CS 189 Introduction to Machine Learning
Fall 2018 HW1

This homework is due **Monday, September 3rd at 10 p.m.**

2 The accuracy of learning decision boundaries

This problem exercises your basic probability (e.g. from 70) in the context of understanding why lots of training data helps to improve the accuracy of learning things.

For each $\theta \in (1/3, 2/3)$, define $f_\theta : [0,1] \to \{0,1\}$, such that

$$f_\theta(x) = \begin{cases} 1 \text{ if } x > \theta \\ 0 \text{ otherwise.} \end{cases}$$

The function is plotted in Figure 1.

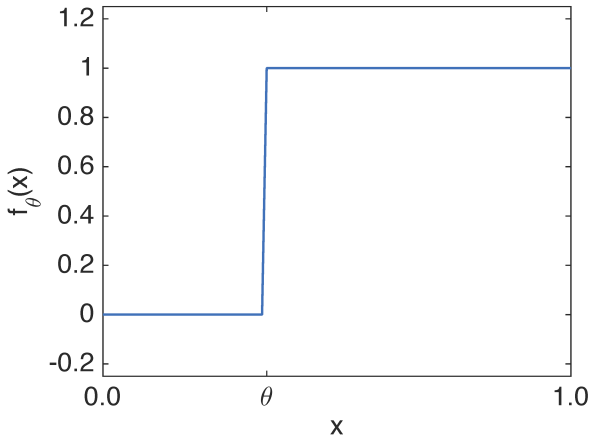

Figure 1: Plot of function $f_\theta(x)$ against x.

We draw samples $X_1, X_2, \ldots, X_n$ uniformly at random and i.i.d. from the interval $[0,1]$. Our goal is to learn an estimate for θ from n random samples $(X_1, f_\theta(X_1)), (X_2, f_\theta(X_2)), \ldots, (X_n, f_\theta(X_n))$.

Let $T_{min} = \max(\{\frac{1}{3}\} \cup \{X_i | f_\theta(X_i) = 0\})$. We know that the true θ must be larger than T_{min}.

Let $T_{max} = \min(\{\frac{2}{3}\} \cup \{X_i | f_\theta(X_i) = 1\})$. We know that the true θ must be smaller than T_{max}.

The gap between T_{min} and T_{max} represents the uncertainty we will have about the true θ given the training data that we have received.

(a) **What is the probability that $T_{max} - \theta > \epsilon$ as a function of ϵ? And what is the probability that $\theta - T_{min} > \epsilon$ as a function of ϵ?**

HW1, ©UCB CS 189, Fall 2018. All Rights Reserved. This may not be publicly shared without explicit permission.

Solution: First note that when $\theta + \epsilon > \frac{2}{3}$ we have that $\mathbf{P}(T_{max} > \theta + \epsilon) \leq \mathbf{P}(T_{max} > \frac{2}{3}) = 0$ by definition of T_{max}.

For the case when $\theta + \epsilon < \frac{2}{3}$, the task is to find the probability of the event of the random variable T_{max} defined by $\mathcal{E} := \{T_{max} > \theta + \epsilon\}$. Note that because $\mathbf{P}(\{T_{max} < \theta\}) = 0$ or equivalently $\{T_{max} < \theta\} = \emptyset$, the probability $\mathbf{P}(\mathcal{E})$ can alternatively be expressed by the probability of a different event $\mathcal{E}_0 = \{\theta \leq T_{max} \leq \theta + \epsilon\}$ in terms of

$$\begin{aligned}\mathbf{P}(\mathcal{E}) &= \mathbf{P}(\{T_{max} > \theta + \epsilon\} \cup \{T_{max} < \theta\}) \\ &= 1 - \mathbf{P}(\{\theta \leq T_{max} \leq \theta + \epsilon\}) = 1 - \mathbf{P}(\mathcal{E}_0).\end{aligned}$$

Now consider the event $\mathcal{E}_1 := \{\text{at least one } X_i \text{ lies in } [\theta, \theta + \epsilon]\}$. You can now show that $\mathcal{E}_0 = \mathcal{E}_1$, i.e.

$$\{\theta \leq T_{max} \leq \theta + \epsilon\} = \{\text{at least one } X_i \text{ lies in } [\theta, \theta + \epsilon]\}.$$

by definition of T_{max}.

Going back to the original event $\mathcal{E}$ we thus find

$$\mathbf{P}(\mathcal{E}) = 1 - \mathbf{P}(\mathcal{E}_0) = 1 - \mathbf{P}(\mathcal{E}_1) = \mathbf{P}(\mathcal{E}_1^c)$$

where the complement $\mathcal{E}_1^c = \{\text{no } X_i \text{ lies in } [\theta, \theta + \epsilon]\} = \bigcap_{i=1}^n \{X_i \notin [\theta, \theta + \epsilon]\}$.

