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.

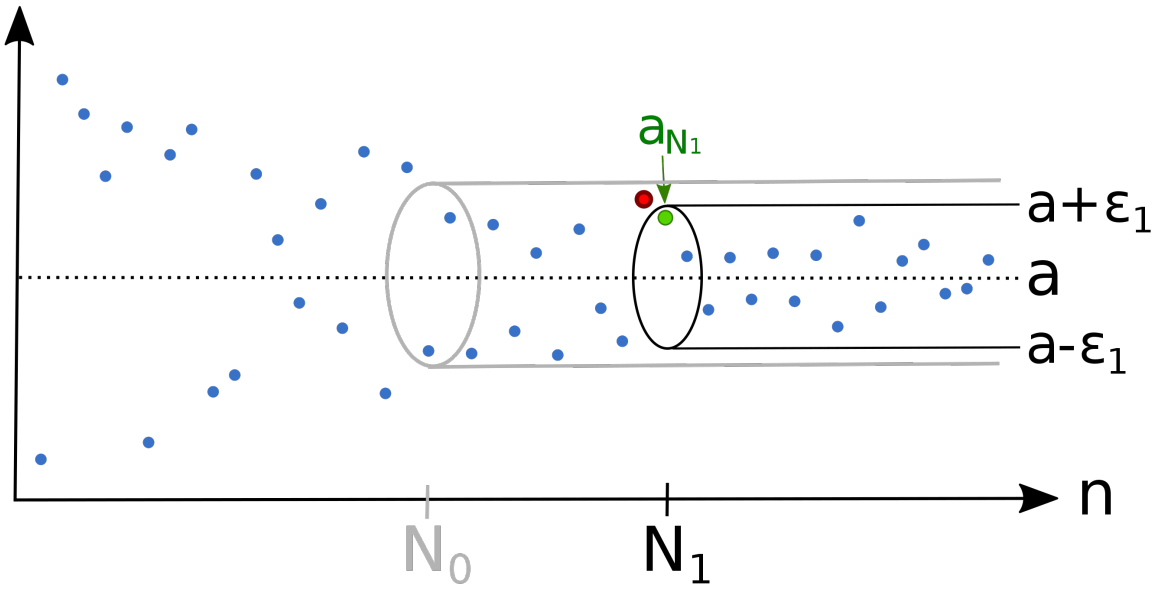

Figure 1: a denotes the limit in this sequence, N_0 denotes $N(\epsilon_0)$ for some ϵ_0, and $N_1 = N(\epsilon_1) < N_0$ for some $\epsilon_1 < \epsilon_0$.

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.

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

□

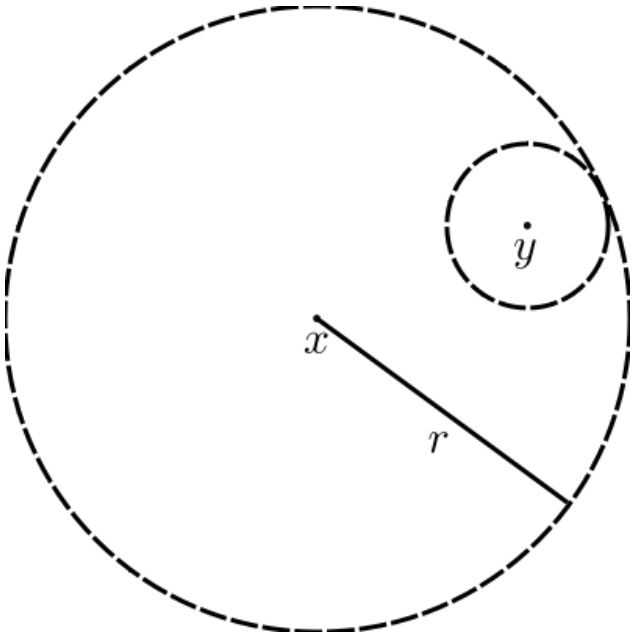

Figure 2: An open r-ball about x. Dotted lines typically denote the unincluded boundary of the set.

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 $x \in A$.

