Machine Learning Theory

Least Squares in Infinite Models i.e. Regression

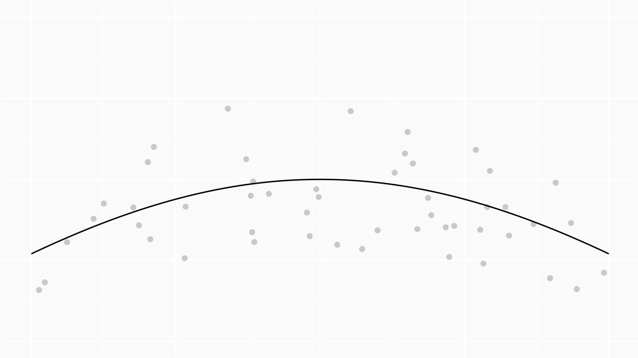
David A. Hirshberg

March 19, 2025

Emory University

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're estimating the curve $\mu(x)$.
Our goal is get close in terms of *sample mean squared distance*.

This is the kind of statement we're after.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{with probability} \quad 1 - \delta$$

The characters

Old Friends

- (X_i, Y_i) for $i = 1 \ldots n$. The data.
- $\mu(x)$, the estimation target. A curve.
- $\mathcal{M}$, the model. A set of curves.
  For today, a *convex set* containing infinitely many curves.
- $\hat{\mu}$, our estimate. Some curve in the model, chosen because it fits the data.
- m, an anonymous curve. Whatever curve we're thinking about at the moment.

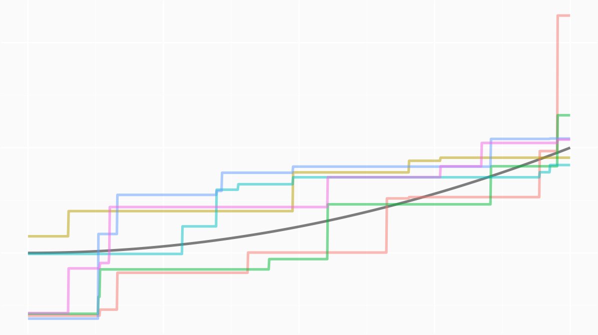
New Ones

- $\mathcal{M}_s$, a *neighborhood* of the target.
  - It's the subset of curves in our model that are close to μ.
  - We're trying to show that $\hat{\mu}$ is in it.
- $\mathcal{M} \setminus \mathcal{M}_s$, its complement.
  - It's the subset of curves in our model that aren't close to μ.
  - It's equivalent to show that $\hat{\mu}$ *is not* one of the curves in it.
- $\mathcal{M}_s^\circ$, the boundary of the neighborhood $\mathcal{M}_s$.
  - This will play a special role in *convex models*.
  - That's what we'll be talking about today.

For now, we'll think of $X_1 \ldots X_n$ as deterministic.
If they are random, we *condition* on them.

What’s changed from last week is our model

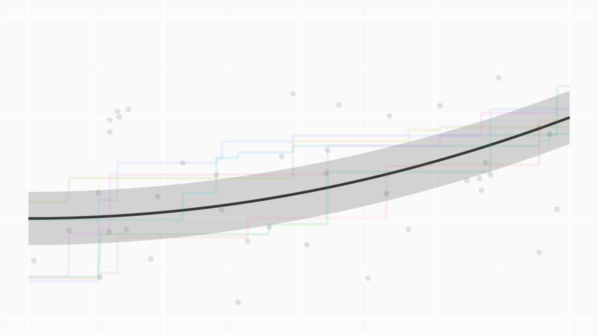
Last week, our model was a finite set of curves.



Like these.

What's changed from last week is our model

Last week, our model was a finite set of curves.



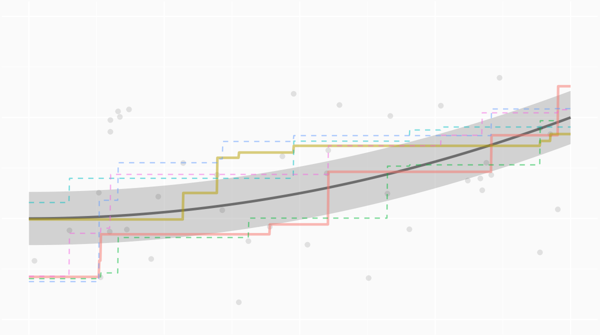
A neighborhood is the subset of these curves that's close enough to μ.
Say within the gray tube.

Caveat.
The gray tube is the set of curves that are close in terms of the infinity norm.

$$\mathcal{M}_s^{\infty} = \{m \in \mathcal{M} : \|m - \mu\|_{\infty} < s\}$$

What's changed from last week is our model

Last week, our model was a finite set of curves.



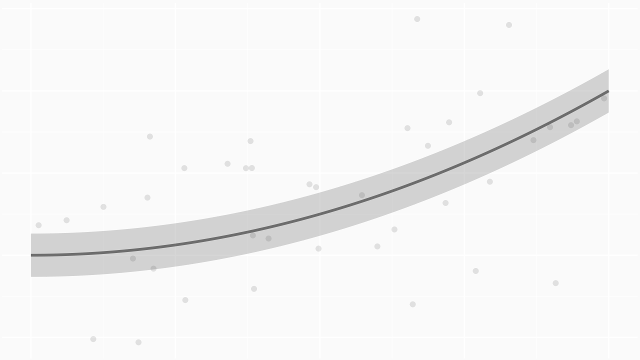
We're talking about the set of curves that are close in terms of the sample two-norm.

$$\mathcal{M}_s = \{ m \in \mathcal{M} : \|m - \mu\|_{L_2(\mathrm{P_n})} < s \}$$

Think of these as curves that are mostly, but not necessarily always, in the tube.
These are plotted as solid lines above. Those in the complement are dashed.

