Machine Learning Theory

Lecture 7: Least Squares and Population Mean Squared Error

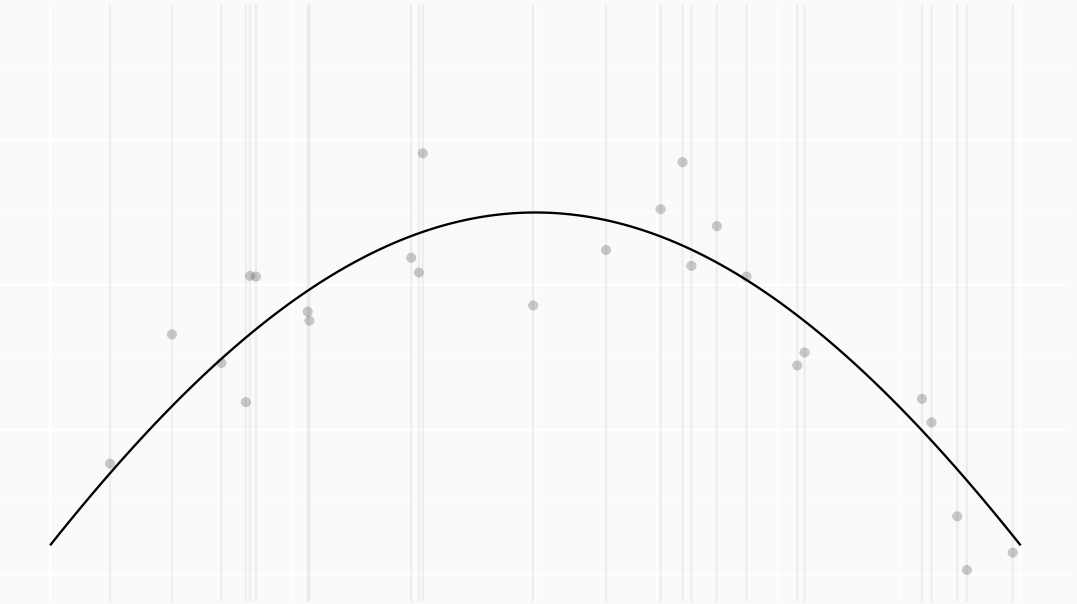
David A. Hirshberg

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 $\mu(x)$.

We've been bounding sample mean squared error.

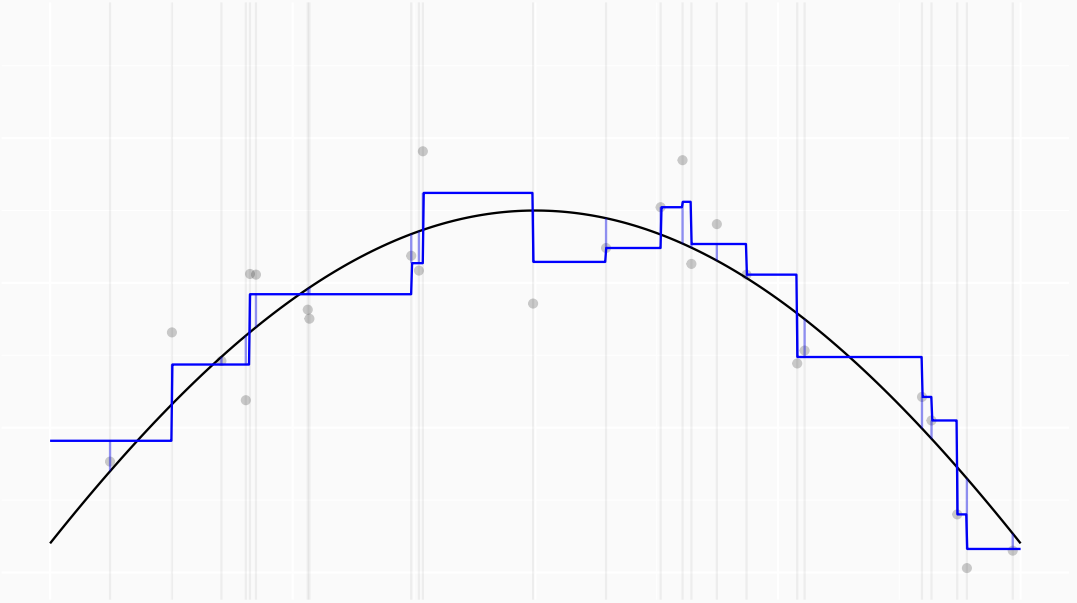
$$\|\hat{\mu} - \mu\|^2_{L_2(\mathrm{P_n})} < s^2 \quad \text{with high probability.}$$

