Machine Learning Theory

Covering Numbers

David A. Hirshberg

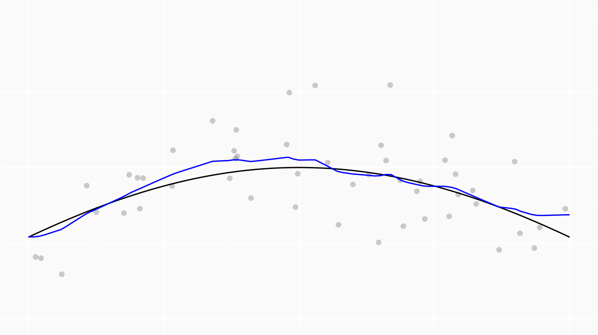
April 19, 2025

Emory University

Review

Least squares with gaussian noise

We observe $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

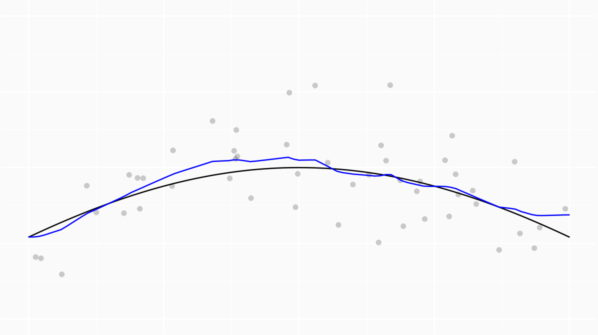


We've focused on least squares estimators. That's the curve in your regression model that minimizes mean squared prediction error.

$$\hat{\mu} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2$$

Least squares with gaussian noise

We observe $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.



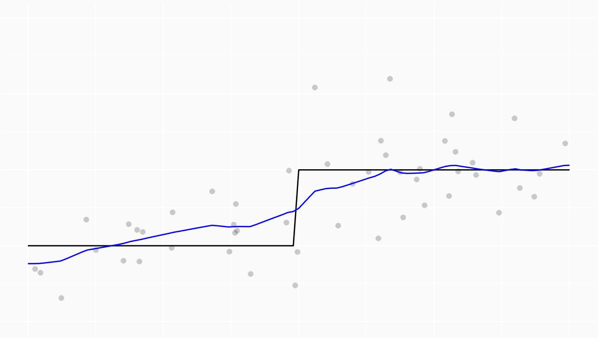
To think about how well this works, we've proven high probability bounds on the error.

$\|\hat{\mu} - \mu\| < s$ with probability $1 - \delta$ where usually $\|v\|^2 = \frac{1}{n}\sum_{i=1}^{n} v(X_i)^2$

We've mostly talked about this error's *sample two norm*.

Least squares with gaussian noise

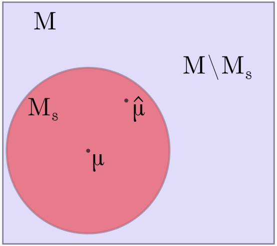
We observe $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

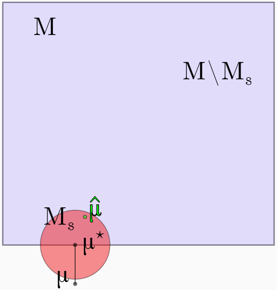


Or, more generally, on norms of the difference between our estimator and the model's best *approximation* to μ.

$\|\hat{\mu} - \mu^\star\| < s$ with probability $1 - \delta$ where $\mu^\star = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \|m - \mu\|$

It's the gaussian width of *neighborhoods* of this best approximation $\mu^\star$.



