# Gaussian Width Homework

QTM 490R: Machine Learning Theory

## 1 Introduction

In this week's lecture, we proved a bound on the error of the least squares estimator $\hat{\mu}$ in a convex model. To keep things simple, we focused on a stylized gaussian-noise model.

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

It's unrealistic to expect real data to look exactly like this, but the conclusions we drew based on it remain essentially valid. Showing this will be the focus of this homework. To do this, we'll prove and then use so-called *symmetrization* and *contraction* inequalities, which tell us how to derive new maximal inequalities from ones we already have. These will come in handy later, too.

Let's start by revisiting our argument from lecture without assuming our stylized model. We will, as in lecture, assume that $\mu$ is in our model $\mathcal{M}$. Without any assumptions at all on our noise vector $\varepsilon$, we were able to bound our estimator's error as follows. The starting point for our bound, proven via a little argument relying on convexity, was this. For the vector $\varepsilon$ with elements $\varepsilon_i = Y_i - \mu(X_i)$,

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{when} \quad s^2 > 2 \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle_{L_2(\mathrm{P_n})}.$$

To derive a statement about the probability that the error $\hat{\mu} - \mu$ satisfies this bound, we can use Markov's inequality. Because the maximum above is non-negative, it cannot be too much larger than it's expected value. In particular, letting $c_\delta = 1/\delta$, Markov's inequality implies that

$$\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle < c_\delta \, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle \quad \text{with probability} \quad 1 - \delta$$

and consequently that, with probability $1 - \delta$,

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{for any } s \text{ satisfying} \quad s^2 > 2c_\delta \, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle_{L_2(\mathrm{P_n})}. \tag{1}$$

In the case that $\varepsilon$ is a standard gaussian vector, i.e. a vector with independent standard normal elements $\varepsilon_i$, this expectation is called the *gaussian width* of

the set $\mathcal{M}_s^\circ - \mu = \{m - \mu : m \in \mathcal{M}_s^\circ\}$. We like to assume $\varepsilon$ *is* a standard gaussian vector, or $\sigma$ times one, because vectors like this have properties (like a spherically symmetric distribution) that make this expectation relatively easy to calculate. To take advantage of this while making more realistic assumptions, we'll develop tools that allow us to compare gaussian widths and widths defined in terms of other random vectors $\varepsilon$, i.e.

$$\mathrm{w}_\varepsilon(\mathcal{V}) := \mathrm{E} \max_{v \in \mathcal{V}} \langle \varepsilon, v \rangle_{L_2(\mathrm{P_n})},$$

We'll get there. But let's use this assignment to get comfortable with gaussian width. We'll calculate the gaussian width of neighborhoods in a few models. This will also give us a little context. Unless we actually have some gaussian widths to compare to, there isn't much point in being able to compare things to them.

# 2 Gaussian Width Calculations

In this section, we're going to be talking about two sets of linear functions of $K$-dimensional covariates, i.e., functions of the form $m(x) = x^T\beta$ for $x \in \mathbb{R}^K$. The first, the kind of linear model we talk about in classes like QTM220, will be the set of all of these. Here's how we write it, both as a set of functions and as the set of vectors $[m(X_1), m(X_2), \ldots, m(X_n)] \in \mathbb{R}^n$ that we get by evaluating it at our observations. We'll let $X$ be the $K \times n$ matrix with columns $X_1 \ldots X_n$.

$$\mathcal{M} = \{m(x) = x^T\beta : \beta \in \mathbb{R}^K\} = \{X^T\beta : \beta \in \mathbb{R}^K\} \tag{2}$$

The second, the set of linear functions we work with when we use *the lasso*, is the subset of these with coefficients satisfying a one-norm bound $\|\beta\|_1 \leq B$.

$$\mathcal{M} = \{m(x) = x^T\beta : \|\beta\|_1 \leq B\} = \{X^T\beta : \|\beta\|_1 \leq B\}. \tag{3}$$

We will assume that $\mu$ is a linear function too, so we can write $\mu(x) = x^T\beta_\mu$ for some vector $\beta_\mu \in \mathbb{R}^K$. Consequently, the neighborhood constraint $\|m - \mu\| \leq s\sqrt{n}$ can be written in the form $\|X^T(\beta - \beta_\mu)\|_2 \leq s\sqrt{n}$. Thus in the case of (2),

$$\mathcal{M}_s = \{X^T\beta : \|X^T(\beta - \beta_\mu)\|_2 \leq s\sqrt{n}\} \tag{4}$$

and in the case of (3),

$$\mathcal{M}_s = \{X^T\beta : \|\beta\|_1 \leq 1 \text{ and } \|X^T(\beta - \beta_\mu)\|_2 \leq s\sqrt{n}\}. \tag{5}$$

If you're short on time, it's ok to do one of the two exercises in this section. The second, on the lasso, is probably quicker. But they do teach different things, so I encourage you to give them both a shot if you can.

