Least Squares and Gaussian Width

March 26, 2025

1 Introduction

1.1 Review

In this week's lectures, we proved a bound on the error of the least squares estimator $\hat{\mu}$ in a convex model.

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 \quad \text{where} \quad \mathcal{M} \text{ is a convex set .} \tag{1}$$

To keep things simple, we focused on a stylized gaussian-noise model.

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

And we proved that the following high probability error bound in terms of the gaussian width of a centered neighborhood of μ.

$$\begin{aligned} &\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{w.p.} \quad 1-\delta \quad \text{if} \quad \frac{s^2}{2\sigma} \geq \mathrm{w}(\mathcal{M}_s^\circ - \mu) + s\sqrt{\frac{2}{\delta n}} \\ &\quad \text{when} \quad \mu \in \mathcal{M}. \end{aligned} \tag{2}$$

Here $\mathcal{M}_s^\circ = \{m - \mu : m \in \mathcal{M} \text{ and } \|m - \mu\| = s\}$. Futhermore, we showed that even if μ is not in the model, we have a bound like this on the distance between our estimator ($\hat{\mu}$) and our model's best approximation to the signal ($\mu_\star$).

$$\begin{aligned} &\|\hat{\mu} - \mu_\star\|_{L_2(\mathrm{P_n})} < s \quad \text{w.p.} \quad 1-\delta \quad \text{if} \quad \frac{s^2}{2\sigma} \geq \mathrm{w}(\mathcal{M}_s^\circ - \mu) + s\sqrt{\frac{2}{\delta n}} \\ &\quad \text{for} \quad \mu_\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\|_{L_2(\mathrm{P_n})}. \end{aligned} \tag{3}$$

Here $\mathcal{M}_s^\circ = \{m - \mu_\star : m \in \mathcal{M} \text{ and } \|m - \mu_\star\| = s\}$. This is a generalization of the previous bound. When $\mu \in \mathcal{M}$, $\mu_\star = \mu$ and (3) is equivalent to (2).

I also claimed (without proof) that this implies the following bound. The advantage is that the inequality characterizing s is a bit simpler.

$$\begin{aligned} &\|\hat{\mu} - \mu_\star\|_{L_2(\mathrm{P_n})} < s + 2\sigma\sqrt{\frac{2}{\delta n}} \quad \text{w.p.} \quad 1-\delta \quad \text{if} \quad \frac{s^2}{2\sigma} \geq \mathrm{w}(\mathcal{M}_s - \mu) \\ &\quad \text{for} \quad \mu_\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\|_{L_2(\mathrm{P_n})}. \end{aligned} \tag{4}$$

Here $\mathcal{M}_s = \{m - \mu_\star : m \in \mathcal{M} \text{ and } \|m - \mu_\star\| \leq s\}$.

1.2 Assignment Summary

In this assignment, we'll take a few steps toward getting concrete, meaningful error bounds. Because it's unrealistic to expect real data to look exactly like signal plus gaussian noise, we'll derive a more meaningful version of (3) (and consequently (4)) in Section 3: a bound that holds when $\epsilon_1 \ldots \epsilon_n$ are independent with mean zero, but don't have to be gaussian or even all have the same distribution. And because our error bound (3) is a little too abstract to make sense directly, we'll bound the gaussian width of neighborhoods $\mathcal{M}_s - \mu_\star$ in a few models and use the result to derive concrete model-specific error bounds. To prepare for all that, we'll start by proving a few properties of gaussian width. And, for good measure, we'll use them to derive our simplified error bound (4) from the one we proved in lecture (3).