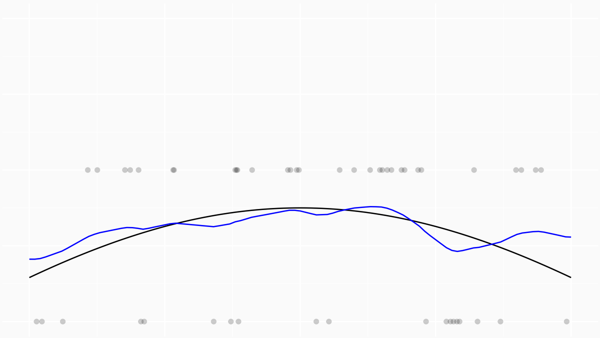


$$\|\hat{\mu} - \mu^\star\| < s \times \sigma \left\{ 1 + \sqrt{\frac{2\Sigma_n}{\delta n}} \right\} \quad \text{w.p.} \;\; 1 - \delta \;\; \text{if } s \text{ satisfies}$$

$$\frac{s^2}{2} \geq \mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad \mathcal{M}_s = \{ m \in \mathcal{M} : \|m - \mu^\star\| \leq s \}.$$

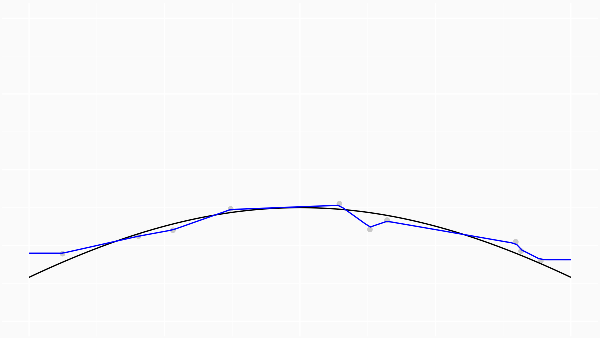
Generalization–Non-Gaussian Noise

These bounds more or less work with non-gaussian noise, too.
For example, bounded noise like what we get in *probabilistic classification*



Generalization–Population Squared Error

Same deal when we're interested in the population two-norm of our error.
Sampling from our population acts like subgaussian noise.



To use this, we need to bound gaussian width

We've done this in a few models using specialized techniques.

1. Finite models using the Union Bound and the Gaussian Tail Bound.

$$s^2 \geq cs\sqrt{\log(K)/n} \quad \text{for} \quad s \geq c\sqrt{\log(K)/n}$$

2. Finite-dimensional models using Projection and the Cauchy-Schwarz Inequality.

$$s^2 \geq s\sqrt{K/n} \quad \text{for} \quad s \geq \sqrt{K/n}$$

3. Sobolev models using Fourier Analysis and the Cauchy-Schwarz Inequality.

$$s^2 \geq cs^{1-d/2p}/\sqrt{n} \quad \text{for} \quad s \geq c' n^{-1/(2+d/p)}$$

There are two essential ideas here.

1. Approximating many curves by combinations of a few.
2. Counting.

This week, we'll talk about a completely general technique for bounding width.
We'll use the same two ideas, but our approximations will be subtler.

Finite Approximations and Gaussian Width

Finite Models

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

$$\mathrm{E}\langle g, m-\mu^\star\rangle^2 = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n g_i\{m(X_i)-\mu^\star(X_i)\}\right)^2$$
$$= \frac{1}{n^2}\sum_{i=1}^n \mathrm{E}\, g_i^2\{m(X_i)-\mu^\star(X_i)\}^2 = \frac{\|m-\mu^\star\|^2}{n}.$$

Q: What happened to the cross terms in the square?

- We can bound this via union bound. We count curves in the model.

$$\mathrm{w}(\mathcal{M}_s) \le cs\sqrt{\frac{\log(K)}{n}} \quad \text{if } \mathcal{M} \text{ contains } K \text{ curves } v_1 \ldots v_K, \text{ all with } \quad \|v-\mu^\star\|_{L_2(\mathrm{P_n})} \le s.$$

- We may be overcounting. This bounds the max of K totally different gaussians.
- That's kind of the worst case, so if there's correlation we're overcounting.
- And our gaussians are as correlated as the curves in our neighborhood.

$$\mathrm{E}\langle g, v_k\rangle\langle g, v_{k'}\rangle = n^{-2}\,\mathrm{E}\, v_k^T g g^T v_{k'} = n^{-2}\ v_k^T(\underbrace{\mathrm{E}\, gg^T}_{I})v_{k'} = n^{-1}\ \langle v_k, v_{k'}\rangle.$$

- This definitely won't work for models with infinitely many curves.
- How do we take advantage of this correlation to tackle infinite models?