It's the average of the squared errors we make at the points $X_1 \ldots X_n$.

What does this tell us?

This statement tells us $\hat{\mu}$ and μ are close, on average, at the observed X_i.

$$\frac{1}{n}\sum_{i=1}^{n}\{\mu(X_i) - \hat{\mu}(X_i)\}^2 < s^2.$$



It doesn't tell us they're close in the gaps between the X_i.

Today, we'll address that. We're going to bound population mean squared error.

$$\|\hat{\mu} - \mu\|^2_{L_2(\mathrm{P})} < s^2 \quad \text{with high probability.}$$

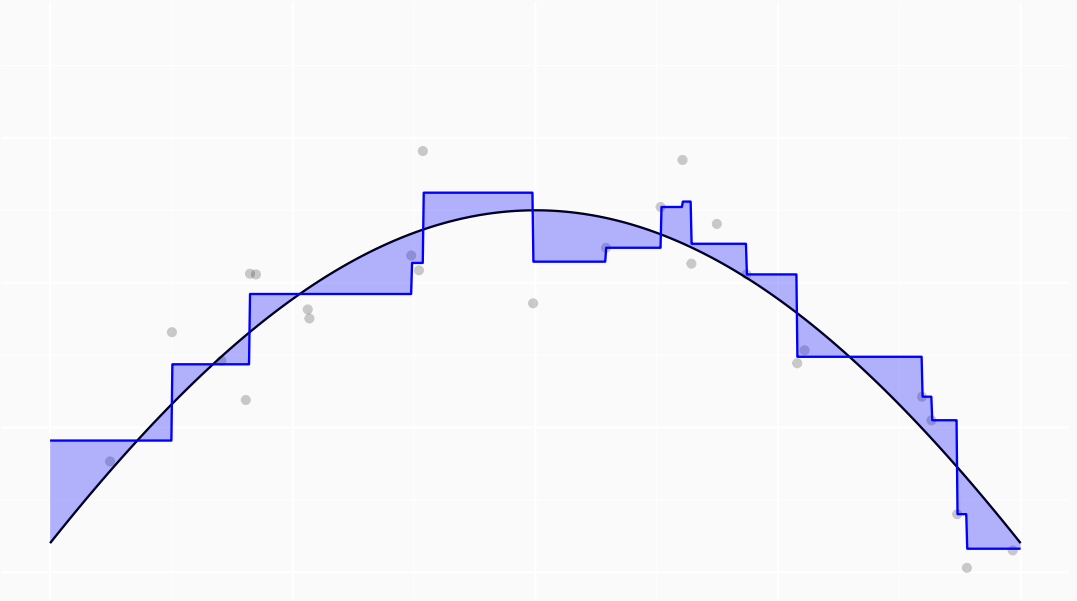
What Population Mean Squared Error Is

It's the mean squared error we make a random point X_{i+1} distributed like $X_1 \ldots X_n$.

$$\|\hat{\mu} - \mu\|^2_{L_2(\mathrm{P})} = \mathrm{E}_{X_{i+1}}\left[\{\hat{\mu}(X_{i+1}) - \mu(X_{i+1}\}^2\right]$$

That's the integral of the squared distance between the two curves, multiplied by the density of X_i.

$$\|\hat{\mu} - \mu\|^2_{L_2(\mathrm{P})} = \int \{\hat{\mu}(x) - \mu(x)\}^2 p(x)\, dx \quad \text{if} \quad X_i \quad \text{has the density} \quad p(x).$$