2 Properties of Gaussian Width

In this section, we'll prove a few properties of gaussian width.

$$\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g,\ v \rangle \quad \text{for} \quad g_i \overset{iid}{\sim} N(0,1).$$

2.1 Basic Properties

- It's *increasing*. It's a maximum over the set $\mathcal{V}$, so it gets bigger if $\mathcal{V}$ does.
$$\mathrm{w}(\mathcal{V}) \le \mathrm{w}(V^+) \quad \text{if} \quad \mathcal{V} \subseteq \mathcal{V}^+$$
- It's *homogeneous*. If we scale the vectors in $\mathcal{V}$, we scale its width.
$$\mathrm{w}(\alpha\mathcal{V}) = \alpha\, \mathrm{w}(\mathcal{V}) \quad \text{where} \quad \alpha\mathcal{V} := \{\alpha v : v \in \mathcal{V}\} \quad \text{for} \quad \alpha \ge 0.$$
- It's *translation invariant*. It doesn't care about how we center our vectors.
$$\mathrm{w}(\mathcal{V} + x) = \mathrm{w}(V) \quad \text{for} \quad \mathcal{V} + x := \{v + x : v \in \mathcal{V}\}.$$

Exercise 1 *Prove that gaussian width* $\mathrm{w}(\mathcal{V})$ *has these three properties.*

Notation. Often, because it's a bit more compact, we'll write $s^2 \ge 2\sigma\, \mathrm{w}(\mathcal{M}_s^\circ)$ instead of $s^2 \ge 2\sigma\, \mathrm{w}(\mathcal{M}_s^\circ - \mu_\star)$ in bounds like (3).

Exercise 2 *Explain why it makes no difference whether we write* $\mathrm{w}(\mathcal{M}_s^\circ - \mu_\star)$ *or* $\mathrm{w}(\mathcal{M}_s^\circ)$. *A sentence should do.*

2.2 Sublinearity

Exercise 3 *Let* $\mathcal{M}$ *be a convex set,* $\mu_\star$ *be a point in* $\mathcal{M}$*, and* ρ *be a seminorm defined on its elements. Prove that, for a neighborhood* $\mathcal{M}_s = \{m - \mu_\star \in \mathcal{M} : \rho(m - \mu_\star) \le s\}$ *of* $\mu_\star$*,* $f(s) = \mathrm{w}(\mathcal{M}_s - \mu_\star)$ *is a sublinear function of* s*. That is, prove that* $f(s)/s$*, a function on the positive real numbers, is (non-necessarily-strictly) decreasing.*

Tips.

1. $f(s)/s$ is decreasing if $f(s)/s \ge f(t)/t$ [or equivalently $f(s) \ge (s/t)f(t)$] whenever $s \le t$. Is $(s/t)f(t)$ the gaussian width of some set? If so, what set? And how is it related to $\mathcal{M}_s - \mu_\star$? Use the properties of gaussian width you proved in Exercise 1.
2. It's important that $\mathcal{M}$ is a convex set containing $\mu_\star$. Why? If m is in $\mathcal{M}$, then so is $m_\lambda = \mu_\star + \lambda(m - \mu_\star)$ for any $\lambda \in [0,1]$. Or equivalently, if $m - \mu_\star$ is in $\mathcal{M} - \mu_\star$, so is $m_t - \mu_\star = \lambda(m - \mu_\star)$.[1]

Now you should have what you need to prove that (3) implies (4).

Exercise 4 *Prove that if* $s^2 \ge 2\sigma\, \mathrm{w}(\mathcal{M}_s - \mu_\star)$*, then* $(s + x)^2 \ge 2\sigma\, \mathrm{w}(\mathcal{M}_{s+x} - \mu_\star) + sx$ *for any* $x \ge 0$*. Then briefly explain why this and* (3) *together imply* (4)*. A sentence or two should be enough for this explanation.*

Tip. You can get a condition equivalent to the one you want to show by dividing both sides by $s+x$.

$$\begin{aligned}(s+x)^2 \geq 2\sigma\,\mathrm{w}(\mathcal{M}_{s+x}-\mu_\star)+sx \qquad &\text{if and only if}\\ s+x \geq 2\sigma\frac{\mathrm{w}(\mathcal{M}_{s+x}-\mu_\star)}{s+x}+\frac{s}{s+x}x.\end{aligned}$$

Looking at this equivalent condition, compare the first term on the left side to the first term on the right and the second term on the left side to the second term on the right.