Neighborhoods in models with infinitely-many curves

Let's take the set of increasing curves to be our regression model.



A neighborhood is the subset of these curves that's close enough to μ.
Say within the gray tube.

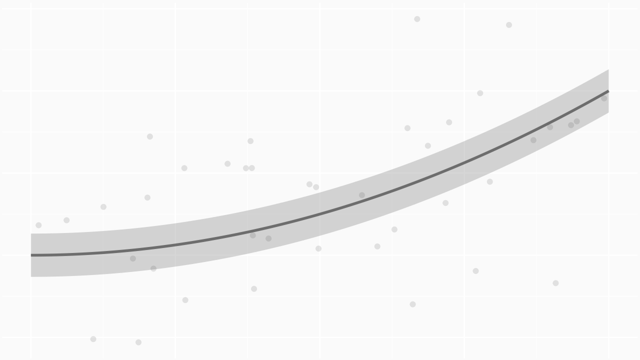
Same caveat.

The gray tube is the set of curves that are close in terms of the infinity norm.

$$\mathcal{M}_s^{\infty} = \{m \in \mathcal{M} : \|m - \mu\|_{\infty} < s\}$$

Neighborhoods in models with infinitely-many curves

Let's take the set of increasing curves to be our regression model.



We're talking about the set of curves that are close in terms of the sample two-norm.

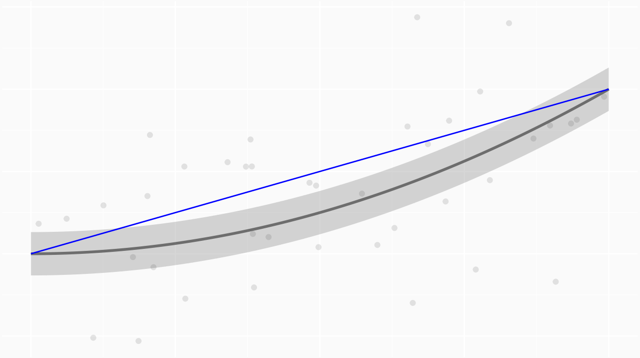
$$\mathcal{M}_s = \{m \in \mathcal{M} : \|m - \mu\|_{L_2(\mathrm{P_n})} < s\}$$

Think of these as curves that are mostly, but not necessarily always, in the tube.

Now that our model has infinitely many curves, we can't draw all of them.
Let's look at a few examples instead.

Neighborhoods in models with infinitely-many curves

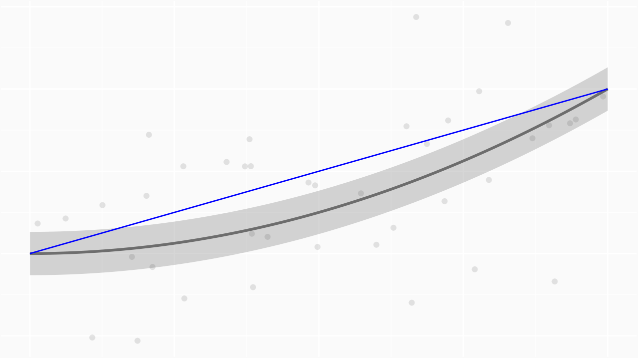
Let's take the set of increasing curves to be our regression model.



Is this in our neighborhood?

Neighborhoods in models with infinitely-many curves

Let's take the set of increasing curves to be our regression model.

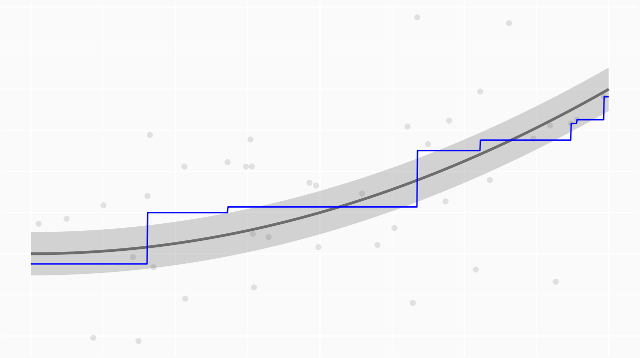


Is this in our neighborhood?

No. It's too far from μ

Neighborhoods in models with infinitely-many curves

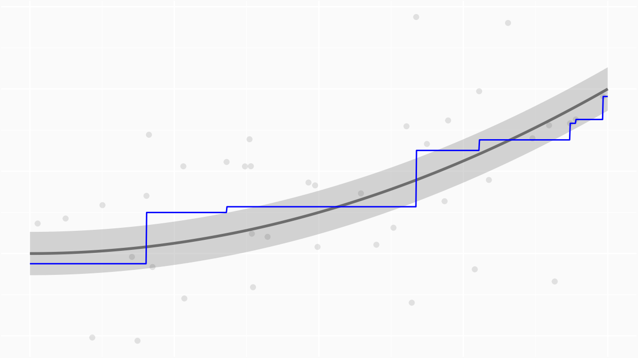
Let's take the set of increasing curves to be our regression model.



Is this in our neighborhood?

Neighborhoods in models with infinitely-many curves

Let's take the set of increasing curves to be our regression model.

