Machine Learning Theory

Lecture 5: Least Squares in Infinite Models i.e. Regression

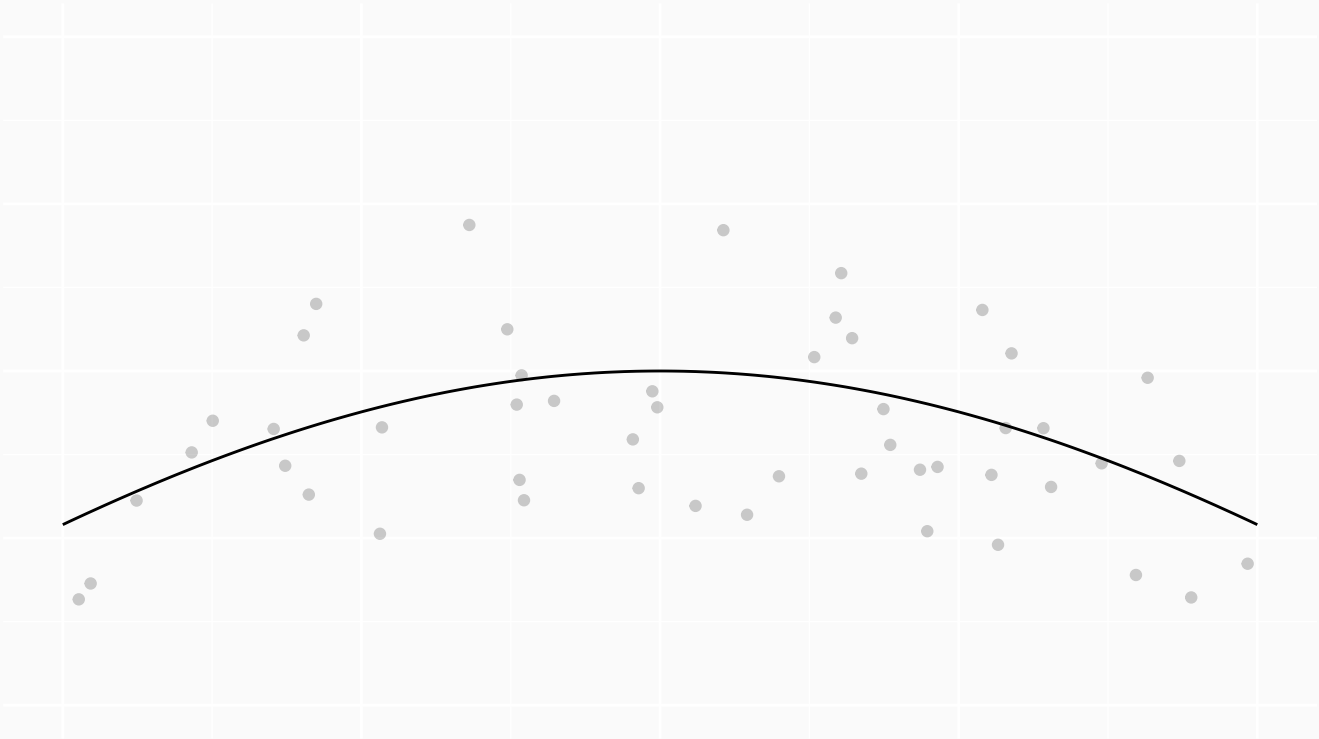
David A. Hirshberg

May 24, 2024

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.

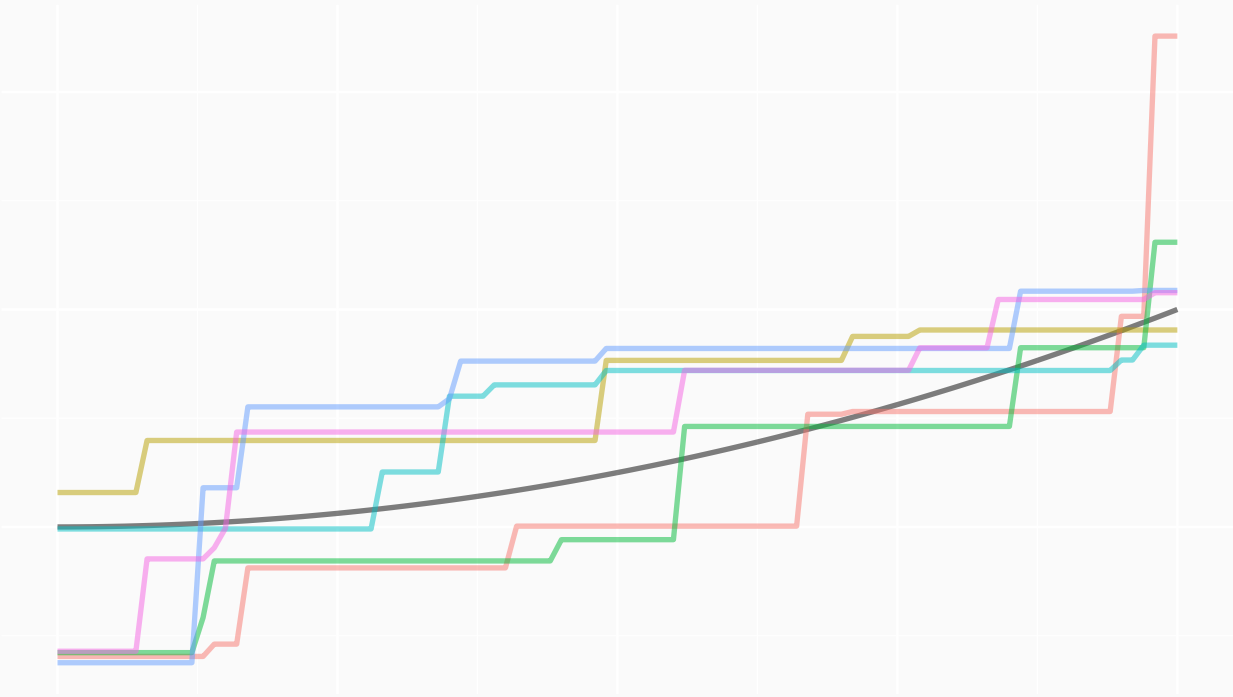
New Ones

- $\mathcal{M}_s$, 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

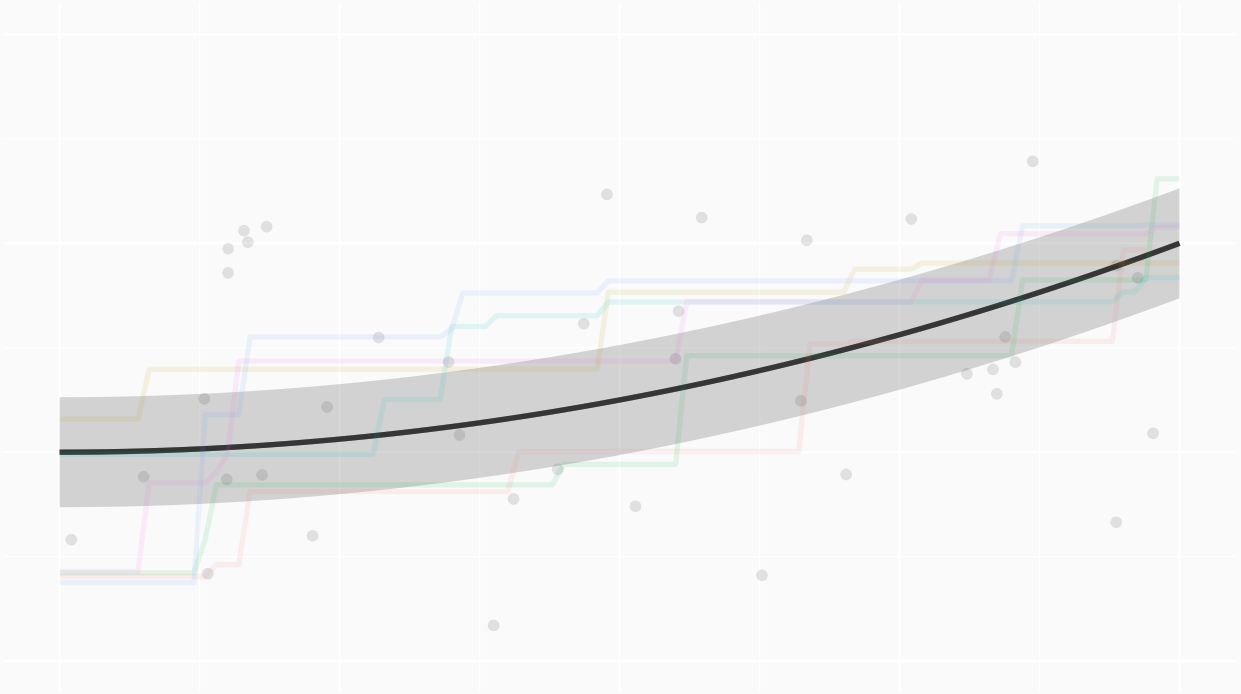
Last week, our model was a finite set of curves.



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.