3 A More Realistic Error Bound

Setting. In this section, we will consider the case that we observe pairs $(X_1, Y_1) \ldots (X_n, Y_n)$ where $X_1 \ldots X_n$ are deterministic and $Y_i = \mu(X_i) + \varepsilon_i$ for $\varepsilon_1 \ldots \varepsilon_n$ that are independent, but not necessarily identically distributed, random variables with $\mathrm{E}\, \varepsilon_i = 0$.

What's more realistic about this? It describes the kind of data we get with an actual usable sampling mechanism. If we draw pairs $(X_1, Y_1) \ldots (X_n, Y_n)$ *uniformly at random with replacement* from a population $(x_1, y_1) \ldots (x_m, y_m)$, then do our analysis *conditioning on* $X_1 \ldots X_n$, this is the setting we find ourselves in. In that case, our signal is $\mu(x) = \mathrm{E}[Y_i \mid X_i = x] = \frac{1}{m_x} \sum_{j: x_j = x} y_j$, the average outcome among people in the population with $x_j = x$. And the high probability error bounds we prove hold with *conditional probability* 1-δ.[1]

In this setting, we can prove the following error bound in terms of the random vector $\varepsilon \in \mathbb{R}^n$ with ith element ε_i.

$$
\begin{aligned}
&\|\hat{\mu} - \mu_\star\|_{L_2(\mathrm{P_n})} < s + 2\sigma\sqrt{\frac{2}{\delta n}} \quad \text{w.p.} \quad 1 - \delta \quad \text{if} \quad \frac{s^2}{2} \geq \mathrm{w}_\varepsilon(\mathcal{M}_s^\circ - \mu_\star) \\
&\quad \text{for} \quad \mu_\star = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \|m - \mu\|_{L_2(\mathrm{P_n})}, \\
&\qquad \mathrm{w}_\varepsilon(\mathcal{V}) := \mathrm{E} \max_{v \in \mathcal{V}} \langle \varepsilon, v \rangle_{L_2(\mathrm{P_n})}, \\
&\quad \text{and} \quad \sigma^2 \geq \mathrm{E}\, \varepsilon_i^2 \text{ for all } i.
\end{aligned} \tag{5}
$$

We'll start with a warm-up.

Exercise 5 *Prove, by plugging in* $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$*, that this error bound* (5) *implies the bound we use in the gaussian case* (2)*. A sentence or two should do.*

Tip. Recall that if $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, then $\varepsilon_i = \sigma g_i$ for $g_i \overset{iid}{\sim} N(0, 1)$.

It's time to prove the new bound. Virtually everything you'll need can be borrowed from lecture, but I'm going to ask you to write a complete proof, copying out the parts you need to. This is meant as encouragement to review the proofs from lecture and understand how they work.

Exercise 6 *Write out a complete proof of the new error bound* (5)*. You don't need to include a proof of the Efron-Stein inequality or anything you've proven above in Section 2, but everything else should be included.*

To make this bound useful, we'll need to bound $\mathrm{w}_\varepsilon(\mathcal{M}_s^\circ - \mu_\star)$ for some models $\mathcal{M}$. Today, we'll do that for the gaussian case $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Later

[1] This implies they hold with unconditional probability $1 - \delta$ too. For any event A and conditioning set B, by the law of iterated expectations, the probability of any event A is the expected value of the conditional probability of A given B: $P(A) = \mathrm{E}[1_A] = \mathrm{E}[\mathrm{E}[1_A \mid B]] = \mathrm{E}[P(A \mid B)]$.

in the semester, we'll prove bounds of the form $\mathrm{w}_\varepsilon(\mathcal{V}) \le \alpha\,\mathrm{w}(\mathcal{V})$ that hold for every set $\mathcal{V}$ with a constant α that depends on the distribution of ε. This'll let us use our gaussian width bounds together with (5) to get concrete, realistic error bounds.