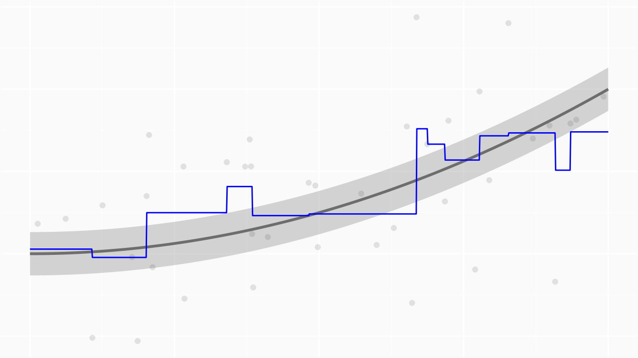


Is this in our neighborhood?

Yes.

Neighborhoods in models with infinitely-many curves

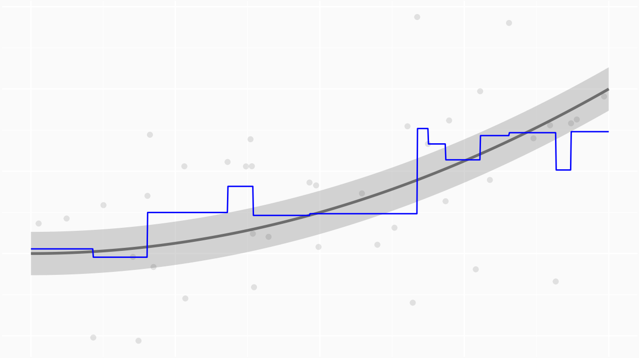
Let's take the set of increasing curves to be our regression model.



Is this in our neighborhood?

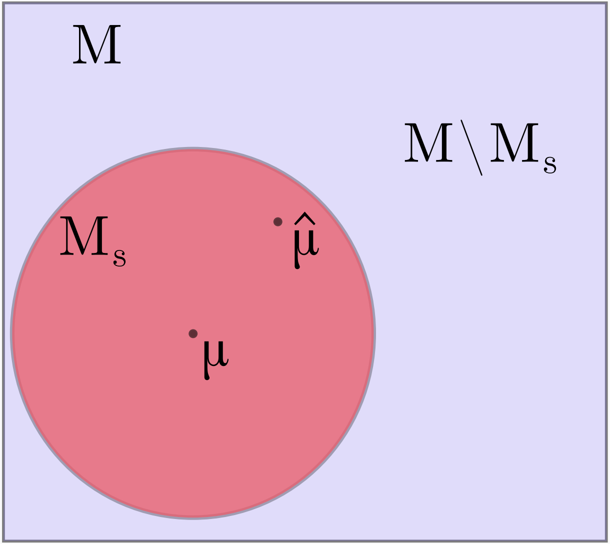
Neighborhoods in models with infinitely-many curves

Let's take the set of increasing curves to be our regression model.

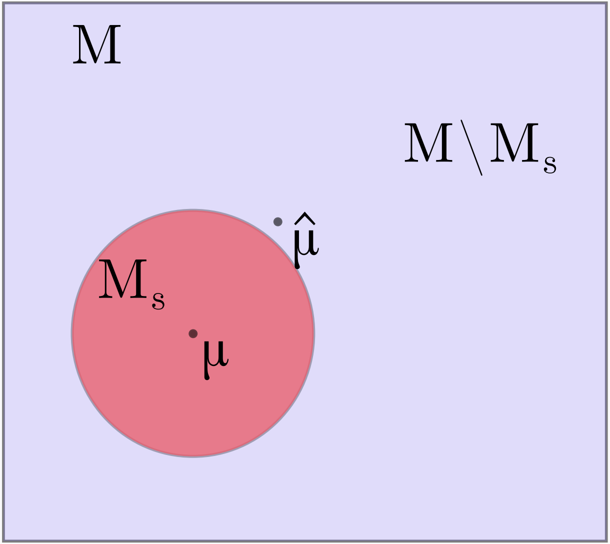


Is this in our neighborhood?

No. It's close to μ, but it's not in our model. It's not increasing.


M
$M \setminus M_s$
M_s
$\hat{\mu}$
μ

What we're going to rule out



The argument in words

What we know is that $\hat{\mu}$ beats or ties every other curve in the model.
That's what a minimizer (argmin) does.

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}} \ell(m) \quad \iff \quad \ell(\hat{\mu}) \le \ell(m) \ \text{ for all } \ m \in \mathcal{M}$$

If our model is right, that means it beats or ties μ.

$$\ell(\hat{\mu}) \le \ell(m) \ \text{ for all } \ m \in \mathcal{M} \ \textbf{ and } \ \mu \in \mathcal{M} \implies \ell(\hat{\mu}) \le \ell(\mu).$$

And if no curve in our neighborhood's complement beats or ties μ,
this means $\hat{\mu}$ isn't in that complement.

$$\ell(\hat{\mu}) \le \ell(\mu) \ \textbf{ and } \ \ell(m) > \ell(\mu) \ \text{ for all } \ m \in \mathcal{M} \setminus \mathcal{M}_s \implies \hat{\mu} \notin \mathcal{M} \setminus \mathcal{M}_s$$

And because $\hat{\mu}$ *is* in the model, that means $\hat{\mu}$ is in the neighborhood.

$$\hat{\mu} \notin \mathcal{M} \setminus \mathcal{M}_s \ \textbf{ and } \ \hat{\mu} \in \mathcal{M} \quad \iff \quad \hat{\mu} \in \mathcal{M}_s$$

When our two if clauses are true, this argument implies $\hat{\mu}$ is in our neighborhood.
So if they're true with some probability, $\hat{\mu}$ is in the neighborhood with that probability.

Today we'll assume we got the model right, so the second if is what we need to prove.

