ML Theory — Homework 0

your NetID here

Version 1

Instructions.

- Everyone must submit an individual write-up.
- This is a **calibration** homework; please work alone and don't hunt for solutions (except in problem 2).
- Homework is due **Monday, September 3, at 3:30pm**; no late homework accepted.
- Please consider using the provided LaTeX file as a template.

1. **(Ordinary least squares and the SVD.)**

   (a) Let $A \in \mathbb{R}^{d\times k}$ be orthonormal with $k < d$, let S denote the span of its columns, and let $B \in \mathbb{R}^{d\times(d-k)}$ be any orthonormal matrix with column span $S^\perp$. Show via direct calculation that every $x \in \mathbb{R}^d$ satisfies

   $$AA^\top x + BB^\top x = x.$$

   (b) Show by direct calculation that every pair $x \in \mathbb{R}^d$ and $w \in \mathbb{R}^k$ satisfy

   $$\|x - AA^\top x\| \le \|x - Aw\|.$$

   Remark. Consequently, the *orthogonal projection* operation $x \mapsto A^\top x$ provides solutions to $\min\left\{\|x - Aw\| : w \in \mathbb{R}^k\right\}$.

   (c) Now let $C \in \mathbb{R}^{d\times k}$ with $k < d$ be a general matrix (not necessarily orthonormal), and let's focus on the *ordinary least squares* problem introduced in the preceding remark, namely

   $$\min_{w\in\mathbb{R}^k} \|x - Cw\|^2.$$

   By differentiating and setting to zero, show that satisfying the *normal equations*

   $$C^\top Cw = C^\top x$$

   is necessary and sufficient for a vector $w \in \mathbb{R}^k$ to be a critical point.

   Remark/bonus. If you feel like it, establish that the Hessian is positive semi-definite, whereby the normal equations are equivalent to global optimality.

   Remark. When C is orthonormal, $C^\top C = I$, meaning the orthogonal projection $w = C^\top x$ from earlier parts satisfies the normal equations and thus can be defined via minimization.

   (d) With an orthonormal matrix A, we had the easy least squares solution $A^\top x$; let's see if we can get something similar in the general case of $C \in \mathbb{R}^{d\times k}$ with $k < d$ as before.

   Let $C = USV^\top$ denote the SVD of C; a sufficient definition for the purposes of this problem is as follows. The matrices $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{k\times k}$ are orthornomal, whereas $S \in \mathbb{R}^{d\times k}$ is zero off the diagonal and along the diagonal has the (nonnegative) reals (the *singular values*) appearing in nonincreasing order; the number of positive singular values in S equals the rank of the matrix C. This decomposition is not unique in general, however it always exists.

   Define the pseudoinverse $S^\dagger \in \mathbb{R}^{k\times d}$ of S by transposing S and inverting the positive entries, and the pseudoinverse $C^\dagger$ of C as $C^\dagger = VS^\dagger U^\top$.

   Show firstly that $C^\dagger x$ is globally optimal for the least squares problem above (take the earlier global optimality remark for granted if you didn't prove it). **Note.** Since $k < d$, the pseudoinverse is not an inverse; make sure your derivation does not imply $CC^\dagger = I$!

   Secondly, show that $C^\dagger = C^\top$ when C is orthonormal, whereby we've recovered the earlier reasoning.

   Note. There is no need to form a Lagrangian (or do any other complicated thing).

Solution.

(Your solution here.)

2. **(Random walks on the integers.)**

Consider a sequence of iid random variables $(X_1, \ldots, X_n)$ where $\Pr[X_i = +1] = \Pr[X_i = -1] = 1/2$ for all $i \in \{1, \ldots, n\}$, and additionally let $S_j := \sum_{i \le j} X_i$ denote their sum.

Remark. While solving homework problems in this class generally should not involve any side references, this problem is partly asking you to look things up! The reason is to build intuition for some of the essential statistical phenomena; it's okay if this is your first exposure.

(a) Use Chebyshev's inequality to fill in the blanks in the following: for any $\delta \in (0, 1]$,

$$\Pr\left[|S_n| < \square\right] \ge 1 - \delta;$$

show your work and give an explicit expression (as small as possible with Chebyshev) for $\square$.
Remark. Statements of this type are common in learning theory, and are stated as "with probability at least $1 - \delta$, $|S_n| < \square$".

(b) Now use Hoeffding's inequality to derive a tighter form of the preceding. Please cite your (favorite) resource for Hoeffding.

Remark. $1 - \delta$ is the "confidence"; we want the bound to scale very mildly with $1/\delta$. The form we get out of Hoeffding will be very useful.

(c) Look up and restate in your own words *the law of the iterated logarithm (LIL)* (give your citation, *not* to wikipedia). You can make your statement mathematical, or you can use plain english, it's up to you.

After that, say something about the random walk S_n which is revealed by the LIL but not captured in the preceding bounds. Ideally, state it mathematically or with a picture.

(d) Provide a plot which demonstrates the above behavior. Specifically, choose a large N (2^{20} should suffice to see LIL), and plot the random walk $(S_1, S_2, S_3, \ldots, S_N)$, along with some curves representing the above bounds. Feel free to be creative, this is homework 0 after all...

Solution.

(Your solution here.)

3. **(Counting oscillations.)**

Given a set of reals $S \subseteq \mathbb{R}$, let $\mathrm{CC}(S)$ denote the *connected components* of S (the sets here will be well-behaved, so don't stress the definition too much; it suffices for our purposes to say C is a connected component of S if (a) $C \subseteq S$, (b) C is an interval, (c) C is not strictly contained within any other interval which is also a subset of S). For example,

$$\mathrm{CC}\left(\left\{x \in \mathbb{R} : x^2 - 1 = 0\right\}\right) = \left\{\{-1\}, \{+1\}\right\}, \qquad \mathrm{CC}\left(\mathbb{R}\right) = \{\mathbb{R}\}.$$

Given any continuous function f, define

$$\mathrm{Osc}(f) := \sup_{a,b \in \mathbb{R}} \left|\mathrm{CC}\left(\left\{x \in \mathbb{R} : f(x) = ax + b\right\}\right)\right|;$$

roughly speaking, $\mathrm{Osc}(f)$ counts the number of times f can intersect any affine function, and is therefore a way to measure oscillations. For example,

$$\mathrm{Osc}(x \mapsto x^2 - 1) = 2, \qquad \mathrm{Osc}(x \mapsto x - 1) = 1.$$

(a) Show that any polynomial f of degree k has $\mathrm{Osc}(f) \leq \max\{1, k\}$.

(b) Show that $g(x) = x^2 - \cos(x)$ has $\mathrm{Osc}(g) = 2$.

(c) As an immediate consequence of the previous parts: show that there exist f, g with $\mathrm{Osc}(f) = 2$ and $\mathrm{Osc}(g) = 2$, but $\mathrm{Osc}(f + g) = \infty$.

(d) Show that if f has $\mathrm{Osc}(f) < \infty$, then there exists x with $|f(x) - \sin(x)| \geq 1$.

(e) Show that for every $\epsilon > 0$, there exists g with $\mathrm{Osc}(g) = \infty$ and f with $\mathrm{Osc}(f) < \infty$ but $\sup_{x \in \mathbb{R}} |g(x) - f(x)| \leq \epsilon$.

Remark. This isn't hard, but together with the preceding part it aims to stress that oscillation counting isn't enough when separating functions.

Solution.

(Your solution here.)