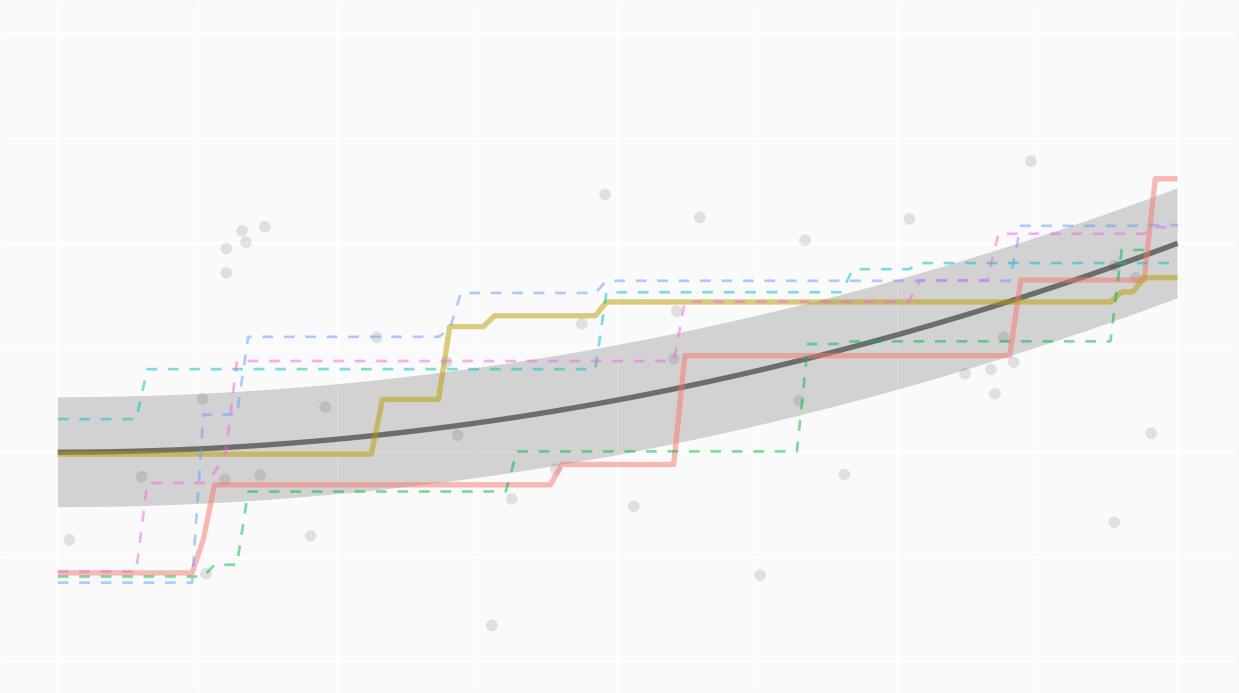
Caveat.

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.



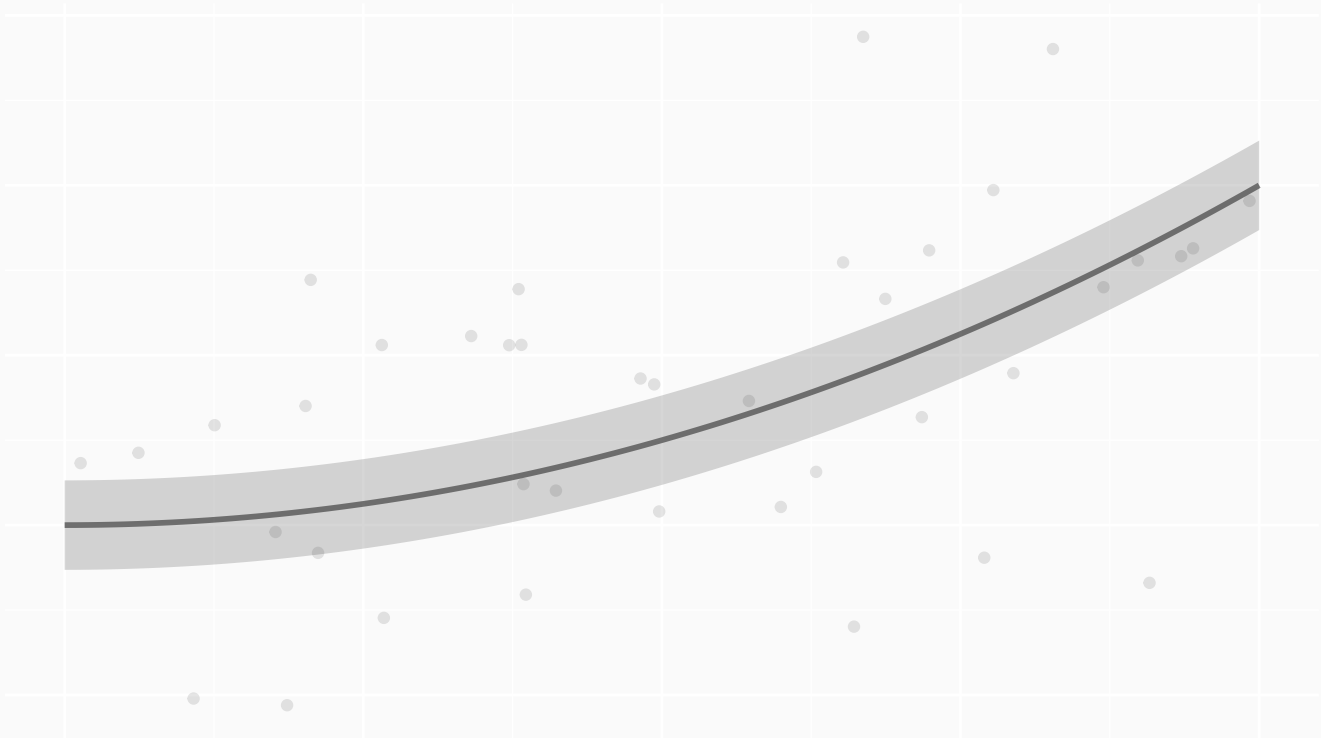
We're talking about the set of curves that are close in terms of the sample two-norm.

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

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.

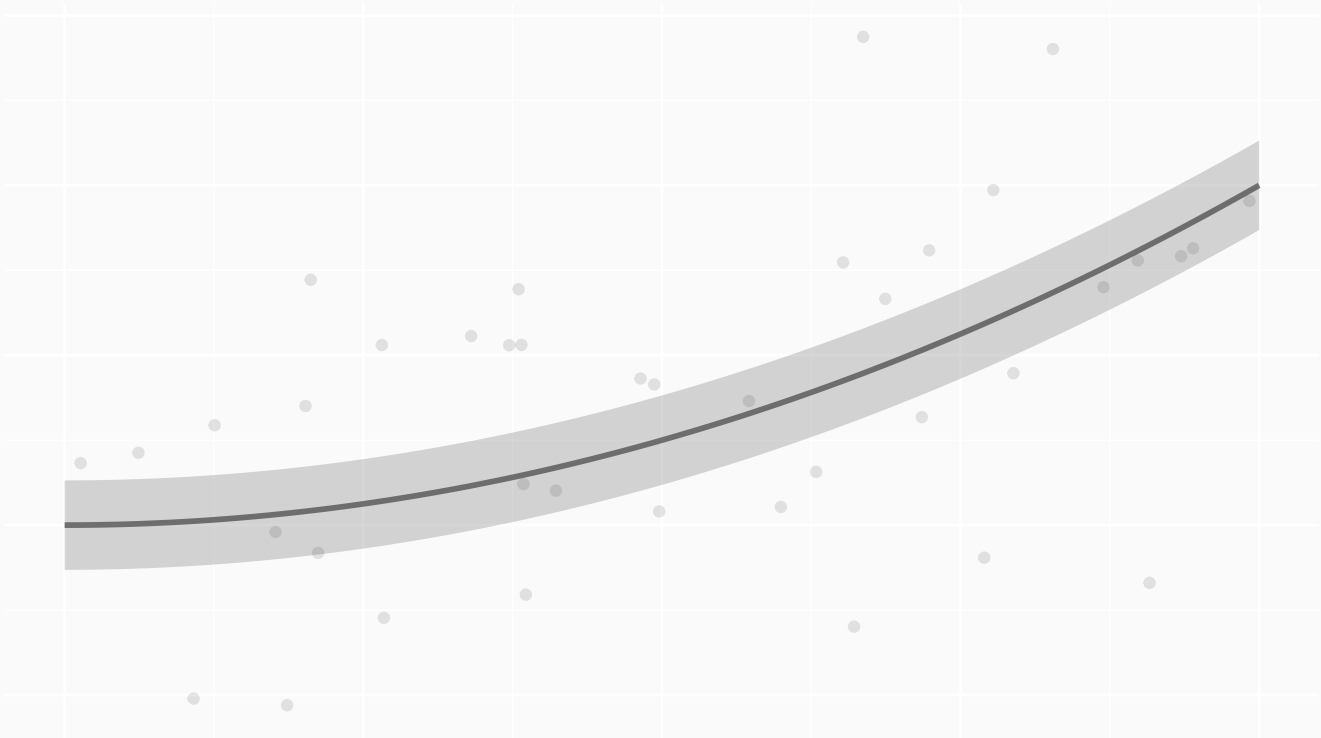
Same caveat.

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.