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

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 $v \in \mathcal{M}$.

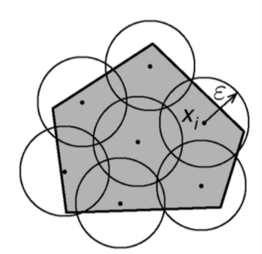
This means we have to talk about how many *meaningfully different* curves we have.

Notation.

From here on, we'll think of $\mathcal{M}_s$ as a neighborhood of **zero** of radius s.

$$\mathcal{M}_s = \{m \in \mathcal{M} \quad : \quad \|m\| \leq s\}.$$

We're using the notation $\mathcal{M}_s$ for what we usually call $\mathcal{M}_s - \mu^\star$. It saves space.



We call a set $\mathcal{V}^\epsilon$ an ϵ-**cover** for the set $\mathcal{V}$ if every element $v \in \mathcal{V}$ is within a distance ϵ of an element $\pi_\epsilon(v) \in \mathcal{V}^\epsilon$.

If we have an ϵ-cover $\mathcal{M}_s^\epsilon$ of size K_ϵ for $\mathcal{M}_s$, then we've got a bound on our width.

$$
\begin{aligned}
\mathrm{w}(\mathcal{M}_s) &= \mathrm{E}\left[\max_{v \in \mathcal{M}_s} \langle g,\ v \rangle\right] \\
&= \mathrm{E}\left[\max_{v \in \mathcal{M}_s} \langle g,\ v - \pi_\epsilon(v) \rangle + \langle g, \pi_\epsilon(v) \rangle\right] \\
&\lesssim \underbrace{\max_{v \in \mathcal{M}_s} \| v - \pi_\epsilon(v) \|}_{\leq \epsilon} + \underbrace{\max_{u \in \mathcal{M}_s^\epsilon} \| u \|}_{\leq 2s} \sqrt{\frac{\log(K_\epsilon)}{n}}.
\end{aligned}
$$

And this works for infinite models just as well as it does for finite ones.
We can think of K_ϵ as the size of the neighborhood $\mathcal{M}_s$ at resolution ϵ.

Let's Stop and Think

Q: Does the ϵ-cover $\mathcal{M}_s^\epsilon$ have to be a subset of $\mathcal{M}_s$ for this?

$$\begin{aligned}
\mathrm{w}(\mathcal{M}_s) &= \mathrm{E}\left[\max_{v \in \mathcal{M}_s} \langle g,\ v\rangle\right] \\
&= \mathrm{E}\left[\max_{v \in \mathcal{M}_s} \langle g,\ v - \pi_\epsilon(v)\rangle + \langle g, \pi_\epsilon(v)\rangle\right] \\
&\lesssim \underbrace{\max_{v \in \mathcal{M}_s} \|v - \pi_\epsilon(v)\|}_{\leq \epsilon} + \underbrace{\max_{u \in \mathcal{M}_s^\epsilon} \|u\|}_{\leq 2s} \sqrt{\frac{\log(K_\epsilon)}{n}}.
\end{aligned}$$

Let’s Stop and Think

Q: Does the ϵ-cover $\mathcal{M}_s^\epsilon$ have to be a subset of $\mathcal{M}_s$ for this?

$$\begin{aligned}
\mathrm{w}(\mathcal{M}_s) &= \mathrm{E}\left[\max_{v\in\mathcal{M}_s} \langle g,\ v\rangle\right] \\
&= \mathrm{E}\left[\max_{v\in\mathcal{M}_s} \langle g,\ v-\pi_\epsilon(v)\rangle + \langle g, \pi_\epsilon(v)\rangle\right] \\
&\lesssim \underbrace{\max_{v\in\mathcal{M}_s} \|v-\pi_\epsilon(v)\|}_{\leq\epsilon} + \underbrace{\max_{u\in\mathcal{M}_s^\epsilon} \|u\|}_{\leq 2s} \sqrt{\frac{\log(K_\epsilon)}{n}}.
\end{aligned}$$

Nope. We pay at most afactor of 2 in our second term when it isn’t.

.