Restating the probability of $\mathcal{E}$ in terms of an intersection of events on X_i now allows us to easily find $\mathbf{P}(\mathcal{E})$ because of independence and uniform distribution of X_i, which reads

$$\mathbf{P}(\bigcap_{i=1}^n \{X_i \notin [\theta, \theta + \epsilon]\}) = \prod_{i=1}^n \mathbf{P}(\{X_i \notin [\theta, \theta + \epsilon]\}) = (1 - \epsilon)^n.$$

In summary, we obtain

$$\mathbf{P}(T_{max} - \theta > \epsilon) = \begin{cases} (1 - \epsilon)^n & \theta + \epsilon < \frac{2}{3} \\ 0 & \text{o.w.} \end{cases}$$

Similar analysis applies to the second part, except our lower bound is $\frac{1}{3}$:

$$\mathbf{P}(\theta - T_{min} > \epsilon) = \begin{cases} (1 - \epsilon)^n & \theta - \epsilon > \frac{1}{3} \\ 0 & \text{o.w.} \end{cases}$$

(b) Suppose that you would like the estimator $\hat{\theta} = (T_{max} + T_{min})/2$ for θ that is ϵ-close (defined as $|\hat{\theta} - \theta| < \epsilon$, where $\hat{\theta}$ is the estimation and θ is the true value) with probability at least $1 - \delta$. Both ϵ and δ are some small positive numbers. **Please bound or estimate how big of an n do you need?** You do not need to find the optimal lowest sample complexity n, an approximation using results of question (a) is fine.

Solution: One way to obtain $\hat{\theta}$ within a window of size 2ϵ is to have both T_{max} and T_{min} be within ϵ of θ. To see this, define random variables $L = \theta - T_{min}, U = T_{max} - \theta$. When $L < \epsilon$

and $U < \epsilon$, we have $\theta - T_{min} < \epsilon$ and $0 < T_{max} - \theta$. Adding the two inequalities, we have $\theta - T_{min} < T_{max} - \theta + \epsilon$, thus $\hat{\theta} - \theta > -\epsilon/2 > -\epsilon$. Similarly, with those conditions, we have $\hat{\theta} - \theta < \epsilon$. Thus $L < \epsilon$ and $U < \epsilon$ is a sufficient condition for $\hat{\theta}$ to be ϵ-close, that is

$$\mathbf{P}(|\hat{\theta} - \theta| < \epsilon) \geq \mathbf{P}(\{L < \epsilon\} \cap \{U < \epsilon\}).$$

Instead of lower bounding $\mathbf{P}(\{L < \epsilon\} \cap \{U < \epsilon\})$ we upper bound the probability of the complement of the event, which reads $\{L > \epsilon\} \cup \{U > \epsilon\}$, via union bound as follows:

$$\mathbf{P}(\{L > \epsilon\} \cup \{U > \epsilon\}) \leq \mathbf{P}(\{L > \epsilon\}) + \mathbf{P}(\{U > \epsilon\}) \leq 2(1-\epsilon)^n$$

using the result in problem (a).

We must ensure that this probability is upper bounded by δ, which ensures that we succeed with probability at least $1 - \delta$. Solving for n, we have

$$2(1-\epsilon)^n < \delta$$
$$n > \frac{\ln\left(\frac{2}{\delta}\right)}{\ln\left(1/(1-\epsilon)\right)}.$$

Again, using the approximation $\ln(1-x) \sim -x$, we have $n > \frac{1}{\epsilon}\ln\left(2/\delta\right)$ for ϵ small.

(c) Let us say that instead of getting random samples $(X_i, f(X_i))$, we were allowed to choose where to sample the function, but you had to choose all the places you were going to sample in advance. **Propose a method to estimate θ. How many samples suffice to achieve an estimate that is ϵ-close as above?** (**Hint:** You need not use a randomized strategy.)

Solution: Pick n points uniformly spaced on the interval $(\frac{1}{3}, \frac{2}{3})$. Then, the ith sample $X_i = \frac{1}{3} + \frac{i}{3n}$. Since we have n points, we create intervals of length $\frac{1}{3n}$. If our intervals are smaller than 2ϵ, we can guarantee that we estimate θ within an interval of 2ϵ. Solving for n, we have $\frac{1}{3n} < 2\epsilon$ and so $n > \frac{1}{6\epsilon}$ samples are sufficient.

Note that using our calculations the sample complexity for this deterministic method is *always lower* than the sample complexity of the probabilistic method in problem (b) $\delta < 1$ since $\ln\left(\frac{2}{\delta}\right) > \frac{1}{6}$ for any $\delta < 1$. Therefore, uniform sampling and allowing for some non-zero probability that we do not obtain an ϵ-close estimator, does not require fewer samples than a deterministic method which always ensures an ϵ-close estimator. In many other settings however, allowing some uncertainty (of finding a good estimator) can help to reduce the sample complexity significantly.

(d) Suppose that you could pick where to sample the function adaptively — choosing where to sample the function in response to what the answers were previously. **Propose a method to estimate θ. How many samples suffice to achieve an estimate that is ϵ-close as above?**

Solution: Use binary search: start with three pointers, $s = 1/3, e = 2/3$ with m as the midpoint. If $f(m) = 0$, set $s = m$ and recompute the midpoint (i.e., search over the second

half of the range). Otherwise, $f(m) = 1$ and set $e = m$ (i.e., search over the first half of the range). For each point sampled, we reduce the size of the range by half, so after n points, the interval we consider is $\frac{1}{3 \cdot 2^n}$. We want this to be less than 2ϵ, and so

$$\frac{1}{3 \cdot 2^n} < 2\epsilon \implies n > \log_2(\frac{1}{3\epsilon}) - 1.$$

(e) In the three sampling approaches above: random, deterministic, and adaptive, **compare the scaling of n with ϵ (and δ as well for the random case).**

Solution:

(a) For random, n is logarithmic in $1/\delta$. For ϵ, we use the approximation $\ln\big(1/(1-\epsilon)\big) \sim \epsilon$ to conclude that n is inversely related to ϵ.

