Matrix Analysis and Linear Algebra **Dr. Donniell Fishkind**
TA: Ronak Mehta Discussion Section Notes

1 September 3, 2019

1.1 Section format

- Go over old homework solutions.
- Reiterate concepts from class.
- (Optional) Make connections to ideas from statistics or optimization.
- Answer any questions that you have.

Note: I will prepare material for each section, but your questions are universally more important than whatever I prepare, so please ask them.

1.2 Algorithm for success

(a) Attend every class.

(b) Take exhaustive notes, typeset them in real-time or otherwise.

(c) View homework immediately after posting.

(d) Memorize lecture notes verbatim prior to exams.

1.3 Real analysis review

Definition 1.1 (Metric space)**.** Let X be a set. Let $d : X \times X \to [0, \infty)$ with the following properties.

(a) $d(x, y) \geq 0$ and $d(x, y) = 0 \iff x = y$ for $x, y \in X$. (Identity of indiscernables)

(b) $d(x, y) = d(y, x)$ (Symmetry)

(c) $d(x, y) \leq d(x, z) + d(z, y)$ for $x, y, z \in X$. (Triangle inequality)

The tuple (X, d) is called a metric space.

Example 1.1. $X = \mathbb{R}$ and $d(x, y) = |x - y|$
Example 1.2. $X = \mathbb{R}^n$ and $d(x, y) = ||x - y||_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, the Euclidean distance.

Definition 1.2 (Sequence)**.** A sequence $(x_n)_{n=1}^{\infty} = x_1, x_2, ...$ is a countably infinitely long list.

Definition 1.3 (Limit of a sequence)**.** A sequence $(x_n)_{n=1}^{\infty}$, where $x_n \in (X, d)$, converges to limit $x \in (X, d)$ if $\forall \epsilon > 0, \exists N(\epsilon)$:

$$\forall n \geq N,\ d(x_n, x) < \epsilon$$

Note that convergence requires the limit to be in the space.

Example 1.3. The sequence $3, 3.1, 3.14, 3.141$ approaches π, but if our metric space was restricted to just the rational numbers $\mathbb{Q}$, then we would not call this a convergent sequence.

Example 1.4. Let $x_n = \frac{1}{n}$. This sequence has limit $x = 0$.

Proof. Given any ϵ, let $N = \lceil \frac{1}{\epsilon} \rceil$. Then, for $n \geq N$:

$$d(x_n, x) = \left| \frac{1}{n} - 0 \right| = \frac{1}{n} \leq \frac{1}{N} \leq \frac{1}{\frac{1}{\epsilon}} = \epsilon$$

□

Definition 1.4 (Cauchy sequence)**.** A sequence $(x_n)_{n=1}^{\infty}$, where $x_n \in (X, d)$, is Cauchy if $\forall \epsilon > 0, \exists N(\epsilon)$:

$$\forall k, l \geq N, \ d(x_k, x_l) < \epsilon$$

Example 1.5. The sequence $x_n = \frac{1}{n}$ is Cauchy.

Proof. Given any ϵ, let $N = \lceil \frac{2}{\epsilon} \rceil$. Then, for $k, l \geq N$:

$$d(x_k, x_l) = \left| \frac{1}{k} - \frac{1}{l} \right| \leq \frac{1}{k} + \frac{1}{l} \leq \frac{2}{N} \leq \frac{2}{\frac{2}{\epsilon}} = \epsilon$$

□

Exercise 1.1. Prove that a sequence converges $\implies$ the sequence is Cauchy.

Definition 1.5. A metric space (X, d) is complete if every Cauchy sequence converges.

Definition 1.6 (Open ball)**.** An open ϵ-ball about c is the set $B_\epsilon(c) = \{x : d(x, c) < \epsilon\}$.

Definition 1.7 (Accumulation point)**.** Let $A \subseteq X$. Point a is an accumulation point of A if

$$\forall \epsilon > 0 \ \exists x \in A : x \neq a \text{ and } x \in B_\epsilon(a)$$

In other words, every open ϵ-ball about a contains a point from A that is different from a.

Example 1.6. Let $X = \mathbb{R}$. The set $A = [0, 1)$ has accumulation point 1, as every interval $B_\epsilon(1) = (1 - \epsilon, e + \epsilon)$ contains a point in $[0, 1)$.

Definition 1.8 (Open set)**.** A set $A \subseteq X$ is called open if for every $x \in A$, $\exists \epsilon > 0 : B_\epsilon(x) \subseteq A$. In other words, for every point in A, a small enough open ball about that point is also in A.

Example 1.7. Let $X = \mathbb{R}$. The set $A = (0, 1)$ open.

Proof. Formally, let $x \in (0, 1)$. Let $\epsilon = \min\{x, 1 - x\}$. Then, $B_\epsilon(x) \subseteq (0, 1)$. □

Example 1.8. Any open ball is open.

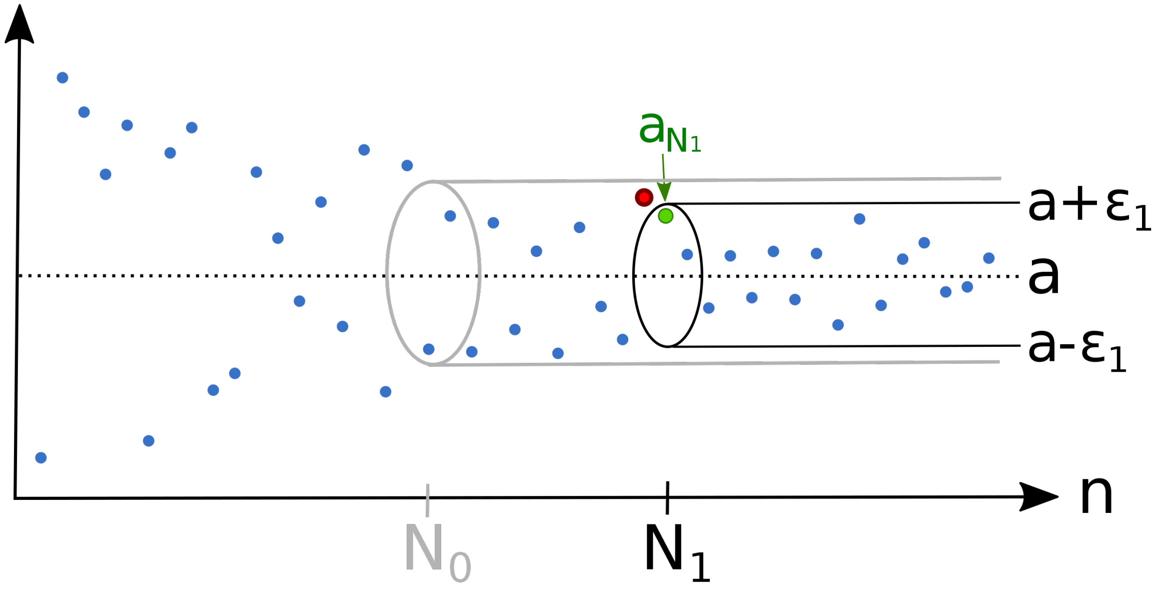

Figure 1: a denotes the limit in this sequence, N_0 denotes $N(\epsilon_0)$ for some ϵ_0, and $N_1 = N(\epsilon_1) < N_0$ for some $\epsilon_1 < \epsilon_0$.