4 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 $X_1 \ldots X_n$.

$$\begin{aligned} \mathcal{M} &= \{m(x) = x^T\beta : \beta \in \mathbb{R}^K\} && \text{as a set of functions} \\ &= \{X^T\beta : \beta \in \mathbb{R}^K\} && \text{as a set of vectors} \end{aligned} \tag{6}$$

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 $\|\beta\|_1 \le B$.

$$\begin{aligned} \mathcal{M} &= \{m(x) = x^T\beta : \beta \in \mathbb{R}^K \text{ and } \|\beta\|_1 \le B\} && \text{as a set of functions} \\ &= \{X^T\beta : \beta \in \mathbb{R}^K \text{ and } \|\beta\|_1 \le B\} && \text{as a set of vectors} \end{aligned} \tag{7}$$

Throughout, we'll focus on the gaussian noise case: $\boldsymbol{\varepsilon_1 \ldots \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)}$.

4.1 The Linear Model

Exercise 7 *Find an upper bound on the gaussian width* $\mathrm{w}(\mathcal{M}_s - \mu_\star)$ *of a centered neighborhood* $\mathcal{M}_s - \mu_\star$ *in the linear model* (6). *Then give a bound on the error* $\|\hat{\mu} - \mu_\star\|_{L_2(\mathrm{P_n})}$ *of the least squares estimator in this model that holds with probability* $1 - \delta$.

Tips.

1. It'll be convenient to work with the Euclidean norm $\|\cdot\|_2$ and inner product $\langle \cdot, \cdot \rangle_2$ instead of the sample two norm and inner product. Rewrite the $s^2 \ge \max\ldots$ condition in (4) in terms of these using the scaling-up relationships

$$\|\cdot\|_2 = \sqrt{n}\|\cdot\|_{L_2(\mathrm{P_n})} \quad \text{and} \quad \langle \cdot, \cdot \rangle_2 = n\langle \cdot, \cdot \rangle_{L_2(\mathrm{P_n})}.$$

   $\mathcal{M}_s - \mu_\star = \{X^Tv : v \in \mathbb{R}^K \;\text{ and }\; \|X^Tv\|_2 \le s\sqrt{n}\}$. Why?

2. You're going to want to use the Cauchy-Schwarz bound, but the bound $\langle \varepsilon, X^Tv \rangle_2 \le \|\varepsilon\|_2 \|X^Tv\|_2 \le \|\varepsilon\|_2 s\sqrt{n}$ 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. What is $\mathrm{E}\|\varepsilon^u\|_2$ where $\varepsilon^u = \sum_{j=1}^K \langle u_i, \varepsilon \rangle_2 u_i$ is the projection of ε onto the span of the columns of X?

3. For any random variable Z including $Z = \|\varepsilon^u\|_2$,

$$\mathrm{E}\, Z^2 = (\mathrm{E}\, Z)^2 + \mathrm{Var}(Z) \quad \text{and therefore} \quad (\mathrm{E}\, Z)^2 \le \mathrm{E}\, Z^2.$$

4.2 The Lasso

Exercise 8 *Find an upper bound of the gaussian width* $\mathrm{w}(\mathcal{M})$ *of the model* (7) *used in the lasso. Then use it to give a bound on the error* $\|\hat{\mu} - \mu_\star\|_{L_2(\mathrm{P_n})}$ *of the least squares estimator in this model that holds with probability* $1 - \delta$.

Tips.

1. How can we bound the dot product $\langle \varepsilon, X^T \beta \rangle_2 = \langle X\varepsilon, \beta \rangle_2$ when $\|\beta\|_1 \leq B$? Look over the Inner Product Spaces Homework and Appendix B.

2. You can get away with using the bound $\mathrm{w}(\mathcal{M}_s - \mu) \leq \mathrm{w}(\mathcal{M})$ when calculating your error bound. It turns out we can't 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 $\|m - \mu\|_{L_2(\mathrm{P_n})} > s$ anyway. Section 7.5 of High Dimensional Probability explains this nicely.[2]

A Projections

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 $u_{\|} + u_{\perp}$ 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 $u_{\perp}$.

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

The best way to do this, in the sense that $\|u_{\|}\|$ is smallest, is to take $u_{\|}$ 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 $u_{\|}$ as a linear combination of these vectors, $u_{\|} = \sum_k u_k \langle u_k, u \rangle$. To check that $\langle u_{\|}, Av \rangle = \langle u, Av \rangle$ 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 $\langle u_{\|}, Av \rangle$ and compare. They're the same.