(b) For deterministic, n is inversely related to ϵ. Note that this is the same scaling as choosing random evaluation points.

(c) For adaptive, n is logarithmic in $\frac{1}{\epsilon}$.

(f) **Why do you think we asked this series of questions? What are the implications of those results in a machine learning application?**

Solution: We ask this question because we want to show how the number of training examples affects the accuracy. Intuitively, more data lead to a more accurate estimator. We quantify this intuition with a simple but concrete example.

The three sampling approaches are some common ways to get the training data. When most of the real world datasets are collected, one doesn't have any control on X_i. That is the random sampling paradigm. The deterministic sampling paradigm refers to the scenario when one could carefully design a set of X_i. One might think that the sample complexity of the deterministic case should be much better than that of the random one, however for this particular model, they are not quite different. There is only a factor of $\log \frac{1}{\delta}$ off. However, when we move to the adaptive paradigm, the sample complexity is exponentially smaller.

For practical machine learning applications, the implication is that you want to have as much control on the samples as you can (such as adaptive sampling) to learn a better model with the same amount of data.

3 Gambling is for deviations

Suppose you go to a casino which has $n \geq 2$ slot machines, where the payouts from the i-th slot machine are i.i.d. random variables with distribution $\mathcal{N}(\theta_i, 1)$, where θ_i are real numbers. Assume that one of the slot machines has mean payout $\theta_{\max}$, and all other machines have mean payout $\theta_{\min} < \theta_{\max}$. For a fixed level of confidence $\delta \in (0, 1)$, your goal is to identify the slot machine with the highest mean payout with probability of error δ. You are allowed to do this by pulling each slot machine T times, guessing the best slot machine based based on the $T \times n$ payouts observed.

At the end, you want to ensure that your guess is correct with probability at least $1-\delta$. In these steps, you will determine an upper bound on the number of times T you neeed to identify the best slot machine.

(a) **Show that if X is a real valued random variable, you can bound $\mathbf{P}[X \geq t] \leq \mathbb{E}[X\mathbf{I}(X \geq t)]/t$ for all $t > 0$, where $\mathbf{I}(X \geq t) = 1$ if $X \geq t$, and 0 otherwise**. If it's easier, you may assume X has a continuous density $p(x)$. **Solution:** Observe that $\mathbf{P}[X \geq t] = \mathbb{E}[\mathbf{I}(X \geq t)]$ and that for all $X \geq t$ and $t > 0$, $\frac{X}{t} \geq 0$. Thus,

$$\begin{aligned}\mathbf{P}[X \geq t] &= \mathbb{E}[\mathbf{I}(X \geq t)] \\ &\geq \mathbb{E}[\frac{X}{t}\mathbf{I}(X \geq t)] \\ &= \frac{1}{t}\mathbb{E}[X\mathbf{I}(X \geq t)]\end{aligned}$$

(b) Let Z be distributed as $\mathcal{N}(0,1)$. **Show that**

$$\forall t > 0, \quad \mathbf{P}[Z \geq t] \leq \frac{1}{\sqrt{2\pi}t}e^{-t^2/2}$$

Use this to bound $\mathbf{P}[|Z| \geq t]$. Solution: By the previous problem, we have for $t > 0$,

$$\begin{aligned}\mathbf{P}[Z \geq t] &\leq \frac{1}{t}\mathbb{E}[Z\mathbf{I}(Z \geq t)] \\ &= \frac{1}{t}\int_t^\infty x\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx \\ &\leq \frac{1}{t\sqrt{2\pi}}\int_{u=t^2/2}^\infty e^{-u}du = \frac{e^{-t^2/2}}{t\sqrt{2\pi}}.\end{aligned}$$

Since the pdf of a gaussian is symmetric, we have for $t > 0$

$$\mathbf{P}[|Z| \geq t] = \mathbf{P}[Z \geq t] + \mathbf{P}[Z \leq -t] = 2\mathbf{P}[Z \geq t] \leq \frac{2e^{-t^2/2}}{t\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}} \cdot \frac{1}{t}e^{-t^2/2}$$

(c) Let $Z_1, \ldots, Z_n$ be distributed $\mathcal{N}(0,1)$ (not necessarily independent!), and let $n \geq 2$. **Show that for any $t \geq 1$**

$$\mathbf{P}[\max_i |Z_i| \geq t] \leq n \cdot \sqrt{\frac{2}{\pi}}e^{-t^2/2}.$$

Solution:

$$\mathbf{P}[\max_i |Z_i| \geq t] = \mathbf{P}\left[\bigcup_i \{|Z_i| \geq t\}\right]$$

$$\leq \sum_{i=1}^{n} \mathbf{P}\left[|Z_i| \geq t\right]$$
$$\leq \frac{n}{t} \cdot \sqrt{\frac{2}{\pi}} e^{-t^2/2}$$

when $t \geq 1$, we can drop the dependence on t.

(d) Suppose $n = 2$. **Show that that, in order to identify the slot machine with the highest payout with probability $1 - \delta$, it suffices to take**

$$T \geq \max\left\{1, \frac{4\log(2/\delta)}{(\theta_{\max} - \theta_{\min})^2}\right\} \text{ samples.}$$

You should use the inequalities developed earlier in the problem. **Solution:** We actually only need $T \geq \frac{4\log(2/\delta)}{(\theta_{\max}-\theta_{\min})^2}$ (The $\max\{1, \cdot\}$ is to ensure T is at least one).

