E0 230
Computational Methods of Optimization
Tutorial 1

Nov. 7, 2020

---

1. Suppose $x,\ y \in \mathbb{R}^n$. Prove the following statements.
   (a) $|\|x\| - \|y\|| \leq \|x - y\| \leq \|x\| + \|y\|$

   Solution:
   We start by writing $\|x-y\|^2 = x^Tx + y^Ty - 2x^Ty = \|x\|^2 + \|y\|^2 - 2x^Ty$. Next, by Cauchy-Schwartz (derived in class, get familiar with this inequality!) we have $-2\|x\|\|y\| \leq -2x^Ty \leq 2\|x\|\|y\|$. From this, it follows that

   $$\begin{aligned}&\|x\|^2 + \|y\|^2 - 2\|x\|\|y\| \leq \|x-y\|^2 \leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \\ \Rightarrow &(|\|x\| - \|y\||)^2 \leq \|x-y\|^2 \leq (\|x\| + \|y\|)^2 \\ \Rightarrow &|\|x\| - \|y\|| \leq \|x-y\| \leq \|x\| + \|y\|\end{aligned}$$

   (b) $2x^Ty = \|x\|^2 + \|y\|^2 - \|x - y\|^2$

   Solution: Recall $\|x-y\|^2 = x^Tx + y^Ty - 2x^Ty = \|x\|^2 + \|y\|^2 - 2x^Ty$. Rearranging terms, we can prove the statement.

   (c) $\|x\|_2 \leq \sqrt{n}\|x\|_\infty$

   Solution: Suppose (wlog) that $|x_1| = \|x\|_\infty$. We have

   $$\|x\|_2 = \sqrt{x_1^2 + ... + x_n^2} = |x_1|\sqrt{1 + (x_2/x_1)^2 + ... + (x_n/x_1)^2} \leq |x_1|\sqrt{n} = \sqrt{n}\|x\|_\infty$$

   (d) $\|x\|_1 \leq n\|x\|_\infty$

   Solution: Suppose (wlog) that $|x_1| = \|x\|_\infty$. We have

   $$\|x\|_1 = |x_1| + ... + |x_n| = |x_1|\left(1 + |x_2/x_1| + ... + |x_n/x_1|\right) \leq |x_1|n = n\|x\|_\infty$$

   (e) $\|x\|_1 \leq \sqrt{n}\|x\|_2$

Solution: We have

$$\begin{aligned}
\|x\|_1 &= |x_1| + ... + |x_n| = 1 * |x_1| + ... + 1 * |x_n| \\
&\leq \left(\sqrt{\sum_{i=1}^{n} 1}\right)\left(\sqrt{\sum_{i=1}^{n} |x_1|^2}\right) \quad \text{(by Cauchy-Schwartz)} \\
&= \sqrt{n}\|x\|_2
\end{aligned}$$

We can also write $\|x\|_1 = x^T s$ where $s = \text{sign}(x)$, and apply Cauchy-Schwartz.

2. Note that $H_f(x)$ denotes the Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$

   (a) Show that $e^x \geq 1 + x$.

Solution:
We start by taking the Taylor series: $e^x = \sum_i \frac{x^i}{i!}$. The Lagrange remainder for the linear approximation gives us $e^x = 1 + x + \frac{e^z}{2}x^2$ for some $z \in [0, x]$. Since $e^z x^2 \geq 0$, it follows that $e^x \geq 1 + x$.
We can also think of this as an optimization problem. Note that $e^x \geq 1 + x \Rightarrow e^x - 1 - x \geq 0 \Rightarrow \min_x e^x - 1 - x \geq 0$. So now, let $g(x) = e^x - 1 - x$. $g'(x) = e^x - 1$, which gives us $e^{x^*} = 1 \Rightarrow x^* = 0$. To check optimality, we employ the second order condtions. For that $g''(x) = e^x$ which is always positive. Thus, $g(x)$ is minimized at $x = 0$, and $g(0) = 0$. Thus, we prove the statement.