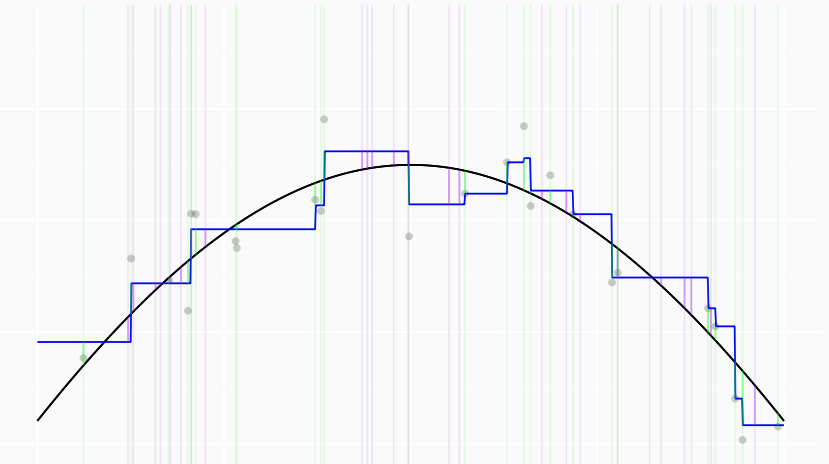


Why we care about Population Mean Squared Error: Generalization

If we're interested in average accuracy for a bunch of new points like $X_1 \dots X_n$, that's more or less exactly what it is.

$$\|\hat{\mu} - \mu\|^2_{L_2(\mathrm{P})} = \mathrm{E}_{X_{i+1}}\left[\{\hat{\mu}(X_{i+1}) - \mu(X_{i+1}\}^2\right] \overset{LLN}{\approx} \frac{1}{n'}\sum_{i=1}^{n'}\{\hat{\mu}(X_{n+i}) - \mu(X_{n+i})\}^2.$$

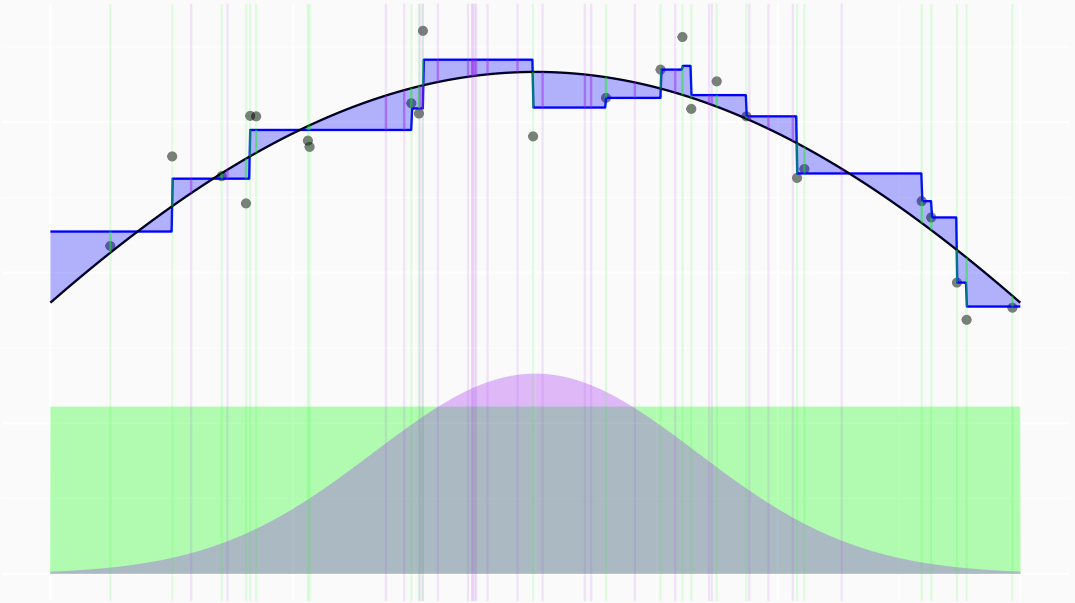
This can be a bit different from accuracy on our original sample $X_1 \dots X_n$.



