Machine Learning Theory

Chaining

David A. Hirshberg

April 19, 2025

Emory University

Notation

Today, we're using the sample mean inner product and sample mean squared error. To keep notation simple, we're going to write this without any special subscripts.

$$\langle u, v \rangle = \langle u, v \rangle_{L_2(\mathrm{P_n})} = \frac{1}{n} \sum_{i=1}^{n} u(X_i) v(X_i)$$

$$\|v\|^2 = \|v\|^2_{L_2(\mathrm{P_n})} = \frac{1}{n} \sum_{i=1}^{n} v(X_i)^2.$$

Keep in mind that for a gaussian vector $g \sim N(0, I_{n \times n})$ and any function v,

$$\langle g,\ v \rangle = \frac{1}{n} \sum_{i=1}^{n} g_i\ v(X_i) \sim N\left(0,\ \frac{\|v\|^2}{n}\right).$$

We'll also write $\mathcal{M}_s$ as a shorthand for what we've called $\mathcal{M}_s - \mu^\star$ before.

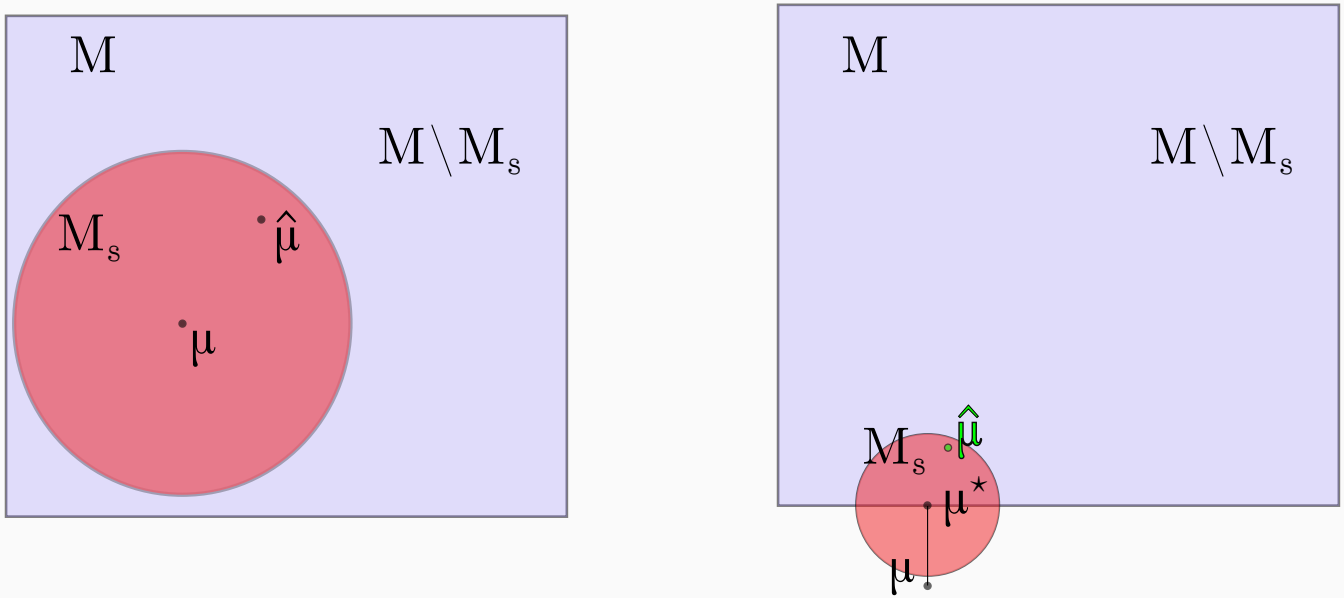
$$\mathcal{M}_s = \{m - \mu^\star : \|m - \mu^\star\| \leq s\}.$$

And we'll ignore some constant factors: $a_n \lesssim b_n$ means $a_n \leq c b_n$ for some constant c.

Review

What determines our error bounds

It's the gaussian width of neighborhoods of $\mu^\star$ in our model.



$$\|\hat{\mu} - \mu^\star\| < s \times \sigma\left\{1 + \frac{2\Sigma_n}{\delta n}\right\} \text{ w.p. } 1-\delta \text{ if } \frac{s^2}{2} \geq \mathrm{w}(\mathcal{M}_s).$$

$$\mathcal{M}_s = \{m - \mu^\star \in \mathcal{M} : \|m - \mu^\star\| \leq s\}.$$

So what we need is a way to bound the gaussian width of these neighborhoods.

Finite Models

- In finite models, bounding width is easy.
- It's the maximum of gaussians with standard deviation $\leq s/\sqrt{n}$.

$$\mathrm{E}\langle g, m - \mu^\star \rangle^2 = \frac{\|m - \mu^\star\|^2}{n}.$$

- We can bound this via union bound. It's down to counting the curves in the model.

$$\mathrm{w}(\mathcal{M}_s) \lesssim s\sqrt{\frac{\log(K)}{n}} \quad \text{if } \mathcal{M} \quad \text{contains } K \text{ curves } m$$

$$\text{all with} \quad \|m - \mu^\star\|_{L_2(\mathrm{P_n})} \leq s.$$

- We may be overcounting. This bounds the max of K totally different gaussians.
- That's the case in which it's largest, so if there's correlation we're overcounting.
- And this definitely won't work for models with infinitely many curves.
- We'll need to take advantage of this correlation to tackle infinite models.

Counting Curves in Infinite Models

Gaussian width is the mean of the *maximum* of a set of gaussians.

$$\mathrm{w}(\mathcal{M}_s) = \mathrm{E} \max_{v \in \mathcal{M}_s} \langle g, v \rangle \quad \text{for} \quad g \sim N(0, I_{n \times n}).$$

And the difference between many of these gaussians $\langle g, v \rangle$ will be small.