   (b) Suppose $n = 1$, and $f^{(k)}$, the kth derivative of f w.r.t x is absolutely continuous. Show that given

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + ...\frac{1}{k!}f^{(k)}(x_0)(x - x_0)^k + R_k$$

we have

$$R_k = \int_{x_0}^{x} \frac{f^{(k+1)}(t)}{k!}(x - t)^k dt.$$

Solution: We prove the statement by induction for all R_k. First, for $k = 1$, we use the fundamental theorem of calculus: $f(x) = f(x_0) + \int_{x_0}^{x} f'(s)ds$. Then, we assume the statement holds for arbitrary k. Then, consider the term

$$R_k = \int_{x_0}^{x} \frac{f^{(k+1)}(t)}{k!}(x - t)^k dt$$

We apply integration by parts, giving us

$$\begin{aligned}
R_k = \int_{x_0}^{x} \frac{f^{(k+1)}(t)}{k!}(x - t)^k dt &= -\left[\frac{f^{(k+1)}(t)}{(k+1)!}(x - t)^{k+1}\right]_{x_0}^{x} - \int_{x_0}^{x} \frac{-f^{(k+2)}(t)}{(k+1)!}(x - t)^{k+1} \\
&= \frac{f^{(k+1)}(x_0)}{(k+1)!}(x - x_0)^{k+1} + \int_{x_0}^{x} \frac{f^{(k+2)}(t)}{(k+1)!}(x - t)^{k+1} = \frac{f^{(k+1)}(x_0)}{(k+1)!}(x - x_0)^{k+1} + R_{k+1}
\end{aligned}$$

Thus, we prove the statement.

(c) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, such that $\max_{x,i} |\lambda_i(H_f(x))| = M < \infty$, where $\lambda_i(H_f(x))$ is the ith eigenvalue of $H_f(x)$. Show that there exists a constant L such that, for each x, y, we have

$$f(y) - f(x) \le \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2$$

Solution: We apply the Taylor series expansion to $f(y)$ centered at x:

$$\begin{aligned}
f(y) &= f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T H_f(x)(y - x) + R_2(x) \\
&= f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T H_f(z)(y - x) \text{ for some } z \\
&\le f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \left(\lambda_{max}(H_f(z))I\right)(y - x) \\
&\le f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} \max_{i,w} |\lambda_i(H_f(w))| \|y - x\|^2 \\
&= f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}\|y - x\|^2 \quad \text{(proof holds with } L = M\text{)}
\end{aligned}$$

3. Suppose we have matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{m \times n}$, $m \neq n$

(a) Suppose A has n orthogonal eigenvectors. Show that we can write $A = V^T \Lambda V$, where the columns of V are the eigenvectors of A, and the diagonal elements of Λ are the eigenvalues of A. When are we unable to use this decomposition?

Solution: Let (λ_i, v_i) be an eigenpair of A. Then, we have $Av_i = \lambda_i v_i$. Taking the matrix $V = [v_1, ..., v_n]$, it follows that $AV = V\Lambda$ where Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i$. Next, recall that $V^T = V^{-1}$ since the columns are orthogonal. Thus, it follows that $AVV^T = V\Lambda V^T = A$.

(b) Suppose B is of rank r. What are the ranks of BB^T and $B^T B$? With this result, what can you say about the row and column ranks of B and B^T?

Solution:
From the rank-nullity theorem, we have $\text{rank}(B) + \text{nullity}(B) = \min\{m, n\}$. Thus, $Bx = 0 \Rightarrow B^T Bx = 0$, since $B^T Bx = 0 \Rightarrow x^T B^T Bx = (Bx)^T Bx = 0$. Thus, $\text{rank}(B) = \text{rank}(B^T B)$. Similarly, we can show that $B^T y = 0 \Rightarrow BB^T y = 0 \Rightarrow \text{rank}(B^T) = \text{rank}(BB^T)$. Next, we know that the columns of $B^T B$ are linear combinations of the columns of $B^T \Rightarrow \text{rank}(B^T B) \le \text{rank}(B^T)$. Similarly, the columns of BB^T are linear combinations of the columns of $B \Rightarrow \text{rank}(BB^T) \le \text{rank}(B)$. Thus, we get $\text{rank}(B) = \text{rank}(B^T B) \le \text{rank}(B^T)$ and $\text{rank}(B^T) = \text{rank}(BB^T) \le \text{rank}(B)$. This only holds if $\text{rank}(B) = \text{rank}(B^T)$.

