Lagrangian as we did in class.

The Lagrangian function is:
$\mathscr{L}(x, \lambda; f) = \frac{1}{2}||Ax||^2 - \lambda^T(Mx - b) = (Ax)^T Ax + \lambda(Mx - b) = x^T A^T Ax$
$+ \lambda^T(Mx - b)$

$\frac{\partial \mathscr{L}}{\partial x} = A^T Ax + \lambda^T M = A^T Ax + M^T \lambda = 0$
$\frac{\partial \mathscr{L}}{\partial \lambda} = Mx - b = 0$

The augmented system is:
$\begin{bmatrix} A^T A & M^T \\ M & 0 \end{bmatrix} \times \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}$

Where: $A^T A = [0_n \;\; I_m]^T [0_n \;\; I_m]$
$M = [C \;\; I_m]$
$M^T = [C \;\; I_m]^T$
$x = [y \;\; r]^T$

So, the augmented system is
$\begin{bmatrix} 0 & 0 & C^T \\ 0 & I_m & I_m \\ C & I_m & 0 \end{bmatrix} \times \begin{bmatrix} y \\ r \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix}$

3. Manipulate this problem to arrive at the normal equations for a least-squares problem:
$C^T Cy = C^T b$**.**
Discuss any advantages of the systems at intermediate steps.

The set of least-squares solutions of $Ax = b$ coincides with the set of solutions of the normal equations: $A^T Ax = A^T b$. The normal equations are always consistent.

We can substitute r by $b - Cy$, such that:

$$\begin{array}{ll} \underset{y}{\text{minimize}} & ||b - Cy||_2 \\ \text{subject to} & b - Cy = 0. \end{array}$$

Let's manilpulate $b - Cy = 0$ to obtain $C^T Cy = C^T b$

$b - Cy = 0 \rightarrow Cy = b$
We multiply both sides by C^T
$C^T Cy = C^T b$

So, we obtain

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & ||b - Cy||_2 \\ \text{subject to} \quad & C^T C y = C^T b. \end{aligned}$$

The advantage of this intermediate step is that it helps simplify the original problem into a more manageable form. The normal equations can be easily solved using various numerical methods, making it possible to obtain an approximate solution to the original problem.