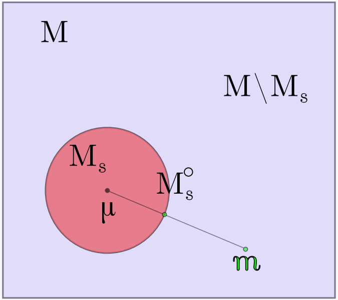
A Reduction

Simplifying our proof for convex models.

What Convexity Buys Us

- When the model is a *convex set*, we needn't worry about most of the complement.
- If there's no curve on the boundary with squared loss less than μ's, there's none in the rest of the complement, either.

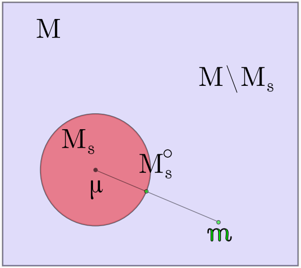
$$\underbrace{\ell(m) > \ell(\mu) \quad \text{for all} \quad m \in \mathcal{M}_s^\circ}_{\text{the thing we're going to prove}} \quad \Longrightarrow \quad \underbrace{\ell(m) > \ell(\mu) \quad \text{for all} \quad m \in \mathcal{M} \setminus \mathcal{M}_s.}_{\text{the thing we said we needed to prove}}$$



- Think of a curve in the complement as having a representative on the boundary.
- To find it, draw a line from the curve toward μ. Stop where you hit the boundary.
- A curve's loss is *always* bigger than μ's if its representative's is.
- So if the representative of every curve in the complement has loss bigger than μ's, so does every curve in the complement.

Representatives do it for their constituents

Proof. A curve's squared error loss is *always* bigger than μ's if its representative's is.



A representative is a point
$m_t = \mu + t(m - \mu)$ for some $t \in [0, 1]$.

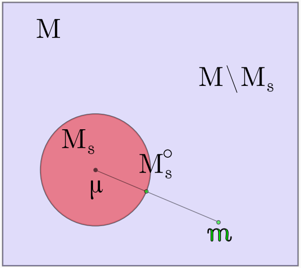
We'll show the loss difference $\ell(m) - \ell(\mu)$ for any curve in the complement is at least t **times** the loss difference $\ell(m_t) - \ell(\mu)$ for its representative.

$$\ell(m_t) - \ell(\mu) \leq t\{\ell(m) - \ell(\mu)\}$$

This means that if the representative's is positive, so is the original curve's.

Representatives do it for their constituents

Proof. A curve's squared error loss is *always* bigger than μ's if its representative's is.

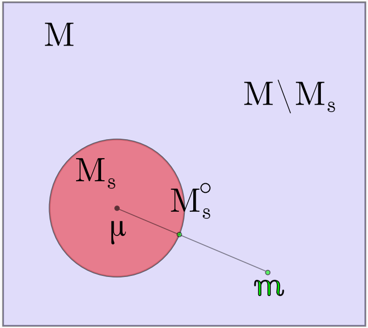


A representative is a point
$m_t = \mu + t(m - \mu)$ for some $t \in [0, 1]$.

$$
\begin{aligned}
\ell(m_t) - \ell(\mu) &= \|m_t - \mu\|^2_{L_2(\mathrm{P_n})} - 2\langle \varepsilon,\ m_t - \mu\rangle_{L_2(\mathrm{P_n})} \\
&= \|\mu + t(m - \mu) - \mu\|^2_{L_2(\mathrm{P_n})} - 2\langle \varepsilon,\ \mu + t(m - \mu) - \mu\rangle_{L_2(\mathrm{P_n})} \\
&= t^2\|m - \mu\|^2_{L_2(\mathrm{P_n})} - 2t\langle \varepsilon,\ m - \mu\rangle_{L_2(\mathrm{P_n})} \\
&\le t\{\ell(m) - \ell(\mu)\} \quad \text{because} \quad t^2 \le t.
\end{aligned}
$$

What is Convexity?

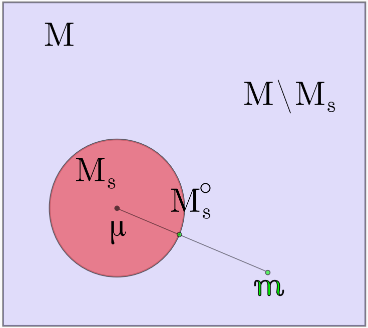
Convexity is a property of a model that guarantees that, no matter what $\mu \in \mathcal{M}$ is, each curve in a neighborhood's complement has a representative on its boundary.



Definition.
A set is convex **if and only if** it contains the line between any two of its points.

Why do we need Convexity?

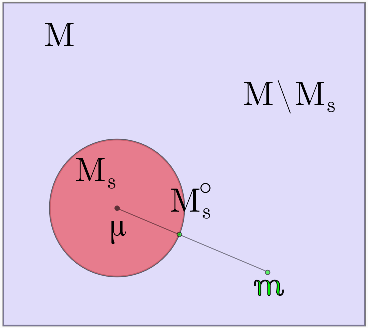
Why isn't ruling out the boundary necessarily enough
to rule out the complement if the model isn't convex?



Hint.
Convexity is a property of a model that guarantees that, no matter what $\mu \in \mathcal{M}$ is, each curve in a neighborhood's complement has a representative on its boundary.

Why do we need Convexity?

Why isn't ruling out the boundary necessarily enough
to rule out the complement if the model isn't convex?



In a nonconvex model, there may be curves in the complement without
representatives on the boundary. Ruling out the boundary doesn't cover them.

Following through.

Proving the simplified claim.