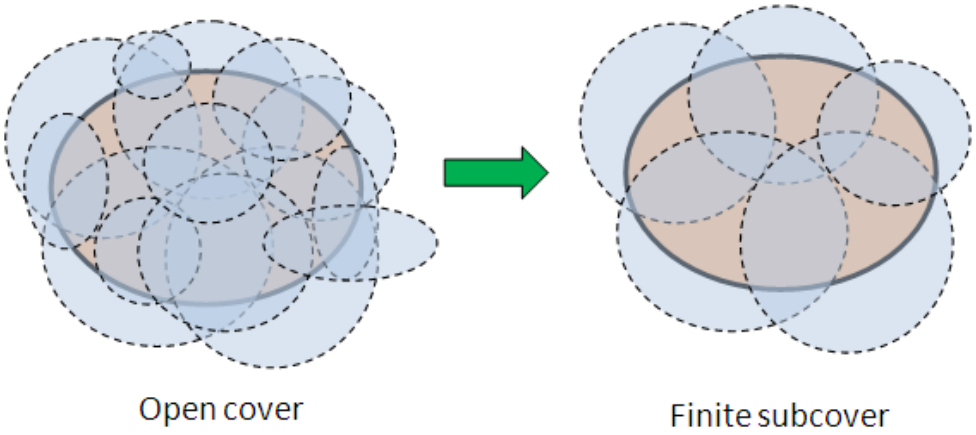

Figure 3: The brown ellipse represents a set, and the blue ellipses represent an open cover. Note that compactness does not require the existence of an open cover, but for every given (even infinite) open cover, one can extract a finite subcover.