## 2.1 Linear Model

**Exercise 1** *Find an upper bound on the gaussian width* $\mathrm{w}(\mathcal{M}_s)$ *of a neighborhood* $\mathcal{M}_s$ *in the linear model* (2)*. Then give a bound on the error* $\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})}$ *of the least squares estimator in this model that holds with probability* $1 - \delta$*. You may assume* $\mu$ *is in the model, i.e., that* $\mu(x) = x^T\beta_\mu$ *for some vector* $\beta_\mu$*.*

**Tip.** The centered neighborhood $\mathcal{M}_s - \mu := \{X^T(\beta - \beta_\mu) : \beta \in \mathbb{R}^K \text{ and } \|X^T(\beta - \beta_\mu)\| \leq s\sqrt{n}\}$ may be a little easier to work with than $\mathcal{M}_s$ itself, as we can make a change of variables $v = \beta - \beta_\mu$ and characterize it as $\mathcal{M}_s - \mu = \{X^T v : v \in \mathbb{R}^K \text{ and } \|X^T v\| \leq s\sqrt{n}\}$. How does the gaussian width of $\mathcal{M}_s - \mu$ compare to that of $\mathcal{M}_s$?

**Tip.** You're going to want to use the Cauchy-Schwarz bound, but the bound $\langle \varepsilon, X^T v\rangle_2 \leq \|\varepsilon\|_2 \|X^T v\|_2 \leq \|\varepsilon\|_2 s\sqrt{n}$ isn't going to be good enough. The bound we get this way is the same one we got in lecture for the completely general model. Take a look at Appendix A. If $u_1 \dots u_K$ are orthonormal vectors, what is the distribution of the vector $\varepsilon^u$ with $\varepsilon^u_i = \langle \varepsilon, u_i\rangle$?

## 2.2 The Lasso

**Exercise 2** *Find an upper bound of the gaussian width* $\mathrm{w}(\mathcal{M}_s)$ *of a neighborhood* $\mathcal{M}_s$ *in the model* (3) *used in the lasso. Then use it to give a bound on the error* $\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})}$ *of the least squares estimator in this model that holds with probability* $1 - \delta$*. You may assume* $\mu$ *is in the model, i.e., that* $\mu(x) = x^T\beta_\mu$ *for some vector* $\beta_\mu$*.*

**Tip.** You can get away with being a little loose here. See how far you can get with the observation that because $\mathcal{M}_s \subseteq \mathcal{M}$, $\mathrm{w}(\mathcal{M}_s) \leq \mathrm{w}(\mathcal{M})$. We can't, it turns out, do much better than this. The reason is, in essence, that this model is so 'pointy' that unless $s$ is very small, it contains very few functions with $\|m - \mu\|_{L_2(\mathrm{P_n})} \geq s$ anyway. Section 7.5 of our textbook explains this nicely.[1]

**Tip.** How can we bound the dot product $\langle \varepsilon, X^T\beta\rangle = \varepsilon^T X^T \beta$ when $\|\beta\|_1 \leq B$?

## 2.3 Solutions

**Setting up** Let's start somewhere obvious. Gaussian width is, fundamentally, a property of a set of vectors. So let's characterize the sets of vectors we're talking about in Exercises 1 and 2. When our model is a set of linear functions $m(x) = x^T\beta$ for coefficients $\beta$ in some set $\mathcal{B}$, those are vectors $v$ with components $v_i = x_i^T\beta$, i.e., vectors $X^T\beta$ where $X$ is a matrix with columns $x_i$. And the gaussian width is ($1/n$ times) the *expected value of the maximum dot product* of a vector like this and a gaussian vector. That is, it's

$$\begin{aligned} \frac{1}{n}\,\mathrm{E}\max_{\beta\in\mathcal{B}} \sum_i g_i x_i^T\beta &= \frac{1}{n}\,\mathrm{E}\max_{\beta \ in \mathcal{B}} g^T X^T\beta \\ &= \frac{1}{n}\,\mathrm{E}\max_{\beta \ in \mathcal{B}} (Xg)^T\beta. \end{aligned} \tag{6}$$