(c) Let $p(l) = \det(lI - A)$. For any A, we have $p(A) = 0$ (this is the Cayley-Hamilton theorem). Show that $p(A) = p(P^{-1}AP)$ for any invertible matrix P. What does this say about the eigenvalues of A and $P^{-1}AP$?

Solution: Recall that $p(l) = \sum_i a_i l^i$, and $p(\lambda) = 0$ for any eigenvalue λ. We have $p(A) = \sum_i a_i A^i$. Next, note that $(P^{-1}AP)^n = (P^{-1}AP)(P^{-1}AP)... = P^{-1}A^nP$. Thus,

$$p(P^{-1}AP) = \sum_i a_i (P^{-1}AP)^i = \sum_i a_i P^{-1}A^iP = P^{-1}\left(\sum_i a_i A^i\right)P.$$

What this shows is that $p(A) = 0$ implies $p(P^{-1}AP) = 0$, which in turn implies that the characteristic polynomials share roots.

Another way to think about this is as follows. Let $C = P^{-1}AP \Rightarrow A = PCP^{-1}$. Then, $Av = \lambda v = PCP^{-1}v \Rightarrow CP^{-1}v = \lambda P^{-1}v$. This implies the eigenvalues are maintained, and the eigenvectors are transformed by P^{-1}.

(d) Show that we can decompose $B = U\Sigma V^T$, where Σ is diagonal and positive-semidefinite, and U and V have orthogonal columns.

Solution: Let V be a matrix containing the eigenvectors of nonzero eigenvalues of B^TB and let U be a matrix containing the eigenvectors of nonzero eigenvalues of BB^T. Let the eigenvalues be σ_i^2. Furthermore, we can write $u_i = \frac{1}{\sigma_i}Bv_i$ (why? $BB^Tu_i = BB^T(Bv_i)/\sigma_i = \sigma_i^2 Bv_i/\sigma_i$). From this, it folows we can write $U = BV\Sigma^{-1} \Rightarrow U\Sigma = BV \Rightarrow B = U\Sigma V^T$ since $VV^T = I$.

(e) Show that A, B are equivalent if and only if, for all vectors $v \in \mathbb{R}^n$, $Av = Bv$.

Solution: If $A = B$, then $Av = Bv$ holds trivially. Now, suppose $Av = Bv$ We need to show that $A = B$. Let $v = e_i$, where e_i is the ith coordinate vector. Then, $Ae_i = Be_i$, which implies the ith column of A is equivalent to the ith column of B for all i. Thus, we prove the statement.

(f) The Frobenius norm of a matrix is given by $\|B\|_F = \sqrt{\text{Tr}(B^TB)}$. Show that $\|B\|_F = \sqrt{\sum_k \sigma_k(B)^2}$, where $\sigma_k(B)$ is the kth largest singular value of B.

Solution: We can write $B = U\Sigma V^T$. Then, we have

$$\begin{aligned}\|B\|_F &= \sqrt{\text{Tr}\left(B^TB\right)} = \sqrt{\text{Tr}\left(V\Sigma^TU^TU\Sigma V^T\right)}\\ &= \sqrt{\text{Tr}\left(V\Sigma^T\Sigma V^T\right)} = \sqrt{\text{Tr}\left(\Sigma^T\Sigma V^TV\right)} = \sqrt{\text{Tr}\left(\Sigma^2\right)}\\ &= \sqrt{\sum_k \sigma_k^2}.\end{aligned}$$

More simply, the singular values of B are the eigenvalues of B^TB; thus, by the definition of trace, the expression holds.