Proof. Using the notation from the figure below, let $A = B_r(x)$ be an open ball in (X, d). Choose $y \in A$. Let $\epsilon = r - d(x, y)$. To show that $B_\epsilon(y) \subseteq B_r(x)$, take any point $z \in B_\epsilon(y)$. By construction, we have:

$$d(y, z) < \epsilon = r - d(x, y) \implies d(y, z) + d(x, y) < r$$

To show that $z \in B_r(x)$, we bound $d(x, z)$ by r.

$$d(x, z) \leq d(y, z) + d(x, y) < r$$

□

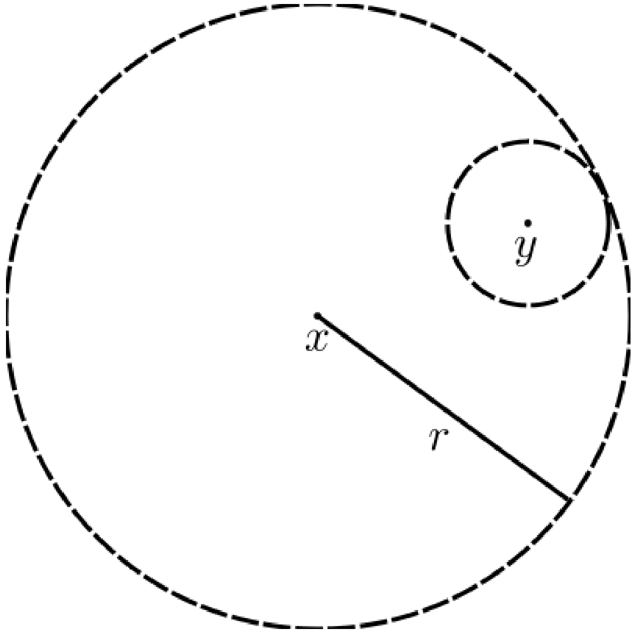

Figure 2: An open r-ball about x. Dotted lines typically denote the unincluded boundary of the set.

Definition 1.9 (Closed set)**.** A set A is closed if it contains all of its accumulation points.

Example 1.9. As seen above, $[0, 1)$ has (only) accumulation point 1. Thus the set $[0, 1]$ is closed.

Theorem 1.1. *A is open if and only if A^c is closed.*

Remark 1.1. In $\mathbb{R}$, the sets $\mathbb{R}$ and $\emptyset = \{\}$ are both open and closed.

Remark 1.2. A finite set $A = \{x_1, x_1, ..., x_n\}$ is closed. With distinct elements, no point is an accumulation point, thus all are contained.

Theorem 1.2. *The following hold in (X, d).*

(a) An arbitrary number of unions of open sets is open.

(b) A finite number of intersections of open sets is open.

(c) A finite number of unions of closed sets is closed.

(d) An arbitrary number of intersections of closed sets is closed.

Example 1.10. Let $A_n = (-\frac{1}{n}, \frac{1}{n})$. Each A_n is open. However, $A = \bigcap_{n=1}^{\infty} A_n = \{0\}$ which is not open.

Example 1.11. Let $A_n = \{\frac{1}{n}\}$. Each A_n is closed. However, $A = \bigcup_{n=1}^{\infty} A_n = \{\frac{1}{n} : n = 1, 2, ..\}$, which has accumulation point 0, and is thus not closed.

2 September 10, 2019

2.1 Real analysis review, cont'd

Definition 2.1 (Open cover)**.** An open cover of set A is a (possibly infinite) collection of open sets $\mathcal{S}$ such that $A \subseteq \bigcup_{O \in \mathcal{S}} O$. A subcover is a subset of $\mathcal{S}$ that is still a cover for A

Definition 2.2 (Compact set)**.** The following are equivalent statements.

(a) A set A is compact.

(b) Every open cover of A has a finite subcover.

(c) Every sequence $x_1, x_2, ...$ with $x_n \in A$ has subsequence that converges to $x \in A$.

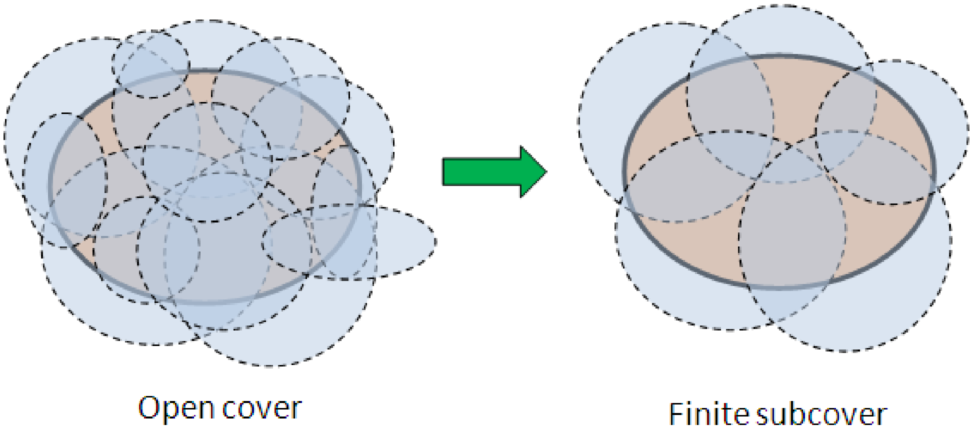

Figure 3: The brown ellipse represents a set, and the blue ellipses represent an open cover. Note that compactness does not require the existence of an open cover, but for every given (even infinite) open cover, one can extract a finite subcover.

These definitions can be difficult to verify. In $\mathbb{R}^n$, there is a simpler characterization.

Definition 2.3 (Boundedness)**.** A set A is bounded if there exists a point $c \in X$ with finite radius r such that $A \subseteq B_r(c)$. In other words, A is bounded if an open ball can contain it fully.

Theorem 2.1 (Heine-Borel)**.** *$A \in \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

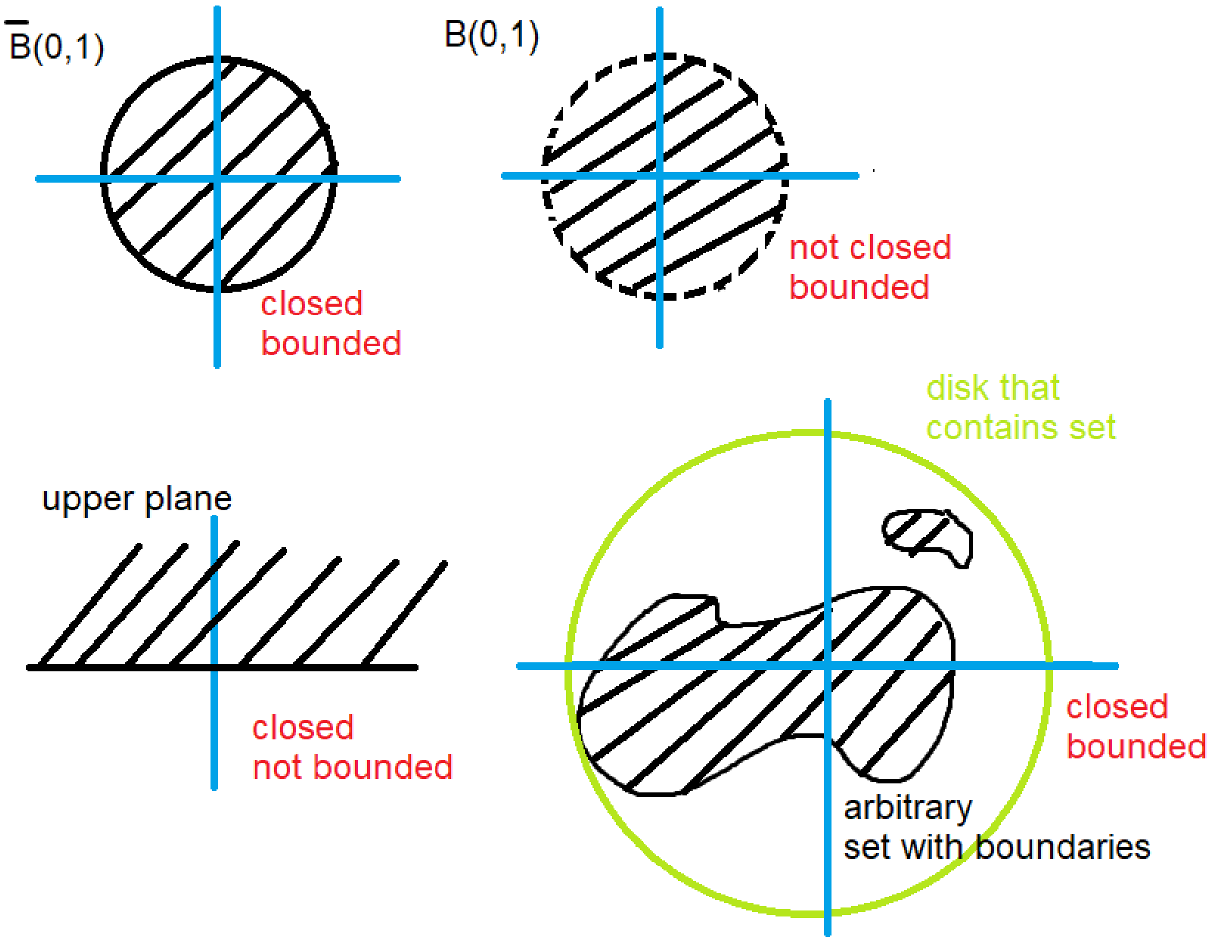

Figure 4: The shaded regions represent elements of the set, with solid and dotted boundaries denoting inclusion and exclusion, respectively. The top-left and bottom-right are compact sets in $\mathbb{R}^2$.

