Tail Bounds and Maximal Inequalities

QTM 385-1: Machine Learning and Nonparametric Regression

1 Introduction

THIS NEEDS FIXING

In this homework, we're going to focus on showing that, with high probability, this is true.

$$\|m-\mu\| > 2\left\langle \varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle \quad \text{for all} \quad m \in \mathcal{M} \setminus \mathcal{M}_s. \tag{1}$$

Here ε is our vector of noise, i.e. the vector with elements $\varepsilon_i = Y_i - \mu(X_i)$ where $\mu(x) = \mathrm{E}[Y_i \mid X_i = x]$. In lecture, we did that in the stylized case that these elements ε_i are independent and normal, each with the same variance σ^2. Now we're going to be a bit more real. We can get pretty far without changing our argument from lecture much. Then we'll rephrase our argument a little to prepare for the coming weeks.

Notation. Here, as in lecture, we'll take $\langle u, v\rangle$ to be the dot product $\sum_i u_i v_i$ and $\|v\|$ to be the two-norm $\sqrt{\langle v, v\rangle}$.

1.1 Why we're doing this

Think about the argument we used to show that our least squares estimator $\hat{\mu}$ is close to μ in this week's lecture. We focused on a property the least squares estimator $\hat{\mu}$ has when μ is in our model—that it's always one of the curves m that satisfies $\ell(m) \le \ell(\mu)$. And we showed that with high probability, every curve in the model that's far from μ *doesn't have this property*, implying that $\hat{\mu}$ cannot be one of these curves. The last step boiled down to (1) above. Why?

$$\begin{aligned}
\ell(m) \underset{\not\leq}{>} \ell(\mu) \quad & \text{if} \quad \|m-\mu\|^2 > 2\langle \varepsilon, m-\mu\rangle \\
& \text{because[1]} \quad n \times \{\ell(m) - \ell(\mu)\} = \|m-\mu\|^2 - 2\langle \varepsilon, m-\mu\rangle \\
& \text{so} \\
\ell(m) \underset{\not\leq}{>} \ell(\mu) \ \text{ for all } \ m \in \mathcal{M}\setminus\mathcal{M}_s \quad & \text{if} \quad \|m-\mu\|^2 > 2\langle \varepsilon, m-\mu\rangle \ \text{ for all } \ m \in \mathcal{M}\setminus\mathcal{M}_s.
\end{aligned}$$

The approach we used to bound this dot product $\langle \varepsilon, m-\mu\rangle$ relied heavily on a very implausible assumption. When read enough machine learning papers, you get at bit numb to it. In mathematical notation, it doesn't bother you.

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2).$$

But it's clear that it's not an assumption you should make when you say it, or some consequence of it, in lay english. Here's an example.

> The fraction of people that earn $100,000$ more than the average
> for people with the same job doesn't depend on what that job is.

That may be big for investment bankers, but it won't for school teachers.

This argument has implications only when μ is in our model, which doesn't happen. But the generalized version we introduced in the Model Selection Lab has implications without that assumption, and relies on a bound like (1) too.

1.2 Subgaussianity

To keep things simple, we'll focus on the case that the elements ε_i of our noise vector are independent and *subgaussian.* In essence, a mean-zero random variable Z is called subgaussian if it's no more likely to be large than a mean-zero normal (a.k.a. gaussian) random variable. Formally, we'll say Z is σ^2-subgaussian if $P(|Z| \geq t) \leq 2e^{-t^2/2\sigma^2}$. Here are the best-known examples.

1. A gaussian random variable Z with mean zero and variance σ^2 is σ^2-subgaussian.[2]

2. A bounded random variable Z with $|Z| \leq \sigma$ is σ^2-subgaussian.

This quantity σ^2, which we tend to call a *variance proxy* for Z, transforms a lot like variance. In particular, it behaves like variance when we scale and sum independent subgaussian random variables. For non-random weights $\alpha_1 \ldots \alpha_n$,

$$\sum_{i=1}^n \alpha_i Z_i \ \text{ is } \ \left(\sum_{i=1}^n \alpha_i^2 \sigma_i^2\right) - \text{subgaussian} \ \text{ if each } \ Z_i \ \text{ is } \ \sigma_i^2 - \text{subgaussian.} \tag{2}$$

We call this Hoeffding's Inequality. It may not look like an inequality, but it is. It's giving us a variance proxy for our weighted average, which is equivalent to a bound on the probability that it's large, in terms of variance proxies for the random variables we're averaging.