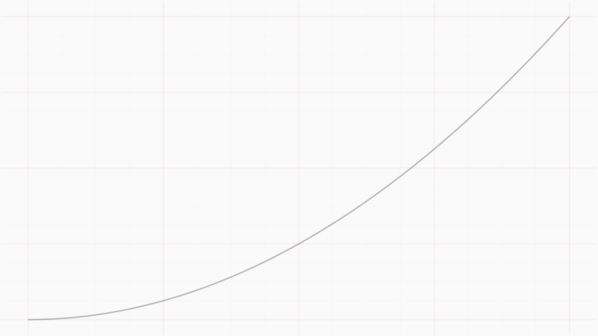
$$\|\pi_\epsilon(v) - v + v\| \leq \|\pi_\epsilon(v) - v\| + \|v\| \leq \epsilon + s \overset{*}{\leq} 2s \quad \text{for} \quad v \in \mathcal{M}_s.$$

(*) Because $\log(K_\epsilon) = \log(1) = 0$ for $\epsilon \geq s$, we can bound $s+\epsilon$ by $2s$.

Example: The Lipschitz Regression Model

- Think of an ϵ-cover of $\mathcal{M}$ as the set of ϵ-approximations $\pi_\epsilon(m)$ for each m in $\mathcal{M}$.
- Often we base these approximations on a grid. π_ϵ snaps to that grid.

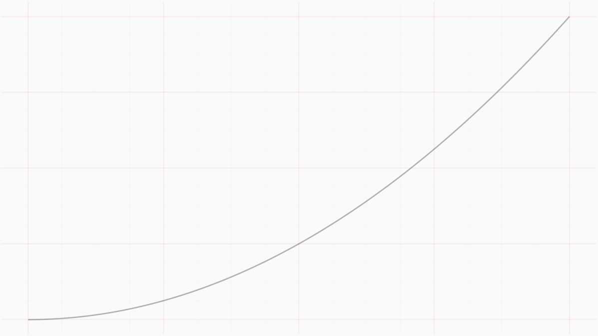
$$\mathcal{M} = \{m : |m(x') - m(x)| \le |x' - x|,\ |m(x)| \le 1\}.$$



Example: The Lipschitz Regression Model

- Think of an ϵ-cover of $\mathcal{M}$ as the set of ϵ-approximations $\pi_\epsilon(m)$ for each m in $\mathcal{M}$.
- Often we base these approximations on a grid. π_ϵ snaps to that grid.

$$\mathcal{M} = \{m : |m(x') - m(x)| \le |x' - x|,\ |m(x)| \le 1\}.$$



1. Draw an ϵ-spaced grid.
2. At each x-coordinate on the grid, snap to the closest grid point.
3. Because our function is 1-Lipschitz, it can't jump by more than ϵ between points.

How many of these are there? Consider $\epsilon = 1/M$ for an integer M.

(starting points) $\cdot$ (options per step)$^{\text{steps}} = 1/\epsilon \cdot 2^{1/\epsilon}$.

Consequences

Our log covering number grows like $1/\epsilon$.

$$\log(K_\epsilon) \le \epsilon^{-1}$$

We know that $\hat{\mu}$ is in a neighborhood of $\mu^\star$ of radius proportional to s satisfying

$$s^2/2 \ge \sigma\,\mathrm{w}(\mathcal{M}_s) \quad \text{for} \quad \mathrm{w}(\mathcal{M}_s) \le c\epsilon + s\sqrt{\log(K_\epsilon)/n} \approx \epsilon + sn^{-1/2}\epsilon^{-1/2}$$

This width bound holds for all $\epsilon > 0$, so we can choose ϵ to minimize it.

$$0 = \frac{d}{d\epsilon}\left(\epsilon + sn^{-1/2}\epsilon^{-1/2}\right) = 1 - sn^{-1/2}\epsilon^{-3/2}/2 \quad \text{for} \quad \epsilon = \left(\frac{s}{2\sqrt{n}}\right)^{2/3} \approx s^{2/3}n^{-1/3}$$

And this tells us we're in a neighborhood of radius s like this.

$$s^2 \ge \underbrace{c\sigma s^{2/3} n^{-1/3}}_{\ge c\sigma\,\mathrm{w}(\mathcal{M}_s^\circ)} \quad \text{for} \quad s^{4/3} \ge \sigma n^{-1/3} \quad \text{i.e.} \quad s \ge \sigma^{3/4} n^{-1/4}.$$

Dissatisfying Results

- We'll show, momentarily, that $\log(K_\epsilon) \approx 1/\epsilon$ for the Lipschitz model.