Example 2.1. In $\mathbb{R}$, a closed interval $[a, b]$ is compact.

Example 2.2. In $\mathbb{R}^n$, the unit sphere $A = \{x : ||x||_2 = 1\}$ is compact (where $||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}$).

Definition 2.4 (Continuity)**.** Let $f : (X, d_X) \to (Y, d_Y)$. f is continuous at x_0 if for every $\epsilon > 0$, there is a $\delta(\epsilon, x_0)$ such that $\forall x \in X$:

$$d(x, x_0) < \delta \implies d(f(x), f(x_0)) < \epsilon$$

f is continuous if it is continuous at all $x_0 \in X$.

The definition of continuity captures the notion that small changes in x should result in small changes in $f(x)$. Specifically, the change in $f(x)$ should be made arbitrarily small by controlling the change in x. Note that δ depends on both ϵ and x_0.

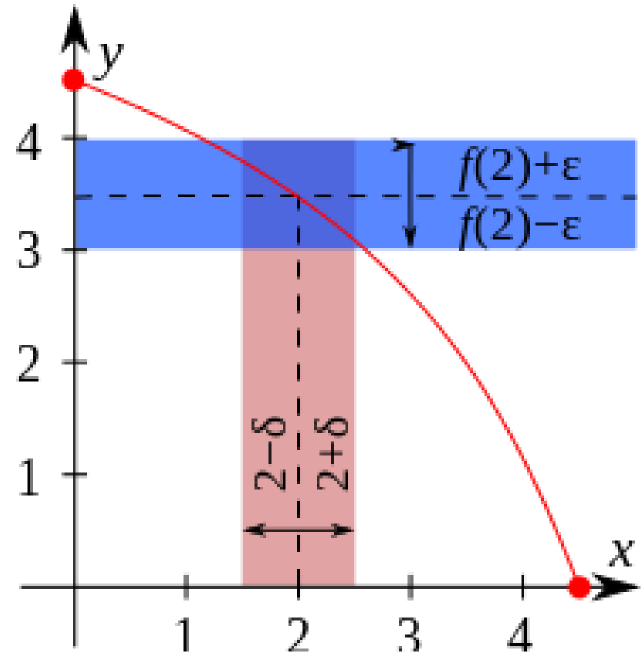

Figure 5: This function is continuous at $x = 2$.

Definition 2.5 (Uniform continuity)**.** Let $f : (X, d_X) \to (Y, d_Y)$. f is uniformly continuous if for every $\epsilon > 0$, there is a $\delta(\epsilon)$ such that $\forall x_0, x_1 \in X$:

$$d(x_0, x_1) < \delta \implies d(f(x_0), f(x_1)) < \epsilon$$

Note that the dependence of δ on the point in X is not gone.

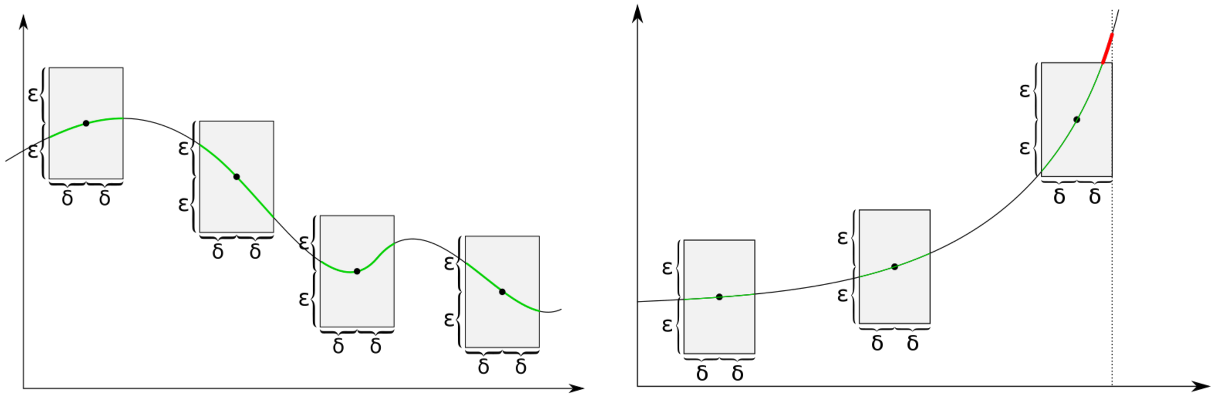

Figure 6: A pictorial characterization of continuity whether a ring of height 2ϵ and width 2δ could slide along the entire function, without turning, for every ϵ. If δ must change as this happens, the the function is only continuous. In the right example, δ must get smaller as x gets large in order for the ring to continue sliding. In the case that δ can stay constant, as in the left example, then the function is uniformly continuous.

3 September 17, 2019

In response to questions from last time:

Exercise 3.1. Show that compactness implies closure in a metric space.

Example 3.1. Show that compactness implies boundedness in a metric space.

Proof. Let (X, d) be the metrix space, and A be the compact set of interest. Chose any $x_0 \in X$, and write $\mathcal{S} = \{B_r(x_0) : r > 0\}$. Clearly, $A \subset \mathcal{S}$, and $\mathcal{S}$ is an open cover of A. Thus, there exists a finite subcover

$$F = \{B_{r_1}(x_0), ..., B_{r_p}(x_0)\}$$

Take $r = \max_{i=1,...,p} r_i$, and $A \subseteq B_r(x_0)$. □

Exercise 3.2. Given an example of a metric space that is closed and bounded, but not compact. Hint: Use the discrete metric. That is,

$$d(x, y) = \begin{cases} 0 \text{ if } x = y \\ 1 \text{ if } x \neq y \end{cases}$$

Example 3.2. Prove that $f(x) = \frac{1}{x}$ on $x > 0$ is not uniformly continuous.

Proof. Let $\epsilon = 1$. For any δ, we must choose x_0 and x_1 such that $|x_0 - x_1| < \delta$ does imply that $|f(x_0) - f(x_1)| < 1$.

$$|f(x_0) - f(x_1)| = \frac{\delta}{x_0 x_1}$$

Letting $x_1 = x_0 + \frac{\delta}{2}$.

$$|f(x_0) - f(x_1)| = \frac{\delta}{x_0(x_0 + \frac{\delta}{2})}$$

Choosing x_0 small enough can make this quantity larger than $\epsilon = 1$. □

Continuing with the review, there are many useful properties that result from a continuous function on a compact set.

Theorem 3.1. *Let $f : (X, d_X) \to (\mathbb{R}, d_Y)$ be continuous over compact set X. Then:*

(a) f is uniformly continuous.

(b) $f(X) = \{f(x) : x \in X\}$ is compact.

(c) f achieves $\max_{x \in X}\{f(x)\}$ and $\min_{x \in X}\{f(x)\}$, as in there exists $x^*_{max}, x^*_{min} \in X$ such that

$$f(x^*_{min}) \leq f(x) \leq f(x^*_{max})$$

for all x.

(d) Let $\lim_{n\to\infty} x_n = x$, where $x_n, x \in X$. Then

$$\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right) = f(x)$$

Example 3.3. Let $f : (X, d_X) \to (Y, d_Y)$ and $g : (Y, d_Y) \to (Z, d_Z)$ both be continuous functions. Show that the composition $f \circ g$ is continuous.