[1] Be aware that the book's definition of gaussian width is $n$ times ours. It's $\mathrm{w}(\mathcal{V}) = \mathrm{E}\max_{v\in V}\langle \varepsilon, v\rangle_2$ and ours is $\mathrm{w}(\mathcal{V}) = \mathrm{E}\max_{v\in V}\langle \varepsilon, v\rangle_{L_2(\mathrm{P_n})}$.

What is this set of coefficients $\mathcal{B}$? In Exercise 1, it's the set of coefficients $\beta \in \mathbb{R}^K$ satisfying that

$$s^2 \geq \|m - \mu^\star\|^2_{L_2(\mathrm{P_n})} = \frac{1}{n}\sum_{i=1}^n \left\{x_i^T\beta - x_i^T\beta^\star\right\}^2 = \frac{1}{n}\|X^T(\beta - \beta^\star)\|_2^2 \qquad (7)$$

In Exercise 2, it's the set of coefficients $\beta \in \mathbb{R}^K$ satisfying this *and* $\|\beta\|_1 \leq B$.

**Exercise 2** . Let's start with this one. We'll take the hint and *ignore* the 'neighborhood constrant' (7). That is, we'll use $\mathrm{w}(\mathcal{M})$ as a bound on $\mathrm{w}(\mathcal{M}_s)$. Recall that it is a bound because $\mathcal{M}_s$ is a subset of $\mathcal{M}$ ($\mathcal{M}_s \subseteq \mathcal{M}$) and $\mathrm{w}(A) \leq \mathrm{w}(B)$ if $A \subseteq B$; if we maximize over $\mathcal{M}$ we consider every vector in $\mathcal{M}_s$ as well as some others, so our maximum has to be at least as big.

Now let's do it. It won't be hard. We'll use Hölder's inequality: $(Xg)^T\beta \leq \|\beta\|_1\|Xg\|_\infty$ for all $\beta$ and therefore

$$\mathrm{w}(\mathcal{M}) = \frac{1}{n}\,\mathrm{E}\max_{\beta:\|\beta\|_1\leq B}(Xg)^T\beta \leq \frac{1}{n}\,\mathrm{E}\max_{\beta:\|\beta\|_1\leq B}\|\beta_1\|\|Xg\|_\infty \leq \frac{B}{n}\,\mathrm{E}\|Xg\|_\infty. \quad (8)$$

All we need to do is bound the maximum of $K$ gaussian random variables: $X_{1\cdot}g \ldots X_{K\cdot}g$ where $X_{k\cdot}$ is the $k$th row of $X$. In our Least Squares in Finite Models Lecture, we saw that is these gaussians have variance less than $\sigma^2$, this maximum will be roughly $\sigma\sqrt{\log(K)}$. We could be more precise, but let's just say that for some constant $c$, the expected maximum will be no larger than $c\sigma\sqrt{\log(K)}$.[2] That means all we're got left is a variance calculation.

$$\mathrm{w}(\mathcal{M}) \leq c\frac{B}{n}\sigma \times c\sqrt{\log(K)} \text{ for } \sigma^2 \geq \mathrm{Var}(X_{k\cdot}g) \text{ for all } k = 1 \ldots K$$

What is this variance? Here's a trick. If we have a dot product $u^Tv$, $(u^Tv)^2 = u^Tvv^Tu$. Thus

$$\mathrm{Var}(X_{k\cdot}g) = \mathrm{E}(X_{k\cdot}g)^2 = \mathrm{E}\,X_{k\cdot}^Tgg^TX_{k\cdot} = X_{k\cdot}^T(\mathrm{E}\,gg^T)X_{k\cdot} = X_{k\cdot}^TX_{k\cdot} = \|X_{k\cdot}\|_2^2 \quad \text{because} \quad \mathrm{E}\,gg^T \text{ is the identity}$$

So we can take $\sigma = \max_k\|X_{k\cdot}\|_2$, which is the norm of a vector of length $n$ and perhaps more naturally expressed in terms of its root-mean-squared element. That is, we can say $\sigma = \sqrt{n}\max_k\|X_{k\cdot}\|_{L_2(\mathrm{P_n})}$. And here, in its final form, is our bound.