- So small, sometimes, that we don't need to 'pay probability' to bound them all using the union bound. They needn't contribute to K.
- We can just use the Cauchy-Schwarz inequality to bound differences.

$$|\langle g, u \rangle - \langle g, v \rangle| = |\langle g, u - v \rangle| \leq \|g\| \|u - v\| \approx \|u - v\|.$$

If the curves u and v are *close enough*, by bounding $\langle g, u \rangle$, we bound $\langle g, v \rangle$ *for free*.

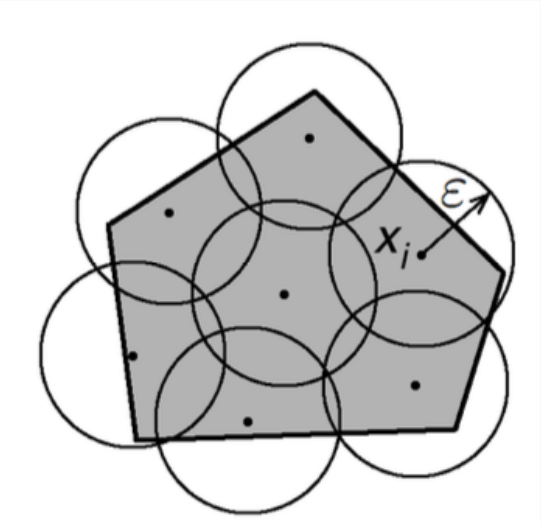
- This means we can take K above to be smaller than the total number of curves.
- It's enough that some set $u_1 \ldots u_K$ gets close enough to all curves $v \in \mathcal{M}$.

This means we have to talk about how many *meaningfully different* curves we have.

ϵ-covers and snapping

We can quantify this using a set's **ϵ-covering number** K_ϵ.
That's the number of balls of size ϵ of radius ϵ it takes to cover the set.
That is, it's the size of the set's smallest ϵ-cover.

We call a set $\mathcal{V}^\epsilon$ an **ϵ-cover** for the set $\mathcal{V}$ if every curve in the set $\mathcal{V}$ is within a distance ϵ of some curve in $\mathcal{V}^\epsilon$.



We can think of this as the set of curves we get by *snapping* each curve in $\mathcal{V}$ to one of finitely many curves—one that's an approximation with error $\leq \epsilon$.

$$\mathcal{V}_\epsilon = \{\pi_\epsilon(v) \ : \ v \in \mathcal{V}\} \quad \text{where} \quad \|\pi_\epsilon(v) - v\| \leq \epsilon$$

I'll call the function π_ϵ that does this an *ϵ-snapping map*.
That's not standard terminology. As far as I know there isn't a standard name for this.

Q

If we've got an ϵ-snapping map, we've got an ϵ-cover.

$$\mathcal{V}_\epsilon = \{\pi_\epsilon(v) \ : \ v \in \mathcal{V}\} \quad \text{where} \quad \|\pi_\epsilon(v) - v\| \leq \epsilon$$

We can go the other way, too.
If we've got an ϵ-cover, we can define an ϵ-snapping map. How?

If we've got an ϵ-snapping map, we've got an ϵ-cover.

$$\mathcal{V}_\epsilon = \{\pi_\epsilon(v) \ : \ v \in \mathcal{V}\} \quad \text{where} \quad \|\pi_\epsilon(v) - v\| \leq \epsilon$$

We can go the other way, too.
If we've got an ϵ-cover, we can define an ϵ-snapping map. How?

We snap to the closest curve in our cover.

$$\pi_\epsilon(v) = \operatorname*{argmin}_{v_\epsilon \in \mathcal{V}^\epsilon} \|v_\epsilon - v\|$$

This means snapping maps and covers are more-or-less interchangeable.

Terminology.
I'll refer to the *size* of a snapping map as the size of the cover induced by it, i.e., the number of different curves it outputs.

Snapping and Gaussian Width

If we have an ϵ-snapping map of size K_ϵ for a set $\mathcal{V}$, then we've got a bound on its gaussian width. We use ϵ-closeness together with with our bound for finite sets.

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &= \mathrm{E} \max_{v \in \mathcal{V}} \langle g,\ v \rangle \\
&= \mathrm{E} \max_{v \in \mathcal{V}} \{ \langle g,\ v - \pi_\epsilon(v) \rangle + \langle g, \pi_\epsilon(v) \rangle \} \\
&\lesssim \underbrace{\|v - \pi_\epsilon(v)\|}_{\leq \epsilon} + \max_{v \in \mathcal{V}} \|\pi(v)\| \sqrt{\frac{\log(K_\epsilon)}{n}}
\end{aligned}$$

When we're talking about a centered neighborhood $\mathcal{V} = \mathcal{M}_s - \mu$, this second term is small because $\|\pi(v)\|$ is small for every $v \in \mathcal{V}$.

$$\|\pi(v)\| \leq \|v - \pi(v)\| + \|v\| \leq \underbrace{\epsilon + s \leq 2s}_{(\text{or } \log(K_\epsilon)=0)} \quad \text{by the triangle inequality}$$

and therefore

$$\mathrm{w}(\mathcal{M}_s - \mu) \lesssim \epsilon + 2s \sqrt{\frac{\log(K_\epsilon)}{n}}$$

Gaussian width doesn't change when we center, so the same bound holds for the neighborhood itself.