Proof. Given any $x_0 \in X$ and any $\epsilon > 0$, let $y_0 = f(x_0)$. Choose $\delta_g(\epsilon, y_0)$ such that:

$$d_Y(y_0, y) < \delta_g \implies d_Z(g(y_0), g(y)) < \epsilon$$

Then choose $\delta_f(\delta_g, x_0)$ such that:

$$d_X(x_0, x) < \delta_f \implies d_Y(f(x_0), f(x)) < \delta_g$$

Thus

$$d_X(x_0, x) < \delta_f \implies d_Y(f(x_0), f(x)) < \delta_g \implies d_Z(g(f(x_0)), g(f(x))) < \epsilon$$

□

Exercise 3.3. Give a function that is bounded, i.e. there is some $B \geq 0$ such that $|f(x)| \leq B$ for all $x \in X$, and continuous, but not uniformly continuous.

Definition 3.1 (Lipshitz continuity)**.** A function $f : (X, d_X) \to (Y, d_Y)$ is called Lipshitz continuous with Lipshitz constant L for all $x_0, x_1 \in X$:

$$d_Y(f(x_0), f(x_1)) \leq L \cdot d_X(x_0, x_1)$$

This means that the changes in $f(x)$ are sublinear in the changes in x. The constant L "quantifies" the continuity of f.

Observation 3.1. *Lipshitz continuity implies uniform continuity.*

Proof. Given any $e > 0$, let $\delta = \frac{\epsilon}{L}$.

$$d_Y(f(x_0), f(x_1)) \leq L \cdot d_X(x_0, x_1) \leq L \cdot \frac{\epsilon}{L} = \epsilon$$

□

Pointers for Homework 1:

(a) Similarity implies sameness of rank, spectrum/charactaristic polynomial and determinant. None of the reverse implications hold.

(b) Same eigenvectors does not imply similarity. What does it imply (assuming there are n of them that are linearly independent)?

(c) What can you say about a matrix with distinct eigenvalues?

(d) What can you say about the eigenvalues of diagonal/triangular matrix?

(e) A matrix is invertible - what can you say about its eigenvalues?

4 September 24, 2019

Homework 1 solutions.

5 October 1, 2019

Homework 2 solutions.

6 October 8, 2019

Copies of Homework 3 solutions distributed.

When I say "application", I mean that these topics are not tested in the course, but topics mostly from mathematical data science that are interesting and make use of material from the course.

Application: Projection Matrices

$A^2 = A \in M_n(\mathbb{R})$. What can you say about this matrix? Given the additional information that $A = A^T$, what else can be said?

The matrix has annihilating polynomial $p(t) = t(t-1)$, therefore any eigenvalues $\lambda \in \{0, 1\}$. q_A divides this polynomial, so q_A is the product of distinct linear factors, implying that A is diagonalizable. In fact $Ax = SDS^{-1}x$ is a projection of x onto the subspace spanned by the eigenvectors of A associated with eigenvalue 1. This happens in three steps.

(a) $y_1 = S^{-1}x$ gives the coordinates of x in the basis described by the columns of A.

(b) $y_2 = Dy_1$ scales the components by 0 or 1, eliminating certain dimensions and keeping other.

(c) $y_3 = Sy_2$ brings us back into the original basis.

If $A = A^T$, then it is real orthogonally diagonalizable, with the eigenvectors forming an orthonormal basis. This idea is depicted in Figure 7. The transformation A is called a projection matrix. The next topic will make use of ideas from Chapters 0 through 4.

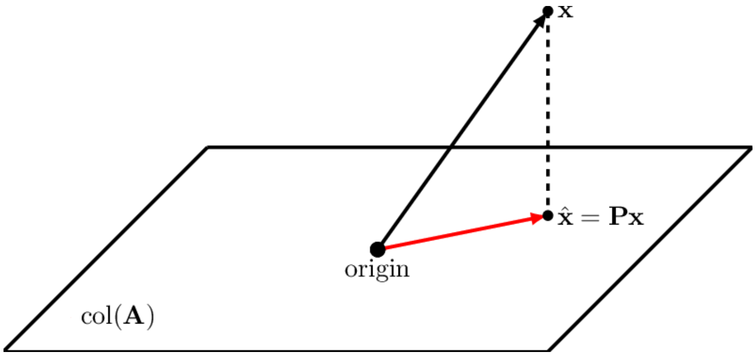

Figure 7: Here $\mathbf{x} \in \mathbb{R}^3$ is the vector of interest and $\mathbf{P} \in M_3(\mathbb{R})$ is the projection matrix. If we $\mathbf{P} = SDS^{-1}$ let $\mathbf{A}$ be the $M_{3,2}(\mathbb{R})$ matrix that column binds the eigenvectors of $\mathbf{P}$ associated to eigenvalue 1, then the subspace column space of $\mathbf{A}$.

Application: numerical optimization

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function, with

$$\nabla f(x) =: g(x) \in \mathbb{R}^d$$
$$\nabla^2 f(x) =: H(x) \in M_d(\mathbb{R})$$

Definition 6.1 (Local minimum)**.** x^* is a local minimum of f if there exists an $\epsilon > 0$ such that for all $x \in B_\epsilon(x)$, $f(x^*) \le f(x)$.

Our goal is to find such a local minimum, and we write this problem as

$$\min_{x \in \mathbb{R}^d} f(x)$$

There are many ways to go about this, but one such was is to iteratively propose candidate solutions $x_1, x_2, ...$ that get "closer" to a local minimum x^*. A subset of these methods is line-search methods, where at iterate x_t, we choose a search direction $p \in \mathbb{R}^d$ and step size $\alpha \in [0, 1)$, and let $x_{t+1} = x_t + \alpha p$. For example, in gradient descent, we let $p = -g(x_t)$, as in Figure 8.

Some basic optimality conditions:

$$x^* \text{ is a local minimum.} \implies g(x^*) = 0 \text{ and } H(x) \text{ is P.S.D.}$$
$$g(x^*) = 0 \text{ and } H(x) \text{ is P.D.} \implies x^* \text{ is a local minimum.}$$

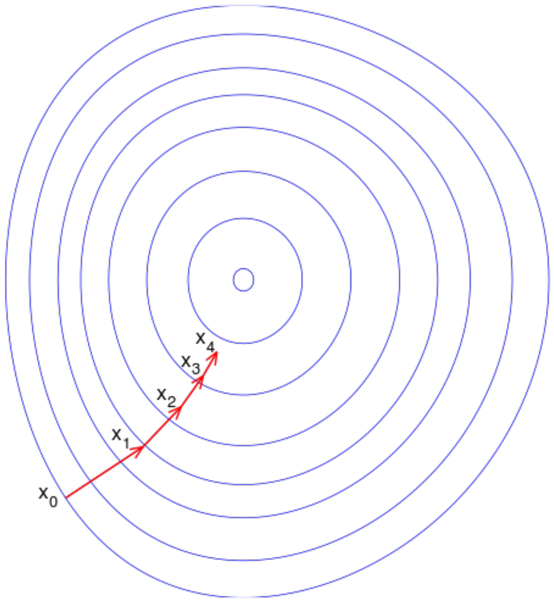

Figure 8: The center of the contour plot is the local minimum.

While we will not prove them, we will attempt to understand them via the Hessian matrix $H(x^*)$. Taylor expand $f(x_0 + \alpha p)$ about x_0.

$$f(x_0 + \alpha p) = f(x_0) + \alpha g(x_0)^T p + \frac{1}{2}\alpha^2 p^T H(x_0) p + o(\alpha^2)$$

Assume that we are at a stationary point x_0, i.e. $g(x_0) = 0$. Note that $H = H^T$ for any x, thus H is real orthogonally diagonalizable. Let $u_1, ..., u_d$ be the eigenvectors of $H(x_0)$. Start by considering $p = u_k$.

$$\begin{aligned} f(x_0 + \alpha u_k) &= f(x_0) + \alpha g(x_0)^T u_k + \frac{1}{2}\alpha^2 u_k^T H(x_0) u_k + o(\alpha^2) \\ \implies f(x_0 + \alpha u_k) - f(x_0) &= \frac{1}{2}\alpha^2 \lambda_k + o(\alpha^2) \end{aligned}$$

where λ_k is the eigenvalue associated to u_k. It's clear that for small α, if λ_k is positive, then the function will increase, where as for negative λ_k, the function will decrease. For $\lambda_k = 0$, the higher-order terms determine the sign of the change in function value.