Define $\widehat{\theta}_i$ to be the average of the payouts for the i-th slot machine. We will define the estimator

$$\widehat{i} := \arg\max_{i \in [n]} \widehat{\theta}_i$$

We will assume without loss of generality that $\theta_1 = \theta_{\max}$ and $\theta_2 = \theta_{\min}$. Since $\widehat{\theta}_i$ has the distribution $\mathcal{N}(\theta_i, 1/T)$, we see that

$$\widehat{\theta}_1 - \widehat{\theta}_2 \sim \mathcal{N}(\theta_{\max} - \theta_{\min}, 2/T).$$

Let

$$Z = -\sqrt{T/2}\left\{\left(\widehat{\theta}_1 - \widehat{\theta}_2\right) - (\theta_{\max} - \theta_{\min})\right\}.$$

Then, we see that Z has the distribution $\mathcal{N}(0, 1)$, and our estimator is only incorrect when $\widehat{\theta}_1 - \theta_2 \leq 0$, which occurs only when

$$Z \geq \sqrt{T/2}(\theta_{\max} - \theta_{\min}).$$

Let $t = \sqrt{T/2} \cdot (\theta_{\max} - \theta_{\min})$. Since $T \geq 4(\theta_{\max} - \theta_{\min})^{-2}\log(2/\delta)$, we see that $t \geq \sqrt{2\log(2/\delta)} \geq \sqrt{2\log 2} \geq 1$. Therefore, by part (c) with $n = 1$,

$$\mathbf{P}[Z \geq t] \leq \mathbf{P}[|Z| \geq t] \leq \sqrt{\frac{2}{\pi}} e^{-t^2/2} \leq e^{\log(1/\delta)} = \delta.$$

(e) **Generalize the above result to $n \geq 2$ slot machines.** When δ is a constant (say $\delta = 1/2$), there should be a $\log n$ somewhere in your answer. **Solution:** We can reduce this to the previous problem. Again, assume that $\theta_1 = \theta_{\max}$, and $\theta_i = \theta_{\min}$ for all $i \geq 2$. Moreover, let $\widehat{\theta}_i$

denote the mean estimator described above. What we showed is that, for any *fixed* $i\{2,\dots,n\}$, that if

$$T \geq \frac{4\log\big(1/\delta\big)}{(\theta_{\max}-\theta_{\min})^2},$$

then

$$\mathbf{P}[\widehat{\theta}_i \geq \widehat{\theta}_1] \leq \delta.$$

Hence, by a union bound,

$$\mathbf{P}[\exists i \in \{2,\dots,n\} \text{ such that } \widehat{\theta}_i \geq \widehat{\theta}_1] \leq (n-1)\delta.$$

Hence, with probability at most $(n-1)\delta$, choosing the slot machine with the highest empirical mean will correctly identity the machine with the greatest mean payoff. If we want a failure probability of exactly δ we may just replace δ with $\delta/(n-1)$, yielding,

$$T \geq \frac{4\log\big((n-1)/\delta\big)}{(\theta_{\max}-\theta_{\min})^2}.$$

4 Much ado about norms

Recall that a norm $\|\cdot\|$ is a function from $\mathbb{R}^d \to \mathbb{R}$ which satisfies the following properties:

(a) For all $x \in \mathbb{R}^d$, $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$

(b) For any real number α, $\|\alpha x\| = |\alpha|\|x\|$

(c) For any $x, y \in \mathbb{R}^d$, $\|x+y\| \leq \|x\| + \|y\|$

(a) For $p \in (0,\infty)$, let $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, and let $\|x\|_\infty = \max_i |x_i|$. **Please prove the following inequalities:**

(a) Show that $\|x\|_2$, $\|x\|_1$, and $\|x\|_\infty$ are norms. **Solution:** We verify the triangle inequality; the other properties are trivial. For $\|x\|_2$, the triangle inequality is Pythagoras' theorem. Alternatively, one can use Cauchy-Schwartz:

$$\begin{aligned}\|x+y\|_2^2 &= \langle x+y, x+y\rangle^2 \\ &= \|x\|_2^2 + \|y\|_2^2 + 2\langle x, y\rangle \\ &\leq \|x\|_2^2 + \|y\|_2^2 + 2\|y\|_2\|_2 x\| = (\|x\|_2 + \|y\|_2)^2.\end{aligned}$$

For $\|\cdot\|_\infty$, we have

$$\|x+y\|_\infty = \max_i |x_i + y_i| \leq \max_i |x_i| + |y_i| \leq \max_i |x_i| + \max_j |y_j| = \|x\|_\infty + \|y\|_\infty.$$

For $\|\cdot\|_1$,

$$\|x+y\|_1 = \sum_i |x_i + y_i| \leq \sum_i |x_i| + |y_i| = \sum_i |x_i| + \sum_j |y_j| = \|x\|_1 + \|y\|_1.$$

(b) Show that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$. **Solution:**

$$\|x\|_2 = \sqrt{\sum_i x_i^2} \geq \sqrt{\max_i x_i^2} = \max_i |x_i| = \|x\|_\infty.$$