Dissatisfying Implications

- We showed last class that $\log(K_\epsilon) \approx 1/\epsilon$ for the Lipschitz model.
- If we choose the resolution ϵ to minimize our bound, it's roughly $\sqrt[3]{s^2/n}$.

$$\mathrm{w}(\mathcal{M}_s) \lesssim \epsilon + s\sqrt{\frac{\log(K_\epsilon)}{n}} \approx \epsilon + \frac{s}{\sqrt{\epsilon n}} \approx s^{2/3} n^{-1/3} \quad \text{at optimal} \quad \epsilon \approx s^{2/3} n^{-1/3}.$$

- This tells us that our estimator converges at a fourth-root rate.

$$s^2 \geq \mathrm{w}(\mathcal{M}_s) \quad \text{if} \quad s^2 \gtrsim s^{2/3} n^{-1/3} \quad \text{i.e. if} \quad s \approx n^{-1/4}.$$

- But we know it converges faster.
- The Lipschitz model is contained in the Sobolev model of order 1.
- And we proved the rate of convergence $s \approx n^{-1/3}$ for that using Fourier series.

We can do better by looking at covering numbers *at multiple resolutions*.

$$\mathrm{w}(\mathcal{M}_s) \lesssim \frac{1}{\sqrt{n}} \int_0^s \sqrt{\log(K_\epsilon)}\, d\epsilon$$

This is called *Dudley's Integral Bound*. Today we'll prove it.

Dudley's Integral Bound

$$\mathrm{w}(\mathcal{M}_s) \lesssim \frac{1}{\sqrt{n}} \int_0^s \sqrt{\log(K_\epsilon)}\, d\epsilon$$

- It's based on an idea called *chaining*.
- The idea is to use approximations $\pi_0(m), \pi_1(m), \ldots$ at increasing resolutions $\epsilon_0, \epsilon_1, \ldots$.
- We write each function as a sum of differences between finer and finer approximations.

$$m = \pi_0(m) + \sum_{j=0}^{\infty} \pi_{j+1}(m) - \pi_j(m)$$

- We call these differences *links* in a chain that goes
  - from the coarsest approximation, $\pi_0(m)$, which is the same for all functions.
  - to the finest approximation, $m = \pi_\infty(m)$ itself.
- Before we dig into this too much, let's warm up.

Warm-up

Our One-Link Bound

Think about the width bound implied by an ϵ-snapping map π_ϵ for very small ϵ.

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &\leq \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v - \pi_\epsilon(v) \rangle + \mathrm{E} \max_{v \in \mathcal{V}} \langle g, \pi(v) \rangle \\
&\qquad \text{our link} \\
&\leq \mathrm{E}\|g\| \max_{v \in \mathcal{M}_s} \|v - \pi_\epsilon(v)\| + \mathrm{E} \max_{v \in \mathcal{V}} \langle g, \pi(v) \rangle \\
&\lesssim \epsilon + \mathrm{rad}(\mathcal{V}) \sqrt{\frac{\log(K_\epsilon)}{n}} \qquad \text{where} \qquad \mathrm{rad}(\mathcal{V}) = \max_{v \in \mathcal{V}} \|v\|
\end{aligned}$$

- This is what we've been doing. But we have a sense that we're being wasteful.
- When our ϵ-cover is fine, it'll contain vectors that are close to one another.
- The corresponding gaussians will be highly correlated, so our $\sqrt{\log(K)}$ bound on their maximum will be loose. Our second term will be bigger than we want.

We could reduce K_ϵ by snapping to *coarser approximations*—taking ϵ to be large. But that makes our first term big.

We can do better by using two approximations—one coarse and one fine.

$$\langle g, \pi_\epsilon(v) \rangle = \langle g, \pi_\epsilon(v) - \pi_{\epsilon'}(v) \rangle + \langle g, \pi_{\epsilon'}(v) \rangle$$

a new link

where $\pi_{\epsilon'}(v)$ is a snapping map that gives *coarser approximations*. One with coarser resolution $\epsilon' \geq \epsilon$ and therefore smaller size $K'_\epsilon \leq K_\epsilon$. We bound the pieces as before.

A Two-Link Bound

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &\lesssim \mathrm{E}\max_{v\in\mathcal{V}} \langle g, v - \pi(v)\rangle + \mathrm{E}\max_{v\in\mathcal{V}} \langle g, \pi(v) - \pi'(v)\rangle + \mathrm{E}\max_{v\in\mathcal{V}} \langle g, \pi'(v) - \mu^\star\rangle \\
&\qquad \text{old link} \qquad\qquad\qquad \text{new link} \\
&\lesssim \max_{v\in\mathcal{V}} \|v - \pi(v)\| + \underbrace{\max_{v\in\mathcal{V}} \|\pi(v) - \pi'(v)\|}_{\leq \epsilon+\epsilon'} \sqrt{\frac{\log(K_\epsilon K_{\epsilon'})}{n}} + \underbrace{\max_{v\in\mathcal{V}} \|\pi'(v)\|}_{\leq \mathrm{rad}(\mathcal{V})+\epsilon'} \sqrt{\frac{\log(K'_\epsilon)}{n}} \\
&\approx \epsilon + \epsilon' \sqrt{\frac{\log(K_\epsilon)}{n}} + \mathrm{rad}(\mathcal{V}) \sqrt{\frac{\log(K'_\epsilon)}{n}}.
\end{aligned}$$

Q: Where do we get this second bound with $\log(K_\epsilon K_{\epsilon'})$?

