098311 Optimization 1 Spring 2018
HW 3

Chen Tessler 305052680
Orr Krupnik 302629027

April 30, 2018

Problem 1. Let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, L \in \mathbb{R}^{p\times n}$ and $\lambda \in \mathbb{R}_{++}$. Consider the function:

$$f(x) = ||Ax - b||_2^2 + \lambda ||Lx||_1$$

1. Show that if f is coercive then $\text{Null}(A) \bigcap \text{Null}(L) = \{0\}$.
2. **Bonus:** Show that the contrary also holds, i.e., if $\text{Null}(A) \bigcap \text{Null}(L) = \{0\}$ then f is coercive.

Solution

1. Let us assume that $\text{Null}(A) \bigcap \text{Null}(L) \neq \{0\}$. Then, there exists some vector $x_* \neq 0 \in \mathbb{R}^n$ such that $x_* \in \text{Null}(A)$ and $x_* \in \text{Null}(L)$. We can multiply x_* by some scalar $\alpha \in \mathbb{R}$, and still have $\alpha x_* \in \text{Null}(A) \bigcap \text{Null}(L)$ (since $A(\alpha x) = \alpha A x = 0$). Now, taking $\alpha \to \infty$, we have $||\alpha x_*|| \to \infty$, however $\lim_{\alpha\to\infty} f(\alpha x_*) = ||b||_2^2 \neq \infty$, contrary to the fact that $f(x)$ is coercive. Therefore, our assumption is incorrect, and $\text{Null}(A) \bigcap \text{Null}(L) = \{0\}$.

2. In order to prove coerciveness, we need to show that for any M there exists some R such that for all $||x||_2 > R$ then $f(x) > M$. We will prove this by splitting into two parts, either $x \in \text{Null}(A)$ or it is not. The combination of both options will cover all possible x values.

   (i) We start by showing that for any M there exists some R such that for all $||x||_2 > R$ and $x \notin \text{Null}(A)$ the following holds: $||Ax - b||_2^2 > M$.
   Note that as $x \notin \text{Null}(A)$, $Ax \neq \mathbf{0}$. We define $\alpha_y y = x$ where $||y|| = 1$ and $\alpha_y = ||x|| > 0$. Hence $\alpha_y A y = Ax \neq \mathbf{0}$ for all y. Observe the norm $||Ax - b||_2^2 = ||\alpha_y Ay - b||_2^2 = \sum_{i=1}^n (\alpha_y Ay - b)_i^2$; as the vector Ay is non zero, there exists at least one element j which is not zero.

$$\Rightarrow ||Ax - b||_2^2 = ||\alpha_y Ay - b||_2^2 = \sum_{i=1}^n (\alpha_y Ay - b)_i^2 \geq (\alpha_y Ay - b)_j^2$$

By selecting $\alpha_y > \frac{\sqrt{M}+|b_j|}{(Ay)_j}$ we have that $||\alpha_y Ay - b|| > M$. Furthermore, by selecting $\bar{\alpha} = \max_y \alpha_y$ we note that the above holds $\forall y \notin \text{Null}(A)$.
We have shown that for $R_1 = \bar{\alpha}$, $\forall x : ||x|| > R_1$ the following holds $||Ax - b|| > M$.

(ii) We now continue by showing that for any M there exists some R such that for all $||x||_2 > R$ and $x \in \text{Null}(A)$ the following holds: $\lambda||Lx||_1 > M$.
Since $\text{Null}(A) \bigcap \text{Null}(L) = \{0\}$ we have that if $y \in \text{Null}(A)$ then $y \notin \text{Null}(L)$. We define similarly to the above $\beta_y y = x$ where $||x|| = \beta_y$ and $||y|| = 1$. Again $x \notin \text{Null}(L)$, Lx has at least one non-zero indice (we denote that indice as j).

$$\lambda||Lx||_1 = \lambda||\beta_y Ly||_1 = \lambda\beta_y \sum_{i=1}^{n} |(Ly)_i| \geq \lambda\beta_y|(Ly)_j|$$

By selecting $\beta_y > \frac{M}{\lambda|(Ly)_j|}$ we have that $\lambda||\beta_y Ly||_1 > M$. Furthermore, by selecting $\bar{\beta} = \max_y \beta_y$ we note that the above holds $\forall y \in \text{Null}(A)$.
We have shown that for $R_2 = \bar{\beta}$, $\forall x : ||x|| > R_2$ the following holds $\lambda||Lx||_1 > M$.

As (i) and (ii) hold for any $R > R_{1,2}$, it also holds for $R = \max\{R_1, R_2\}$.