On the other hand,

$$\|x\|_1^2 = (\sum_i |x_i|)^2 = \sum_i x_i^2 + 2\sum_{i<j} |x_i||x_j|$$
$$\geq \sum_i x_i^2 = \|x\|_2^2$$

(c) Show that $\|x\|_2^2 \leq \|x\|_1\|x\|_\infty$. **Solution:**

$$\|x\|_2^2 = \sum_i x_i^2 = \sum_i |x_i| \cdot |x_i| \leq \sum_i |x_i| \cdot (\max_j |x_j|) = \|x\|_1\|x\|_\infty$$

(d) Show that $\|x\|_1 \leq \sqrt{d}\|x\|_2$ and $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$. **Solution:**

$$\|x\|_2 = \sqrt{\sum_i x_i^2} \leq \sqrt{d \max_i x_i^2} = d\|x\|_\infty.$$

For the second inequality, let $\text{sign}(x)$ denote the vector of signs of x. By Cauchy Schwartz,

$$\|x\|_1 = \langle \text{sign}(x), x \rangle \leq \|x\|_2 \|\text{sign}(x)\|_2 = \|x\|_2\sqrt{d}.$$

(b) **For each the following functions $f(x) : \mathbb{R}^d \to \mathbb{R}$, state if f is always, sometimes, or never a norm. Give a proof. If the answer is 'sometimes', give a necessary and sufficient condition for f to be a norm**

(a) $f(x) = \log\cosh \|x\|_2$. **Solution:** Never a norm. Let $\alpha > 0$, and $\|x\|_2 = 1$. Then $f(\alpha x) = \log\cosh \alpha$, which is not linear in α.

(b) $f(x) = \sum_{i=1}^n x_i^2$. **Solution:** Never a norm. Let $\alpha > 0$, and $\|x\|_2 > 0$. Then $f(\alpha x) = \alpha^2 f(x)$, which is not linear in α.

(c) $f(x) = \|Ax\|$, where $\|x\|$ is a norm on $\mathbb{R}^d$, and $A \in \mathbb{R}^{d\times d}$. **Solution:** $f(x)$ is a norm if and only if $\text{rank}(A) = d$.

(a) The triangle inequality always holds

$$f(x+y) = \|A(x+y)\| = \|Ax + Ay\| \overset{(i)}{\leq} \|Ax\| + \|Ay\| = f(x) + f(y)$$

where (i) uses the fact that $\|\cdot\|$ satisfies the triangle inequality.

(b) Scaling always holds: $f(\alpha x) = \|A(\alpha x)\| = \|\alpha(Ax)\| = |\alpha|\|Ax\|$, again using that $\|\cdot\|$ is a norm.

(c) $\|Ax\| \geq 0$ since $\|\cdot\|$ is a norm. However, $\|Ax\| > 0$ for all $x \neq 0$ holds if and only if $\text{rank}(A) = d$. Indeed, if and only if $Ax = 0$. If $\text{rank}(A) = d$, then $Ax = 0$ if and only $x = 0$. But if $\text{rank}(A) < d$, then there necessarily exists an $x \neq 0$ in $\mathbb{R}^d$ for which $Ax = 0$, and thus $\|Ax\| = \|0\| = 0$.

(d) $f(x) = \sqrt{x^\top \Sigma x}$ where Σ is symmetric and has strictly positive eigenvalues. **Solution:** Always a norm. We will use the following fact:

Lemma 1. *If Σ is symmetric and has strictly positive eigenvalues, then there exists a full rank matrix Z such that $Z^\top Z = \Sigma$.*

Note that this completes the problem, since then

$$\sqrt{x^\top \Sigma x} = \sqrt{x^\top Z^\top Z x} = \sqrt{\|Zx\|_2^2} = \|Zx\|_2,$$

and we can use the answer to the previous problem with the fact that $\text{rank}(Z) = d$. Lemma 1 is standard, but we include a proof for completeness

Proof of Lemma 1. If Σ has strictly positive eigenvalues, then we can diagonalize $\Sigma = UDU^\top$, where D is diagonal and has positive entries on its diagonal, and where U is orthogonal. Let $D^{1/2}$ denote the matrix whose entries are the square roots of the entries of D. Then one can check that the matrix $Z := UD^{1/2}U^\top$ satisfies $Z^\top Z = \Sigma$. Moreover, Z has full rank, since its eigenvalues are the entries of $D^{1/2}$, which are all positive. □

(e) $f(x) = \sqrt{x^\top \Sigma x}$, where $\Sigma = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$, and $A \in \mathbb{R}^{n \times m}$ where $m + n = d$. **Solution:** Never a norm. Let e_1 be the first cannonical basis vector. Then $f(e_1) = 0$.

(f) $f(x) = \sum_i \alpha_i |x_i|$, $\alpha_i \in \mathbb{R}$. **Solution:** Sometimes a norm, if and only if $\alpha_i > 0$ for all $i \in [n]$. Indeed, suppose there exists an i such that $\alpha_i \leq 0$. Then $f(e_i) \leq 0$, where $e_i \neq 0$ is the cannonical basis vector with a 1 and the i-th position. If $\alpha_i > 0$ for all $i \in [n]$, then f is a norm by essentially the same argument that shows that $\|x\|_1$ is a norm.