A Two-Link Bound

$$
\begin{aligned}
\mathrm{w}(\mathcal{V}) &\lesssim \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v - \pi(v) \rangle + \mathrm{E} \max_{v \in \mathcal{V}} \langle g, \pi(v) - \pi'(v) \rangle + \mathrm{E} \max_{v \in \mathcal{V}} \langle g, \pi'(v) - \mu^\star \rangle \\
&\qquad \text{old link} \qquad\qquad\qquad\qquad \text{new link} \\
&\lesssim \max_{v \in \mathcal{V}} \|v - \pi(v)\| + \underbrace{\max_{v \in \mathcal{V}} \|\pi(v) - \pi'(v)\|}_{\leq \epsilon + \epsilon'} \sqrt{\frac{\log(K_\epsilon K_{\epsilon'})}{n}} + \underbrace{\max_{v \in \mathcal{V}} \|\pi'(v)\|}_{\leq \mathrm{rad}(\mathcal{V}) + \epsilon'} \sqrt{\frac{\log(K'_\epsilon)}{n}} \\
&\approx \epsilon + \epsilon' \sqrt{\frac{\log(K_\epsilon)}{n}} + \mathrm{rad}(\mathcal{V}) \sqrt{\frac{\log(K'_\epsilon)}{n}}.
\end{aligned}
$$

Q: Where do we get this second bound with $\log(K_\epsilon K_{\epsilon'})$?

- There are $K_\epsilon K_{\epsilon'}$ pairs of the K_ϵ values of π and the $K_{\epsilon'}$ values of π'.
- We could probably find a better bound.
- Probably not many more than K_ϵ occur as $\pi(v)$ and $\pi'(v)$ for some point v.
- But the difference between K_ϵ and $K_\epsilon K_{\epsilon'}$ doesn't matter here.

A Two-Link Bound

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &\lesssim \mathrm{E}\max_{v\in\mathcal{V}} \langle g, v - \pi(v)\rangle + \mathrm{E}\max_{v\in\mathcal{V}} \langle g, \pi(v) - \pi'(v)\rangle + \mathrm{E}\max_{v\in\mathcal{V}} \langle g, \pi'(v) - \mu^\star\rangle \\
&\qquad \text{old link} \qquad\qquad\qquad \text{new link} \\
&\lesssim \max_{v\in\mathcal{V}} \|v - \pi(v)\| + \underbrace{\max_{v\in\mathcal{V}} \|\pi(v) - \pi'(v)\|}_{\leq \epsilon+\epsilon'} \sqrt{\frac{\log(K_\epsilon K_{\epsilon'})}{n}} + \underbrace{\max_{v\in\mathcal{V}} \|\pi'(v)\|}_{\leq \mathrm{rad}(\mathcal{V})+\epsilon'} \sqrt{\frac{\log(K'_\epsilon)}{n}} \\
&\approx \epsilon + \epsilon' \sqrt{\frac{\log(K_\epsilon)}{n}} + \mathrm{rad}(\mathcal{V}) \sqrt{\frac{\log(K'_\epsilon)}{n}}.
\end{aligned}$$

Q: Why is the last approximation valid?

A Two-Link Bound

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &\lesssim \mathrm{E}\max_{v\in\mathcal{V}} \langle g, v - \pi(v)\rangle + \mathrm{E}\max_{v\in\mathcal{V}} \langle g, \pi(v) - \pi'(v)\rangle + \mathrm{E}\max_{v\in\mathcal{V}} \langle g, \pi'(v) - \mu^\star\rangle \\
&\qquad \text{old link} \qquad\qquad\qquad \text{new link} \\
&\lesssim \max_{v\in\mathcal{V}} \|v - \pi(v)\| + \underbrace{\max_{v\in\mathcal{V}} \|\pi(v) - \pi'(v)\|}_{\leq \epsilon+\epsilon'} \sqrt{\frac{\log(K_\epsilon K_{\epsilon'})}{n}} + \underbrace{\max_{v\in\mathcal{V}} \|\pi'(v)\|}_{\leq \mathrm{rad}(\mathcal{V})+\epsilon'} \sqrt{\frac{\log(K'_\epsilon)}{n}} \\
&\approx \epsilon + \epsilon' \sqrt{\frac{\log(K_\epsilon)}{n}} + \mathrm{rad}(\mathcal{V}) \sqrt{\frac{\log(K'_\epsilon)}{n}}.
\end{aligned}$$

Q: Why is the last approximation valid?

- Triangle inequality.

$$\begin{aligned}
\big\|\pi(v) - \pi'(v)\big\| &= \big\|\pi(v) - v + v - \pi'(v)\big\| \\
&\leq \|\pi(v) - v\| + \big\|\pi'(v) - v\big\| \leq \epsilon + \epsilon' \leq 2\epsilon'.
\end{aligned}$$

- Log of products is sum of logs.

$$\log\big(K_\epsilon K'_\epsilon\big) \leq \log(K_\epsilon) + \log(K_{\epsilon'}) \leq 2\log(K_\epsilon).$$

Better Results

Let's think about the Lipschitz model again.
$\log(K_\epsilon) \approx 1/\epsilon$.

Old Bound

$$\mathrm{w}(\mathcal{M}_s) \lesssim \epsilon + s\sqrt{\frac{\log(K_\epsilon)}{n}} \approx s^{2/3}n^{-1/3} \quad \text{at optimal} \quad \epsilon \approx s^{2/3}n^{-1/3}$$

$$\implies \quad s^2 \geq \mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad s^{4/3} \approx n^{-1/3} \quad \text{i.e.} \quad s \approx n^{-1/4}.$$

New Bound

$$\mathrm{w}(\mathcal{M}_s) \lesssim \epsilon + \epsilon'\sqrt{\frac{\log(K_\epsilon)}{n}} + s\sqrt{\frac{\log(K'_\epsilon)}{n}} \approx s^{4/7}n^{-3/7} \quad \text{at optimal} \quad \epsilon \approx s^{4/7}n^{-3/7}$$

$$\epsilon' \approx n^{1/2}\epsilon^{3/2}$$

$$\implies \quad s^2 \geq \mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad s^{10/7} \approx n^{-3/7} \quad \text{i.e.} \quad s \approx n^{-3/10}.$$

This isn't the $s \approx n^{-1/3}$ bound we got using Fourier series, but it's closer.
Let's see what happens when we use a longer chain of approximations.