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.*

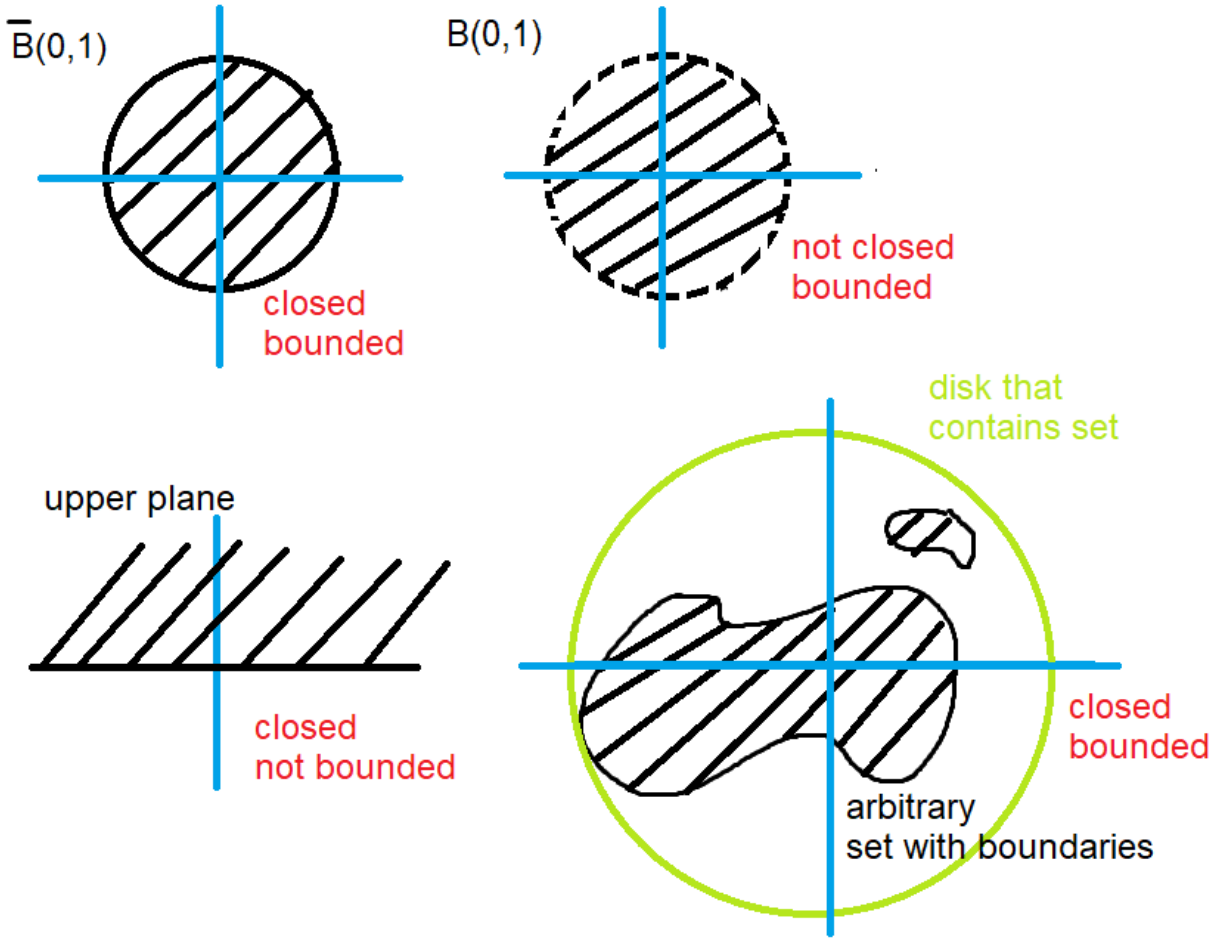

Figure 4: The shaded regions represent elements of the set, with solid and dotted boundaries denoting inclusion and exclusion, respectively. The top-left and bottom-right are compact sets in $\mathbb{R}^2$.

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.

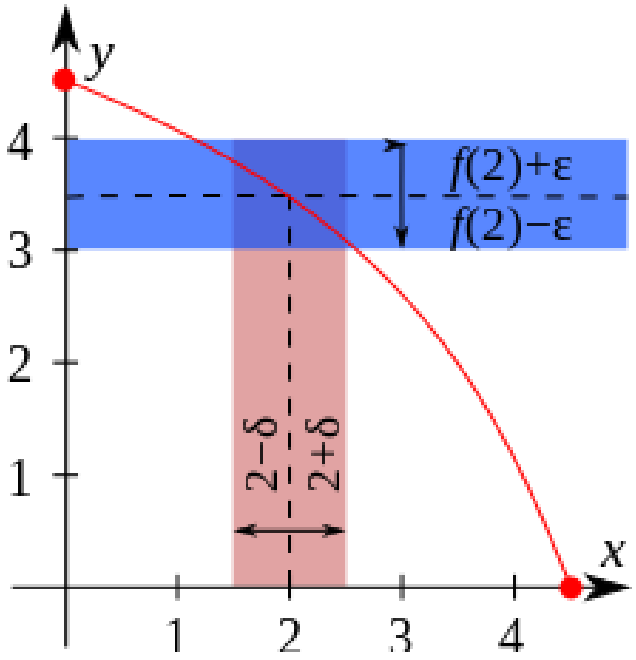

Figure 5: This function is continuous at $x = 2$.

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.