(g) $f(x) = \sup_{w \in \mathcal{C}} \langle w, x \rangle$, where $\mathcal{C} \subset \mathbb{R}^d$ has the following properties:

- $x \in \mathcal{C}$ if and only if $-x \in \mathcal{C}$.
- $\mathcal{C}$ is bounded; that is $\sup_{x \in \mathcal{C}} \|x\| < \infty$.
- There exists an orthonormal basis $\{e_1, e_2, \ldots, e_d\} \subset \mathcal{C}$.

Solution: This is a fun one. Not only is f always a norm, but it turns out that if you replace the third condition $\{e_1, e_2, \ldots, e_d\} \subset \mathcal{C}$ with something slightly more general, *any* norm in $\mathbb{R}^d$ can be expressed as $f(x) = \sup_{w \in \mathcal{C}} \langle w, x \rangle$ for some set $\mathcal{C}$.

Anyways, let's prove that this is a norm. **Triangle Inequality:**

$$f(x + y) = \sup_{w \in \mathcal{C}} \langle w, x + y \rangle$$

$$= \sup_{w \in \mathcal{C}} \langle w, x \rangle + \langle w, y \rangle$$
$$= (\sup_{w \in \mathcal{C}} \langle w, x \rangle) + (\sup_{w \in \mathcal{C}} \langle w, y \rangle) = f(x) + f(y)$$

Nonnegativity: Note that since $\mathcal{C}$ is symmetric and contains $\{e_1, \ldots, e_d\}$,

$$f(x) = \sup_{w \in \mathcal{C}} \langle w, x \rangle \geq \sup_{\pm e_i} \langle e_i, x \rangle = \sup_i |x_i| = \|x\|_\infty,$$

which implies that $f(x) \geq 0$ and $f(x) = 0 \iff x = 0$.

Scaling: First note that $f(x) = f(-x)$, since by symmetry of $\mathcal{C}$

$$f(-x) = \sup_{w \in \mathcal{C}} \langle w, -x \rangle = \sup_{w \in \mathcal{C}} \langle -w, -x \rangle = \sup_{w \in \mathcal{C}} \langle w, x \rangle = f(x)$$

Further, for $\alpha \geq 0$, define $\mathcal{C}_x := \{w \in \mathcal{C} : \langle w, x \rangle \geq 0\}$. Note then that $\mathcal{C}_x$ is nonempty, and thus, for $\alpha \geq 0$.

$$f(\alpha x) = \sup_{w \in \mathcal{C}} \langle w, \alpha x \rangle = \sup_{w \in \mathcal{C}_x} \langle w, \alpha x \rangle \tag{1}$$
$$f(x) = \sup_{w \in \mathcal{C}} \langle w, x \rangle = \sup_{w \in \mathcal{C}_x} \langle w, x \rangle. \tag{2}$$

Hence,

$$\begin{aligned} f(\alpha x) &= \sup_{w \in \mathcal{C}_x} \langle w, \alpha x \rangle && \text{(by (1))} \\ &= \sup_{w \in \mathcal{C}_x} \alpha \langle w, x \rangle \\ &\overset{(ii)}{=} \alpha \sup_{w \in \mathcal{C}_x} \langle w, x \rangle \\ &= \alpha f(x)., && \text{(by (2))} \end{aligned}$$

where (ii) uses the fact that $\langle w, x \rangle \geq 0$ for all $w \in_x$. Lastly if $\alpha < 0$, then $f(\alpha x) = f(-\alpha x) = f(|\alpha| x) = |\alpha| f(x)$ by the $\alpha \geq 0$ case.

5 Projecting your problems

Given $1 \leq d \leq n$, a matrix $P \in \mathbb{R}^{n \times n}$ is said to be a rank-d orthogonal projection matrix if $\text{rank}(d) = P$, $P = P^\top$ and $P^2 = P$.

(a) **Prove that P is a rank-d projection matrix if and only if there exists a $U \in \mathbb{R}^{n \times d}$ such that $P = UU^\top$ and $U^\top U = I$ Solution:** Since P is symmetric, $P = V \Sigma V^\top$ for some orthogonal $V \in \mathbb{R}^{n \times n}$ and real, diagonal $\Sigma \in \mathbb{R}^{n \times n}$. Let v be an eigenvector of P with eigenvalue λ. Then, $\lambda^2 v = P^2 v = Pv = \lambda v$, so that $\lambda \in \{0, 1\}$. Hence, the diagonals of Σ are binary-valued. Letting U denote the matrix whose columns correspond to the indicies i for which $\Sigma_{ii} = 1$, we have $P = V \Sigma V^\top = UU^\top$. Since P has rank d, there that there are d such 1-valued indices.

Since the columns of U are a subsect of those of V, they are orthonormal, whence $U^\top U = I$.

Conversely, if $P = UU^\top$, then $P = P^\top$ trivially, and $P^2 = UU^\top UU^\top = UU^\top$. Moreover, P has rank at most d since $P = UU^\top$, and rank at least $\text{rank}(PU) = \text{rank}(UU^\top U) = \text{rank}(U) = d$.

Because the original HW did not inlude rank of P in the problem statement, points were not deducted for failing to show that the dimension of U and rank of P coincide. However, the solution includes the proof of these facts.

(b) **Prove that if P is a rank d projection matrix, then $\text{tr}(P) = d$. Solution:**

Approach 1: Using the trace trick $\text{tr}(AB) = \text{tr}(BA)$, $\text{tr}(P) = \text{tr}\big(UU^\top\big) = \text{tr}\big(U^\top U\big) = \text{tr}(I_d) = d$.