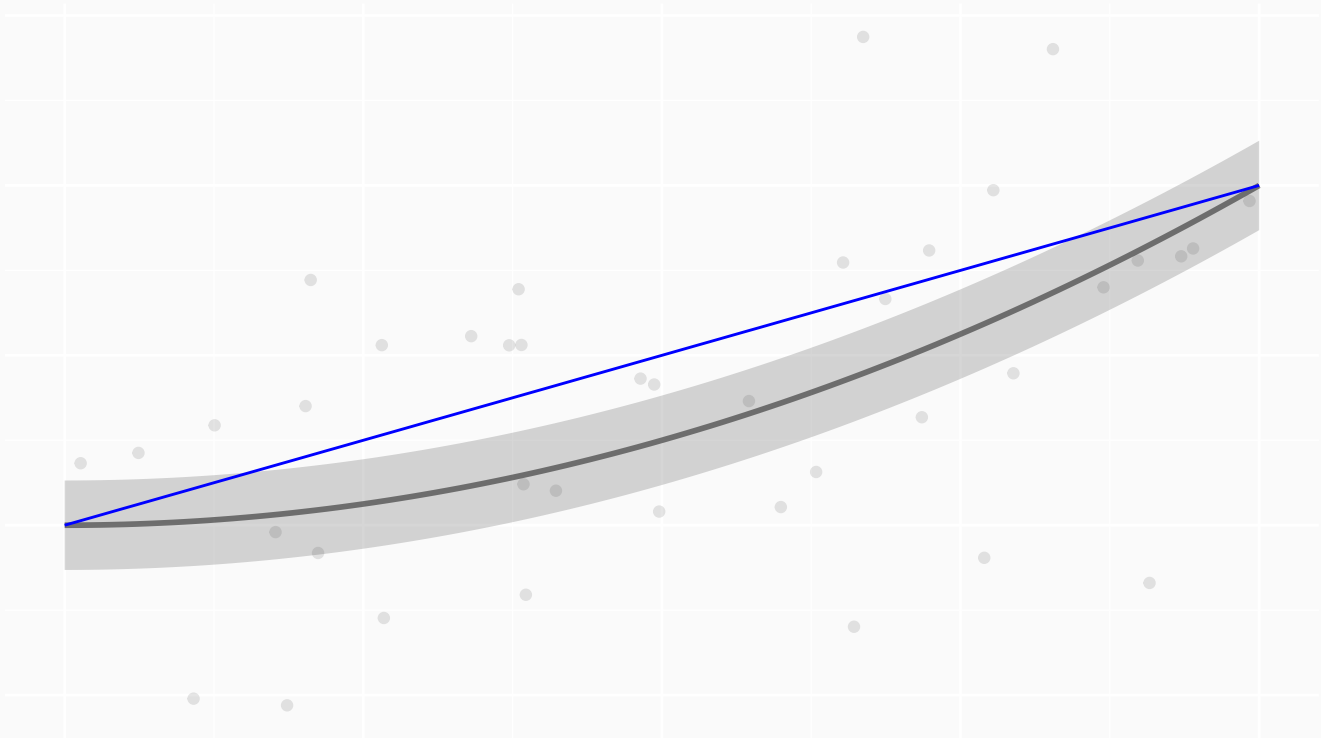
$$\mathcal{M}_s = \{m \in \mathcal{M} : \|m - \mu\|_{L_2(\mathrm{P_n})} < s\}$$

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

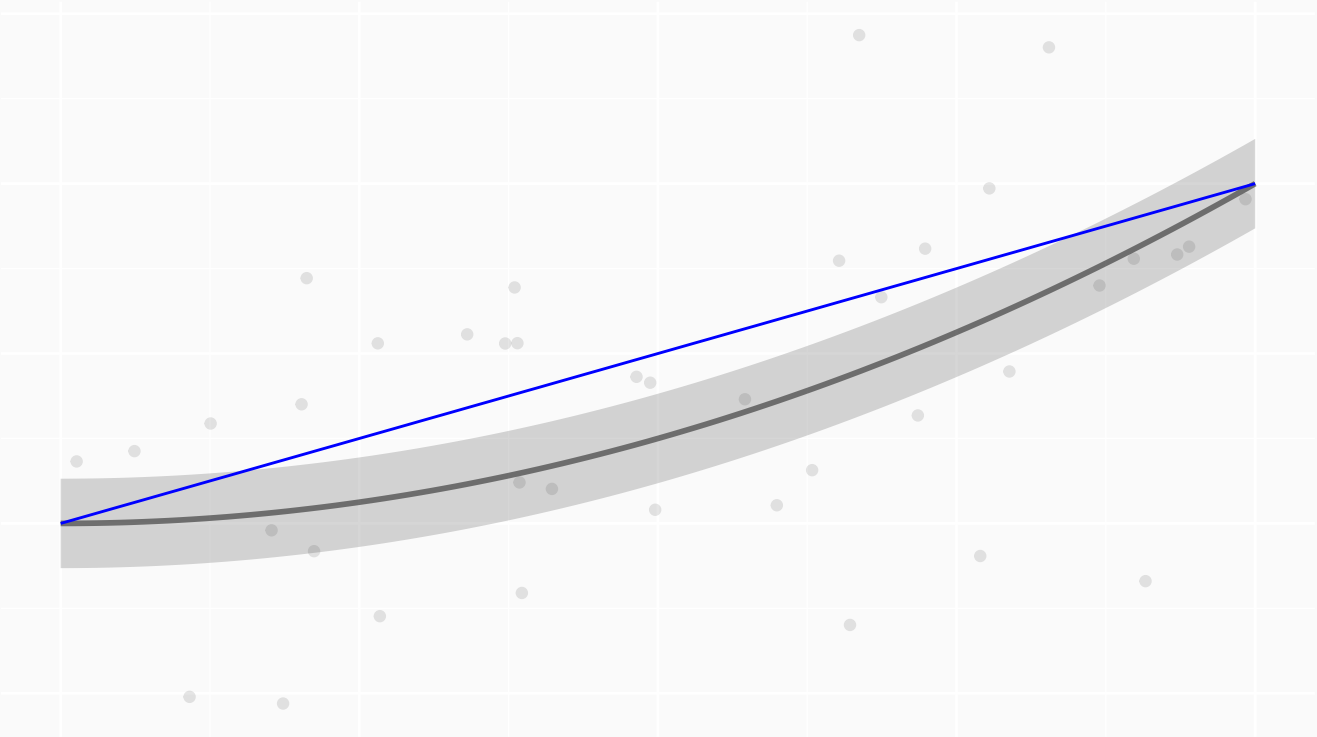
Let’s take the set of increasing curves to be our regression model.



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.

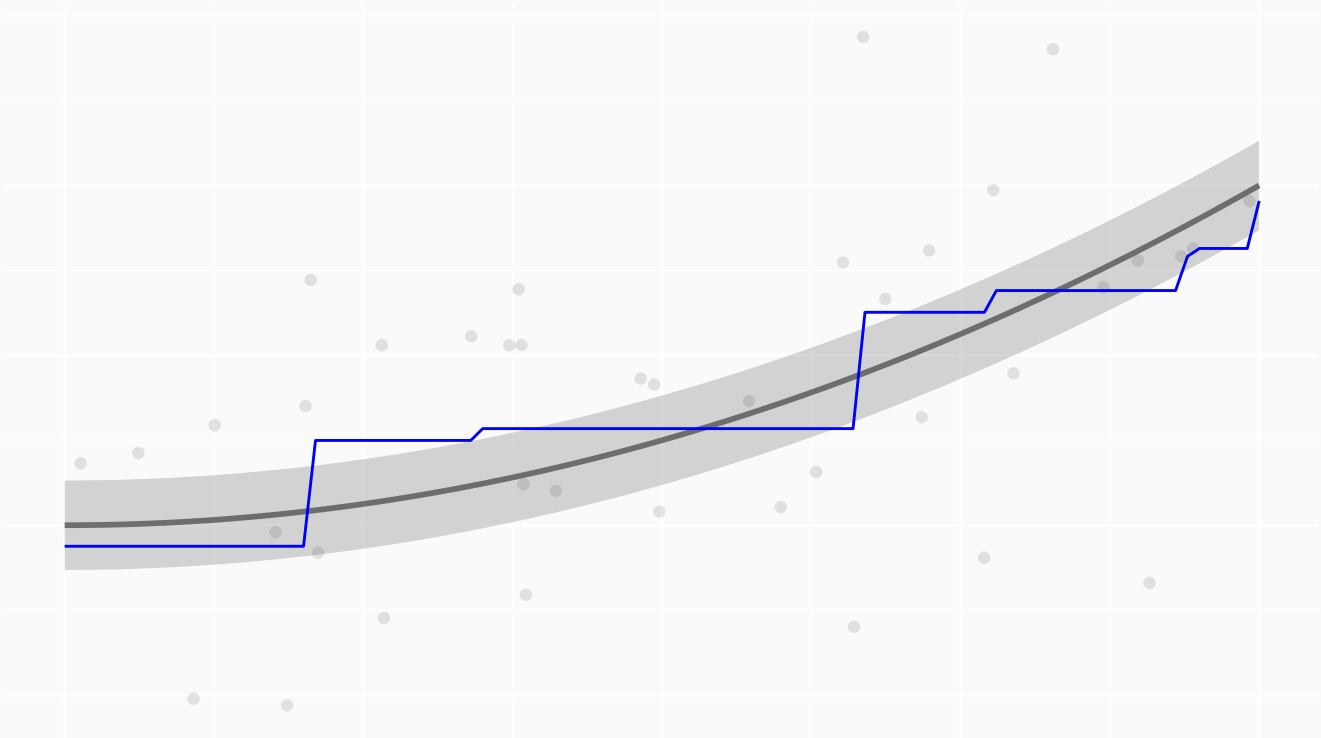


Is this in our neighborhood?