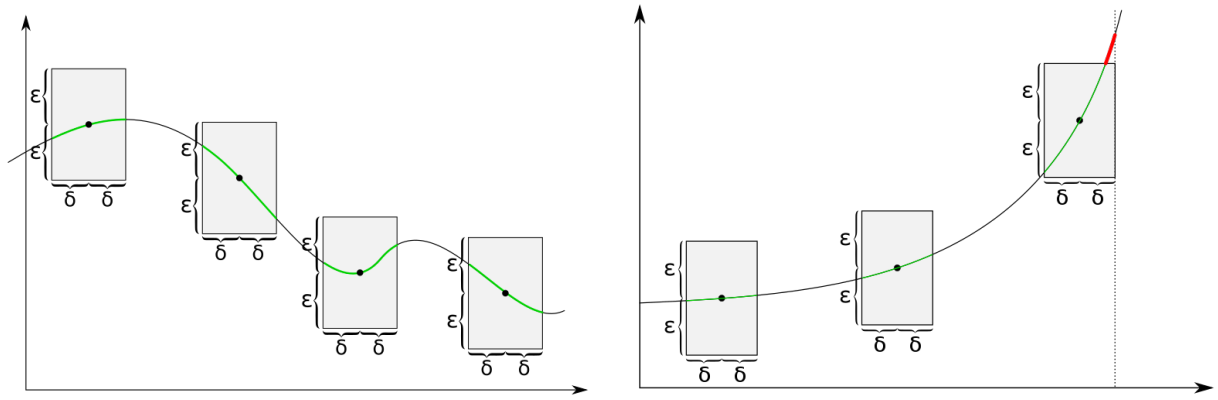

Figure 6: A pictorial characterization of continuity whether a ring of height 2ϵ and width 2δ could slide along the entire function, without turning, for every ϵ. If δ must change as this happens, the the function is only continuous. In the right example, δ must get smaller as x gets large in order for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

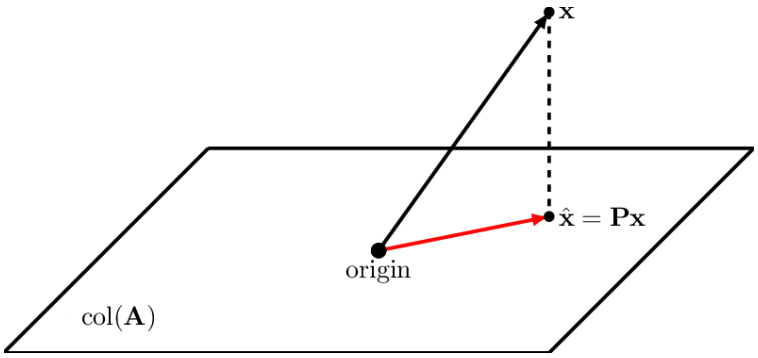

Figure 7: Here $\mathbf{x} \in \mathbb{R}^3$ is the vector of interest and $\mathbf{P} \in M_3(\mathbb{R})$ is the projection matrix. If we $\mathbf{P} = SDS^{-1}$ let $\mathbf{A}$ be the $M_{3,2}(\mathbb{R})$ matrix that column binds the eigenvectors of $\mathbf{P}$ associated to eigenvalue 1, then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

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:

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

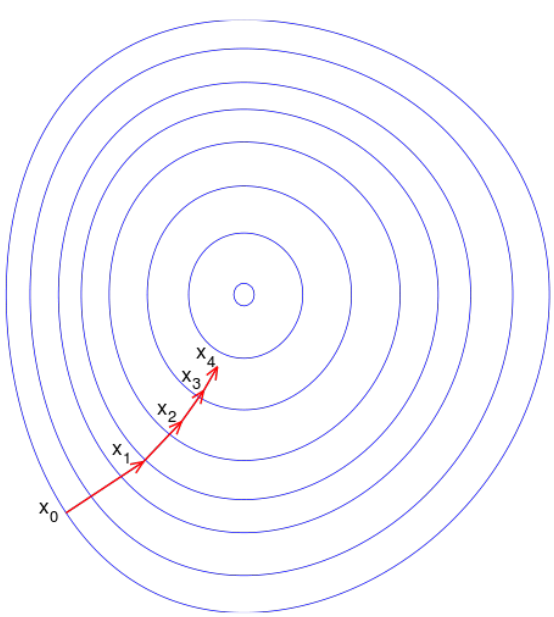

Figure 8: The center of the contour plot is the local minimum.

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 $p = u_k$.

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

$$\implies f(x_0 + \alpha u_k) - f(x_0) = \frac{1}{2}\alpha^2 \lambda_k + o(\alpha^2)$$

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 $f(x) = b + g^T x + \frac{1}{2} x^T H x$ 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.

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

□

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

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

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

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 $\text{span}\{y_1,...,y_{k-1}\}^\perp \cap \text{span}\{u_1,...,u_k\}$ 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,

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

This also means that

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

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

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

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

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

So the intersection $\text{span}\{y_{k+1}, ..., y_n\}^\perp \cap \text{span}\{u_k, ..., u_n\}$ 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 $X^T X = \hat{U}\hat{D}\hat{U}^T$ 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 $X_r^T X_r = \hat{U}_r \hat{D}_r \hat{U}_r^T$. 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.)