Better Results

Let's think about the Lipschitz model again.
$\log(K_\epsilon) \approx 1/\epsilon$.

Old Bound

$$\mathrm{w}(\mathcal{M}_s) \lesssim \epsilon + s\sqrt{\frac{\log(K_\epsilon)}{n}} \approx s^{2/3} n^{-1/3} \quad \text{at optimal} \quad \epsilon \approx s^{2/3} n^{-1/3}$$

$$\implies \quad s^2 \geq \mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad s^{4/3} \approx n^{-1/3} \quad \text{i.e.} \quad s \approx n^{-1/4}.$$

New Bound

$$\mathrm{w}(\mathcal{M}_s) \lesssim \epsilon + \epsilon'\sqrt{\frac{\log(K_\epsilon)}{n}} + s\sqrt{\frac{\log(K'_\epsilon)}{n}} \approx s^{4/7} n^{-3/7} \quad \text{at optimal} \quad \epsilon \approx s^{4/7} n^{-3/7}$$

$$\epsilon' \approx n^{1/2} \epsilon^{3/2}$$

$$\implies \quad s^2 \geq \mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad s^{10/7} \approx n^{-3/7} \quad \text{i.e.} \quad s \approx n^{-3/10}.$$

No magic here. We optimize as usual.

1. Set the derivative with respect to ϵ' to zero and solve for ϵ' in terms of ϵ.
2. Set the derivative with respect to ϵ to zero and solve for ϵ.

Chaining

The Idea

Suppose we want to bound the gaussian width of a set $\mathcal{V}$.

$$\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g,\ v \rangle.$$

- And we have, for each $v \in \mathcal{V}$, increasingly fine approximations $\pi_0(v) \ldots \pi_M(v)$.
- These are the closest vectors to v in ϵ-covers for increasingly small $\epsilon_0 \ldots \epsilon_M$.
- Then we write each $v \in \mathcal{V}$ as the sum over *links* in a *chain* from $\pi_0(v)$ to $\pi_M(v)$.
- Plus a final link from the finest approximation, $\pi_M(v)$, to v itself.

$$v = v - \pi_M(v) + \sum_{j=1}^{M} \underbrace{\pi_j(v) - \pi_{j-1}(v)}_{\text{a link in the chain}}.$$

- We can expand v this way when we write our gaussian width.
- And we can bound it by maximizing each term separately.
- Just like we did in our warm-up, but with more terms.

The Bound

The thing we're bounding.

$$\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g,\ v \rangle.$$

The decomposition we're working with.

$$v = v - \pi_M(v) + \sum_{j=1}^{M} \pi_j(v) - \pi_{j-1}(v).$$

a link in the chain

The bound we get.

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &= \mathrm{E}\left[\max_{v \in \mathcal{V}} \langle g, v - \pi_M(v) \rangle + \sum_{j=1}^{M} \big\langle g,\ \pi_j(v) - \pi_{j-1}(v) \big\rangle \right] \\
&\leq \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v - \pi_M(v) \rangle + \sum_{j=1}^{M} \mathrm{E} \max_{v \in \mathcal{V}} \big\langle g,\ \pi_j(v) - \pi_{j-1}(v) \big\rangle \\
&\lesssim \epsilon_M + \sum_{j=1}^{M} \epsilon_{j-1} \sqrt{\frac{\log\left(K_{\epsilon_j}\right)}{n}}.
\end{aligned}$$

Now all we've got to do is choose $\epsilon_0 \ldots \epsilon_M$.

Considerations

$$\mathrm{w}(\mathcal{V}) \le \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v - \pi_M(v) \rangle + \sum_{j=1}^{M} \mathrm{E} \max_{v \in \mathcal{V}} \big\langle g,\ \pi_j(v) - \pi_{j-1}(v) \big\rangle$$

$$\lesssim \epsilon_M + \sum_{j=1}^{M} \epsilon_{j-1} \sqrt{\frac{\log\big(K_{\epsilon_j}\big)}{n}}.$$

- We want K to be small.
  - That is, we want there to be few distinct values of each link $\pi_j(v) - \pi_{j-1}(v)$ for $v \in \mathcal{V}$.
  - The more values, the more gaussians $\langle g, \pi_j(v) - \pi_{j-1}(v) \rangle$ we have to deal with in our union bound.
- We want ϵ to be small.
  - That is, we want all the links to be short in the sense that their variance $\|\pi_j(v) - \pi_{j-1}\|^2/n$ is small.
  - The longer the links, the bigger the individual gaussians we need to bound.

We can't get both at any one resolution.

- The finer our resolution ϵ_j, the more vectors we need in our cover.
- To balance these considerations, we use a lot of short links and a few large ones.
- Since ϵ_j and $\sqrt{\log\big(K_{\epsilon_{j-1}}\big)}$ are multiplied, this can make the product small.

Considerations

$$\mathrm{w}(\mathcal{V}) \leq \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v - \pi_M(v) \rangle + \sum_{j=1}^{M} \mathrm{E} \max_{v \in \mathcal{V}} \big\langle g,\ \pi_j(v) - \pi_{j-1}(v) \big\rangle$$

$$\lesssim \epsilon_M + \sum_{j=1}^{M} \epsilon_{j-1} \sqrt{\frac{\log\left(K_{\epsilon_j}\right)}{n}}.$$

- We want K to be small.
  - That is, we want there to be few distinct values of each link $\pi_j(v) - \pi_{j-1}(v)$ for $v \in \mathcal{V}$.
  - The more values, the more gaussians $\langle g, \pi_j(v) - \pi_{j-1}(v) \rangle$ we have to deal with in our union bound.
- We want ϵ to be small.
  - That is, we want all the links to be short in the sense that their variance $\|\pi_j(v) - \pi_{j-1}\|^2 / n$ is small.
  - The longer the links, the bigger the individual gaussians we need to bound.