$$\mathrm{w}(\mathcal{M}) \leq \frac{B}{n} \times (\sqrt{n}\max_k\|X_{k\cdot}\|_{L_2(\mathrm{P_n})}) \times c\sqrt{\log(K)} = cB\max_k\|X_{k\cdot}\|_{L_2(\mathrm{P_n})}\sqrt{\frac{\log(K)}{n}}$$

In particular, if the elements of $X$ are in $[-1, 1]$, we can say our bound is simply $cB\sqrt{\log(K)/n)}$. And we can solve for a neighborhood radius $s$ that contains $\hat{\mu}$ with high probability by solving $s^2 \geq cB\sqrt{\log(K)/n)}$, i.e., by taking $s = \sqrt{cB}\sqrt[4]{\log(K)/n}$. This is, for what it's worth, called the *slow rate* or *assumptionless* analysis of the lasso. There's a fancier argument we can use when $X$ is a very special matrix and most components of $\beta_\star$ are zero that leads to a what is called a *sparsity-dependent fast rate bound*, which can be better.

[2] You can take a look at the posted-but-not-assigned Tail Bounds Homework for a more precise version.

**Exercise 1** I was a bit reluctant to introduce the right tool for this one and that made it way harder than it should be. What we want to use here is the *singular value decomposition* of the matrix $X$. For any matrix $X$, there is a decomposition $X = \sum_{k=1}^{\text{rank}(X)} \sigma_k u_k v_k^T$ where $u_1, u_2, \ldots$ and $v_1, v_2, \ldots$ are orthonormal. You can think of this as saying that there's a natural set of directions to use to think of the 'inputs' and 'outputs' of the matrix multiplication $Xg$: multiplication $X$ takes the input vector $v_k$ to the vector $\sigma_k u_k$ of length $\sigma_k$.

It'll be convenient to think of $u_1, u_2, \ldots$ as an orthogonal basis for $R^n$, and since we only have $\text{rank}(X)$ of them, they won't necessarily be. However, we can go ahead and find $n - \text{rank}(X)$ more orthogonal vectors to *get* a basis $u_1 \ldots u_n$.

Now let's think about the product $g^T X(\beta - \beta_\star)$ that occurs in $\text{w}(\mathcal{M}_s - \mu_\star)$ in these terms. This is the dot product of $g$ with the *output* of a multiplication like this, so it makes sense to decompose $g$ as a combination of the basis vectors $u_1 \ldots u_n$, i.e., $g = \sum_{k=1}^{n} (g^T u_k) u_k$. Something interesting happens: for $k > \text{rank}(X)$ the corresponding term doesn't really affect our dot product because $u_k^T X = 0$. Let's see why.

$$\begin{aligned} g^T X &= \left\{ \sum_{j=1}^{n} (g^T u_j) u_j^T \right\} \left\{ \sum_{k=1}^{\text{rank}(X)} \sigma_k u_k v_k^T \right\} \\ &= \sum_{j=1}^{n} \sum_{k=1}^{\text{rank}(X)} (g^T u_j) \sigma (u_j^T u_k) v_k^T \qquad (9) \\ &= \sum_{k=1}^{\text{rank}(X)} (g^T u_k) \sigma_k v_k^T \quad \text{because} \quad u_j^T u_k = 0 \quad \text{unless} \quad j = k \end{aligned}$$

and therefore

$$g^T X(\beta - \beta_\star) = g_\parallel^T X(\beta - \beta_\star) \quad \text{for} \quad g_\parallel = \sum_{k=1}^{\text{rank}(X)} (g^T u_k) u_k. \qquad (10)$$

Now we're just going to use the Cauchy-Schwarz bound.

$$g^T X(\beta - \beta_\star) = g_\parallel^T X(\beta - \beta_\star) \le \|g_\parallel\|_2 \|X(\beta - \beta_\star)\|_2. \qquad (11)$$

For the vectors $\beta$ we're interested in, i.e. ones satisfying (7), we have $\|X(\beta - \beta_\star)\|_2 \le s\sqrt{n}$, and therefore

$$\text{w}(\mathcal{M}_s) = \frac{1}{n} \text{E} \max_{\beta \ in \mathcal{B}} g^T X^T \beta \le \frac{1}{n} \text{E} \|g_\parallel\|_2 \times s\sqrt{n} = s \frac{E\|g_\parallel\|_2}{\sqrt{n}} \qquad (12)$$

Because $\text{E}\, Z^2 = (\text{E}\, Z)^2 + \text{Var}(Z) \ge (\text{E}\, Z)^2$ for any random variable $Z$ and in particular $Z = \|g_\parallel\|_2$, we'll bounds this by substituting $\sqrt{\text{E}\|g_\parallel\|_2^2} = {}^3\sqrt{\text{rank}(X)}$ for $\text{E}\|g_\parallel\|_2$. This gives us our final bound.