Finally, by combining all of the above, we have that $\forall ||x|| > R$:

$$f(x) = ||Ax - b||_2^2 + \lambda||Lx||_1 \geq \begin{cases} ||Ax - b||_2^2, & x \notin \text{Null}(A) \\ \lambda||Lx||_1, & x \in \text{Null}(A) = \text{Null}^C(A) \end{cases} > M$$

where the first inequality is as all elements are non-negative and the second one is from (i) and (ii).

As $\text{Null}(A) \bigcup \text{Null}^C(A) = \mathbb{R}^n$ we have shown that for any M there exists an $R > 0$ such that $\forall x : ||x|| > R$ the following holds $f(x) > M$ and thus $f(x)$ is coercive which concludes our proof $\quad \square$.

Problem 2. The principal minors criterion states that a symmetric matrix A is positive definite if and only if all the leading principal minors of A are positive. The purpose of this problem is to give an elegant proof of this result using optimization techniques.
Let A be an $(n+1) \times (n+1)$ symmetric matrix in the form $A = \begin{bmatrix} B & b \\ b^T & c \end{bmatrix}$ where B is a positive definite $n \times n$ matrix, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

1. Consider the quadratic function

$$p(x) = (x^T, 1) \begin{bmatrix} B & b \\ b^T & c \end{bmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = x^T Bx + 2b^T x + c$$

on $\mathbb{R}^n$. Show that the point $x^* = -B^{-1}b$ is the unique global minimizer of p on $\mathbb{R}^n$, and $p(x^*) = c - b^T B^{-1} b$. Thus, p is positive on $\mathbb{R}^n$ if and only if $c > b^T B^{-1} b$.

2. Show that $\det A = \det B \cdot (c - b^T B^{-1} b)$.
   <u>Hint</u>: Find a suitable vector $d \in \mathbb{R}^n$ such that

$$\begin{bmatrix} I & 0 \\ d^T & 1 \end{bmatrix} \begin{bmatrix} B & b \\ b^T & c \end{bmatrix} \begin{bmatrix} I & d \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} B & 0 \\ 0 & c - b^T B^{-1} b \end{bmatrix}$$

3. Prove the principal minors criterion by induction on the dimension of A using parts 1. and 2. of the question.

Solution

1. As B is PD it has an inverse (since all eigenvalues are positive), thus we can calculate the gradient of $p(x)$ and compare to zero in order to find the stationary points.

$$\nabla p(x) = 2Bx + 2b = 0 \Rightarrow x = -B^{-1}b$$

   The hessian of $p(x)$ is:

$$\nabla^2 p(x) = 2B \succ 0$$

   As $\nabla p(-B^{-1}b) = 0$ (and $\nabla p(x) \neq 0, \forall x \neq -B^{-1}b$) and $\nabla^2 p(-B^{-1}b) \succ 0$, we have that $x^* = -B^{-1}b$ is a unique global minimum of p.

   B is the Hessian of $p(x)$ and hence a symmetric matrix, now plugging in x^* we have:

$$\begin{aligned} p(x^*) &= x^* B x^* + 2b^T x^* + c = (B^{-1}b)^T B (B^{-1}b) - 2b^T (B^{-1}b) + c \\ &= b^T B^{-1} b - 2b^T B^{-1} b + c = -b^T B^{-1} b + c \end{aligned}$$

   hence $p(x^*) > 0 \Rightarrow c > -b^T B^{-1} b$ and as x^* is a unique global minimum of $p(x)$ we have that if $c > b^T B^{-1} b$ then $p(x) > 0$ for any $x \in \mathbb{R}^n$.

2. Given we have some $q = \begin{bmatrix} I & d \\ 0 & 1 \end{bmatrix}$:

$$\det(q^T A q) = \det(q^T)\det(A)\det(q) = \det(q)^2 \det(A) = \det(A)$$

   where the final equality is since $\det(q) = 1$ (as q is a triangular matrix).

   We define $d \triangleq -B^{-1}b$:

$$\begin{aligned} &\begin{bmatrix} I & 0 \\ -B^{-1}b & 1 \end{bmatrix} \begin{bmatrix} B & b \\ b^T & c \end{bmatrix} \begin{bmatrix} I & -b^T B^{-1} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} B & Bd + b \\ d^T B + b^T & d^T B d + 2d^T b + c \end{bmatrix} \\ &= \begin{bmatrix} B & B(-B^{-1}b) + b \\ d^T B + b^T & (-B^{-1}b)^T B (-B^{-1}b) + 2(-B^{-1}b)^T b + c \end{bmatrix} \\ &= \begin{bmatrix} B & 0 \\ 0 & c - b^T B^{-1} b \end{bmatrix} \end{aligned}$$