One thing to keep in mind is that variance proxies are defined in terms of *upper bounds.* As a result, if a random variable Z is σ^2-subgaussian it's also σ_+^2-subgaussian for any $\sigma_+ \geq \sigma$. This means that if we have a collection of random variables $Z_1 \ldots Z_n$, where Z_i is σ_i^2-subgaussian, we can think of them as a collection of σ^2-subgaussian random variables with common variance proxy $\sigma^2 = \max_i \sigma_i^2$.

[1] We multiply by n here so we can work with dot products rather than sample inner products, as $\ell(m) - \ell(\mu) = \|m - \mu\|_{L_2(\mathrm{P_n})}^2 - 2\langle \varepsilon, m - \mu \rangle_{L_2(\mathrm{P_n})} = \|m - \mu\|^2/n - 2\langle \varepsilon, m - \mu \rangle/n$.

[2] In this week's lecture, we came close to proving a standard normal is 1-subgaussian.

References. If you want to know more, including proofs for the claims above, this blog post isn't a bad place to look. Our optional textbook High Dimensional Probability works too, but the presentation there is a bit less self-contained.

2 Generalizing our Error Bound

Often, when we've proven a result assuming some random variables are gaussian with variance σ^2, it's true when they're σ^2-subgaussian too. Not always. In this exercise, we'll look into what we can say about the error $\|\hat{\mu} - \mu\|$ when we have σ-subgaussian noise. That is, we'll think about observations like this.

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{for} \quad i = 1 \ldots n \qquad \text{where} \qquad \varepsilon_1 \ldots \varepsilon_n \quad \text{are independent and } \sigma\text{-subgaussian}$$

And as in this week's lecture, we'll talk about what happens when we use a finite model $\mathcal{M}$ containing no more than K curves, one of which (miraculously) is μ.

Exercise 1 *State a radius s for which* (1) *holds with probability* $1 - \delta$. *And briefly describe how you'd prove that it does. If you want to reuse steps from the proof given in lecture, you don't need to write them all out; just summarize what you're using. What does this tell us about $\|\hat{\mu} - \mu\|$?*

If you like, take δ to be a convenient function $\delta(K)$ of K, as long as you can make $\delta(K)$ arbitrarily small by making K large.

Tip. Read this week's lecture carefully.

Exercise 2 *The bound on $\|\hat{\mu} - \mu\|$ we get in the subgaussian case is* **much** *more applicable than the version we proved in lecture. Briefly explain why. Give an example or two of applications in which the assumption of subgaussianity is justifiable.*

Tip. Read the introduction section above carefully.

3 Tail Bounds and Maximal Inequalities

Exercise 3 *Why is it important that we divided by $\|m - \mu\|$? What could we say using an analogous sufficient condition based more directly on* (1)*, involving the single random variable* $\max_{m \in \mathcal{M} \setminus \mathcal{M}_s} \langle \varepsilon, m - \mu \rangle$*?*

The more you know. We can see this as an instance of a fairly general technique. If we're basing an argument on the condition that some function $\mathcal{L}(m)$ is positive, it makes sense to ask whether we'd be better off thinking about the ratio of $\mathcal{L}$ and some positive function of m. This is equivalent, in the sense that one will be positive if and only if the other is, but the ratio we're talking about might be easier to work with. In this case, what we're doing is working with the ratio $\{\ell(m) - \ell(\mu)\}/\|m - \mu\|$ instead of $\ell(m) - \ell(\mu)$.

3.1 Maximal Inequalities

What we're seeing is that our proof worked by reducing the claim $\|m-\mu\| \le s$ we wanted to prove to a *maximal inequality*, i.e. a bound on a maximum of random variables, that implies it. This is probably the most common proof technique in modern statistics, which makes maximal inequalities very important. What's nice about this approach is that there are books full of maximal inequalities out there. If you're looking for one, our optional textbook has plenty and Michel Talagrand's Upper and Lower Bounds on Stochastic Processes has even more.

There are two main types of maximal inequalities. What we've been using, the bound $P(Z_K > 2\sigma\sqrt{\log(K)}) \le 1/K$ on a maximum $Z_K = \max_{i \in 1 \ldots K} Z_i$ of gaussian random variables with variance σ^2, is a *tail bound*. It's a bound on a tail probability—the probability that something is big. Later on, we'll often be using bounds on the expected value of a maximum like Z_K—an *expectation bound*. Often, these are a bit easier to work with.