(g) Consider the function $f : U \to \mathbb{R}$, where $f(x) = x^TAx$ and U is the set of all unit vectors of dimension n. Show that, if A is symmetric, range$(f) \subseteq [\lambda_{min}(A), \lambda_{max}(A)]$.

Solution:
Note that since A is symmetric, we can write $A = Q\Lambda Q^T$, where Q is an orthogonal matrix. Furthermore, let $y = Qx$, and note that $\|y\| = \sqrt{x^TQ^TQx} = \sqrt{x^Tx} = \|x\| = 1$. Thus, $f(x) = x^TQ^T\Lambda Qx = y^T\Lambda y = \lambda_1 y_1^2 + ... + \lambda_n y_n^2$. Then, it follows that $\lambda_{\min} \leq \lambda_1 y_1^2 + ... + \lambda_n y_n^2 \leq \lambda_{\max}$

(h) What is the solution to

$$s^* = \max_x \frac{\|Bx\|_2}{\|x\|_2}?$$

*R*emark: s^* is the Spectral norm of B, denoted by $\|B\|_2$.

Solution:
First, note that

$$\max_x \frac{\|Bx\|_2}{\|x\|_2} = \max_x \frac{\|Bx\|_2^2}{\|x\|_2^2} = \max_x \frac{x^T B^T B x}{x^T x}.$$

Now, we can write $B^T B = Q^T \Lambda Q$. Then, we denote $y = Qx$, and get

$$\frac{x^T B^T B x}{x^T x} = \frac{x^T Q^T \Lambda Q x}{x^T Q^T Q x} = \frac{y^T \Lambda y}{y^T y} = \frac{\sum_i \lambda_i y_i^2}{\sum_i y_i^2} \leq \lambda_{\max}.$$

Here, $\lambda_{\max}$ is the largest eigenvalue of $B^T B$, and is thus the largest singular value $\sigma_{\max}(B)$.

(i) Suppose $\sum_{i=1}^n \sigma_i u_i v_i^T = B \in \mathbb{R}^{m\times n}$. Show that $B_k = \sum_{i=1}^k \sigma_i u_i v_i^T$ satisfies $\|B - B_k\|_F \leq \|B - C\|_F$ for any $C \in \mathbb{R}^{m\times n}$ of rank k.

Solution: Let $B = \sum_{i=1}^n \sigma_i u_i v_i^T$ and $C = \sum_{j=1}^K \rho_j x_j y_j^T$, and $\sigma_1 \geq ... \geq \sigma_n$ and $\rho_1 \geq ...\rho_n$. Using the definition of the Frobenius norm, we get

$$\begin{aligned}
\|B - C\|_F^2 &= \mathrm{Tr}((B-C)^T(B-C)) = \mathrm{Tr}(B^TB) + \mathrm{Tr}(C^TC) - 2\mathrm{Tr}(B^TC) \\
&= \sum_i \sigma_i^2 + \sum_j \rho_j^2 \mathrm{Tr}(y_j x_j^T x_j y_j^T) - 2\mathrm{Tr}(\sum_i \sum_j \sigma_i \rho_j v_i u_i^T x_j y_j^T) \\
&= \sum_i \sigma_i^2 + \sum_j \rho_j^2 \mathrm{Tr}(y_j x_j^T x_j y_j^T) - 2\sum_i \sum_j \sigma_i \rho_j \mathrm{Tr}(v_i u_i^T x_j y_j^T) \\
&= \sum_i \sigma_i^2 + \sum_j \rho_j^2 x_j^T x_j y_j^T y_j - 2\sum_i \sum_j \sigma_i \rho_j u_i^T x_j \mathrm{Tr}(v_i y_j^T) \\
&= \sum_i \sigma_i^2 + \sum_j \rho_j^2 x_j^T x_j y_j^T y_j - 2\sum_i \sum_j \sigma_i \rho_j u_i^T x_j v_i^T y_j
\end{aligned}$$