   Therefore, we have:

$$\det(A) = \det(q^T A q) = \det \begin{bmatrix} B & 0 \\ 0 & c - b^T B^{-1} b \end{bmatrix} = \det(B) \cdot (c - b^T B^{-1} b)$$

3. We will show by induction that $A \succ 0 \iff PMP$ (All **P**rincipal **M**inors are **P**ositive).

Base: for $n = 1$, we have:

$PMP \Rightarrow B > 0, c > \dfrac{b^2}{B} > 0 \Rightarrow \det(A) > 0, \operatorname{tr}(A) > 0 \Rightarrow A \succ 0$

$A \succ 0 \Rightarrow \det(A) > 0,\ x^T A x > 0,\ \forall x \in \mathbb{R}^2 \Rightarrow \begin{pmatrix} 1 & 0 \end{pmatrix} A \begin{pmatrix} 1 \\ 0 \end{pmatrix} > 0 \Rightarrow B > 0 \Rightarrow PMP$

Step: Assume for some $n \geq 1$, we have: $A_{n\times n} \succ 0 \iff PMP$. Then, we shall prove the same holds for $A_{(n+1)\times(n+1)}$:

Direction 1: $A_{(n+1)\times(n+1)} \succ 0 \Rightarrow PMP$:

$$A_{(n+1)\times(n+1)} \succ 0 \Rightarrow x^T \begin{bmatrix} B_{n\times n} & b_{n\times 1} \\ b^T_{1\times n} & c \end{bmatrix} x > 0,\ \forall x \in \mathbb{R}^{n+1}$$

This holds specifically for any $\bar{x}^T = \begin{pmatrix} y^T & 0 \end{pmatrix}$ for all $y \in \mathbb{R}^n$. Therefore:

$$0 < \bar{x}^T \begin{bmatrix} B_{n\times n} & b_{n\times 1} \\ b^T_{1\times n} & c \end{bmatrix} \bar{x} = y^T B y,\ \forall y \in \mathbb{R}^n \Rightarrow B \succ 0$$

Since B is an $n \times n$ PD matrix, by the induction assumption, all the principal minors of B are positive. These principal minors are also all the principal minors of A, except $\det(A)$. It now remains to show $\det(A) > 0$. From section 2, we have: $\det(A) = \det(B) \cdot (c - b^T B^{-1} b)$. Since $A \succ 0$, we have $x^T A x > 0$ for any $x \in \mathbb{R}^{n+1}$, and specifically for $\bar{x}^T = \begin{pmatrix} y^T & 1 \end{pmatrix}$, for any $y \in \mathbb{R}^n$. This means $p(y) > 0$, $\forall y \in \mathbb{R}^n$ (where $p(y)$ as defined in section 1), and from section 1, this means $c - b^T B^{-1} b > 0$. Since $B \succ 0$, we have $\det(B) > 0$, and together we get $\det(A) > 0$, hence all the principal minors of A are positive.

Direction 2: $PMP \Rightarrow A_{(n+1)\times(n+1)} \succ 0$:

For $A = \begin{bmatrix} B_{n\times n} & b_{n\times 1} \\ b^T_{1\times n} & c \end{bmatrix}$ notice that the first n principal minors are the principal minors of $B_{n\times n}$. All the principal minors of $B_{n\times n}$ are positive $\Rightarrow$ by the induction assumption $B_{n\times n} \succ 0$.
We shall now show that: $x^T A_{(n+1)\times(n+1)} x > 0$, $\forall x \in \mathbb{R}^{n+1}$.

Let us look at two separate cases.

- $x_{n+1} = 0$: We can write x as $\begin{pmatrix} y^T & 0 \end{pmatrix}^T$, and then: $x^T A x = y^T B y > 0$ since $B \succ 0$.
- $x_{n+1} \neq 0$: We can write x as $\frac{1}{x_{n+1}} \begin{pmatrix} z^T & 1 \end{pmatrix}^T$ and then:

$$(*)\ x^T A x = \frac{1}{(x_{n+1})^2} \begin{pmatrix} z^T & 1 \end{pmatrix} A \begin{pmatrix} z \\ 1 \end{pmatrix} = \frac{1}{(x_{n+1})^2} p(z)$$

Since all the principal minors of A are positive, we have $\det(A) > 0$ (the largest principal minor) and from section 1: $\det(B) \cdot (c - b^T B^{-1} b) = \det(A) > 0$. Since all the principal minors of B are positive, specifically $\det(B) > 0$, which implies $c > b^T B^{-1} b$. Now, from section 2, we have $p(z) > 0$, $\forall z \in \mathbb{R}^n$, and therefore from $(*)$ we have $x^T A x > 0$, $\forall x$ for which this case applies.