A sensible choice: halve ϵ each time. $\epsilon_j = 1/2^j$.

Assuming all elements of $\mathcal{V}$ are $\epsilon = 1$-close, i.e. $\epsilon_0 = 1$ is big enough that $K_1 = 1$.

$$\begin{aligned} \|\pi_j(v) - \pi_{j-1}(v)\| &\leq \|\pi_j(v) - v\| + \|v - \pi_{j-1}(v)\| \\ &\leq \epsilon_j + \epsilon_{j-1} = 1/2^j + 2/2^j = 3/2^j. \end{aligned}$$

Considerations

$$\mathrm{w}(\mathcal{V}) \le \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v - \pi_M(v) \rangle + \sum_{j=1}^{M} \mathrm{E} \max_{v \in \mathcal{V}} \big\langle g,\ \pi_j(v) - \pi_{j-1}(v) \big\rangle$$

$$\lesssim \epsilon_M + \sum_{j=1}^{M} \epsilon_{j-1} \sqrt{\frac{\log\big(K_{\epsilon_j}\big)}{n}}.$$

- We want K to be small.
  - That is, we want there to be few distinct values of each link $\pi_j(v) - \pi_{j-1}(v)$ for $v \in \mathcal{V}$.
  - The more values, the more gaussians $\langle g, \pi_j(v) - \pi_{j-1}(v) \rangle$ we have to deal with in our union bound.
- We want ϵ to be small.
  - That is, we want all the links to be short in the sense that their variance $\|\pi_j(v) - \pi_{j-1}\|^2 / n$ is small.
  - The longer the links, the bigger the individual gaussians we need to bound.

Plugging these in yields a bound in terms of cover sizes K_{ϵ_j}

$$\mathrm{w}(\mathcal{V}) \lesssim 2^{-M} + \sum_{j=1}^{M} \frac{3}{2^j} \sqrt{\frac{\log\Big(K_{1/2^j}\Big)}{n}}.$$

Chaining all the way

$$\mathrm{w}(\mathcal{V}) \lesssim 2^{-M} + \sum_{j=1}^{M} \frac{3}{2^j} \sqrt{\frac{\log\left(K_{1/2^j}\right)}{n}}.$$

- If $\mathcal{V}$ is small enough in the right sense, the terms of the sum get small quickly.
- And if terms get small quickly enough, the sum doesn't really depend much on M.
- This happens if $\mathcal{V}$ has ϵ-covers of size $K_\epsilon \lesssim 2^{1/\epsilon^\alpha}$ for $\alpha < 2$.

$$\sum_{j=1}^{M} \frac{1}{2^j} \sqrt{\log\left(K_{1/2^j}\right)} \lesssim \sum_{j=1}^{M} \frac{1}{2^j} \sqrt{2^{\alpha j}} = \sum_{j=1}^{M} 2^{(\alpha/2-1)j} \leq \frac{2^{\alpha/2-1}}{1-2^{\alpha/2-1}}.$$

This means we can drop the special term for our *final link* from $\pi_M(v) \to v$.

- If it doesn't matter how big M is, we can have this link be arbitrarily short.
- That is, we can use the limit of this bound as $M \to \infty$.

Integral approximation

Often people approximate this sum by an integral

$$\begin{aligned}
\mathrm{w}(\mathcal{V}) &\lesssim \frac{1}{\sqrt{n}} \sum_{j=1}^{M} \frac{1}{2^j} \sqrt{\log\left(K_{1/2^j}\right)} && \overset{(a)}{=} && \frac{1}{\sqrt{n}} \sum_{j=1}^{M} \int_{1/2^{j+1}}^{1/2^j} 2\sqrt{\log\left(K_{1/2^j}\right)} \\
&\overset{(b)}{\leq} \frac{1}{\sqrt{n}} \sum_{j=1}^{M} \int_{1/2^{j+1}}^{1/2^j} 2\sqrt{\log(K_\epsilon)}\, d\epsilon && = && \frac{2}{\sqrt{n}} \int_{1/2^{M+1}}^{1} \sqrt{\log(K_\epsilon)}\, d\epsilon \\
&\overset{(c)}{\leq} \frac{2}{\sqrt{n}} \int_{0}^{1} \sqrt{\log(K_\epsilon)}\, d\epsilon
\end{aligned}$$

(a) We're integrating a constant.

$$\int_{1/2^{j+1}}^{1/2^j} 2c = \left(\frac{1}{2^j} - \frac{1}{2^{j+1}}\right) 2c = \frac{1}{2^j}\left(1 - \frac{1}{2}\right) 2c$$

(b) Smaller ϵ, bigger ϵ-cover.

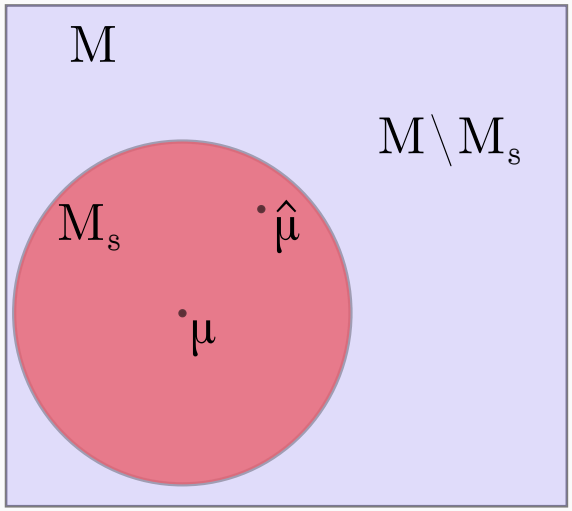
$$K_\epsilon \geq K_{1/2^j} \quad \text{for} \quad \epsilon \leq 1/2^j.$$

(c) Bigger range, bigger integral — our integrand is non-negative.