No. It's too far from μ

Neighborhoods in models with infinitely-many curves

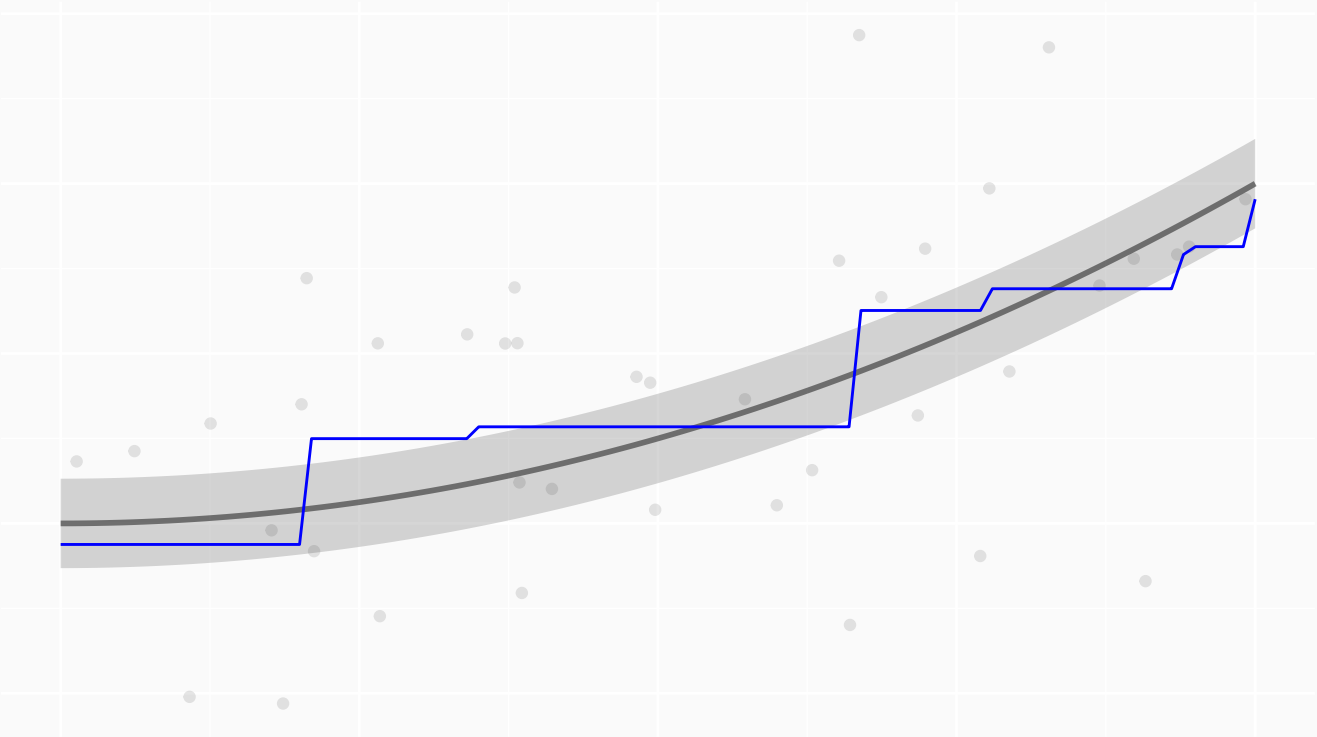
Let’s take the set of increasing curves to be our regression model.



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.

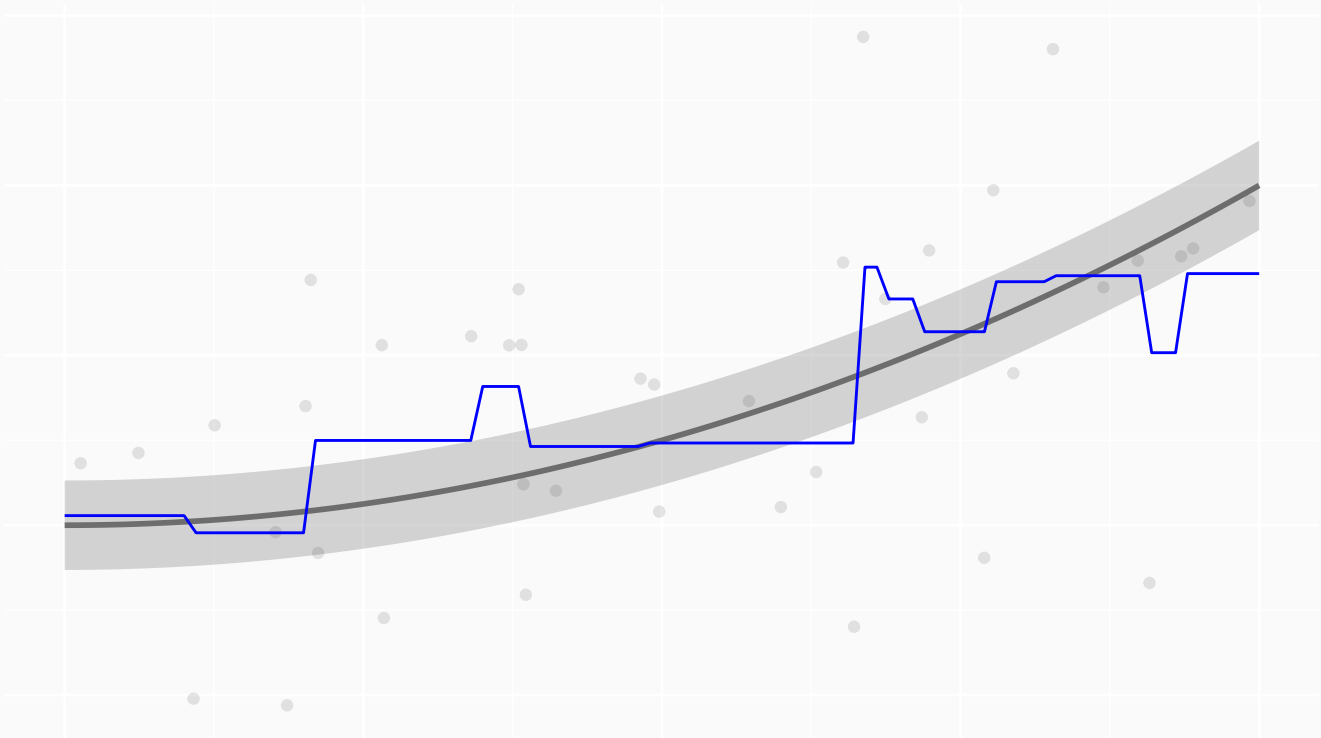


Is this in our neighborhood?

Yes.

Neighborhoods in models with infinitely-many curves

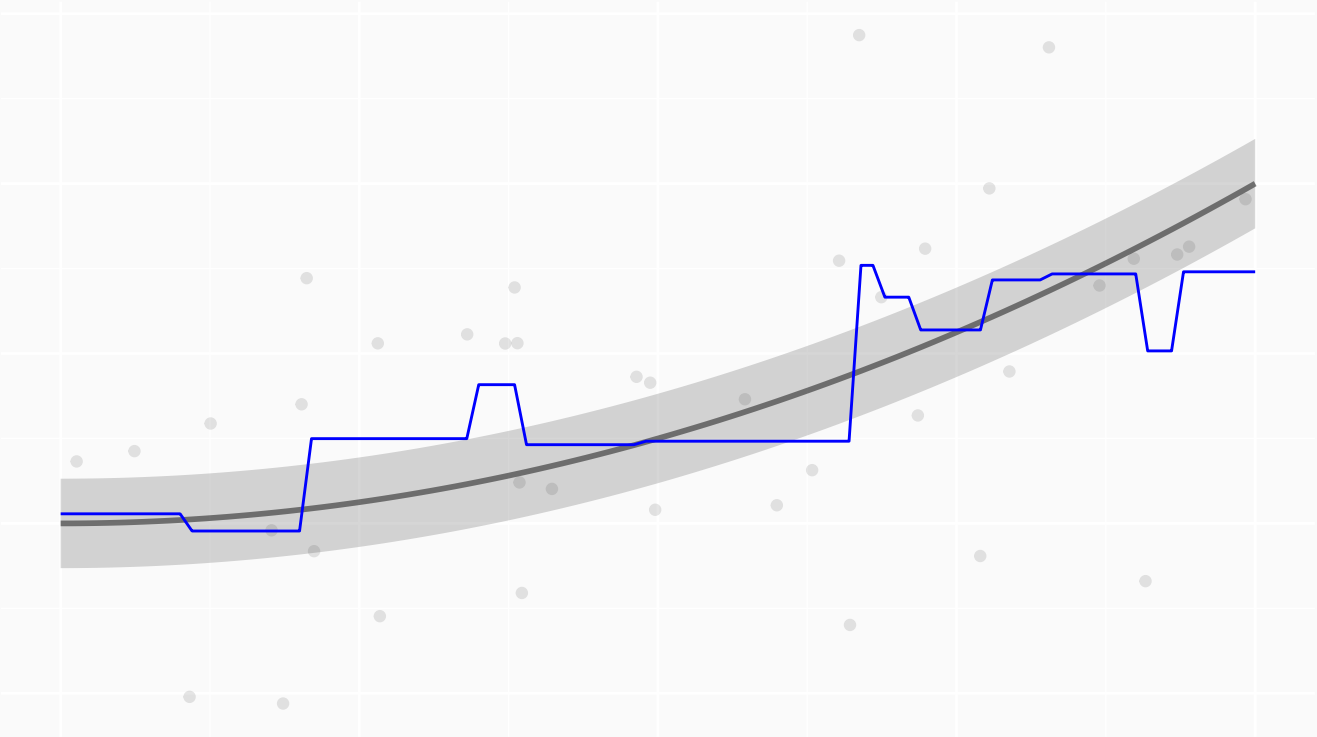
Let’s take the set of increasing curves to be our regression model.