Approach 2: $\text{tr}(P)$ is the sum of the eigenvalues of P. As verified in Part (a), these lie in $\{0, 1\}$, and since $\text{rank}(P) = d$, we must have that P has d eigenvalues equal to 1, and all others zero. Thus, $\text{tr}(P) = 1 \cdot d = d$.

(c) **Prove that, for all $v \in \mathbb{R}^n$,**

$$Pv = \arg \min_{w \in \text{range}(P)} \|v - w\|_2^2.$$

Solution:

Approach 1: From lecture, it suffices to show that Pv is the orthogonal projection onto the range of P. Let $u_1, \ldots, u_d$ denote the columns of U. Then, $u_1, \ldots, u_d$ are orthonormal because $U^\top U = I$. Moreover, , since $Pu_i = U^\top U u_i = \sum_{j=1}^d \langle u_j, u_i \rangle u_i = u_i$, $\{u_1, \ldots, u_d\}$ lie in the $\text{range}(P)$. Since $\dim(\text{range}(P)) = d$, we conclude that $\{u_1, \ldots, u_d\}$ are an orthonormal eigenbasis for the range of P. It follows that for any vector v, $Pv = UU^\top v = \sum_{i=1}^d \langle u_i, v \rangle u_i$ is the orthogonal projection onto the range of P.

Approach 2: As above, it suffices Pv is the orthogonal projection onto the range of P. Let $v = Pv + (v - Pv)$. Since $Pv \in \text{range}(P)$, it suffices to show that $v - Pv \in \text{range}(P)^\perp$. $P(v - Pv) = Pv - P^2 v = Pv - Pv = 0$. Thus, $v - Pv \in \ker(P)$. Hence, $v - Pv \in \text{range}(P)^\perp$, since the kernel and range of a symmetric matrix are orthogonal complements.

Approach 3: From lecture, the $\|w - Pv\|_2^2$ is minimized when $w - v \perp \text{range}(P)$. Since P is symmetric, it suffices that $P(w - v) = 0$. $w = Pv$ satisfies this condition.

Approach 4: Let $w = Pz$, and expand out $\|Pz - v\|_2^2 = \|Pz - Pv + (v - Pv)\|_2^2$, using properties of P ($P^2 = P$, $P = P^2$) to verify that the cross term is zero.

(d) **Prove that if $X \in \mathbb{R}^{d \times d}$ and $\text{rank}(X) = d$, then $X(X^\top X)^{-1} X^\top$ is a rank-d orthogonal projection matrix. What is the corresponding matrix U? Solution:** Let $X = U\Sigma V$ denote the SVD of X, with $V \in \mathbb{R}^{d \times d}$, and $\Sigma \in \mathbb{R}^{d \times d}$, and $U \in \mathbb{R}^{d \times n}$. Then, $X^\top X = V^\top \Sigma^2 V$, and since X is full rank, Σ^2 is invertible, with $X^\top X = V^\top \Sigma^{-2} V$. Hence,

$$X(X^\top X)^{-1} X^\top = U\Sigma V^\top (V^\top \Sigma^{-2} V) V \Sigma U^\top = U \Sigma \Sigma^{-2} \Sigma U = UU^\top,$$

which shows that $X(X^\top X)^{-1} X^\top$ is the projection matrix $UU^\top$, where U is the left singular-vectors matrix of X.

(e) Let $y = X\theta_* + z$, where $X \in \mathbb{R}^{n\times d}$, $\theta_* \in \mathbb{R}^d$, $y \in \mathbb{R}^n$, and $z = \mathcal{N}(0, I) \in \mathbb{R}^n$, and suppose $\text{rank}(X) = d$. **Prove that if $\widehat{\theta} = \arg\min_\theta \|X\theta - y\|_2^2$, then**

$$\mathbb{E}[\|\theta_* - \widehat{\theta}\|_2^2] = \text{tr}\Big((X^\top X)^{-1}\Big)$$

Solution: Recall from lecture that

$$\begin{aligned}\widehat{\theta} &= (X^\top X)^{-1}X^\top y \\ &= (X^\top X)^{-1}X^\top(X\theta_* + z) = \theta_* + (X^\top X)^{-1}X^\top z\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\|\widehat{\theta} - \theta_*\|_2^2 &= \mathbb{E}\|(X^\top X)^{-1}X^\top z\|_2^2 \\ &= \mathbb{E}z^\top X(X^\top X)^{-2}X^\top z \\ &= \text{tr}\Big(X(X^\top X)^{-2}X^\top\Big),\end{aligned}$$

where we use the simple computation that if $A \in \mathbb{R}^{d,d}$ is any matrix matrix and $z \sim \mathcal{N}(0, I)$, then

$$\begin{aligned}\mathbb{E}z^\top Az &= \mathbb{E}\sum_{1\le i,j\le n} A_{ij}z_iz_j \\ &= \sum_{1\le i,j\le n} A_{ij}\mathbb{E}z_iz_j \\ &= \sum_{1\le i,j\le n} A_{ij}\mathbf{I}(i=j) = \text{tr}(A).\end{aligned}$$

This computation can be also performed with traces, via $\mathbb{E}z^\top Az = \mathbb{E}\,\text{tr}\big(z^\top Az\big) = \mathbb{E}\,\text{tr}\big(zz^\top A\big) = \text{tr}\big((\mathbb{E}zz^\top)A\big) = \text{tr}(IA) = \text{tr}(A)$, where we used linearity of the trace.