- TV regression spends its 'variation budget' jumping to fit on the original sample.
- Between observations, it doesn't know whether it should jump or not.
- So we get larger error at our new points.
- It's usually not much larger, but sometimes it is. We'll see why.

Why we care about Population Mean Squared Error: Generalization

If we're interested in average accuracy for new points from a different distribution Q, we can bound this by comparing this distribution's density to that of our observations.

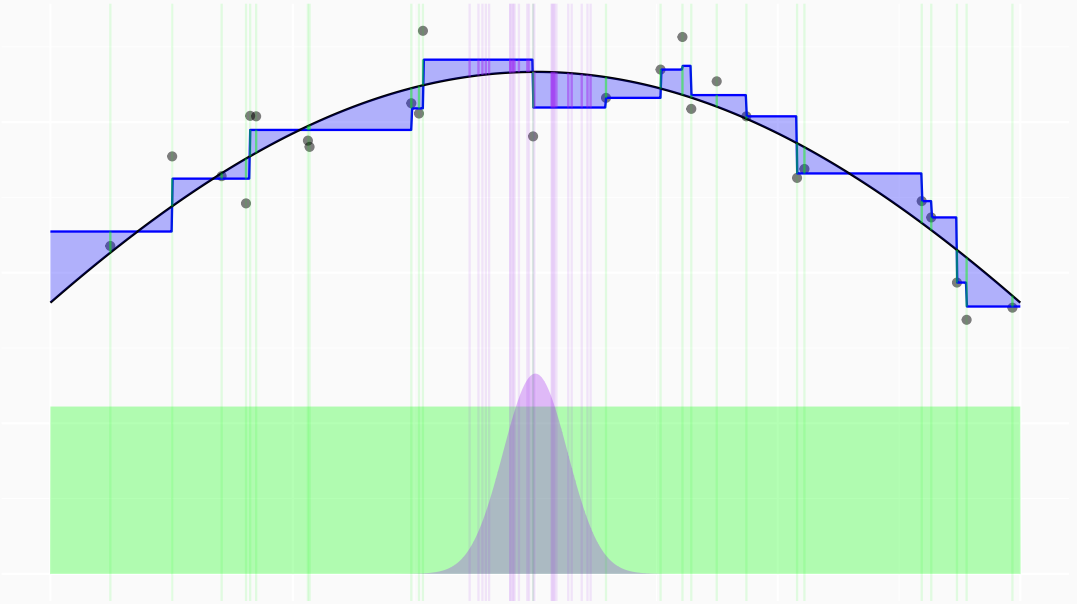


$$\frac{1}{n'}\sum_{i=1}^{n'}\{\hat{\mu}(X_{n+i}) - \mu(X_{n+i})\}^2 \approx \|\hat{\mu} - \mu\|^2_{L_2(\mathrm{Q})} = \int \{\hat{\mu}(x) - \mu(x)\}^2 \frac{q(x)}{p(x)} p(x)\,dx$$

$$\leq \max_x \frac{q(x)}{p(x)} \|\hat{\mu} - \mu\|^2_{L_2(\mathrm{P})}.$$

Why we care about Population Mean Squared Error: Generalization

If we're interested in accuracy at a specific point x', we can think of this new distribution Q as a little bump around x'.



$$\{\hat{\mu}(x') - \mu(x')\}^2 \approx \|\hat{\mu} - \mu\|_{L_2(Q_\epsilon)} \quad \text{for} \quad Q = N(x', \epsilon^2).$$

This is what we want, more or less, to estimate a treatment effect using a discontinuity. We want the value of the curve just left and right of the discontinuity.