Is this in our neighborhood?

Neighborhoods in models with infinitely-many curves

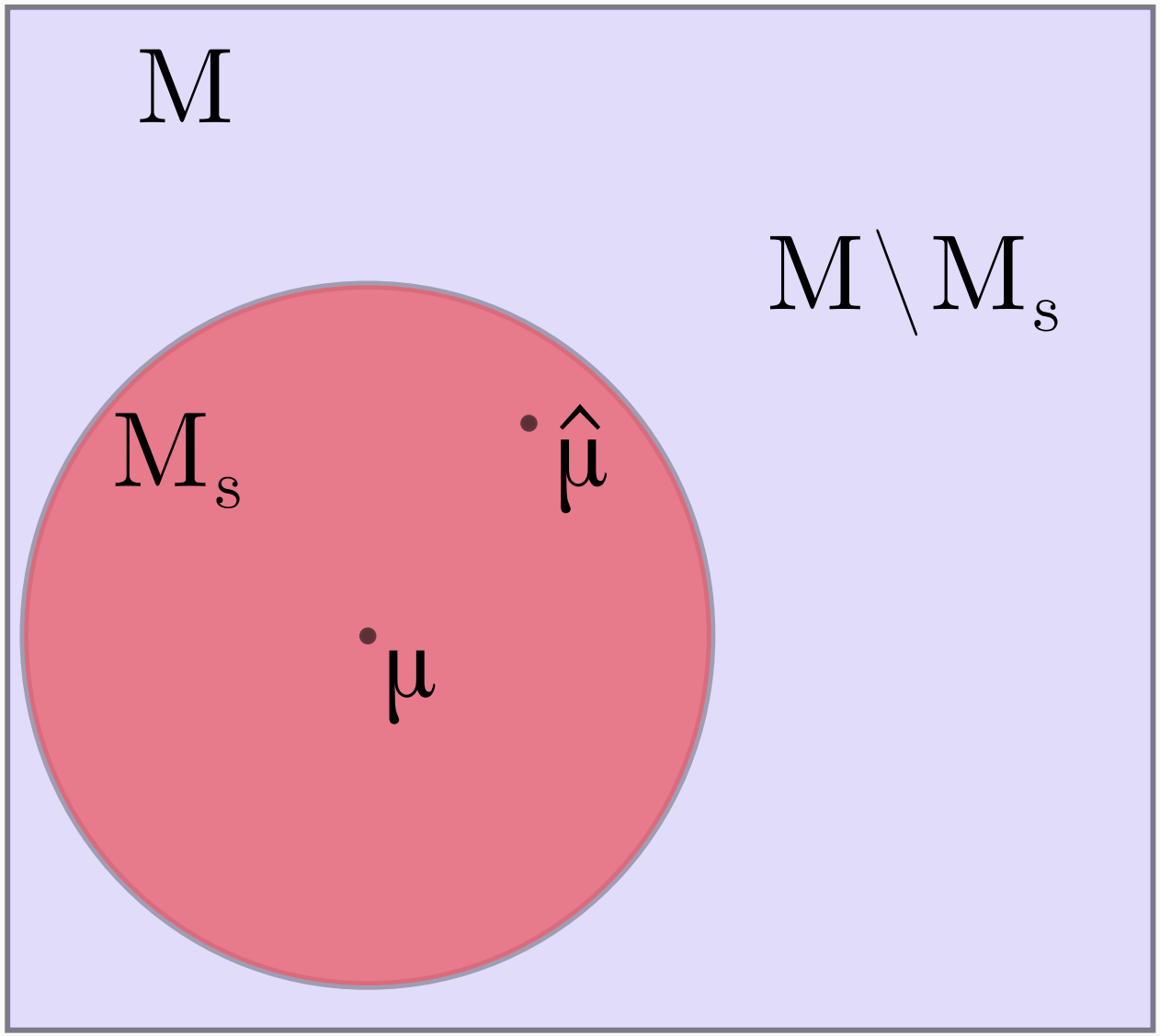
Let's take the set of increasing curves to be our regression model.



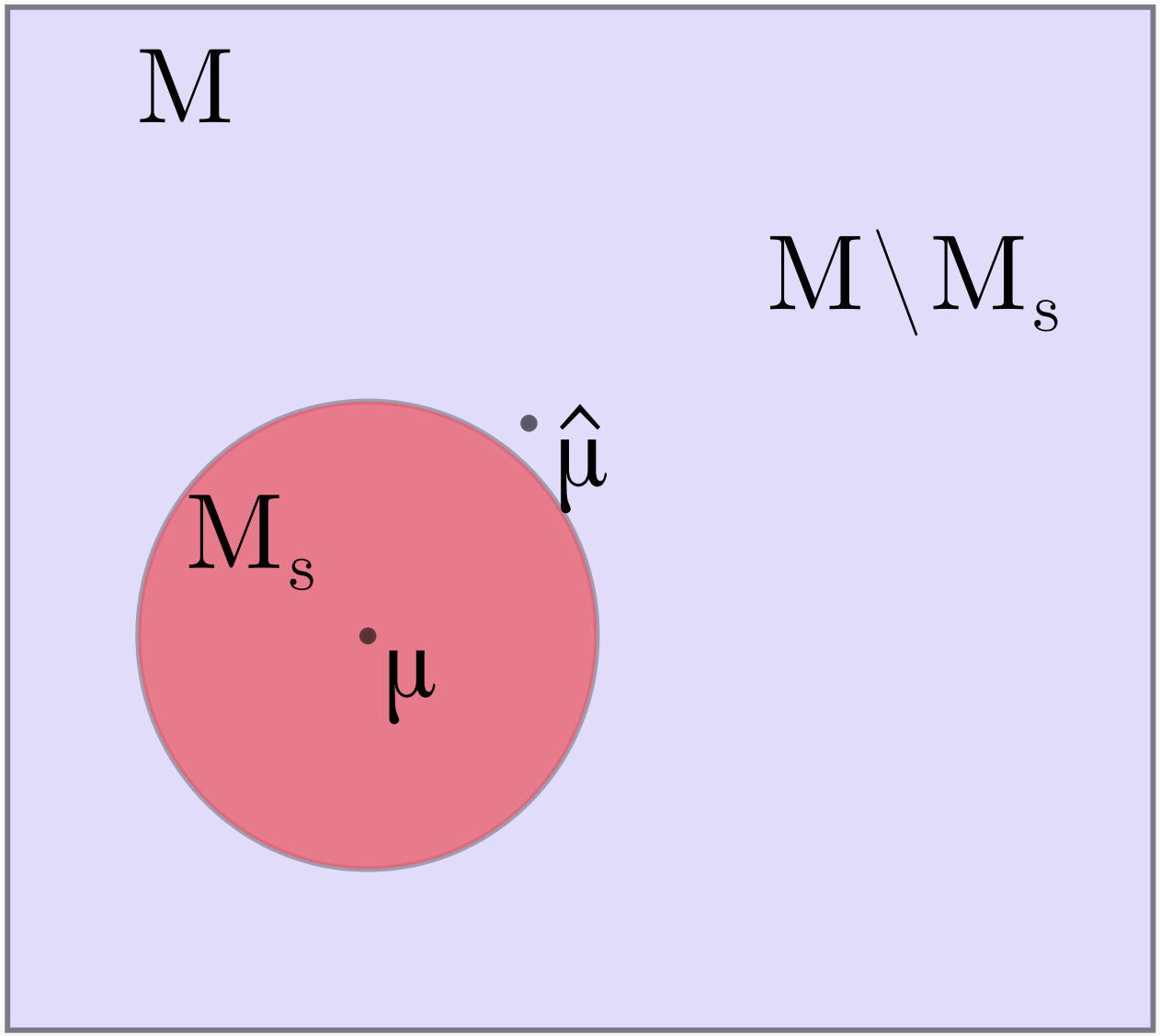
Is this in our neighborhood?