Neighborhoods

$$\mathrm{w}(\mathcal{V}) \lesssim \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log(K_\epsilon)}\, d\epsilon$$

- If all $v \in \mathcal{V}$ are small, we don't have to integrate all the way to one.
- If we can cover $\mathcal{V}$ with one ball of radius s, we're integrating zero for $\epsilon \geq s$.
- For example, for our centered neighborhood $\mathcal{V} = \mathcal{M}_s$ or its boundary.



$$\mathrm{w}(\mathcal{V}) \lesssim \frac{1}{\sqrt{n}} \int_0^s \sqrt{\log(K_\epsilon)}\, d\epsilon \quad \text{for} \quad s := \max_{v \in \mathcal{V}} \|v\|.$$

The Lipschitz Regression Case

$$\log(K_\epsilon) \lesssim 1/\epsilon \quad \text{for} \quad \mathcal{M} = \left\{f : \rho_{Lip}(f) \leq 1,\ |f| \leq 1\right\}.$$

Integrating, we can bound the width of a neighborhood

$$\mathrm{w}(\mathcal{M}_s) \lesssim \frac{1}{\sqrt{n}} \int_0^s \sqrt{\log(K_\epsilon)}\, d\epsilon \lesssim \frac{1}{\sqrt{n}} \int_0^s \sqrt{\frac{1}{\epsilon}}\, d\epsilon = \frac{1}{\sqrt{n}}\, 2\sqrt{\epsilon}\,|_0^s = 2\sqrt{\frac{s}{n}}.$$

And solve for the radius s for a least squares estimator

$$s^2 \gtrsim \mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad s^{-3/2} \approx n^{-1/2} \quad \text{i.e.} \quad s \approx n^{-1/3}.$$

This agrees with what we see based on Fourier series.

Chaining and Gaussian Width in General

- This isn't just another bound — it's pretty tight.
- This bound — with K_ϵ the size of the smallest ϵ-cover — can barely be improved.
- It's off by at most a factor of $\log(n)$. Proving it isn't so hard.
- See Vershynin [2018, Chapter 8.1.2] if you're interested.

$$\frac{1}{\sqrt{n}\log(n)}\int_0^1 \sqrt{\log(K_\epsilon)}d\epsilon \lesssim \mathrm{w}(\mathcal{V}) \lesssim \frac{1}{\sqrt{n}}\int_0^1 \sqrt{\log(K_\epsilon)}d\epsilon$$

- In fact, if we're a bit more careful about how we choose $\pi_k(v)$, chaining gives us a bound that's off by no more than a constant factor.
- This fancier chaining is pretty straightforward conceptually.
- We just do the bound thinking of $\pi_k(v)$ as an arbitrary function taking on 2^{2^k} distinct values, then minimize the chaining bound over all the π_k.
- It's easy to prove this is no worse than what we've talked about.
- But proving it's tight up to constants is a feat. See Talagrand [2014].

Chaining and Fourier Series

Chaining is, in a sense, approximating our analysis using Fourier series.

- Using Fourier series, we were able to decompose the functions in Sobolev models into combinations of orthogonal functions.
- There were infinitely many such functions, but only a few were allowed to be big.

$$\{m = \sum_j m_j \phi_j : \sum_j m_j^2 \lambda_j \leq B\} \implies \|m_j \phi_j\|_{L_2} = m_j \leq B/\sqrt{\lambda_j}.$$

The links in our chains play the role of the Fourier basis functions ϕ_j.

- These links, $\phi_{j,v}(x) = \{\pi_j(v) - \pi_{j-1}(v)\}(x)$, are approximately orthogonal.
  - for different resolutions j
  - for the same resolution and different v — unless they're the same curve.
  - i.e. unless $\pi_j(v) = \phi_j(v')$ and $\pi_{j-1}(v) = \pi_{j-1}(v')$, so $\phi_{j,v} = \phi_{j,v'}$.
- And as a result, the corresponding gaussians are approximately uncorrelated.

$$\mathrm{E}\langle g,\ u\rangle\ \langle g,\ v\rangle = \frac{1}{n^2} \sum_{ij} u_i v_j \,\mathrm{E}\, g_i g_j = \frac{1}{n^2} \sum_{i=1}^{n} u_i v_j = \frac{\langle u,\ v\rangle}{n}.$$

and therefore

$$\begin{aligned} &\mathrm{Cov}\big\{\big\langle g,\ \pi_j(u) - \pi_{j-1}(u)\big\rangle,\ \langle g,\ \pi_{j'}(v) - \pi_{j'-1}(v)\rangle\big\} \\ &= \frac{\big\langle \pi_j(v) - \pi_{j-1}(v),\ \pi_{j'}(u) - \pi_{i-1}(u)\big\rangle}{n}. \end{aligned}$$