It's not hard to go back and forth between tail bounds and expectation bounds. Going from tail bounds to expectation bounds is basically just a matter of integration: $\mathrm{E}\, Z = \int_0^\infty P(Z > z)dz$ for any non-negative random variable Z. And going from expectation bounds to tail bounds, if we're ok with losing a bit of precision, is a simple application of Markov's inequality: $P(Z \ge z) \le \mathrm{E}\, Z/z$ for any non-negative random variable Z and $z > 0$. If these are unfamiliar, it may be worth taking a look at Appendix A, where I'll prove them. In a few weeks, we'll talk about a way of showing that some interesting random variables like Z_K are much closer to their expectation than Markov's inequality implies.

Let's try to derive an expectation bound from our tail bound on the maximum Z_K. Or to be precise, from tail bound $P(Z_K \ge z) \le e^{\log(K) - z^2/2\sigma^2}$ that we substituted $z = 2\sqrt{\log(K)}$ into to get the bound above.

Exercise 4 *First, as a warm-up, find an upper bound on the* median *of Z_K: the value of z for which $P(Z_K \ge z) = 1/2$. Then find one for the mean $\mathrm{E}\, Z_K$.*

Hint. Break the integral $\mathrm{E}\, Z_K = \int_0^\infty P(Z_K > z)dz$ into a sum of two integrals, one up to z_0 and one from there to ∞. See what you can do with the facts that $z/z_0 \ge 0$ on the domain of the second integral and that $(d/dz)e^{-z^2/2} = -ze^{-z^2/2}$. Then come up with a reasonable way to choose z_0.[3]

3.2 Gaussian Width

The maximum we're working with in our argument, $Z_K = \max_{m \in \mathcal{M} \setminus \mathcal{M}_s} \langle \varepsilon, m - \mu \rangle / \|m - \mu\|$, can be characterized as the maximum $\max_{\delta \in \Delta} \langle \varepsilon, \delta \rangle$ of the dot product of the gaussian vector $\varepsilon \in \mathbb{R}^n$ and a vector δ in the set $\Delta = \{(m(X) - \mu(X))/\|m - \mu\| : m \in \mathcal{M} \setminus \mathcal{M}_s\}$. Here $m(X) \in \mathbb{R}^n$ and $\mu(X) \in \mathbb{R}^n$ are vectors with ith elements $m(X_i)$ and $\mu(X_i)$ respectively. We see maxima like this so

[3] Here I'm hinting at the Proof of Lemma 2.3.2 in Talagrand's Upper and Lower Bounds on Stochastic Processes.

often we have a special name for them and their expected values. If ε is a vector of independent standard normal random variables,

$$\begin{aligned} \hat{\mathrm{w}}(\Delta) &= \max_{\delta\in\Delta}\langle \varepsilon, \delta\rangle_{L_2(\mathrm{P_n})} && \text{is the } \textit{sample gaussian width} \text{ of } \Delta \\ \mathrm{w}(\Delta) &= \mathrm{E}\max_{\delta\in\Delta}\langle \varepsilon, \delta\rangle_{L_2(\mathrm{P_n})} && \text{is the } \textit{gaussian width} \text{ of } \Delta \end{aligned} \tag{3}$$

We can talk about the gaussian width of infinite sets as well as finite ones. It turns out to be a very meaningful way to measure the size of a set. Let's do a few exercises to get used to the idea. In the first, we'll see where working with gaussian width gets us when we try to prove a version of our error bound from this week's lecture. In the second, we'll bound the gaussian width of an infinite set just to get used to the idea.

Scaling. Because the sample inner product is n times the dot product, the maximal dot product $\max_{\delta\in\Delta}\langle\varepsilon,\delta\rangle$ that we discussed above is $n\hat{\mathrm{w}}(\Delta)$. Going back and forth between sample inner products and dot products is an inconvenience that we'll have to get used to, as there are circumstances in which each is more natural.

Exercise 5 *Suppose that, instead of using the bound $P\{Z_K > 2\sigma\sqrt{\log(K)}\} \le 1/K$ on the maximum in (**??**), we wanted to use a bound in terms of the gaussian width $\mathrm{w}(\Delta)$. First, using Markov's inequality, write a bound on $\lVert\hat{\mu}-\mu\rVert$ that involves $\mathrm{w}(\Delta)$ and holds with probability $1-u$ for arbitrary u. Then, using the result of Exercise 4, bound $\mathrm{w}(\Delta)$ in terms of K. Is the resulting bound on $\lVert\hat{\mu}-\mu\rVert$ better or worse than the bound we proved in this week's lecture?*

You may assume $\sigma = 1$.