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

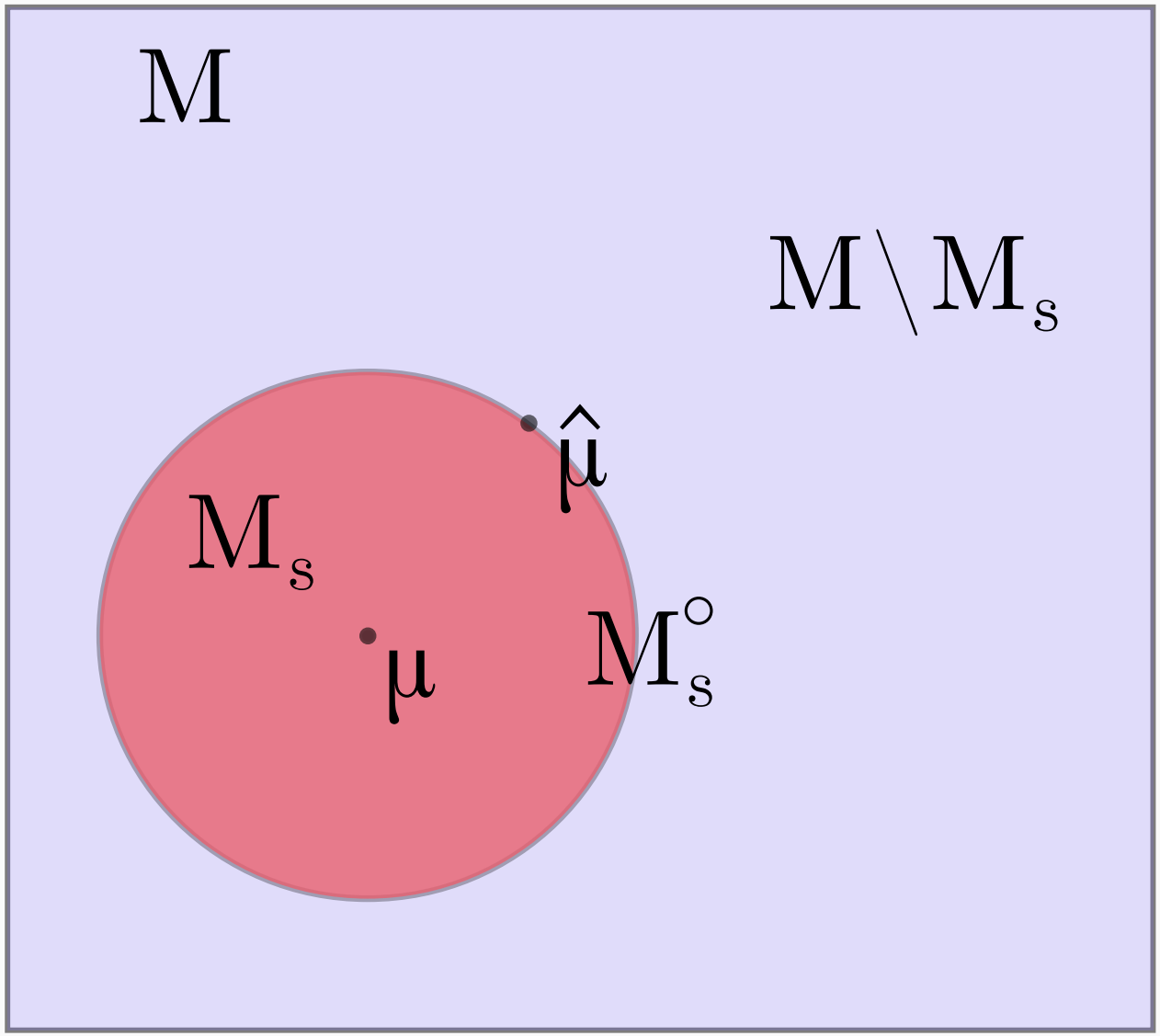
What we're going to prove today



What we're going to rule out



In fact, we'll show it's enough to rule this 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.

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

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

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

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

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

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

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

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.

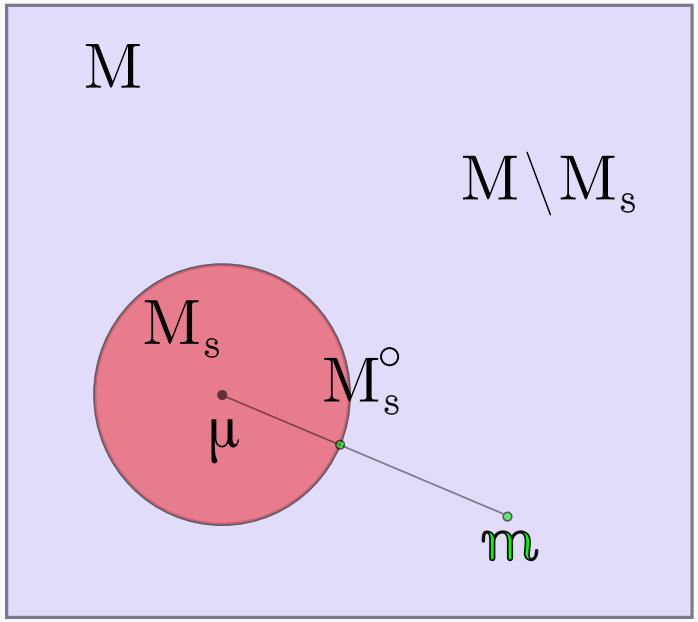
A Reduction

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.

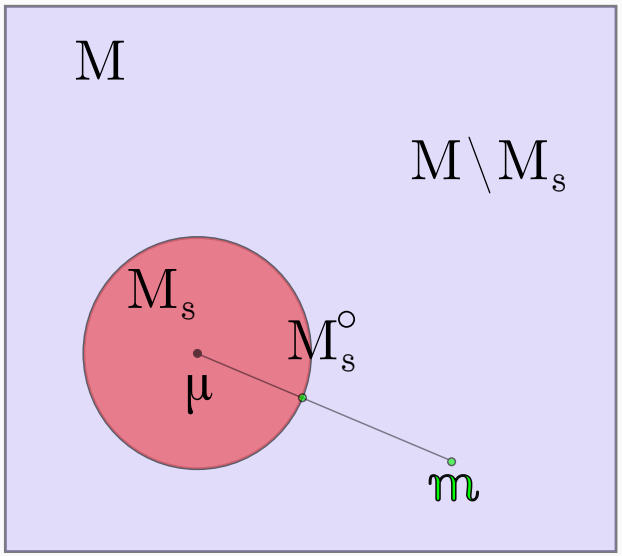
$$\underbrace{\ell(m) > \ell(\mu) \quad \text{for all} \quad m \in \mathcal{M}_s^\circ}_{\text{the thing we're going to prove}} \implies \underbrace{\ell(m) > \ell(\mu) \quad \text{for all} \quad m \in \mathcal{M} \setminus \mathcal{M}_s.}_{\text{the thing we said we needed to prove}}$$



- 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 $t \in [0, 1]$.

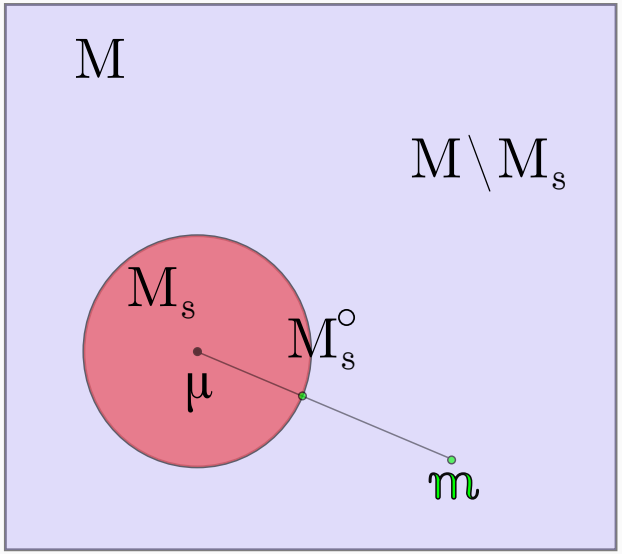
We'll show the loss difference $\ell(m) - \ell(\mu)$ for any curve in the complement is at least t **times** the loss difference $\ell(m_t) - \ell(\mu)$ for its representative.

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

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.

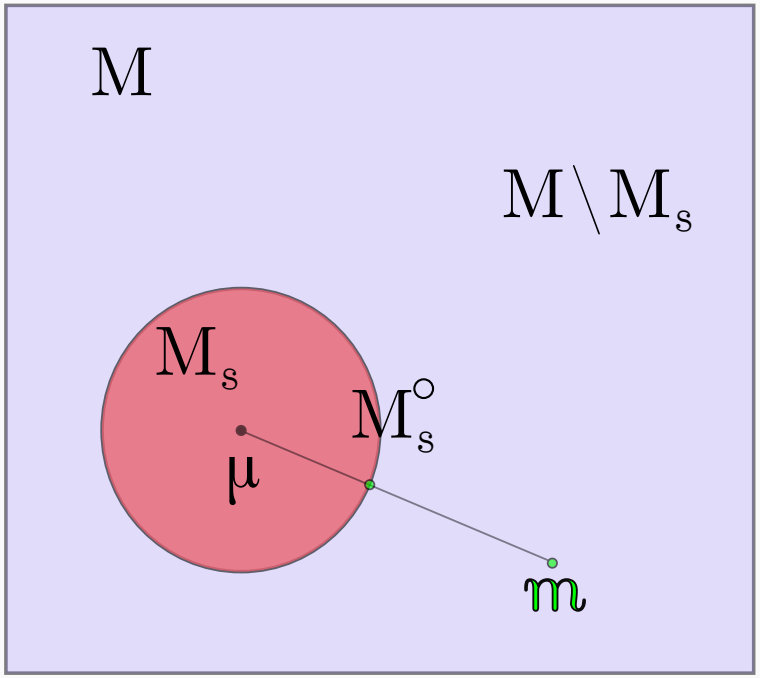


A representative is a point
$m_t = \mu + t(m - \mu)$ for some $t \in [0, 1]$.

$$
\begin{aligned}
\ell(m_t) - \ell(\mu) &= \frac{1}{n}\sum_{i=1}^{n}\{m_t - \mu\}^2 - \frac{2}{n}\sum_{i=1}^{n}\{Y - \mu\}\{m_t - \mu\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\{\mu + t(m - \mu) - \mu\}^2 - \frac{2}{n}\sum_{i=1}^{n}\{Y - \mu\}\{\mu + t(m - \mu) - \mu\} \\
&= \frac{t^2}{n}\sum_{i=1}^{n}\{m - \mu\}^2 - \frac{2t}{n}\sum_{i=1}^{n}\{Y - \mu\}\{m - \mu\} \\
&\leq t\{\ell(m) - \ell(\mu)\} \quad \text{because} \quad t^2 \leq t.
\end{aligned}
$$

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.