Notation

Throughout, $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ will be the $L_2(\mathrm{P_n})$ norm and inner product.

$$\|f\|^2 = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2$$

$$\langle f, g\rangle = \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i).$$

The Deterministic Part

What we're proving is a *lower bound* on differences in mean squared error.

$$\ell(m) > \ell(\mu) \quad \text{or equivalently} \quad \ell(m) - \ell(\mu) > 0 \quad \text{for all} \quad m \in \mathcal{M}_s^\circ.$$

And we only need to bother with curves on the neighborhood's boundary.

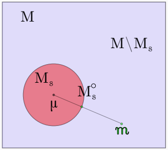
$$\ell(m) - \ell(\mu) = \underbrace{\|m - \mu\|^2}_{=s^2} - 2\langle \varepsilon,\ m - \mu \rangle \quad \text{for} \quad m \in \mathcal{M}_s^\circ$$

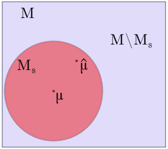
Nice! All we have to do is bound the mean zero term for all curves on the boundary.

The difference is positive if $\quad s^2/2 > \langle \varepsilon,\ m - \mu \rangle \quad$ for all $\quad m \in \mathcal{M}_s^\circ$

or equivalently if $\quad s^2/2 > \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle.$

This implies that every curve in the complement $\mathcal{M} \setminus \mathcal{M}_s$ has bigger loss than μ.
When it's satisfied, we know that $\hat{\mu} \in \mathcal{M}_s$, i.e. that $\|\hat{\mu} - \mu\| < s$.





Taking Advantage of Approximate Constancy

- Our error bound holds with high probability if our $s^2/2 \ge \max\ldots$ inequality does.

$$\|\hat{\mu} - \mu\| < s \ \text{ when } \ \frac{s^2}{2} > \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle \qquad \text{so}$$

$$P(\|\hat{\mu} - \mu\| < s) \ge P\left(\frac{s^2}{2} > \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle\right)$$

- We want to choose s so this happens with probability (at least) $1-\delta$.
- It helps that this maximum is *approximately constant*.

$$\left|\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle - \mathrm{E}\left[\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle\right]\right| \le \mathbf{s}\sigma\sqrt{\frac{2}{\delta\mathbf{n}}} \quad \text{w.p.} \quad 1-\delta.$$

- It's almost always close to its expected value. And the way it differs is simple.
  - We can bound the difference without thinking about the model $\mathcal{M}$.
  - And our bound is small unless δ is very small, i.e. unless we want too much certainty.
- We'll use this to *sandwich* a bound between $s^2/2$ and $\max\ldots$ above.

$$\frac{s^2}{2} \overset{(a)}{\ge} \mathrm{E}\left[\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle\right] + s\sigma\sqrt{\frac{2}{\delta n}} \overset{(b)}{\ge} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle \quad \text{w.p.} \quad 1-\delta.$$

- (a). We choose s so this inequality is satisfied.
  - We have to do this every time we consider a new model.
  - But we don't have to worry about randomness. Both sides are deterministic.
- (b). This inequality follows from our 'approximate constancy' result.
  - We'll only have to prove that once. It's true for every model.
  - We'll do that next class using, and proving, the Efron-Stein inequality.

Intuition on Approximate Constancy

Why is $\left|\max_{m\in\mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu\rangle - \mathrm{E}\left[\max_{m\in\mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu\rangle\right]\right| \le \mathbf{s}\sigma\sqrt{\frac{2}{\delta\mathbf{n}}}$ w.p. $1-\delta$?

- The difference won't be too big compared to the maximum's standard deviation.

  $|Z - \mathrm{E}\, Z| < \frac{\mathrm{sd}(Z)}{\sqrt{\delta}}$ w.p. $1-\delta$ for any random variable Z. [Chebyshev's Inequality]
- So what's that standard deviation? Or what's the variance?
  - It's the variance of a maximum of sample means: one for each function m in $\mathcal{M}_s^\circ$.
  - So, as a starting point, let's bound the variance of a single one.

Intuition on Approximate Constancy

Why is $\left| \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu \rangle - \mathrm{E}\left[\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu \rangle \right] \right| \le \mathbf{s}\sigma\sqrt{\frac{2}{\delta \mathbf{n}}}$ w.p. $1-\delta$?

- The difference won't be too big compared to the maximum's standard deviation.

  $|Z - \mathrm{E}\, Z| < \frac{\mathrm{sd}(Z)}{\sqrt{\delta}}$ w.p. $1-\delta$ for any random variable Z. [Chebyshev's Inequality]

- So what's that standard deviation? Or what's the variance?
  - It's the variance of a maximum of sample means: one for each function m in $\mathcal{M}_s^\circ$.
  - So, as a starting point, let's bound the variance of a single one.

$$
\begin{aligned}
\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left\{ m(X_i) - \mu(X_i) \right\} \right] &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{E}\, \varepsilon_i^2 \left\{ m(X_i) - \mu(X_i) \right\}^2 \\
&= \frac{\sigma^2}{n} \times \|m-\mu\|^2 \\
&= \frac{\sigma^2}{n} \times s^2 \quad \text{for} \quad m \in \mathcal{M}_s^\circ.
\end{aligned}
$$

- Next class, we'll show that the variance of the maximum is *at most twice as large*.

$$
\mathrm{Var}\left[\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu \rangle \right] \le \frac{2\sigma^2}{n} \times s^2 \quad \text{so} \quad \mathrm{sd}(\ldots) \le s\sigma\sqrt{\frac{2}{n}}
$$

- Plugging it into Chebyshev's Inequality, we get our approximate constancy result.

Summary

$$\|\hat{\mu} - \mu\| < s \quad \text{w.p.} \quad 1-\delta \quad \text{if} \quad \frac{s^2}{2} > \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle + s\sigma\sqrt{\frac{2}{\delta n}}.$$

Let's introduce some notation to say this a bit more compactly.
Here's our characterization of s rephrased in terms of $g_i \overset{iid}{\sim} N(0,1)$.