Exercise 6 *Bound the gaussian width $\mathrm{w}(\mathcal{B}_1)$ of the sample one-norm's unit ball, $\mathcal{B}_1 = \{v : \lVert v\rVert_{L_1(\mathrm{P_n})} \le 1\}$.*

Tip. Use Hölder's inequality. We proved that in the Vector Spaces Homework.

A Appendix: Probability Stuff

Much of this is borrowed from a homework for another class, so it'll have some additional stuff and take the form of problems and solutions. If you're comfortable thinking of probabilities as expected values and don't want the extra stuff, go ahead and skip to Section A.2 for the formula $\mathrm{E}\,Z = \int_0^\infty P(Z > z)dz$ and Section A.4 for Markov's Inequality.

A.1 Probabilities as Expected Values

In this problem, we'll prove a relationship between probabilities and expected values. We'll show that the probability of an event is the expected value of an indicator that it happens. This is true generally, but for the sake of concreteness, we'll focus on a particular kind of event: the event that a random variable $X \in \mathbb{R}^n$ is in a rectangle $[a_1, b_1] \times [a_2, b_2] \times \ldots [a_n, b_n]$ which, to keep our notation a bit more compact, we'll call $[A, B]$ where convenient. That is, we'll show that

$$P(X \in [A,B]) = \mathrm{E}\, 1(X \in [A,B]) \quad \text{where} \quad 1(x \in [A,B]) = \begin{cases} 1 & \text{if } x \in [A,B] \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

We'll do this for two types of random variables X: discrete ones, which have probability mass functions, and continuous ones, which have probability density functions.

An alternate notation, which may make it a bit easier for you to think about the problem, is to write $1_{[A,B]}(x)$ with the same meaning as $1(x \in [A,B])$. This highlights that the indicator $1(x \in [A,B]) = 1_{[A,B]}(x)$ is a function of x, and $1(X \in [A,B]) = 1_{[A,B]}(X)$ is that function evaluated at a random variable.

A.1.1 The Discrete Case

For a discrete random variable X with probability mass function $f_X(x) = P(\{X = x\})$, $\mathrm{E}\, g(X) = \sum_x g(x) f_X(x)$ for any function g and, in addition, that for disjoint events $E_1, E_2, \ldots$ that $P(E_1 \cup E_2 \ldots) = P(E_1) + P(E_2) + \ldots$. Prove that Equation 4 above holds for a discrete random variable X and any rectangle $[A, B]$.

Solution .

$$\mathrm{E}\, 1_{[A,B]}(X) = \sum_x 1_{[A,B]}(x) f_X(x) = \sum_{x \in [A,B]} P(\{X = x\}) \overset{(a)}{=} P(\cup_{x \in [A,B]} \{X = x\}) \overset{(b)}{=} P(X \in [A,B]).$$

Here the notation $\cup_{x \in [A,B]} \{X = x\}$ means the union of the events $\{X = x_1\} \cup \{X = x_2\} \cup \ldots$ where $x_1, x_2, \ldots$ are the set of elements in $[A, B]$. The equivalence (a) holds because events $\{X = x\}$ for different x are disjoint and the probability of a union of disjoint events is the sum of their individual probabilities. The equivalence (b) holds because the union of these individual events is the event that $\{X \in [A, B]\}$.

A.1.2 The Continuous Case

For a continuous random variable X with probability density function $f_X(x)$, $\mathrm{E}\, g(X) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(x) f_X(x) dx_1 \ldots dx_n$ for any function g and, in addition, $P(X \in [A,B]) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f_X(x) dx_1 \ldots dx_n$.[4]

[4] Where this is discussed in your book, in Definition 1.4.4, they focus on the probability that X is in a rectangles of the form $(-\infty, B] = (-\infty, b_1] \times \ldots (-\infty, b_2]$. Can you show that this implies the result for a general rectangle $[A, B]$ by decomposing the event $\{X \in [-\infty, B]\}$ into disjoint events, one of which is $\{X \in [A, B]\}$? This is totally optional, but if you want to do it, it may be worth trying the one-dimensional case first, where the events $E_1 = \{X \in (-\infty, a)\}$

Solution.

$$\mathrm{E}\, 1_{[A,B]}(X) = \int 1_{[A,B]}(x) f_X(x)dx \overset{(a)}{=} \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_X(x)dx_1 \ldots dx_n = P(X \in [A,B]).$$

Here the equivalence (a) holds because the integrand is $1 \cdot f_X(x)$ inside the rectangle $[A,B]$ and zero elsewhere.