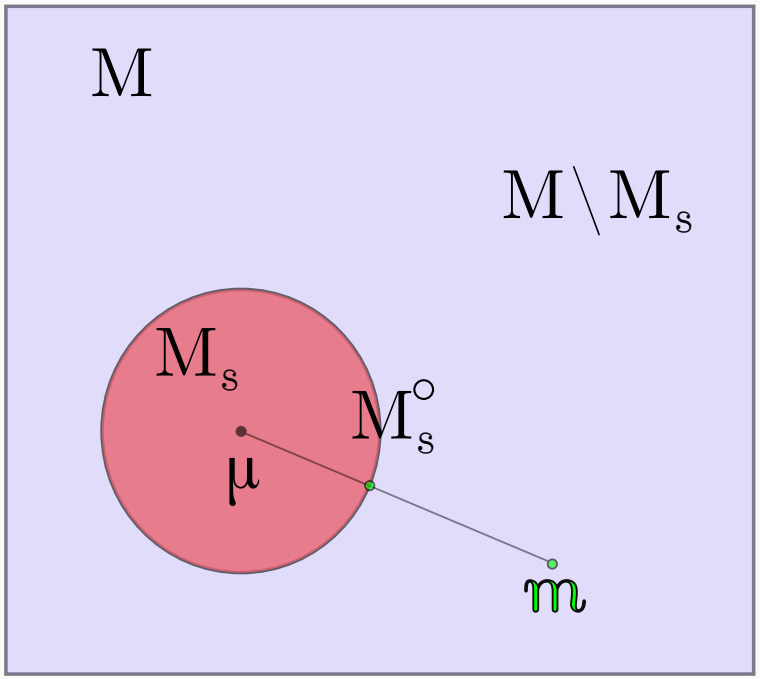


Definition.

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?

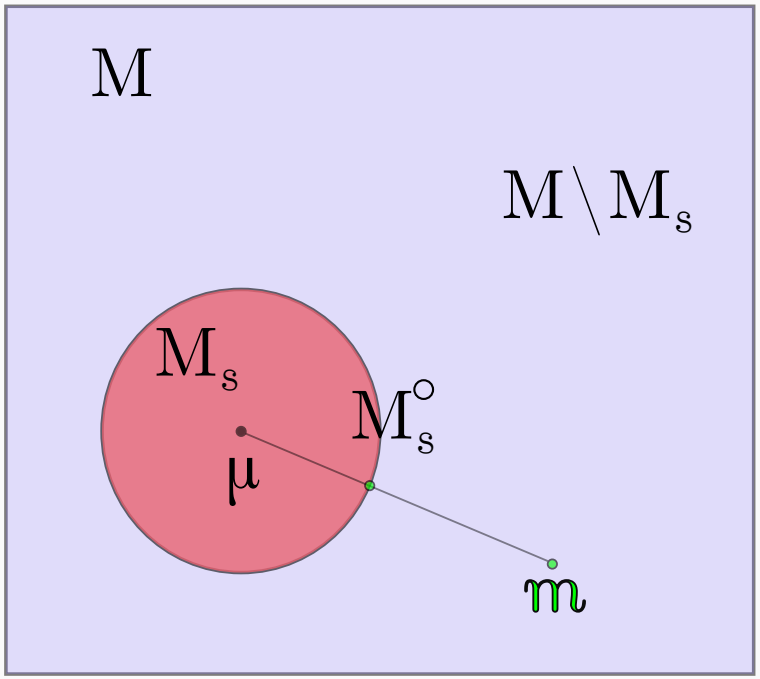


Hint.

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.

Differences in squared error

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.

$$\begin{aligned}\ell(m) - \ell(\mu) &= \frac{1}{n}\sum_{i=1}^{n}\{m(X_i) - \mu(X_i)\}^2 - \frac{2}{n}\sum_{i=1}^{n}\{ Y_i - \mu(X_i)\}\{m(X_i) - \mu(X_i)\} \\ &= s^2 - \frac{2}{n}\sum_{i=1}^{n}\{ Y_i - \mu(X_i)\}\{m(X_i) - \mu(X_i)\}.\end{aligned}$$

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

$$s^2 > \frac{2}{n}\sum_{i=1}^{n}\{ Y_i - \mu(X_i)\}\{m(X_i) - \mu(X_i)\} \quad \text{for all} \quad m \in \mathcal{M}_s^\circ$$

or equivalently

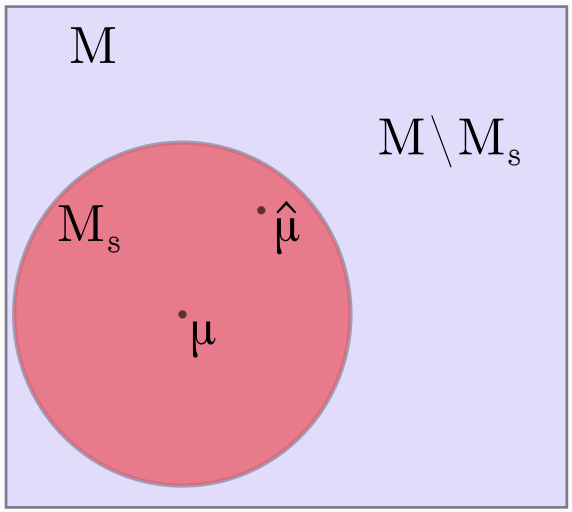
$$s^2 > \max_{m \in \mathcal{M}_s^\circ} \frac{2}{n}\sum_{i=1}^{n}\{ Y_i - \mu(X_i)\}\{m(X_i) - \mu(X_i)\}.$$

All we've got to do is bound maximal correlation with noise

We want to find s satisfying this with high probability.

$$
\begin{aligned}
s^2 &> \max_{m \in \mathcal{M}_s^\circ} \frac{2}{n} \sum_{i=1}^{n} \{ Y_i - \mu(X_i) \} \{ m(X_i) - \mu(X_i) \} \\
&= \max_{m \in \mathcal{M}_s^\circ} \frac{2\sigma}{n} \sum_{i=1}^{n} g_i \{ m(X_i) - \mu(X_i) \} && \text{when} \quad Y_i = \mu(X_i) + \sigma g_i \\
&= 2\sigma \max_{m \in \mathcal{M}_s^\circ} \langle g,\ m - \mu \rangle_{L_2(\mathrm{P_n})} && \text{in more compact notation .}
\end{aligned}
$$

This implies that every curve in the complement has bigger loss than μ.
When it's satisfied, we know that $\hat{\mu} \in \mathcal{M}_s$, i.e., $\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s$



Bounding its mean is good enough

We want a radius s with s^2 bigger than this maximal correlation.

$$s^2 > 2\sigma \max_{m \in \mathcal{M}_s^\circ} \langle g,\ m - \mu \rangle_{L_2(\mathrm{P_n})} \text{ with probability } \quad 1 - \delta$$

Suppose we can show that this maximal correlation isn't much bigger than its mean.

$$\max_{m \in \mathcal{M}_s^\circ} \langle g,\ m - \mu \rangle_{L_2(\mathrm{P_n})} \leq \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle g,\ m - \mu \rangle_{L_2(\mathrm{P_n})} + \epsilon_\delta \qquad \text{w.p.} \quad 1 - \delta$$

Then it'd be enough that s^2 is a bit bigger than the mean maximal correlation.

$$s^2 > 2\sigma\, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle g,\ m - \mu \rangle_{L_2(\mathrm{P_n})} + \epsilon_\delta$$

Chebyshev's Inequality

To do this, we'll use Chebyshev's Inequality.

$$P(|Z - \mathrm{E}\, Z| \geq t) \leq \frac{\mathrm{Var}(Z)}{t^2} \quad \text{so} \quad Z \leq \mathrm{E}\, Z + \delta \underbrace{\sqrt{\mathrm{Var}\, Z}}_{\epsilon_\delta} \quad \text{w.p.} \quad 1 - \delta$$

That means all we've got to do is bound the variance of the maximal correlation.

- So we want to know about the variance of a maximum of sample means.
- We know the variance of a single one will be proportional to s^2/n. Why?
- It turns out that this maximum won't be much larger.
- That's a consequence of the Efron-Stein inequality.
- It's very useful and not hard to prove. We'll do it in homework.

Summary

Here's what we want.

$$s^2 \geq 2\sigma \, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle g, \; m - \mu \rangle_{L_2(\mathrm{P_n})} + \epsilon_\delta$$

Let's introduce some notation to say this a bit more compactly.

What we want, rephrased.

$$s^2 \geq 2\sigma \, \mathrm{w}(\mathcal{M}_s^\circ - \mu) + \epsilon_\delta$$

where

$$\mathcal{M}_s^\circ - \mu := \{m - \mu : m \in \mathcal{M}_s^\circ\}$$

is the *centered neighborhood boundary*,

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

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 $X_1 \ldots X_n$.