The Point

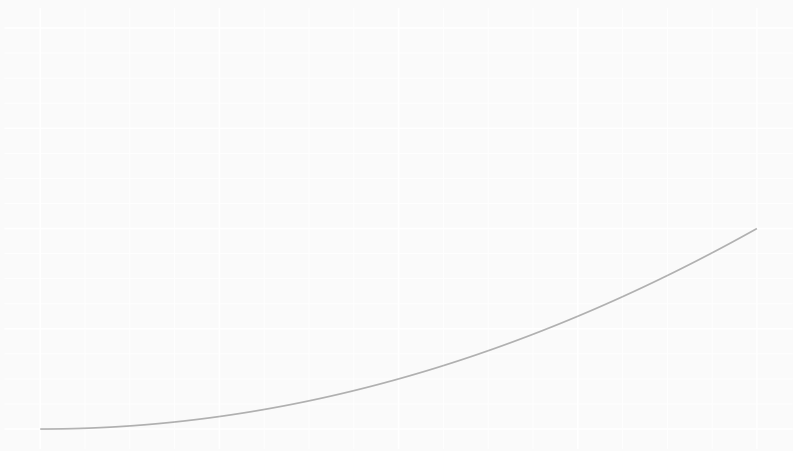
- The point is to make sure that when we use the union bound,
  we're not being wasteful and bounding more-or-less the same thing twice.
- Decomposing the curves in our model into sums of approximately orthogonal functions helps us keep track of what we're bounding more accurately.
- It helps us not overcount when we're bounding gaussian width.

Let's look into how orthogonal our links are.

$\pi_k(v)$ for our Lipschitz cover

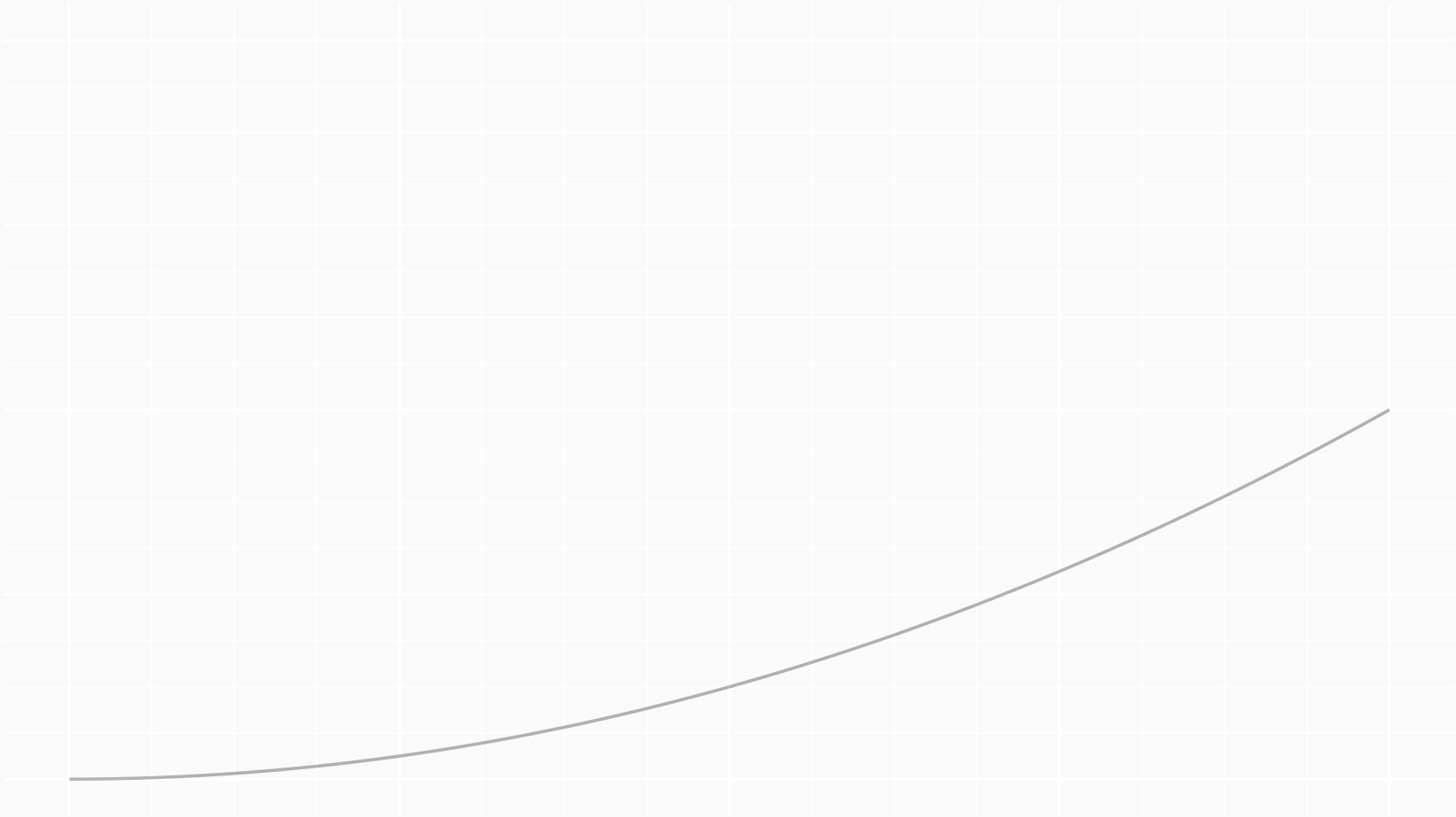
1. Draw an $\epsilon_k \times \epsilon_k$ grid.
2. Snap $v(x)$ to it at each x on the grid.
3. Piecewise-linear between grid points.

Use the small squares for π_{j+1}, two for π_j, and four for π_{j-1}.



Check the inner product between links $\ell_j(v) = \pi_j(v) - \pi_{j-1}(v)$.
Do it both for different j and different curves.

A bigger grid



You can try it for more curves and more resolutions

- Do it by hand on the blank grid on the next slide.
- Or code it up in R so you can try more stuff.

A bigger grid

References

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.