To conclude, we use the fact that $\text{tr}(AB) = \text{tr}(BA)$ with $A = X$, and $B = (X^\top X)^{-2}X^\top$,

$$\text{tr}\Big(X(X^\top X)^{-2}X^\top\Big) = \text{tr}\Big((X^\top X)^{-2}X^\top X\Big) = \text{tr}\Big(X^\top X\Big).$$

(f) **In the setting of the Part (e), show that**

$$\frac{1}{n}\mathbb{E}[\|X(\theta_* - \widehat{\theta})\|_2^2] = \frac{d}{n}.$$

How does the answer change if $\text{rank}(X) < d$? Solution:

If $\text{rank}(X) = d$, one approach is to solve the problem similarly to part (e). Again we have that

$$\widehat{\theta} - \theta_* = (X^\top X)^{-1}X^\top z,$$

and therefore

$$X(\widehat{\theta} - \theta_*) = X(X^\top X)^{-1}X^\top z,$$

Let $P = X(X^\top X)^{-1}X^\top$. By part d, P is a rank-d orthogonal projection matrix, and therefore $P^\top P = P^2 = P$. Hence,

$$\begin{aligned}\mathbb{E}\|X(\widehat{\theta} - \theta_*)\|_2^2 = \mathbb{E}\|Pz\|_2^2 &= \mathbb{E}z^\top(P^\top P)z \\ &= \mathbb{E}z^\top Pz.\end{aligned}$$

In our solution to the previous question, we showed that $\mathbb{E}z^\top Pz = \mathrm{tr}(P)$, which is equal to d by part (d).

If $\mathrm{rank}(X) < d$, there are a couple approaches. **Approach 1:** If X has $\mathrm{rank}(X) = k < d$ rank, one can simply remove columns from X to make it full rank, an ammend θ_* and θ appropriately. Specifically, let $\widetilde{X} \in \mathbb{R}^{n\times k}$, $\widetilde{\theta}_*$ such that $\widetilde{X}\widetilde{\theta}_* = X\theta_*$. Then, any

$$X\widehat{\theta}, \text{ where } \widehat{\theta} \in \arg\min \|X\theta - y\|_2^2$$

can be expressed as a

$$\widetilde{X}\widehat{\widetilde{\theta}}, \text{ where } \widehat{\widetilde{\theta}} \in \arg\min \|\widetilde{X}\widetilde{\theta} - y\|_2^2,$$

reducing to the full rank case.

Approach 2: Let $k = \mathrm{rank}(X)$. Realize that

$$X\widehat{\theta}, \text{ where } \widehat{\theta} \in \arg\min \|X\theta - y\|_2^2$$

can be written as

$$w \text{ where } \widehat{\theta} \in \arg\min_{\mathrm{range}(X)} \|w - y\|_2^2$$

By Problem (c), the solution is just Py, P is is any rank k projection matrix such that $\mathrm{range}(P) = \mathrm{range}(X)$. Such a P exists, since we can take $P = UU^\top$ where $U \in \mathbb{R}^{n\times k}$ has the an orthonormal basis for the range of X as its columns.

To conclude, we see that $y - Py = X\theta - PX\theta_* - Pz$. But $X\theta_* \in \mathrm{range}(X) = \mathrm{range}(P)$, and since P is a projection operator (we checked this in (c)), then $PX\theta_* = X\theta_*$. Thus, $y - Py = -Pz$. Hence, for any

$$X\widehat{\theta}, \text{ where } \widehat{\theta} \in \arg\min \|X\theta - y\|_2^2$$

we have

$$\mathbb{E}[\|X\widehat{\theta} - y\|_2^2] = \mathbb{E}[\|Pz\|_2^2] = \mathbb{E}[z^\top P^\top Pz] = \mathbb{E}[z^\top P^2 z] = \mathbb{E}[z^\top Pz] = \mathrm{tr}(P) = k.$$

Approach 3: Use the Pseudoinverse. Let $\widehat{\theta} = X^\dagger\theta_*$. Using the pseudo-inverse, one can show that $X^\dagger\theta_* - y = -P_{\mathrm{range}(X)}z$. Left as an exercise to the reader.

6 Installing Python

Set up your Python environment for the course by downloading Anaconda 5.2 here: `https://www.anaconda.com/download/#macos` (Mac), `https://www.anaconda.com/download/#windows`(Windows) or `https://www.anaconda.com/download/#linux`(Linux). Make sure to install the *Python 3.6* version.

7 Your Own Question

Write your own question, and provide a thorough solution.

Writing your own problems is a very important way to really learn the material. The famous "Bloom's Taxonomy" that lists the levels of learning is: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using what you know to create is the top-level. We rarely ask you any HW questions about the lowest level of straight-up remembering, expecting you to be able to do that yourself. (e.g. make yourself flashcards) But we don't want the same to be true about the highest level.

As a practical matter, having some practice at trying to create problems helps you study for exams much better than simply counting on solving existing practice problems. This is because thinking about how to create an interesting problem forces you to really look at the material from the perspective of those who are going to create the exams.

Besides, this is fun. If you want to make a boring problem, go ahead. That is your prerogative. But it is more fun to really engage with the material, discover something interesting, and then come up with a problem that walks others down a journey that lets them share your discovery. You don't have to achieve this every week. But unless you try every week, it probably won't happen ever.