Taking the gradients of this function w.r.t. each x_j and y_j, we get

$$\rho_j^2 x_j - \rho_j \sum_i \sigma_i v_i^T y_j u_i = 0 \text{ and } \rho_j^2 y_j - \rho_j \sum_i \sigma_i u_i^T x_j v_i = 0 \text{ for all } j \in [k]$$

Thus, it follows that $y_j = \frac{1}{\rho_j}\sum_i \sigma_i u_i^T x_j v_i$. Substituting the expression back into the gradient w.r.t. x_j, we get

$$\begin{aligned}
\rho_j^2 x_j &= \rho_j \sum_i \sigma_i v_i^T y_j u_i = \rho_j \sum_i \sigma_i v_i^T \left(\frac{1}{\rho_j}\sum_l \sigma_l u_l^T x_j v_l\right) u_i \\
&= \sum_i \sigma_i^2 u_i^T x_j u_i = \sum_i \sigma_i^2 u_i u_i^T x_j \\
\Rightarrow \rho_j^2 \sum_i \alpha_i u_i &= \sum_i \sigma_i^2 u_i u_i^T \left(\sum_l \alpha_l u_l\right) = \sum_i \sigma_i^2 \alpha_i u_i \\
\Rightarrow 0 &= \sum_i (\rho_j^2 - \sigma_i^2)\alpha_i u_i
\end{aligned}$$

The last equality holds if $\alpha_j = 1$, $\alpha_i = 0$ for $i \neq j$, and $\sigma_j = \rho_j$. Thus, $x_j = u_j$ and using a similar technique, we get $y_j = v_j$. This is a minimum because the second derivatives w.r.t. x_j or y_j are positive definite for all j. Thus, we prove the statement. Alternatively. we use the Von Neumann Trace inequality: $|\mathrm{Tr}(B^TC)| \leq \sum_i \sigma_i \rho_i$. We then get

$$\begin{aligned}
\|B - C\|_F^2 &= \sum_i \sigma_i^2 + \sum_j \rho_j^2 - 2\mathrm{Tr}(B^TC) \geq \sum_i \sigma_i^2 + \sum_j \rho_j^2 - 2|\mathrm{Tr}(B^TC)| \\
&\geq \sum_i \sigma_i^2 + \sum_j \rho_j^2 - 2\sum_j \sigma_i \rho_j = \sum_j (\sigma_j^2 + \rho_j^2 - 2\sigma_j\rho_j) + \sum_{i>k} \sigma_i^2 \\
&= \sum_j (\sigma_j - \rho_j)^2 + \sum_{i>k} \sigma_i^2.
\end{aligned}$$

This is minimized when $\sigma_i = \rho_i$ for $i \leq k$. Thus, the optimal $C = B_k$.

For the Von Neumann trace inequality, see *Matrix Analysis* Chapter 8.7.6 by Horn and Johnson. For a geometric proof of this theorem (a.k.a. the Eckart-Young-Mirsky Theorem) see *Foundations of Data Science* Chapter 3.1-3.3 by Hopcroft, Blum, and Kannan.

4. Consider the polynomial

$$p(x, y, z) = x^4y^2 + x^2y^4 + z^6 - 3x^2y^2z^2.$$

Show that

$$f^* = \inf_{x,y,z} p(x, y, z) = 0.$$

Solution: In this problem, we're essentially asked to prove that $p(x, y, z) \geq 0$.To do so, we employ the AM-GM inequality (given positive $\{a_i\}_{i=1}^n$, $\frac{1}{n}\sum_i a_i \geq (a_1...a_n)^{1/n}$) on the first three monomials:

$$\frac{x^4y^2 + x^2y^4 + z^6}{3} \geq \left(x^6y^6z^6\right)^{\frac{1}{3}}$$

from which we see that the statement holds.

5. Suppose A, B are symmetric and that the problems

$$\text{(P1)} \quad \underset{x}{\operatorname{argmin}}\, x^TAx \quad \text{and} \quad \text{(P2)} \quad \underset{x}{\operatorname{argmin}}\, x^TBx$$

have unique solutions $x_{P1} = x_{P2} = 0$. What is the solution to

$$\underset{x}{\operatorname{argmin}}\, x^TABx,$$

and is it unique? (hint: every symmetric PD matrix has a unique, positive definite square root - can you prove this?)