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

- That gives us a $n^{-1/4}$ 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 at a faster rate.
- 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.

Has the covering idea failed us?

Dissatisfying Results

- We'll show, momentarily, that $\log(K_\epsilon) \approx 1/\epsilon$ for the Lipschitz model.

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

- That gives us a $n^{-1/4}$ rate.

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

- But we know it converges at a faster rate.
- 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.

Has the covering idea failed us?

No. We just have to make better use of it. We'll do that next class.
When we do that, we'll see a rough connection to Fourier series.

Refined Bounds in terms of ϵ-covers

By working with ϵ-covers at different resolutions, we can prove a refined upper bound.

$$\mathrm{w}(\mathcal{V}) \lesssim \frac{1}{\sqrt{n}} \int_0^\infty \sqrt{\log(K_\epsilon)}d\epsilon \quad \text{where } K_\epsilon \text{ is the size of the smallest } \epsilon\text{-cover for } \mathcal{V}.$$

This multi-resolution argument is called *chaining*.
The bound, *Dudley's Integral Bound*.

The Lipschitz Regression Case

Refined Bounds in terms of ϵ-covers

By working with ϵ-covers at different resolutions, we can prove a refined upper bound.

$$\mathrm{w}(\mathcal{V}) \lesssim \frac{1}{\sqrt{n}} \int_0^\infty \sqrt{\log(K_\epsilon)}\, d\epsilon \quad \text{where } K_\epsilon \text{ is the size of the smallest } \epsilon\text{-cover for } \mathcal{V}.$$

This multi-resolution argument is called *chaining*.
The bound, *Dudley's Integral Bound*.

The Lipschitz Regression Case

$\log(K_\epsilon) = 0$ for $\epsilon > s$. Why? And $\log(K_\epsilon) \lesssim \epsilon^{-1}$ generally.

$$\begin{aligned} \mathrm{w}(\mathcal{M}_s) &\lesssim \frac{1}{\sqrt{n}} \int_0^\infty \epsilon^{-1/2}\, d\epsilon \\ &= \frac{1}{\sqrt{n}} \int_0^s \epsilon^{-1/2}\, d\epsilon \\ &= \frac{1}{\sqrt{n}} 2\epsilon^{1/2} \,|_0^s = 2n^{-1/2} s^{-1/2}. \end{aligned}$$

and consequently

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

Optimality

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

This approach to bounding gaussian width is almost optimal.

- There's also a lower bound, *Sudakov's Minoration Inequality*
- It depends on the size K_ϵ of the set's *smallest* ϵ-cover.
- These bounds are close: the upper bound is no more than $\log(n)$ times the lower.

$$\mathrm{w}(\mathcal{V}) \gtrsim \frac{1}{\sqrt{n}} \max_{\epsilon > 0} \epsilon \sqrt{\log(K_\epsilon)}.$$

Summary

The accuracy of our estimator is determined by the rate at which the gaussian width of our model's neighborhood boundary grows.

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

That gaussian width is a measure of the boundary's size at multiple resolutions.

$$\frac{1}{\sqrt{n}} \max_{\epsilon > 0} \epsilon \sqrt{\log(K_\epsilon)} \underset{\lesssim}{\approx} \mathrm{w}(\mathcal{M}_s^{\circ}) \underset{\lesssim}{\approx} \frac{1}{\sqrt{n}} \int_0^{\infty} \sqrt{\log(K_\epsilon)} d\epsilon.$$

Credit

Some things borrowed from Vershynin's *High Dimensional Probability*.

- The presentation of the refined bounds
- The ϵ-net picture.

Its chapters 7-8 are a good, although relatively sophisticated, reference for this stuff.