Bounding Population Mean Squared Error

It's like bounding Sample Mean Squared Error

To bound Sample MSE

- We showed that with high probability the least squares estimator $\hat{\mu}$ must be in a neighborhood of the curve μ.
- Or, more generally, the model's best approximation to it, $\hat{\mu}^\star$.
- Both of these things were defined in terms of distance on the sample.

$$\mathcal{M}_s = \left\{ m \in \mathcal{M} : \|m - \mu^\star\|_{L_2(\mathrm{P_n})} \leq s \right\}$$
$$\mu^\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\|_{L_2(\mathrm{P_n})}.$$

To bound Population MSE

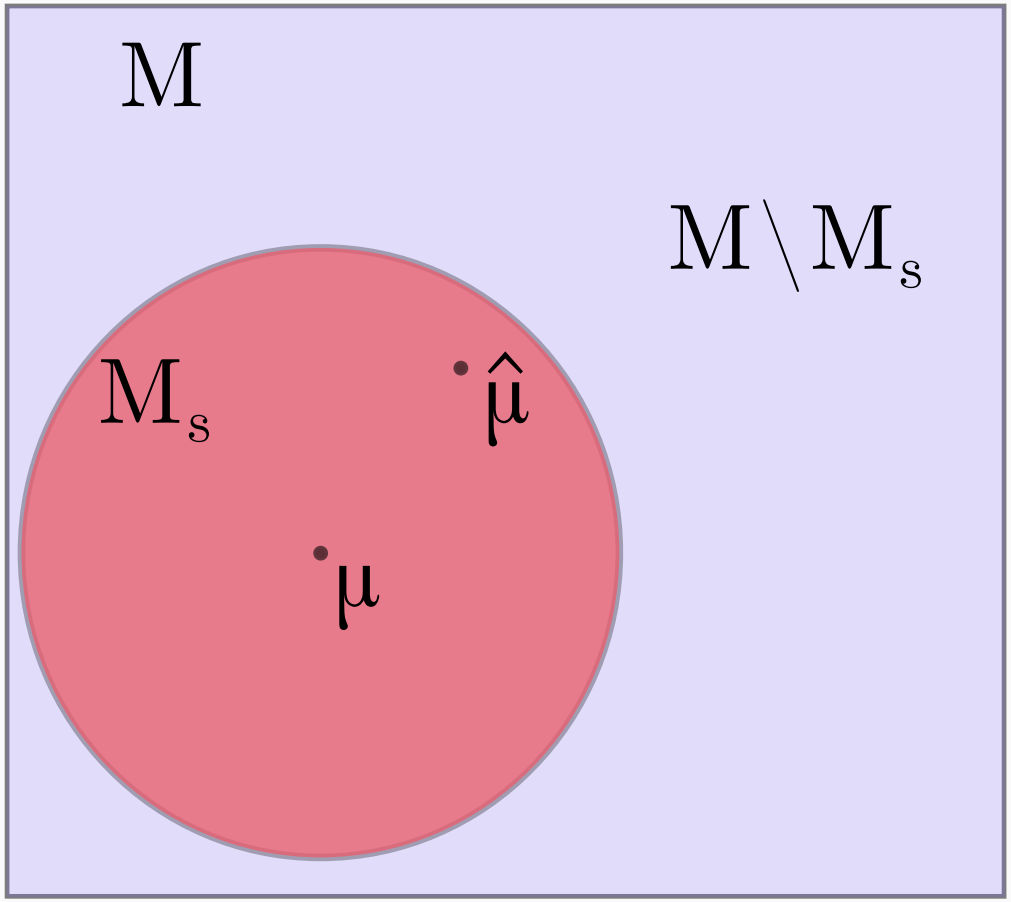
- We do exactly the same thing with different definitions.
- We define our neighborhood and best approximation using population distance.

$$\mathcal{M}_s = \left\{ m \in \mathcal{M} : \|m - \mu^\star\|_{L_2(\mathrm{P})} \leq s \right\}$$
$$\mu^\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\|_{L_2(\mathrm{P})}.$$