Now consider arbitrary search direction p. Because $H = H^T$, the eigenvectors of H for an othonormal basis for $\mathbb{R}^d$, and we can write

$$\begin{aligned} p &= UU^T p \\ &= (u_1^T p) u_1 + ... + (u_d^T p) u_d \\ &= \sum_{k=1}^{d} (u_k^T p) u_k \end{aligned}$$

Then, the the Taylor expansion gives us

$$f(x_0 + \alpha u_k) - f(x_0) = \frac{1}{2}\alpha^2 \sum_{k=1}^{d} (u_k^T p)^2 \lambda_k + o(\alpha^2)$$

Thus, the change in the function is a weighted sum of the eigenvalues of $H(x_0)$, weighted by how parallel the search direction is with the associated eigenvector. If all $\lambda_k > 0$ (i.e. $H(x_0)$ is P.D.) then there is nothing to worry about, and every direction will increase the function value for small enough α. This is just another way of saying that we are at a local minimum. However, any negative eigenvalue reveals a descent direction, and if an eigenvalue is 0, then the higher-order term $o(\alpha^2)$ could still decrease the function (which is why $H(x_0)$ P.S.D. is not sufficient).

Finally, we answer the following question: what is the interpretation of D? Assume that $f(x)$ is a quadratic form, in that $f(x) = b + g^T x + \frac{1}{2}x^T Hx$ for $b \in \mathbb{R}$, $x, g \in \mathbb{R}^d$, and $H \in M_d(\mathbb{R})$ symmetric (why can we make H symmetric without loss of generality?). Let $f_U(x) = f(U^T x)$.

$$\begin{aligned} \nabla^2 f_U(x) &= U^T \nabla^2 f(U^T x) U \\ &= U^T H U \\ &= D \end{aligned}$$

For a quadratic function, letting $H = UDU^T$, D is the Hessian of the function evaluated in a rotated basis.

Another important point is that eigenvalues of large magnitude have eigenvectors which are direction of steep increase or decrease, as suggested by the Taylor expansion. We will come back to this point when we discuss condition number, specifically how different ratios of eigenvalues affects the success of optimization algorithms.

As a final message: in applications related to mathematical data science, when we see (say square) matrices, if you gain anything from this class it will be the willingness and ability to answer the following two questions.

(a) If the matrix is symmetric (Hermitian, normal), what is the interpretation of its eigenvalues and eigenvectors? What is the interpretation of the diagonal matrix D?

(b) Is the matrix full-rank, low-rank, or approximately low-rank? What are the implications of each case? What is its condition number, and what are its implications?

By approximately low-rank, we mean eigenvalues (or more generally, singular values) small in magnitude. The second question we will be able to attack more after Chapter 7 and 5, but the first we can start understanding now.

7 October 15, 2019

Announcements

Exam 1 just passed, and I unfortunately do not have the answer to the following questions.

- I got grade x on Exam 1, will I still be able to get grade y in the class?
- Are Exams 2 and 3 easier than Exam 1?

The first question I will direct to Dr. Fishkind, while for the second, the best answer I can offer is that I looked at the exam scores from last years, and students performed statistically significantly better on Exam 2 and 3 than on Exam 1. However, operationally, the average was 4 points higher on Exam 2 and 8 points higher on Exam 3, so take that as you will. It is likely that students' expectations are better calibrated for future exams, rather than them being easier.

Some chapter 4 results

Theorem 7.1. *Let $A \in M_n$ be normal. Let $F(A) = \{\frac{x^*Ax}{x^*x} : x \in \mathbb{C}^n \backslash \{0\}\}$. Then $F(A) = \mathcal{H}(A)$, where $\mathcal{H}(\cdot)$ denotes the convex hull.*

Theorem 7.2 (Rayleigh-Ritz)**.** *Let $A \in M_n$ be Hermitian, with (orthonormal) eigenvectors $u_1, ..., u_n$, associated with $\lambda_1 \leq ... \leq \lambda_n$. For $k = 1, ..., n$, we have:*

$$\min_{\substack{x \neq 0 \\ x \perp u_1, ..., u_{k-1}}} \frac{x^*Ax}{x^*x} = \lambda_k$$
$$\max_{\substack{x \neq 0 \\ x \perp u_{k+1}, ..., u_n}} \frac{x^*Ax}{x^*x} = \lambda_k$$

Proof. We'll prove the maximum case. Represent any such x as

$$x = \delta_1 u_1 + ... + \delta_k u_k = \sum_{i=1}^k \delta_i u_i$$

with not all δ_i equal to 0. Let $A = UDU^*$ be the unitary diagonalization.

$$\begin{aligned}
\max_{\substack{x \neq 0 \\ x \perp u_{k+1}, ..., x_n}} \frac{x^*Ax}{x^*x} &= \max_{\delta_1, ..., \delta_k} \frac{(\sum_{i=1}^k \delta_i u_i)^* UDU^* \sum_{i=1}^k \delta_i u_i}{(\sum_{i=1}^k \delta_i)^* (\sum_{i=1}^k \delta_i)} \\
&= \max_{\delta_1, ..., \delta_k} \frac{(\sum_{i=1}^k \overline{\delta_i} u_i^* U) D (\sum_{i=1}^k \delta_i U^* u_i)}{\sum_{i=1}^k \overline{\delta_i} \delta_i} \\
&= \max_{\delta_1, ..., \delta_k} \frac{\begin{bmatrix} \overline{\delta_1} & \dots & \overline{\delta_k} & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_k & & & \\ & & & \lambda_{k+1} & & \\ & & & & \ddots & \\ & & & & & \lambda_n \end{bmatrix} \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}}{\sum_{j=1}^k |\delta_j|^2} \\
&= \max_{\delta_1, ..., \delta_k} \sum_{i=1}^k \frac{|\delta_i|^2}{\sum_{j=1}^k |\delta_j|^2} \lambda_i \\
&= \lambda_k
\end{aligned}$$

□

Theorem 7.3 (Courant-Fischer). *Let $A \in M_n$ be Hermitian. Then*

$$\max_{y_1,\dots,y_{k-1}} \phi_{\min}(y_1,...,y_{k-1}) = \max_{y_1,\dots,y_{k-1}} \min_{\substack{x\neq 0 \\ x\perp y_1,\dots,y_{k-1}}} \frac{x^*Ax}{x^*x} = \lambda_k$$

$$\min_{y_{k+1},\dots,y_n} \phi_{\max}(y_{k+1},...,y_n) = \min_{y_{k+1},\dots,y_n} \max_{\substack{x\neq 0 \\ x\perp y_{k+1},\dots,y_n}} \frac{x^*Ax}{x^*x} = \lambda_k$$

Proof. We show the first equality. Noting that $y_1,...,y_{k-1}$ are not necessarily linearly independent, and letting $u_1,...,u_k,...,u_n$ be the eigenvectors of $\lambda_1 \leq ... \leq \lambda_k \leq ... \leq \lambda_n$, we have

$$\begin{aligned} &\dim \text{span}\{y_1,...,y_{k-1}\} \leq k-1 \\ \implies &\dim \text{span}\{y_1,...,y_{k-1}\}^\perp \geq n-(k-1) = n-k+1 \\ &\dim \text{span}\{u_1,...,u_k\} = k \end{aligned}$$

Thus

$$\dim \text{span}\{y_1,...,y_{k-1}\}^\perp + \dim \text{span}\{u_1,...,u_k\} \geq n+1$$

So the intersection span$\{y_1,...,y_{k-1}\}^\perp \cap$span$\{u_1,...,u_k\}$ must have some nonzero vectors. Let $w \neq 0$ be one such vector. We have

$$w \perp u_{k+1},...,u_n,$$

and so by Rayleigh-Ritz,

$$\frac{w^*Aw}{w^*w} \leq \lambda_k \implies \min_{\substack{w\neq 0 \\ w\perp y_1,\dots,y_{k-1} \\ w\perp u_{k+1},\dots,u_n}} \frac{w^*Aw}{w^*w} \leq \lambda_k$$