Solution:

Since P1 and P2 have unique solutions, it follows that A, B are positive definite. We know that symmetric PD matrices have a PD square root, since each symmetric PD matrix C can be decomposed as $C = V\Lambda V^T = V\Lambda^{1/2}V^TV\Lambda^{1/2}V^T$. Thus, $C^{1/2} = V\Lambda^{1/2}V^T$. Then, we know that the spectrum of B is the same as the spectrum of PAP^{-1} for any invertible matrix P. Thus, the eigenvalues of B are the same as the eigenvalues of $C = A^{1/2}BA^{-1/2} \succ 0$. Thus, we see that $AB = A^{1/2}CA^{1/2}$, and $x^TABx = y^TCy$ where $y = A^{1/2}x$. Since C is PD, $\operatorname{argmin}_y y^TCy = x_{P2} = 0$.

6. Suppose we have m scalar data points $\{x_i\}_{i=1}^m$. What is the solution to

$$z_2 = \underset{z}{\operatorname{argmin}} \sum_i (x_i - z)^2.$$

Solution: Let $f(z) = \sum_i (x_i - z)^2$. To find the extreme points $f'(z) = \sum_i 2(z - x_i) = 0 \Rightarrow 2nz = 2\sum_i x_i \Rightarrow z = \frac{1}{n}\sum_i x_i$. Furthermore, we have $f''(z) = 2 > 0$, so the minimum is unique.

7. Suppose we have m pairs of data points (x_i, y_i), where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. Solve

$$w_* = \underset{w}{\operatorname{argmin}} \sum_i (y_i - w^Tx_i)^2.$$

What if there are only $r < \min\{m, n\}$ linearly independent data points? Will you still have a unique solution (show why or why not).

Solution:

Define $Y \in \mathbb{R}^n$, where $Y_i = y_i$, and $X \in \mathbb{R}^{m \times n}$ where the ith column of X is x_i. Then, $f(w) = \sum_i (y_i - w^T x_i)^2 = \|Y - Xw\|^2 = w^T X^T X w + Y^T Y - 2Y^T X w$. The Hessian of $f(w)$ is $X^T X$, which is positive definite if there are m linearly independent data points (in which case, a unique minimum exists), or PSD if there are $r < m$ LI datapoints (in which there may be infinitely many equally good minima). To solve this problem, we set $\nabla f(w^*) = 0 \ \Rightarrow 2X^T X w^* = 2X^T Y \ \Rightarrow w^* = (X^T X)^{-1} X^T Y$, or, if $\text{rank}(X) = r < m$, we can use the Psuedoinverse $w^* = (X^T X)^\dagger X^T Y$, where $A^\dagger = V \hat{\Sigma} U^T$, where $\hat{\Sigma}_i i = \frac{1}{\Sigma_i i}$ if $\Sigma_i i \neq 0$.

8. Suppose we have positive definite $A \in \mathbb{R}^{n \times n}$, and linearly independent vectors $\{v_i\}_{i=1}^m$, where $m < n$. How would you convert the problem

$$\underset{x}{\text{argmin}}\, x^T A x \ \text{ such that } x \in \text{span}\,(v_1, ..., v_m)$$

into an unconstrained problem? Does this problem have a unique solution? If so, under what conditions would this problem not have a unique solution?

Solution:

We have $f(x) = x^T A x$. If we write $x = \sum_i \alpha_i v_i = V\alpha$, where the ith column of V is v_i, we get

$$f(x) = g(\alpha) = (V\alpha)^T A\,(V\alpha) = \alpha^T V^T A V \alpha.$$

Next, since $A \succ 0$, it follows that $V^T A V \succ 0$ as well, since $f(x) > 0 \Leftrightarrow g(\alpha) > 0$. Thus, $g(\alpha)$ has a unique minimum as well. The problem would not have a unique solution if A is PSD and if at least 1 v_i is in the nullspace of A.