Combining both cases, we have $x^T A x > 0 \; \forall x \in \mathbb{R}^{n+1}$, and therefore $A \succ 0$. This completes the second direction of our proof. ■

Problem 3. See HW exercise description.

Solution The coefficients without random perturbations are:

$$\begin{pmatrix} 1.0076 \\ -0.4666 \\ 1.0106 \\ -1.0961 \\ -2.0112 \end{pmatrix}$$

The coefficients with random perturbations change as follows:

$$\begin{pmatrix} 1.0135 \\ -0.4415 \\ 1.0193 \\ -1.1701 \\ -2.0205 \end{pmatrix}$$

The resulting orbits can be seen in Figure 1. The fact that the problem is almost rank deficient allows the small perturbations in the data points to create a relatively large perturbation in the resulting coefficients (and consequentially, orbit). This makes the difference clearly visible in 1, even though the perturbations are small enough to not be visible (and still seem to fit the same curve).

Problem 4.

1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called strongly coercive if

$$\lim_{||x||_2 \to \infty} \frac{f(x)}{||x||_2} = \infty$$

.

   (a) Show that a strong coercive function is also a coercive function. Give an example that demonstrates that the opposite is not necessarily true.

   (b) Prove that if f is strongly coercive and differentiable then the gradient operator $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is onto $\mathbb{R}^n$. That is, given $y \in \mathbb{R}^n$ there exists a point $x \in \mathbb{R}^n$ such that $\nabla f(x) = y$. Consider, for example, the following one-dimensional functions:

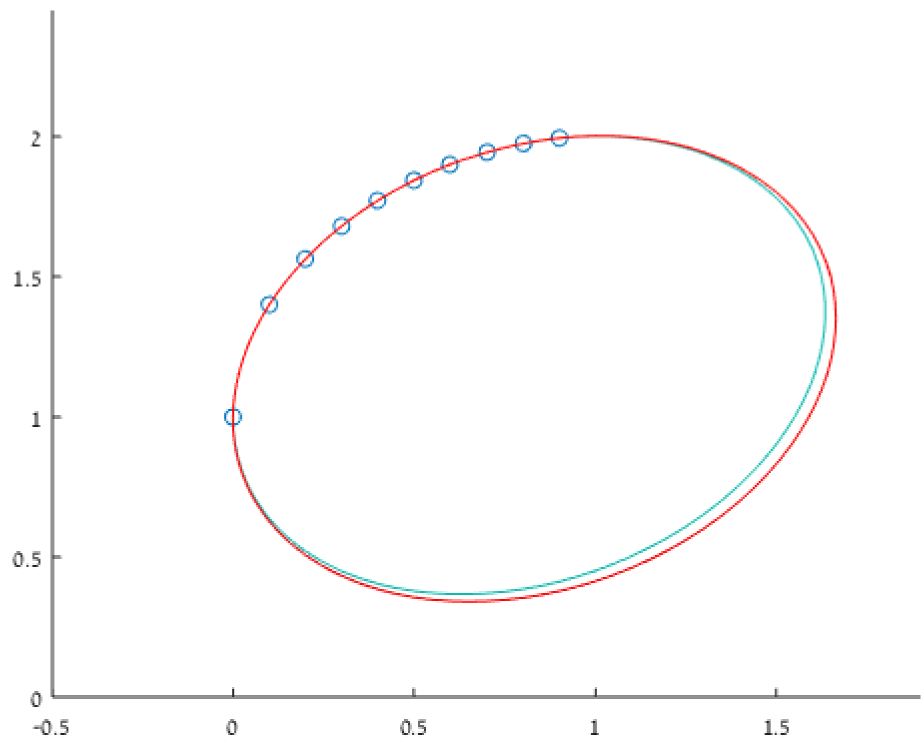

Figure 1: Plot of the calculated orbits. Green is without random perturbations; red is after adding the random noise to the data points. The original data points are marked with circles.

- $f(x) = \frac{1}{2}x^2$ which is strong coercive and its derivative $f'(x) = x$ range is all of $\mathbb{R}$.
- $f(x) = \frac{1}{3}x^3$ is not strong coercive and its derivative $f'(x) = x^2$ does not have all of $\mathbb{R}$ for its range.
  Hint: Recall that a continuous coercive function admits a minimum.

2. For each of the following functions **prove** whether it is coercive or not.

   (a) $f(x_1, x_2) = x_1^2 - 2x_1x_2^2 + x_2^4$.