This also means that

$$\phi_{\min}(y_1,...,y_{k-1}) = \min_{\substack{x\neq 0 \\ x\perp y_1,\dots,y_{k-1}}} \frac{x^*Ax}{x^*x} \leq \min_{\substack{w\neq 0 \\ w\perp y_1,\dots,y_{k-1} \\ w\perp u_{k+1},\dots,u_n}} \frac{w^*Aw}{w^*w} \leq \lambda_k$$

and thus we have found an upper bound for $\phi_{\min}(y_1,...,y_{k-1})$. Setting $y_1 = u_1,...,y_{k-1} = u_{k-1}$, we achieve this upper bound, and thus

$$\max_{y_1,\dots,y_{k-1}} \phi_{\min}(y_1,...,y_{k-1}) = \max_{y_1,\dots,y_{k-1}} \min_{\substack{x\neq 0 \\ x\perp y_1,\dots,y_{k-1}}} \frac{x^*Ax}{x^*x} = \lambda_k$$

as desired. As for the second equality (which is entirely analogous), we have

$$\begin{aligned} &\dim \text{span}\{y_{k+1},...,y_n\} \leq n-k \\ \implies &\dim \text{span}\{y_{k+1},...,y_n\}^\perp \geq n-(n-k) = k \\ &\dim \text{span}\{u_k,...,u_n\} = n-k+1 \end{aligned}$$

Thus

$$\dim \operatorname{span}\{y_{k+1}, ..., y_n\}^{\perp} + \dim \operatorname{span}\{u_k, ..., u_n\} \geq n+1$$

So the intersection $\operatorname{span}\{y_{k+1}, ..., y_n\}^{\perp} \cap \operatorname{span}\{u_k, ..., u_n\}$ must have some nonzero vectors. Let $w \neq 0$ be one such vector. We have

$$w \perp u_1, ..., u_{k-1},$$

and so by Rayleigh-Ritz,

$$\frac{w^*Aw}{w^*w} \geq \lambda_k \implies \max_{\substack{w \neq 0 \\ w \perp y_{k+1}, ..., y_n \\ w \perp u_1, ..., u_{k-1}}} \frac{w^*Aw}{w^*w} \geq \lambda_k$$

This also means that

$$\phi_{\max}(y_{k+1}, ..., y_n) = \max_{\substack{x \neq 0 \\ x \perp y_{k+1}, ..., y_n}} \frac{x^*Ax}{x^*x} \geq \max_{\substack{w \neq 0 \\ w \perp y_{k+1}, ..., y_n \\ w \perp u_1, ..., u_{k-1}}} \frac{w^*Aw}{w^*w} \geq \lambda_k$$

and thus we have found a lower bound for $\phi_{\max}(y_{k+1}, ..., y_n)$. Setting $y_{k+1} = u_{k+1}, ..., y_n = u_n$, we achieve this lower bound, and thus

$$\min_{y_{k+1}, ..., y_n} \phi_{\max}(y_{k+1}, ..., y_n) = \min_{y_{k+1}, ..., y_n} \max_{\substack{x \neq 0 \\ x \perp y_{k+1}, ..., y_n}} \frac{x^*Ax}{x^*x} = \lambda_k$$

as desired. □

Theorem 7.4 (Weyl). *Let $A, B \in M_n$ be Hermitian. Then, for all $k = 1, ..., n$*

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_n(B)$$

where $\lambda_i(A)$ is the i-th smallest eigenvalue of A.

Corollary 7.1. *Let $A, B \in M_n$ be P.S.D.. Then for all $k = 1, ..., n$,*

$$\lambda_k(A) \leq \lambda_k(A+B)$$

Remark 7.1. For $a \in \{-1, 1\}, y \in \mathbb{C}^n \backslash \{0\}$, ayy^* is a rank 1 Hermitian matrix. Any rank 1 Hermitian matrix can be written as such.

Theorem 7.5 (Interlacing I). *For A Hermitian, $a \in \{-1, 1\}, y \in \mathbb{C}^n \backslash \{0\}$, then for all k,*

$$\lambda_k(A + ayy^*) \leq \lambda_{k+1}(A)$$

Theorem 7.6 (Interlacing II). *Let $A \in M_n$ be Hermitian, $B \in M_r$ be a principal submatrix of A. Then for all k such that $1 \leq k \leq r$,*

$$\lambda_k(A) \leq \lambda_k(B) \leq \lambda_{k+n-r}(A)$$

Corollary 7.2. *If $A \in M_n$ is Hermitian, and a_{ii} is on the diagonal, then*

$$\lambda_1(A) \leq a_{ii} \leq \lambda_n$$

Application: multivariate statistics

In the interest of time, and the fact that there are many resources on this subject, we may not go over the statistics example fully in class. However, here is the setup and the questions, and you can answer them on your own and discuss in office hours!

Let $x : \Omega \to \mathbb{R}^d$ and $x \in \mathbb{R}^d$ both represent a d-dimensional random variable and its realization. The distinction should be clear from context. Let $\mathbb{E}[x] = \mu$ be the mean and $\text{Cov}[x] = \mathbb{E}[(x - \mu)(x - \mu)^T] = \Sigma$ be the (Hermitian) covariance matrix.

(a) Prove that Σ is P.S.D.. Why is it not P.D. in general?

(b) Interpret the eigenvectors, eigenvalues, and diagonal matrix D.

(c) Let $x \sim \mathcal{N}(\mu, \Sigma)$. Prove that Σ is in fact P.D.. Plot the probability density function for $d = 2$ and use the optimization example to interpret the shape.

- Why does a large eigenvalue λ_k correspond to a large spread in u_k (meaning large value of $\text{Var}(||\text{proj}_{u_k}(x)||)$)?
- For what values of the spectrum of Σ are the level sets spherical versus oblong?
- When are the major axes of the level sets aligned with the coordinate axes?

(d) If $x \sim \mathcal{N}(\mu, \Sigma)$ and $\Sigma = UDU^T$, what is the distribution of $z = U^T x$? Interpret z.

(e) Let $X \in M_{n,d}(\mathbb{R})$ be a data matrix, where each row x_i^T is an independent observation of x. Assume for simplicity that $\mathbb{E}[x] = 0$. Finally, let $\frac{1}{n}X^T X = \hat{U}\hat{D}\hat{U}^T$ be an estimate of the unitary diagonalization of Σ. Consider two dimension reduction routines.

- Represent the data as $Z = X\hat{U}$, i.e. the principal components of X. Then, drop $d - r$ columns by some method.
- Drop columns from the *original* data matrix X to X_r, then orthogonalize the reduced covariance matrix $\frac{1}{n}X_r^T X_r = \hat{U}_r \hat{D}_r \hat{U}_r^T$. Represent the reduced data as $Z_r = X_r \hat{U}_r$.

What is the conceptual difference between these two methods, that is, orthogonalizing and then dropping columns (of which PCA is an example) or vice versa? (Hint: use Interlacing II.)

8 October 22, 2019

Exercises in section

(a) What can be said about the diagonal elements of a skew Hermitian matrix?

(b) Let $A \in M_n$ be Hermitian. Prove that $A = A_+ + A_-$, where

- A_+ is P.S.D. and A_- is N.S.D.,
- $\text{rank}(A) = \text{rank}(A_+) + \text{rank}(A_-)$, and
- $A_+ A_- = A_+ A_- = 0$.

(c) Prove that

- if A is Hermitian, then A^2 is P.S.D.
- if A is skew Hermitian, then A^2 is N.S.D.

(d) Let $\mathcal{H}_n$ be the Hermitian matrices in M_n. $A \succeq B$ for $A, B \in \mathcal{H}_n$ if $A - B$ is P.S.D.. Prove that this is a partial ordering on $\mathcal{H}_n$, in that

- $A \succeq A$,
- $A \succeq B$ and $B \preceq A \iff A = B$, and
- $A \succeq B$ and $B \succeq C \implies A \succeq C$.

Why is $\mathcal{H}_n$ not totally ordered?

(e) Let $A \in M_n$ be Hermitian with exactly one positive and one negative eigenvalue. Explain why $\lambda_2(A) \geq 0$ and $\lambda_{n-1}(A) \leq 0$ with equality if and only if $n > 2$.