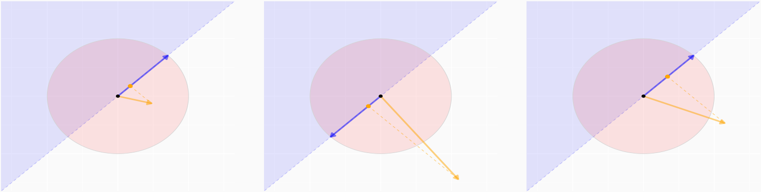
$\frac{s^2}{2\sigma} \geq \mathrm{w}(\mathcal{M}_s^\circ - \mu) + s\sqrt{\frac{2}{\delta n}}$	where
$\mathcal{M}_s^\circ - \mu := \{m - \mu : m \in \mathcal{M}_s^\circ\}$	is the *centered neighborhood boundary*,
$\mathrm{w}(\mathcal{V}) := \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v \rangle$	is the Gaussian width of the set $\mathcal{V}$.

The *Gaussian width* $\mathrm{w}(\mathcal{V})$ of a set of vectors $\mathcal{V} \subseteq \mathbb{R}^n$ is the expectation of the largest sample inner product between

- any vector $v \in \mathcal{V}$
- a vector g of independent standard normals

As a shorthand, we'll talk about the Gaussian width of a set of functions.
We'll mean the set of vectors we get when we evaluate them at $X_1 \ldots X_n$.

Visualizing Gaussian Width



- The gaussian width $\mathrm{w}(\mathcal{M}_s^\circ)$ is the expected value of something.
  - The maximum inner product between a gaussian vector g and a vector $v \in \mathcal{M}_s^\circ$.
  - We can think of this as the average of this maximum when we sample g over and over.

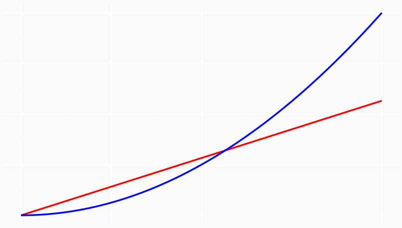
$$\mathrm{w}(\mathcal{M}_s^\circ) = \mathrm{E} \max_{v \in \mathcal{M}_s^\circ} \langle g, v \rangle \approx \frac{1}{m} \sum_{j=1}^{m} \max_{v \in \mathcal{M}_s^\circ} \langle g^{(j)}, v \rangle \quad \text{for} \quad g_i^{(j)} \overset{iid}{\sim} N(0,1).$$

- That's something we can draw and look at. I've drawn it three times above.
- In each, $\mathcal{M}$ is the blue region and $\mathcal{M}_s^\circ - \mu$ is the rim of the red semicircle. And …
  - I've sampled a gaussian vector g.
  - Then found the vector v in $\mathcal{M}_s^\circ$ maximizing the inner product $\langle g, \ v \rangle$.
  - It's $\|v\| = s$ times the length $\langle g, \ v/\|v\| \rangle$ of the projection indicated by the orange dot.
- The width of $\mathcal{M}_s^\circ$ is the average inner product $\langle g, \ v \rangle$ in infinitely many plots like these.

A Simplification

$$\|\hat{\mu} - \mu\| < s \quad \text{w.p.} \quad 1 - \delta \quad \text{if} \quad \frac{s^2}{2\sigma} > \mathrm{w}(\mathcal{M}_s^\circ - \mu) + s\sqrt{\frac{2}{\delta n}}.$$

The right side's second term tends to be small. We can ignore it and be vague.



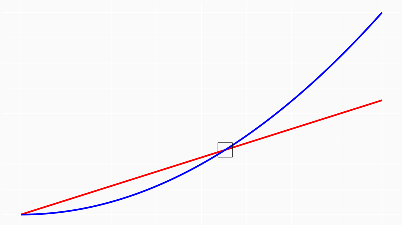
$\|\hat{\mu} - \mu\|$ isn't much bigger than s with high probability if $s^2 \geq 2\sigma\,\mathrm{w}(\mathcal{M}_s^\circ - \mu)$.

What is the smallest s for which this is true?

A Simplification

$$\|\hat{\mu} - \mu\| < s \quad \text{w.p.} \quad 1 - \delta \quad \text{if} \quad \frac{s^2}{2\sigma} > \mathrm{w}(\mathcal{M}_s^\circ - \mu) + s\sqrt{\frac{2}{\delta n}}.$$

The right side's second term tends to be small. We can ignore it and be vague.



$\|\hat{\mu} - \mu\|$ isn't much bigger than s with high probability if $s^2 \geq 2\sigma\,\mathrm{w}(\mathcal{M}_s^\circ - \mu)$.

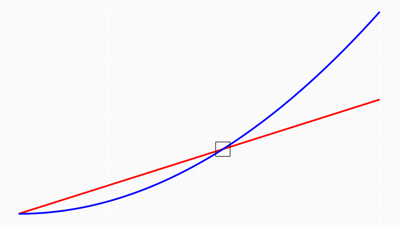
What is the smallest s for which this is true?

The point at which the red and blue curves intersect.

A Simplification

$$\|\hat{\mu}-\mu\| < s \quad \text{w.p.} \quad 1-\delta \quad \text{if} \quad \frac{s^2}{2\sigma} > \mathrm{w}(\mathcal{M}_s^\circ - \mu) + s\sqrt{\frac{2}{\delta n}}.$$

The right side's second term tends to be small. We can ignore it and be vague.