   (b) $f(x_1, x_2, x_3) = x_1^4 + x_2^4 + x_3^2 - 3x_1x_2x_3$.

   (c) $f(x) = \frac{x^T Ax}{||x||_2+1}$, where $A \in \mathbb{R}^{n\times n}$ is positive definite.

   (d) $f(x) = x^T Ax - 3||x||_2^2 + 3\sum_{i=1}^n x_i$, where $A \in \mathbb{R}^{n\times n}$ is positive definite.

Solution

1.

   (a) By definition of a strong coercive function we have that $\forall M > 0$ there exists some $R > 0$ such that $\forall x : ||x||_2 > R$ the following holds $\frac{f(x)}{||x||_2} > M$. This is true for any x with norm greater than R and specifically greater than $\bar{R} = \max(1, R)$.

$$\frac{f(x)}{||x||_2} > M \Rightarrow f(x) > M||x||_2 \geq M$$

   The opposite does not hold, consider the function $f(x) = ||x||_2$. Coerciveness holds trivially by definition yet $\frac{f(x)}{||x||_2} = 1$, hence $f(x)$ is not strongly coercive.

   (b) Observe the function $g(x) = f(x) - a^T x$ where $a \in \mathbb{R}^n$. Notice that $\lim_{||x||_2\to\infty} \frac{g(x)}{||x||_2} = \lim_{||x||_2\to\infty} \frac{f(x)-a^T x}{||x||_2} \to \infty$ as $\frac{a^T x}{||x||_2} \leq \max_i a_i < \infty$ (is bounded), and $\lim_{||x||_2\to\infty} \frac{f(x)}{||x||_2} \to \infty$ by definition.
   Hence $g(x)$ is also strongly coercive and hence coercive (using the proof from Q4 1.a).
   As $g(x)$ is coercive, it receives a global minimum on x.

$$\nabla g(x) = \nabla f(x) - a = \mathbf{0} \Rightarrow \nabla f(x) = a$$

   We have found that for any $a \in \mathbb{R}^n$, the function $g(x)$ has a global minimum at any x which satisfies $\nabla f(x) = a$.

2.

   (a)

$$f(x_1, x_2) = x_1^2 - 2x_1x_2^2 + x_2^4 = (x_1 - x_2^2)^2 \geq 0$$

Yet selecting $x_1 = x_2^2$ we have:

$$f(x_1, x_2) = (x_2^2 - x_2^2)^2 = 0, \forall x_2$$

Hence for $x_2 \to \infty$ then $||x||_2 \to \infty$ yet $f(x) = 0$. We have found a direction in which $f(x)$ is bounded for any vector in that direction, thus the function is not coercive.

(b) $f(x)$ is not coercive. To show this, note the vector $\bar{x} = (t, t, t^2)$:

$$f(x_1, x_2, x_3) = x_1^4 + x_2^4 + x_3^2 - 3x_1x_2x_3 \Rightarrow f(t, t, t^2) \quad = t^4 + t^4 + t^4 - 3t^4 \equiv 0$$

For any norm of the vector x the value of $f(x)$ is 0 which concludes our proof that $f(x)$ is non-coercive.

(c) For any $x : ||x||_2 \geq 1$:

$$f(x) = \frac{x^T Ax}{||x||_2 + 1} \geq \frac{x^T Ax}{2||x||_2} = \frac{||x||_2^2 \hat{x}^T A\hat{x}}{2||x||_2} \geq ||x||_2 \lambda_{min}(A) \geq ||x||_2$$

where $\hat{x}^T A\hat{x} \geq \lambda_{min}(A)$ as we've seen in the recitation. As $f(x)$ is larger than $||x||_2$, we have $f(x) \xrightarrow{||x||_2 \to \infty} \infty$, hence $f(x)$ is coercive.

(d)

$$f(x) = x^T Ax - 3||x||_2^2 + 3\sum_i x_i$$

For $f(x)$ to be coercive, it should be such for any $A \succ 0$. Yet, for $A = I \succ 0$ we have that:

$$\begin{aligned} f(x) &= x^T Ix - 3||x||_2^2 + 3\sum_i x_i = ||x||_2^2 - 3||x||_2^2 + 3\sum_i x_i \\ &= -2||x||_2^2 + 3\sum_i x_i = \sum_i (-2x_i^2 + 3x_i) \end{aligned}$$

for any x where $x_i < 0$ for all i, $f(x) < 0$ hence $f(x)$ is not coercive (since we can find some path for which $x_i < 0$ for all i, for which $||x|| \to \infty$, such as $\phi(t) = (-|t|, ..., -|t|))$.