(f) Prove that the inner minimization and maximization of Courant-Fishcher always exists. That is, for any $A \in M_n$, $y_1, ..., y_m$ with $m < n$,

$$\inf_{x \neq 0,\ x \perp y_1, ..., y_m} \frac{x^* A x}{x^* x} \quad \text{and} \quad \sup_{x \neq 0,\ x \perp y_1, ..., y_m} \frac{x^* A x}{x^* x}$$

are achieved and can be replaced by minimum and maximum, respectively. You can take for granted that the matrix $z^* A z$ is continuous in z, with respect to the metric $d(z_0, z_1) = ||z_0 - z_1||$ where $||\cdot||$ is the Euclidean norm. We will be able to show this rigorously in Chapter 5.)

We'll give a proof for the infimum in part (f), as to start getting used to the real analysis style proofs.

Proof. First we note that

$$\frac{x^* A x}{x^* x} = \frac{x^* A x}{||x||^2} = \left(\frac{x}{||x||} \right)^* A \left(\frac{x}{||x||} \right)$$

$\frac{x}{||x||}$ is a unit vector, so we can just search values of $z^* A z$ over unit vectors z. Similarly, $x \perp y_1, ..., y_m$ if and only if $\frac{x}{||x||} \perp y_1, ..., y_m$, as scaling the length does not affect orthogonality (for non-zero vectors). Thus,

$$\inf_{x \neq 0,\ x \perp y_1, ..., y_m} \frac{x^* A x}{x^* x} = \inf_{||z||_2 = 1,\ z \perp y_1, ..., y_m} z^* A z$$

Using the hint that $z^* A z$ is continuous, we can show the existence of the infimum if the search space is compact. It is clear that the unit sphere $\{z : ||z|| = 1\}$ is compact. On the other hand, the set $\{z : z \perp y_1, ..., y_m\}$ is a subspace of dimension $l \leq m$. The search space is the intersection of these two sets. See Figure 9 for an example. The subspace is represented by a hyperplane that crosses the origin. The intersection of an l-dimensional subspace and an n-dimensional unit sphere is an l-dimensional unit sphere. This is still a compact set, completing the proof that the inf (and sup) exist, which can now be replaced by min (and max). □

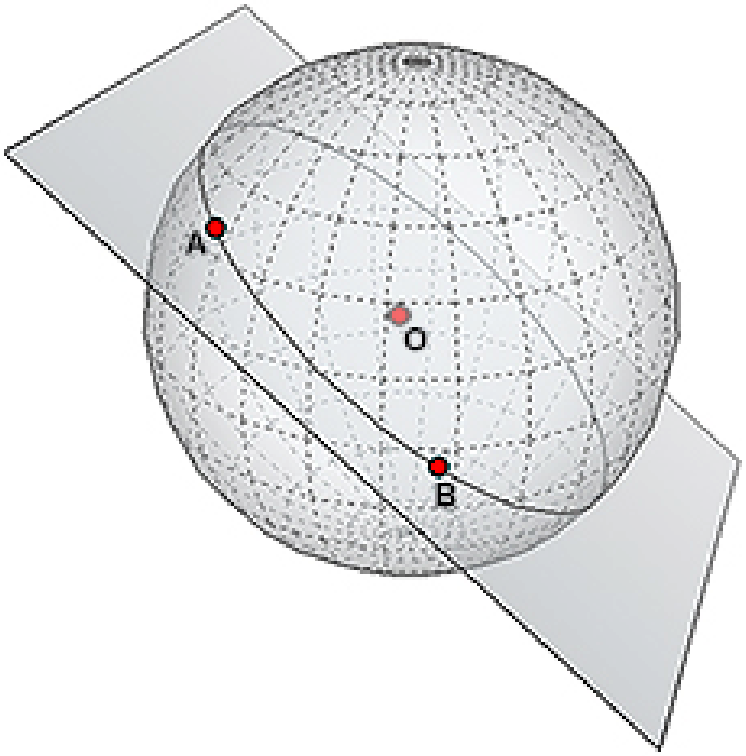

Figure 9: Let O be the origin, and y (not shown) be a vector normal to the plane. Then, A and B are two points in the intersection. They are both of unit length, and both orthogonal to y.

Pointers for Homework 4:

(a) What relationship can you advance about the set of Hermitian matrices and the set of skew Hermitian matrices?

(b) If $x \in \mathbb{R}^n$ majorizes $y \in \mathbb{R}^n$, what can you say about x_1 and y_1 (by definition) and about x_n and y_n (by inspection)?

(c) What can you say about any commuting matrices? If they are normal?

(d) Remember that in constructing principle submatrices, the undeleted rows/columns need not be contiguous. That is, for $n = 10$, I can delete rows/columns 1 - 3 and 5 - 7. This generates a submatrix of rows/columns 4, 8, 9, and 10.

(e) The leading principle submatrices $A_1, A_2, ..., A_n$ of A are by definition principle submatrices of A. What else is true about them?

(f) For Problem 7, you should have a clear and specific analogue for Rayleigh-Ritz, Courant-Fischer, Interlacing I and II (and consequences), Weyl's theorem, and the majorization results.

9 October 29, 2019

The theme of this section is understanding the singular value decomposition (SVD) of common data matrices. In statistics and machine learning, the data matrix (or design matrix) $X \in M_{n,d}(\mathbb{R})$ has **rows** $x^{(1)}, ..., x^{(n)} \in \mathbb{R}^d$ where each $x^{(i)}$ is an observation of some random vector

$$x : \Omega \to \mathbb{R}^d$$

with mean

$$\mu = \mathbb{E}[x] \in \mathbb{R}^d$$

and covariance matrix

$$\Sigma = \text{Cov}(x) = \mathbb{E}[(x-\mu)(x-\mu)^T] = \mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x^T]$$

This is analogous to the variance formula for univariate x, i.e. $\text{Var}(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$. We will just be thinking about the real version of these objects, but for complex numbers the analysis is the same, replacing transpose with conjugate transpose. Another data matrix we will visit is a grayscale image $P \in M_{m,n}(\mathbb{R})$.

We first will compare the SVD and eigendecomposition of an arbitrary Hermitian matrix A. If $A = UDU^*$ and $A^* = UD^*U = UDU^*$, then

$$AA^* = UDU^*UD^* = UD^2U^*$$

meaning that A and AA^* share the same eigenvectors, with AA^* having the same eigenvalues, but squared. The "U" matrix of the eigendecomposition of AA^* is the same as the "U" of the SVD of A, by definition.

$$A = U\Sigma V^*$$

What about Σ? Also by definition, we have that

$$\Sigma = (D^2)^{\frac{1}{2}} = |D| = D \cdot \text{sgn}(D)$$

where the $|\cdot|$ denotes element-wise absolute value, and $\text{sgn}(\cdot)$ is the element-wise sign function.

$$\text{sgn}(t) = \begin{cases} +1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0 \end{cases}$$

Thus we can write

$$UDU^* = A = U\Sigma V^* = UD \cdot \text{sgn}(D)V^*$$

If A is invertible, then D is invertible, and we can conclude by cancelling out UD that

$$U^* = \text{sgn}(D)V^*$$

Thus, the rows of V^* are just equal to the rows of U^*, down to a sign. Where A has a positive eigenvalue λ_i, the rows u_i^* and v_i^* will agree. Where A has a negative eigenvalue λ_i, then $u_i^* = -v_i^*$. If D is not invertible, then there will be zeros in $\text{sgn}(D)$, but if that is the case, then it doesn't really matter what is in those rows of U^* and V^* (as long as they are orthonormal), as those rows will get wiped out by the zero singular values in Σ. So we can interpret this idea as "essentially", U^* and V^* will agree down to signs of rows. Additionally, if A is positive semidefinite (P.S.D.), then we can say that the SVD and eigendecomposition are the **same** (we will omit the "essentially" from now on).