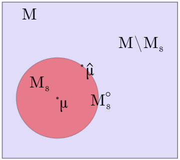


But we can can boil things down to the same picture and still be precise.

$$\|\hat{\mu}-\mu\| < s + 2\sigma\sqrt{\frac{2}{\delta n}} \quad \text{w.p.} \quad 1-\delta \quad \text{if} \quad {\color{blue}s^2} \ge {\color{red}2\sigma\,\mathrm{w}(\mathcal{M}_s - \mu)}$$

- We'll prove it for homework. It's not hard to derive from the first bound.
- The key idea is that $\mathrm{w}(\mathcal{M}_s - \mu)$ is a *sublinear* function of s.

Caveat



$$\|\hat{\mu}-\mu\| < s \qquad \text{w.p.} \quad 1-\delta \quad \text{if} \quad \frac{s^2}{2\sigma} > \mathrm{w}(\mathcal{M}_s^{\circ}-\mu) + s\sqrt{\frac{2}{\delta n}}$$

$$\|\hat{\mu}-\mu\| < s + 2\sigma\sqrt{\frac{2}{\delta n}} \qquad \text{w.p.} \quad 1-\delta \quad \text{if} \quad s^2 \geq 2\sigma\, \mathrm{w}(\mathcal{M}_s-\mu)$$

- In our second bound, we use the width of the centered neighborhood $\mathcal{M}_s-\mu$ instead of its boundary $\mathcal{M}_s^{\circ}-\mu$.
  - We need to do this because the second requires width to grow sublinearly with s.
  - When $\mathcal{M}_s$ is convex, this is true for $\mathrm{w}(\mathcal{M}_s-\mu)$, but not necessarily for $\mathrm{w}(\mathcal{M}_s^{\circ}-\mu)$.
- This width is larger because it's a maximum over a larger (i.e. containing) set.

$$\underbrace{\mathrm{E}\max_{m\in\mathcal{M}_s}\langle g, m-\mu\rangle}_{\mathrm{w}(\mathcal{M}_s-\mu)} \geq \underbrace{\mathrm{E}\max_{m\in\mathcal{M}_s^{\circ}}\langle g, m-\mu\rangle}_{\mathrm{w}(\mathcal{M}_s^{\circ}-\mu)} \quad \text{because} \quad \mathcal{M}_s \supseteq \mathcal{M}_s^{\circ}.$$

  - But usually we don't pay much—if anything—for making this substitution.
  - When we try to bound $\mathrm{w}(\mathcal{M}_s-\mu)$, we'll often get one that applies to $\mathrm{w}(\mathcal{M}_s)$ too.
- If it does matter, and we really want to work with $\mathrm{w}(\mathcal{M}_s^{\circ}-\mu)$, we can.
  - We'll just need to use the first (messier) $s^2 \geq \max\ldots$ bound.
  - Or take a different approach to simplifying it that doesn't require sublinearity.

Implications

Example 1

Suppose we have a width bound that is proportional to s: $\alpha s \geq \mathrm{w}(\mathcal{M}_s - \mu)$.

$$
\begin{aligned}
s^2 \geq 2\sigma\,\mathrm{w}(\mathcal{M}_s - \mu) \text{ if } \quad & s^2 \geq 2\sigma\alpha s \\
\text{and therefore if } \quad & s \geq 2\sigma\alpha.
\end{aligned}
$$

This is what happens when we use a model with K parameters.

$$
c\sqrt{\frac{K}{n}}\, s \geq \mathrm{w}(\mathcal{M}_s - \mu) \quad \text{and therefore} \quad s = 2\sigma c\sqrt{\frac{K}{n}} \quad \text{works.}
$$

- One interesting thing we can do with this is see what happens when we choose model size as a function of sample size.
- Our estimator converges when our model's size is much smaller than sample size.

$$
\begin{aligned}
\lVert \hat{\mu} - \mu \rVert &\leq 2\sigma c\sqrt{\frac{K}{n}} + 2\sigma\sqrt{\frac{2}{\delta n}} \quad \text{w.p.} \quad 1 - \delta \\
&\to 0 \quad \text{if} \quad K/n \to 0
\end{aligned}
$$

This and finite-dimensional approximation get you pretty far with infinite-dimensional models, too.

Example 2

Suppose we have a width bound that doesn't depend on s: $\beta \geq \mathrm{w}(\mathcal{M}_s - \mu)$.

$$s^2 \geq 2\sigma\,\mathrm{w}(\mathcal{M}_s - \mu) \text{ if } \quad s^2 \geq 2\sigma\beta$$
$$\text{and therefore if } \quad s \geq \sqrt{2\sigma\beta}.$$

This is what happens when our model is the set of *weighted averages* of K functions.

$$\mathrm{w}(\mathcal{M}_s - \mu) \leq \sqrt{c\log(K)/n} \quad \text{and therefore} \quad s = \sqrt{2}\,\sqrt[4]{c\sigma\log(K)/n} \quad \text{works.}$$

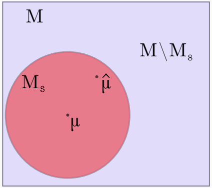
- We get a $n^{-1/4}$ rate essentially independent of the number of functions K.
- This is at the heart of the Assumptionless Analysis of the Lasso.

$$\|\hat{\mu} - \mu\| \leq \sqrt{2}\,\sqrt[4]{c\sigma\log(K)/n} + 2\sigma\sqrt{\frac{2}{\delta n}} \quad \text{w.p.} \quad 1-\delta$$
$$\to 0 \quad \text{if} \quad \log(K)/n \to 0.$$

Summary

With high probability …

the error of the least squares estimator in a convex model $\mathcal{M}$ is smaller than a radius s determined by the Gaussian width of the *centered neighborhood boundary*.