A.2 Integrating Tail Bounds

In this subsection, we're going to characterize the mean of a nonegative random variable Z as the integral of a tail bound. In fact, we're going to do it in this line.

$$\mathrm{E}\, Z = \mathrm{E} \int_0^Z dz = \mathrm{E} \int_0^\infty 1(Z > z)dz = \int_0^\infty \mathrm{E}\, 1(Z > z)dz$$

A.3 The Law of Large Numbers

In this problem, we're going to prove the law of large numbers.

Theorem 1 *Let $X_1, X_2, \ldots$ be independent random variables with $\mathrm{E}\, X_i = 0$ and $\mathrm{E}\, X_i^2 \le B$ for some $B < \infty$ and $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$. Z_n converges to zero in probability.*

This'll be a two-step process. Our first step is to show that Z_n converges to zero in a different sense. We'll show that it converges to zero in mean square, i.e., that $\mathrm{E}\, Z_n^2 \to 0$. Then we'll show that this implies convergence in probability.

A.3.1 Convergence in Mean Square

Now we'll show that $\mathrm{E}\, Z_n^2 \to 0$ as $n \to \infty$. To do this, first calculate Z_n^2. Because because Z_n is a sum, Z_n^2 is a double sum. I'll get you started.

$$Z_n^2 = \left(\frac{1}{n}\sum_{i=1}^n X_i\right)\left(\frac{1}{n}\sum_{j=1}^n X_j\right) = \ldots$$

Then, calculate its expected value $\mathrm{E}\, Z_n^2$. Many terms in this double sum have expected value zero. Show, using the bound $\mathrm{E}\, X_i^2 \le B$, that this expected value goes to zero as $n \to \infty$.

and $E_2 = \{X \in [a,b]\}$ are disjoint and have union $E_1 \cup E_2 = \{X \in (-\infty, b]\}$.

Here, to be formal, the interval $(a,b]$ refers to the set of x with $a < x \le b$, not including the left endpoint; $[a,b)$ to the set of x with $a \le x < b$ not including the right; and $[a,b]$ to the set of x with $a \le x \le b$, including both endpoints. For a continous random variable X, however, the distinction doesn't really matter: $P(\{X \in (a,b]\}) = P(\{X \in (A,B]\}) = P(\{X \in [a,b]\})$, and similarly for continous random vectors. If you like, think about why, if the probability that X is in an interval is the integral of its probability density function over that interval, that is the case.

Solution.

$$\begin{aligned} Z_n^2 &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j \end{aligned}$$

Because $\mathrm{E}(X+Y) = \mathrm{E}\,X + \mathrm{E}\,Y$, it follows that

$$\mathrm{E}\,Z_n^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\,X_i X_j,$$

and because X_i, X_j for $i \neq j$ are independent, $\mathrm{E}\,X_i X_j = \mathrm{E}\,X_i \cdot \mathrm{E}\,X_j = 0 \cdot 0 = 0$. Thus, including only the terms with $i = j$ and using the bound $\mathrm{E}\,X_i^2 \leq B$, we get

$$\mathrm{E}\,Z_n^2 = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}\,X_i^2 \leq \frac{1}{n^2}\sum_{i=1}^{n} B = \frac{1}{n^2} \cdot n \cdot B = B/n \to 0.$$

A.4 Convergence in Probability

Convergence in mean square implies convergence in probability. To show this, we'll use a very simple but useful trick. Let W be a non-negative random variable, e.g. Z_n^2. Then $1(W \geq t) \leq W/t$, and it follows that $\mathrm{E}\,1(W \geq t) \leq \mathrm{E}\,W/t$. Explain why. Hint: think about what we can say about W/t when $1(W \geq t) = 0$ and when $1(W \geq t) = 1$. Then explain why this, and the result from the previous part, imply the law of large numbers (Theorem 1).

Solution. When $W < t$, $1(W \geq t) = 0$ and W/t is nonnegative and therefore no smaller. When $W \geq t$, $W/t \geq 1$ and therefore exceeds the indicator $1(W \geq t) = 1$. Because $\mathrm{E}\,X \leq \mathrm{E}\,Y$ for any random variables X, Y with $X \leq Y$, it follows that $P(W \geq t) = \mathrm{E}\,1(W \geq t) \leq \mathrm{E}\,W/t$. Taking $W = Z_n^2$, it follows that for any t, $P(Z_n^2 \geq t) \leq \mathrm{E}\,Z_n^2/t$, and we've shown the numerator goes to zero in the previous part.