[3] Try the calculation!

$$\mathrm{w}(\mathcal{M}_s) \leq s\sqrt{\frac{\mathrm{rank}(X)}{n}} \tag{13}$$

Solving $s^2 \geq 2s\sqrt{\frac{\mathrm{rank}(X)}{n}}$ gives us the error bound $s = 2\sqrt{\frac{\mathrm{rank}(X)}{n}}$. Note that because the rank of an $n \times K$ matrix is at most $\min(n, K)$, this implies $s = \sqrt{\min(n, K)/n}$.

# A Projections

Here's a way of thinking about the SVD stuff without actually doing SVD.

It's often useful to decompose a vector into relevant and irrelevant parts. For example, if we're interested in an inner product $\langle u, Av\rangle$, it's helpful to decompose $u$ as a sum $u_\parallel + u_\perp$ where $\langle u_\perp, Av\rangle = 0$ for all $v$. This is particularly nice if we're going to use a Cauchy-Schwarz bound, as we can get a better bound by first getting rid of the irrelevant part $u_\perp$.

$$\langle u, Av\rangle = \langle u_\parallel, Av\rangle + \langle u_\perp, Av\rangle = \langle u_\parallel, Av\rangle \leq \|u_\parallel\|\|Av\|.$$

The best way to do this, in the sense that $\|u_\parallel\|$ is smallest, is to take $u_\parallel$ to be the *orthogonal projection* onto the *image* of $A$—the image of $A$ is the set of all vectors we can write as matrix-vector projects $Av$. To do that, we'll want an orthonormal basis for the image of $A$, i.e., a set of vectors $u_1, u_2, \ldots$ with the property that $\langle u_i, u_j\rangle$ is one if $i = j$ and zero otherwise. To get a basis like this, we can run any set of vectors that spans the image of $A$, e.g. the columns of $A$, through the Gram-Schmidt Process. Then we write $u_\parallel$ as a linear combination of these vectors, $u_\parallel = \sum_k u_k\langle u_k, u\rangle$. To check that $\langle u_\parallel, Av\rangle = \langle u, Av\rangle$ for all $v$, observe that because $u_1, u_2, \ldots$ is a basis for the image of $A$, we can express $Av$ as a linear combination $\sum_k \alpha_k u_k$ of these basis vectors. And we can calculate $\langle u, Av\rangle$ and $\langle u_\parallel, Av\rangle$ and compare. They're the same.

$$\langle u, Av\rangle = \left\langle u, \sum_k \alpha_k u_k \right\rangle = \sum_k \alpha_k \langle u, u_k\rangle$$

$$\langle u_\parallel, Av\rangle = \left\langle \sum_j u_j\langle u_j, u\rangle, \sum_k \alpha_k u_k \right\rangle = \sum_j \sum_k \alpha_k \langle u_j, u_k\rangle\langle u_j, u\rangle = \sum_k \alpha_k \langle u_k, u\rangle$$

When we simplified the double sum above, we observed that terms with $j \neq k$ were zero because $\langle u_j, u_k\rangle = 0$ and that $\langle u_j, u_k\rangle = 1$ in terms with $j = k$.

I'll leave it to you to convince yourself that this is the best we can do, i.e., that there is no vector $\tilde{u}_\parallel$ satisfying $\langle \tilde{u}_\parallel, Av\rangle = \langle u, Av\rangle$ for all $v$ with $\|\tilde{u}_\parallel\| < \|u_\parallel\|$.

This all works with any inner product $\langle u, v\rangle$ and associated norm $\|v\| = \sqrt{\langle v, v\rangle}$. In this homework, we'll use it to talk about the dot product between gaussian vectors and vectors of the form $Av$. Note that if $A$ is a $m \times n$ matrix, then our basis $u_1, u_2, \ldots$ contains at most $\min(m, n)$ vectors.