$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \le s + 2\sigma\sqrt{\frac{2}{\delta n}}$$

with probability $1 - \delta$ if

$$s^2 \ge 2\sigma\, \mathrm{w}(\mathcal{M}_s - \mu)$$

- That's true if μ is in the model.
  - If it's not, it's true about distance to the best approximation to μ in the model.
  - We'll talk about that next lecture.
- It's also true that this is about as good a guarantee as you can get.
- This means that the study of least squares estimation is, for the most part, just the study of Gaussian width.

Gaussian Width

Gaussian Width

An Example
The Completely General Model

The General Model

Consider a model $\mathcal{M}$ that contains every curve outright.

$$\mathcal{M} = \{ \text{ all curves } m(x) \}$$

Or, as we're interested in its values on the sample, the corresponding set of vectors.

$$\mathcal{M} = \{ \text{ all vectors } \vec{m} \in \mathbb{R}^n \ : \ \vec{m}_i = \vec{m}_j \quad \text{if} \quad X_i = X_j \} \subseteq \{ \text{ all vectors } \vec{m} \in \mathbb{R}^n \} = \tilde{\mathcal{M}}.$$

To keep things simple, we can enlarge this set by dropping the function constraint.

A neighborhood is just a ball centered on $\vec{\mu} = \mu(X_1) \ldots \mu(X_n)$.

$$\tilde{\mathcal{M}}_s = \left\{ \vec{m} : \|\vec{m} - \vec{\mu}\|_{L_2(\mathrm{P_n})} \le s \right\}$$

And when we center it, we get a ball around the origin.

$$\tilde{\mathcal{M}}_s - \mu = \left\{ \vec{m} - \vec{\mu} : \|\vec{m} - \vec{\mu}\|_{L_2(\mathrm{P_n})} \le s \right\} = \left\{ v \in \mathbb{R}^n : \|v\|_{L_2(\mathrm{P_n})} \le s \right\}.$$

The General Model's Gaussian Width

$$\mathrm{w}(\tilde{\mathcal{M}}_s - \mu) = \mathrm{E} \max_{\substack{v \in \mathbb{R}^n \\ \|v\|_{L_2(\mathrm{P_n})} \le s}} \langle g, v \rangle_{L_2(\mathrm{P_n})}$$

No matter what gaussian vector we get, this set has a vector v with 2-norm s in exactly the same direction. It's got one in every direction.

$$\mathrm{w}(\tilde{\mathcal{M}}_s - \mu) = \mathrm{E} \max_{v \in \mathcal{M}_s} \|g\|_{L_2(\mathrm{P_n})} \|v\|_{L_2(\mathrm{P_n})} = s \,\mathrm{E}\|g\|_{L_2(\mathrm{P_n})}.$$

And the expected sample two-norm of a gaussian vector $g \in \mathbb{R}^n$ is roughly 1. That's the square root of its expected *squared* sample two-norm.

$$\mathrm{E}\|g\|_{L_2(\mathrm{P_n})} \approx \sqrt{\mathrm{E}\|g\|^2_{L_2(\mathrm{P_n})}} = \sqrt{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[g_i^2]} = \sqrt{1}.$$

So the gaussian width we get is roughly just s. Let's call it s. Our bound doesn't tell us that we'll ever get close to μ at all.

$$s^2 \ge 2\sigma \,\mathrm{w}(\mathcal{M}_s - \mu) \quad \text{if} \quad s^2 \ge 2\sigma s \quad \text{i.e. if} \quad s \ge 2\sigma.$$

Wouldn't that be too much to expect—to get close without any assumptions at all?

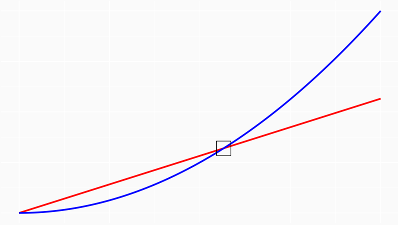
Lab Preview

Computing Gaussian Width

- Gaussian width is the mean of something we can compute samples of.
- That means we can approximate it by a sample average.

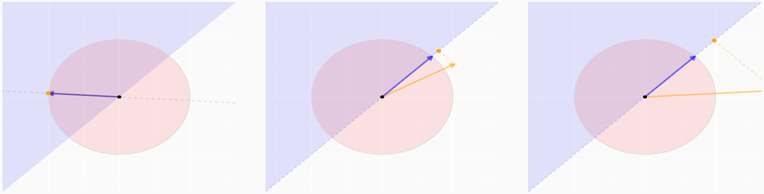
$$\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g, v \rangle \approx \frac{1}{m} \sum_{j=1}^{m} \max_{v \in \mathcal{V}} \langle g^{(j)}, v \rangle \quad \text{for} \quad g_i^{(j)} \overset{iid}{\sim} N(0,1).$$

- This means we can calculate a neighborhood's Gaussian width whether we can work out how to do it analytically or not.
  - Next week, we'll use CVXR to do it.
  - And search over the radius s to find an error bound.



$$\|\hat{\mu} - \mu\| < s + 2\sigma\sqrt{\frac{2}{\delta n}} \quad \text{w.p.} \quad 1 - \delta \quad \text{if} \quad {\color{blue} s^2} \geq {\color{red} 2\sigma\, \mathrm{w}(\mathcal{M}_s - \mu)}$$

A Warm Up Lab



- To prepare, we'll do a drawing exercise to get a feel for gaussian width.
  - We'll work out what vectors are in some models when we have just two observations.
  - And use our drawings, as described earlier, to calculate width in the 2 observation case.
- We'll vary the radius s to search for an error bound, too.