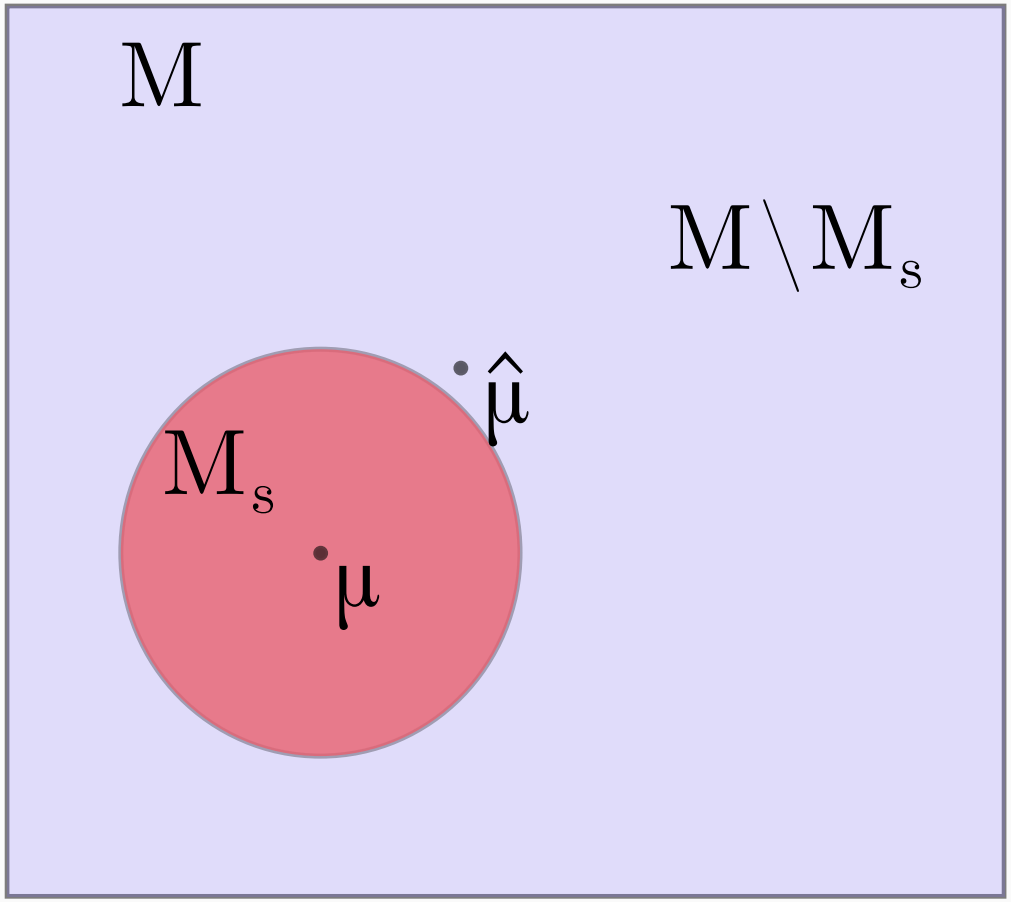
Warning: We're using the same notation for different but analogous things.
Context should tell us what's what.

We'll focus on the case that the curve μ is in the model $\mathcal{M}$.
We can generalize to the case that it isn't just like we did for sample MSE.