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

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

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 $\tilde{u}_{\|}$ satisfying $\langle \tilde{u}_{\|}, Av \rangle = \langle u, Av \rangle$ for all v with $\|\tilde{u}_{\|}\| < \|u_{\|}\|$.

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.

B Bounding Expectations by Integrating Tail Bounds

Going from bounds on tail probabilities to bounds on expectations is basically just a matter of integration, as $\mathrm{E}\, Z = \int_0^\infty P(Z > z) dz$ for any positive random variable Z. Here's a proof.

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

More generally, $\mathrm{E}\, Z \leq \int_0^\infty P(Z > z) dz$ for any random variable Z. To show that, we can use the formula for nonnegative random variables on $Z_+ = \max\{Z,\, 0\}$ and observe that (i) $\mathrm{E}\, Z \leq \mathrm{E}\, Z_+$ and (ii) $P(Z_+ > z) = P(Z > z)$ for $z \geq 0$.

Let's use it to derive a bound on the expected value of the maximum $M_K = \max_{j \in 1 \ldots K} Z_j$ of K mean-zero normals $Z_j \sim N(0, \sigma^2)$. Or to be precise, from tail bound $P(M_K \geq z) \leq K e^{-z^2/2\sigma^2}$ that we substituted $z = 2\sqrt{\log(K)}$ into to get the bound $P(M_K \geq 2\sigma\sqrt{\log(K)}) \leq 1/K$ in our lecture on least squares in finite models. Here's the bound.

$$\mathrm{E}\, M_K \leq 2\sigma\sqrt{2\log(K)}.$$

To prove it, we'll break the integral $\mathrm{E}\, M_K = \int_0^\infty P(M_K > z) dz$ into a sum of two integrals, one up to z_0 and one from there to ∞. We'll bound the first using the simple observation that probabilities are less than one. And we'll bound the second using the fact that $z/z_0 \geq 0$ on the domain of integration $[z_0, \infty)$ and the identity $(d/dz) e^{-z^2/2} = -z e^{-z^2/2}$.

$$\begin{aligned}
\mathrm{E}\, M_K &= \int_0^{z_0} P(M_K \geq z) + \int_{z_0}^\infty P(M_K \geq z) \\
&\leq \int_0^{z_0} 1 + \int_{z_0}^\infty \frac{z}{z_0} K e^{-z^2/2\sigma^2} \\
&\leq z_0 + \frac{\sigma^2 K}{z_0} \int_{z_0}^\infty \frac{z}{\sigma^2} e^{-z^2/2\sigma^2} \\
&= z_0 - \frac{\sigma^2 K}{z_0} \int_{z_0}^\infty \frac{d}{dz} e^{-z^2/2\sigma^2} \\
&= z_0 - \frac{\sigma^2 K}{z_0} e^{-z^2/2\sigma^2} |_{z_0}^\infty \\
&= z_0 + \frac{\sigma^2 K e^{-z_0^2/2\sigma^2}}{z_0}.
\end{aligned}$$

Taking $z_0 = \sigma\sqrt{2\log(K)}$, $e^{-z_0^2/2\sigma^2} = e^{-\log(K)} = 1/K$. And we get this bound.

$$\mathrm{E}\, M_K \leq \sigma\sqrt{2\log(K)} + \frac{\sigma^2 K \times 1/K}{\sigma\sqrt{2\log(K)}} = \sigma\left(\sqrt{2\log(K)} + \frac{1}{\sqrt{2\log(K)}}\right).$$

The bound $\operatorname{E} Z_K \leq 2\sigma\sqrt{2\log(K)}$ is a simplified version of this. Let's focus on the case that $K \geq 2$. Because $2\log(K) \geq 2\log(2) > 1$ for $K \geq 2$, the second term in curly brackets is smaller than the first, so their sum is bounded by twice the first. Thus, $\operatorname{E} Z_K \leq 2\sigma\sqrt{2\log(K)}$ as claimed.

This bound applies for $K = 1$ as well, as in that case $M_K = Z_1$ and $\operatorname{E} M_K = \operatorname{E} Z_1 = 0$.