Coming back to the random variable x, will first show that the covariance matrix Σ is P.S.D.. Give any $a \neq 0 \in \mathbb{R}^d$, we have:

$$\begin{aligned} a^T \Sigma a &= a^T(\mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x^T])a \\ &= \mathbb{E}[a^T x x^T a] - \mathbb{E}[a^T x]\mathbb{E}[x^T a] \\ &= \mathbb{E}[(a^T x)^2] - (\mathbb{E}[a^T x])^2 \\ &= \text{Var}(a^T x) \end{aligned}$$

This value is nonnegative for any a, completing the proof. We cannot say more than that, because $a^T x$ can be constant even when each coordinate of x has variance. Take for example

$$x = \begin{bmatrix} y \\ 1-y \end{bmatrix}$$

where $y \sim \text{Unif}(0, 1)$. Letting $a = [1\ 1]^T$, we have that $a^T x = 1$ with probability 1, i.e. $\text{Var}(a^T x) = 0$. So if the covariance matrix was P.D., what would that imply? It means that no dimensions of x are linearly dependent. For those taking Statistical Theory I with Carey Priebe, this is the concept behind a rank d exponential family.

Similarly, if we expand to the full data matrix X, every column of X contains independent copies of observations from each dimension. If Σ is P.D., then no column of X can be represented as a linear combination of the others. This reveals a goal - if the data matrix did have columns that were expressible as linear combinations of the others, we might want to drop those columns, reducing the dimension of our dataset, which might have benefits for statistical inference. This can be done harmlessly if Σ is not P.D. and has some of its eigenvalues (hence singular values) set to zero. What if there is an eigenvalue λ_j of Σ such that

$$\lambda_j = \epsilon \approx 0$$

for some small $\epsilon > 0$? This means that there is a dimension of x (hence a column of X) that is **approximately** a linear combination of the others. We still might be able to benefit statistically by dropping such a column from out dataset. As it turns out, in any dataset, we can have columns highly correlated with one another. This means that the non-diagonal entries of Σ can be far from zero. The next question is whether there exists a basis to represent the dimensions of the data are uncorrelated. We know that Σ is P.S.D., so we can write

$$\Sigma = UDU^T$$

What is the interpretation of D?

$$\begin{aligned} D &= U^T \Sigma U \\ &= U^T(\mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x^T])U \\ &= \mathbb{E}[U^T x(U^T x)^T] - \mathbb{E}[U^T x]\mathbb{E}[(U^T x)^T] \\ &= \text{Cov}(U^T x) \end{aligned}$$

The random vector $U^T x$ is just x in a rotated basis, but in this basis, all of the dimensions of $U^T x$ are uncorrelated! ($U^T x$ is a transform of x into the basis of $u_1, ..., u_d$, the columns of U, because U is real orthogonal.) Now, we can consider the random variable $z = U^T x$ with linearly independent coordinates. The j-th diagonal of D is λ_j, we have

$$\lambda_j = \text{Var}(z_j) = \text{Var}(u_j^T x)$$

because

$$z = U^T x = \begin{bmatrix} u_1^T x \\ \vdots \\ u_d^T x \end{bmatrix}$$

We can also write

$$x = Uz = z_1 u_1 + ... z_d u_d = (u_1^T x)u_1 + ... + (u_d^T x)u_d = \sum_{j=1}^{d} (u_j^T x)u_j$$

as the orthogonal projection of x onto the orthonormal basis $\{u_1, ..., u_d\}$. We are interested in recovering the matrix $Z = XU \in M_{n,d}(\mathbb{R})$, the orthogonalized representation of X. Other than the benefit of being able to observe the components of x that are uncorrelated, we can drop columns that have low values of λ_j, as they may not describe the main patterns in the data. This linear dimension reduction technique is known as principle component analysis (PCA), and can be read about in detail here.

We generally do not have access to the true distribution generating x, and therefore do not have Σ on hand. Let's assume that $\mathbb{E}[x] = 0$. (We can achieve this by subtracting the sample mean from each point.) Then, a natural estimate of the covariance matrix is

$$\hat{\Sigma} = \frac{1}{n} X^T X$$

As an exercise, prove that this estimate converges in probability to Σ. That is, prove that for all ϵ ¿ 0

$$P[||\hat{\Sigma} - \Sigma||_F > \epsilon] \to 0$$

as $n \to \infty$. You can use the Weak Law of Large Numbers. Similarly, show trivially that the eigenvalue vector $\vec{\lambda}(\hat{\Sigma})$ converges in probability to the true eigenvalue vector $\vec{\lambda}(\Sigma)$. (You can use that the eigenvalues are continuous in the matrix.) Finally, check that $\hat{\Sigma}$ is P.S.D. as well. Thus, $\hat{\Sigma}$ can be eigendecomposed.

$$\hat{\Sigma} = \hat{U}\hat{D}\hat{U}^T$$

and we can estimate the *loading* matrix

$$\hat{Z} = X\hat{U}$$

From here, columns can be dropped by any method. While this checks out theoretically, from a numerical viewpoint, there are some shortcomings. Both the computation of $\hat{\Sigma}$ as well as its eigendecomposition are notoriously expensive operations. Additionally, eigendecomposition is less numerically stable than singular value decomposition. (You should believe that taking the eigendecomposition of AA^* is not how the SVD of A is typically computed by linear algebra libraries.) Is there a way to compute $\hat{Z}$ without either of the above steps? The answer is in the SVD. Consider the SVD of the matrix below.

$$\tfrac{1}{\sqrt{n}}X^T = \hat{U}\hat{S}\hat{V}^T$$

This matrix is chosen so that "$\hat{U}$" is the same "$\hat{U}$" that we discussed before - an estimate of the eigenvectors of Σ, as

$$\tfrac{1}{\sqrt{n}}X^T\left(\tfrac{1}{\sqrt{n}}X^T\right)^T = \tfrac{1}{n}X^TX = \hat{U}\hat{D}\hat{U}^T$$

and this the $\hat{U}$ used in the SVD of $\frac{1}{\sqrt{n}}X^T$. Then we can take the transpose of the SVD $\frac{1}{\sqrt{n}}X^T$, which is the SVD of $\frac{1}{\sqrt{n}}X$.

$$\tfrac{1}{\sqrt{n}}X = \hat{V}\hat{S}^T\hat{U}^T$$

and multiply both sized by $\hat{U}$.

$$\tfrac{1}{\sqrt{n}}X\hat{U} = \hat{V}\hat{S}^T\hat{U}^T\hat{U} = \hat{V}\hat{S}^T$$

Thus, we have

$$\hat{Z} = X\hat{U} = \sqrt{n}\hat{V}\hat{S}^T$$

The orthogonalized $\hat{Z}$ can be computed fully by the SVD of X! You may have not seen the $\sqrt{n}$ factor before, but this is only written so that the M_d-valued matrix in the SVD of X is the same as the eigenvector matrix of $\hat{\Sigma}$. Multiplying the entire dataset by a number will usually not affect the result of statistical inference.

Departing from the statistical example to a very simple image analysis example, we can consider an m-by$n$ grayscale image P. If we compute its SVD, rather than writing in matrix form, we can write P as the sum of scaled rank 1 matrices.

$$P = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T$$

where u_i and v_i are the i-th columns of U and V, respectively, and σ_i is the i-th "diagonal" of Σ. Thus, P is generated by overlaying many rank 1 layers. It is conceivable that the largest singular values correspond to layers containing main effects such as objects and shapes, moderate singular values contain details such as texture, and near-zero singular values correspond to noise layers. The image can be written in a compressed, low-rank format with virtually no loss of discernable information. At this site, you can drag a slider that determines the number of singular values to retain in various images. For many of the images, hundreds of singular values can be dropped without much change in the image. These are just two among the many applications of the SVD.

10 November 5, 2019

Exam 2 is on Friday, November 8, 2019. An enumeration of the topics in Chapter 4 can be found at the end of the seventh lecture module (Hadamard's Inequality). As for Chapter 7, the main topics covered were:

- Construction of the SVD.
- Properties and consequences of the SVD.
  - Range, rank, nullspace from SVD.
  - Uniqueness of singular values, not of singular vectors.
  - $||A||_{2,2} = \sigma_1(A)$
- Polar decomposition.
- Generalized inverse properties.
- Existence and uniqueness of Moore-Pensrose generalized inverse.
- Generalized inverses and linear systems.

Copies of Homework 4 and 5 solutions distributed.