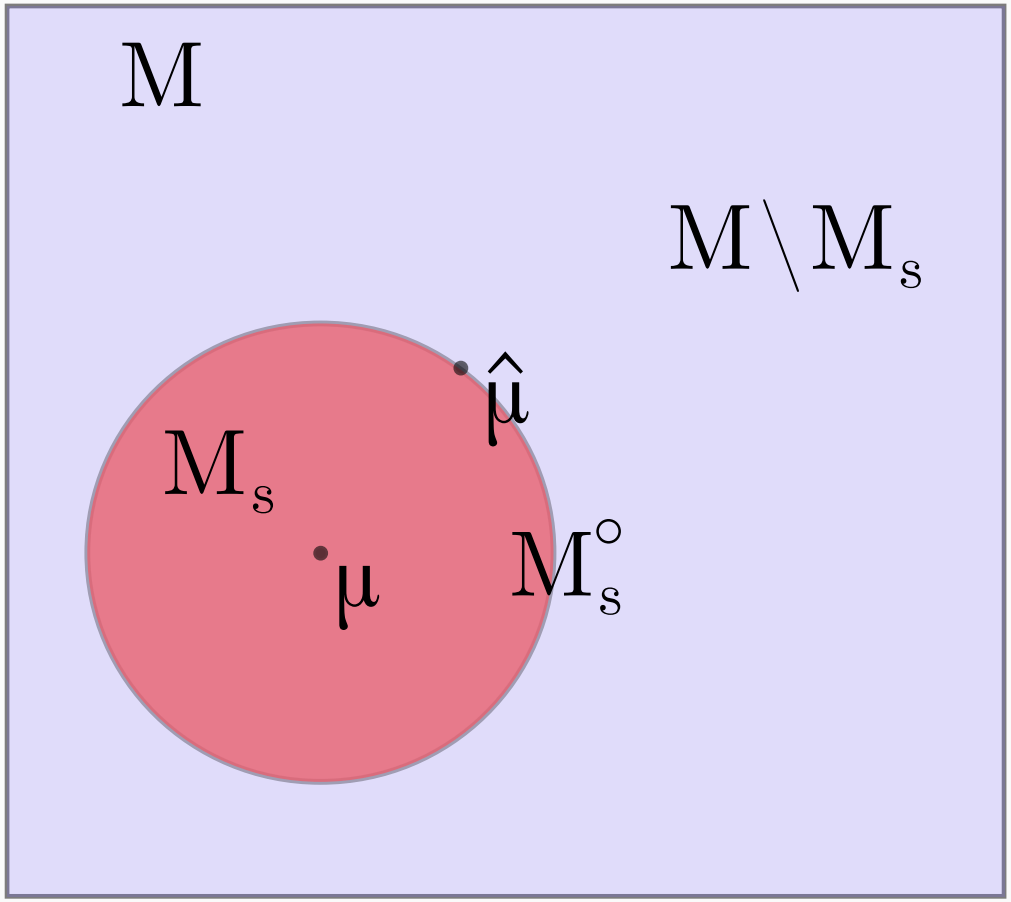
We'll show our estimator is in a neighborhood of μ.



To do it, we'll rule out the possibility that it's outside it.



And it’s enough to rule out the possibility it’s on the boundary.



Why? A Review

There are two forces working against each other when we're doing least squares.

$$\ell(m) - \ell(\mu) = \|m - \mu\|^2_{L_2(\mathrm{P_n})} \quad - \quad \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \left\{m(X_i) - \mu(X_i)\right\} \quad \text{for} \quad \varepsilon_i = Y_i - \mu(X_i).$$

- The first term, squared distance, tells us we don't want something far from μ.
- The second term, which has mean zero and scale $\propto$ distance, obscures that.
- Once we're far enough away, the first term wins out. It grows faster with distance.

The distance involved is Sample MSE, but we can switch to Population MSE.
The difference between the two is just another mean-zero term.

$$\begin{aligned}\|m - \mu\|^2_{L_2(\mathrm{P})} - \|m - \mu\|^2_{L_2(\mathrm{P_n})} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{m(X_i) - \mu(X_i)\}^2 - \{m(X_i) - \mu(X_i)\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mathrm{E} - 1)\{m(X_i) - \mu(X_i)\}^2 \quad \text{in compact notation.}\end{aligned}$$