Implications

Example 1

Suppose we have a gaussian width bound that is proportional to s.

$$\mathrm{w}(\mathcal{M}_s^\circ - \mu) \leq \alpha s.$$

Then

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

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

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

- This is a $1/\sqrt{n}$ rate, like what we see in parametric stats classes (e.g. QTM 220).
- But it's non-asymptotic, so it tell us what happens
  when we use larger models at larger sample sizes.
- Our estimator converges when our model's size is much smaller than sample size.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \leq \left(1 + \frac{\epsilon_\delta}{2s}\right) \cdot 2\sigma c\sqrt{K/n} \quad \text{w.p.} \quad 1 - \delta$$
$$\to 0 \quad \text{if} \quad K/n \to 0$$

Example 2

Suppose we have a gaussian width bound that doesn't depend on s.

$$\mathrm{w}(\mathcal{M}_s^\circ - \mu) \leq \beta.$$

Then

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

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

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

- 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.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \leq \left(1 + \frac{\epsilon_\delta}{2s}\right) \cdot \sqrt{2\sigma}\sqrt[4]{c\log(K)/n} \quad \text{w.p.} \quad 1-\delta$$
$$\to 0 \quad \text{if} \quad \log(K)/n \to 0$$

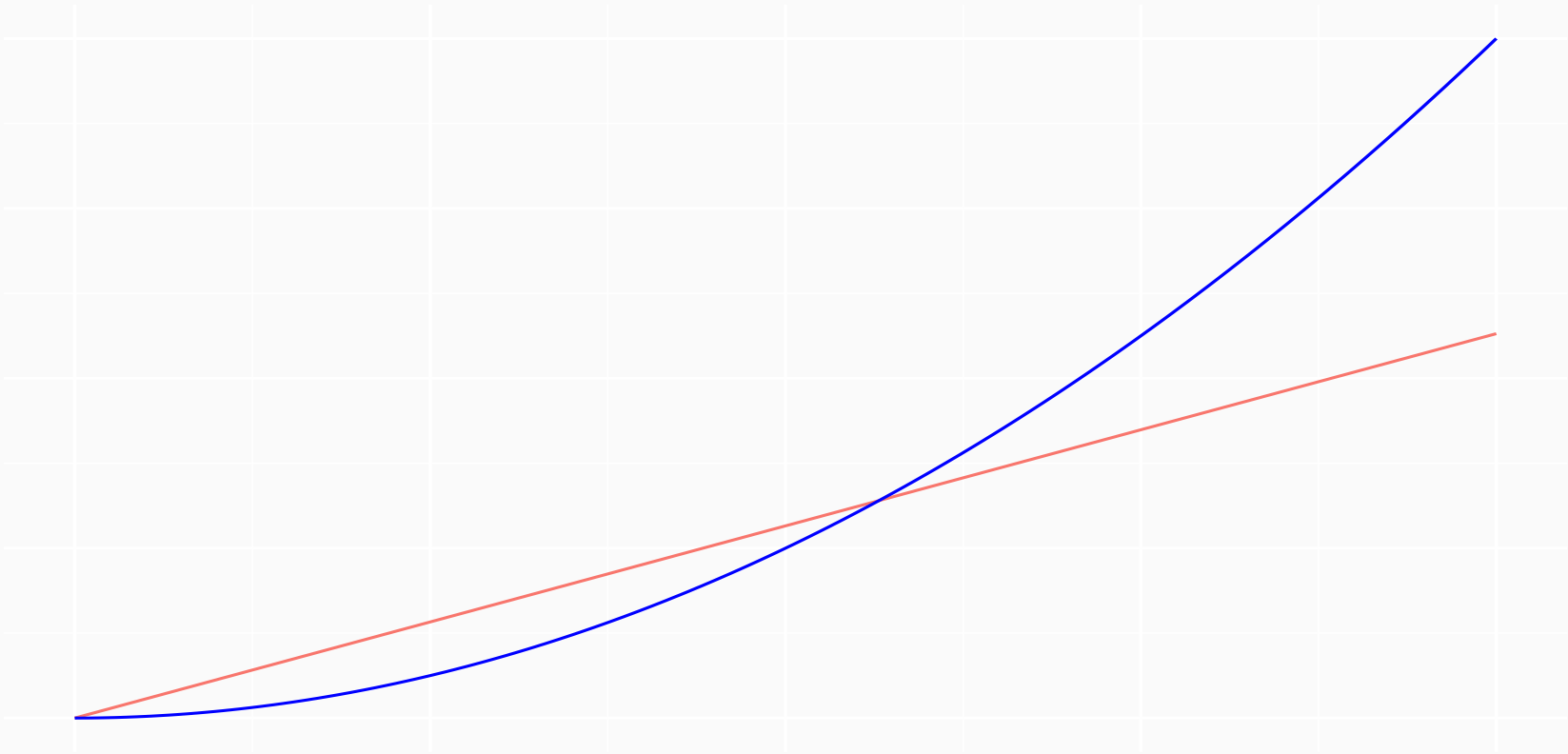
General Recipe

To find a good bound on the distance of our estimator from μ:

1. Bound Gaussian width and substitute the bound into our inequality.
2. Solve for the smallest radius s that satisfies the result.
3. To get a bound that holds with probability $1 - \delta$, we scale that up a bit.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \leq \left(1 + \frac{\epsilon_\delta}{2s}\right) \cdot s \quad \text{if} \quad s^2 \geq 2\sigma\, \mathrm{w}(\mathcal{M}_s^\circ - \mu)$$

A picture



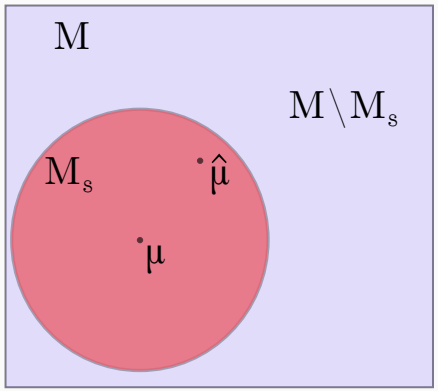
What is the smallest s for which our inequality is satisfied?

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

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*.



$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \leq s_\delta$$

with probability $1 - \delta$ if

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

- 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.
The error will, in some cases, be roughly as large as this bound.

This means that the study of least squares estimation is,
for the most part, just the study of Gaussian width.

Gaussian Width

Some Properties

$$\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g,\ v \rangle_{L_2(\mathrm{P_n})} \quad \text{for} \quad g_i \overset{iid}{\sim} N(0,1).$$

What do we know about it?

- It's a maximum over the set $\mathcal{V}$, so it gets bigger if $\mathcal{V}$ does.

$$\mathrm{w}(\mathcal{V}) \leq \mathrm{w}(V^+) \quad \text{if} \quad \mathcal{V} \subseteq \mathcal{V}^+$$

- It's *homogeneous*: if we scale the vectors in $\mathcal{V}$, we scale its width.

$$\mathrm{w}(\alpha\mathcal{V}) = \alpha\, \mathrm{w}(\mathcal{V}) \quad \text{for} \quad \alpha\mathcal{V} := \{\alpha v : v \in \mathcal{V}\}.$$

- It doesn't care about how we center our vectors.

$$\mathrm{w}(\mathcal{V} + x) = \mathrm{w}(V) \quad \text{for} \quad \mathcal{V} + x := \{v + x : v \in \mathcal{V}\}.$$

That's enough for now. We'll see more in this week's homework.

Gaussian Width

Example 1
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.

The neighborhood boundary is just a sphere in centered on $\vec{\mu} = \mu(X_1) \ldots \mu(X_n)$.

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

And the centered neighborhood boundary is just a sphere around the origin.

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

- In smaller sample sizes, this is true for substantially smaller models.
- i.e., if our model contains curves that take on every set of values $\vec{m}$ on the sample.
- e.g., a linear model with as many covariates as observations.

The General Model's Gaussian Width

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

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.

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

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.

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

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.

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

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

Next Steps

Coming Events

Homework

- We'll talk about the gaussian width of neighborhoods in more models.
- We'll do some by hand and some computationally.
- We'll talk about what happens with non-gaussian noise.

Future Lectures

- We'll talk about what happens when our model is misspecified, i.e., when $\mu \notin \mathcal{M}$.
- We'll do a version of the same argument for the *nonparametric R-learner*.

Lab

- We'll do a drawing exercise to get a feel for gaussian width.
- In essence, our drawings will be width computations for some models in the 2 observation case.

Lab and HW Preview

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

$$\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} g_i v_i \overset{S \to \infty}{\approx} \frac{1}{S} \sum_{j=1}^{S} \max_{v \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} g_{ij} v_i \quad \text{for} \quad g_{ij} \overset{iid}{\sim} N(0,1).$$

- Computationally, this can be hard.
- But it means we can bound Gaussian width,
  even if we can't work out how to do it analytically.

Shorthand

We phrased the inequality determining the radius s in terms of the width of the *centered neighborhood boundary*.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \leq \left(1 + \frac{\epsilon_\delta}{2s}\right) \cdot s \quad \text{w.p.} \quad 1 - \delta \quad \text{if} \quad s^2 \geq 2\sigma\, \mathrm{w}(\mathcal{M}_s^\circ - \mu).$$

Centering doesn't matter, so we don't have to. From now on, we'll say this.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} \leq \left(1 + \frac{\epsilon_\delta}{2s}\right) \cdot s \quad \text{w.p.} \quad 1 - \delta \quad \text{if} \quad s^2 \geq 2\sigma\, \mathrm{w}(\mathcal{M}_s^\circ).$$