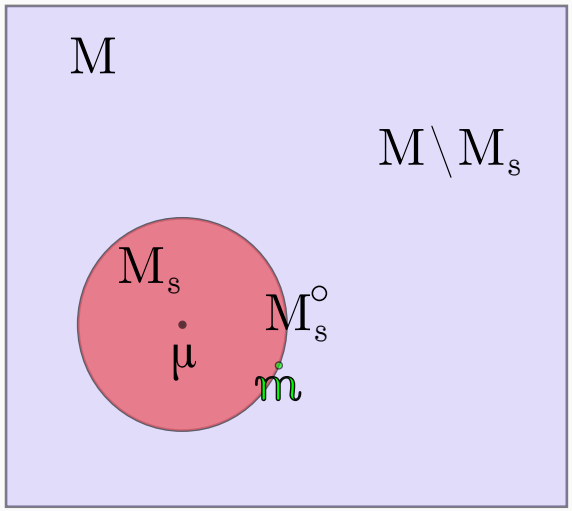
$$\ell(m) - \ell(\mu) = \|m - \mu\|^2_{L_2(\mathrm{P})} - \frac{1}{n}\sum_{i=1}^{n} \Bigg[\underbrace{2\varepsilon_i \left\{m(X_i) - \mu(X_i)\right\}}_{\text{old mean zero term}} + \underbrace{(\mathrm{E} - 1)\{m(X_i) - \mu(X_i)\}^2}_{\text{new mean zero term}}. \Bigg]$$

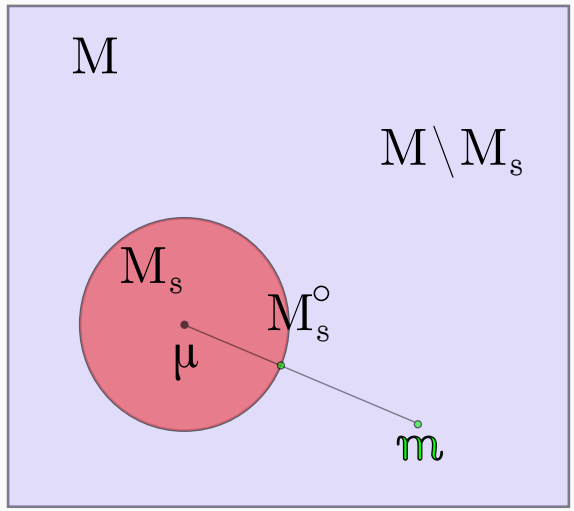
The Argument

$$\ell(m)-\ell(\mu) = \|m-\mu\|^2_{L_2(\mathrm{P})} - \frac{1}{n}\sum_{i=1}^{n}\Big[\underbrace{2\varepsilon_i\,\{m(X_i)-\mu(X_i)\}}_{\text{old mean zero term}} \;-\; \underbrace{(\mathrm{E}-1)\{m(X_i)-\mu(X_i)\}^2}_{\text{new mean zero term}}.\Big]$$

For a certain *distance* s, squared distance exceeds this mean zero term for all curves in the model at that distance, i.e., curves $m \in \mathcal{M}_s^\circ$.

- That rules out the possibility that our minimizer is *any* curve at that distance.
- And any curve at a greater distance. It's representative is ruled out, and so is it.





So all we've got to do is characterize a radius s for which this happens.

The Characterization

$$\ell(m) - \ell(\mu) = s^2 - \frac{1}{n}\sum_{i=1}^{n}\big[2\varepsilon_i\{m(X_i) - \mu(X_i)\} + (\mathrm{E}-1)\{m(X_i) - \mu(X_i)\}^2\big]$$

exceeds zero for all curves m in the neighborhood boundary $\mathcal{M}_s^\circ$ if ...

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

and

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

Here we're using half of s^2 to bound each mean zero term.

We know what we need for the old one.
It's enough that s^2 exceeds a multiple of its mean.

$$\frac{s^2}{2} \geq 2c_\delta\, \mathrm{w}_\varepsilon(\mathcal{M}_s^\circ) = c_\delta\, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\{m(X_i) - \mu(X_i)\}.$$

To find a radius s satisfying this, we bound the width of our neighborhood boundary.

Dealing with the New Part

To deal with the new part, we do the same. We need to show this.

$$\frac{s^2}{2} \gtrsim c_\delta \, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \frac{1}{n} \sum_{i=1}^{n} (\mathrm{E} - 1)\{m(X_i) - \mu(X_i)\}^2$$

Can we find a radius s satisfying this essentially the same way?

Dealing with the New Part

To deal with the new part, we do the same. We need to show this.

$$\frac{s^2}{2} \gtrsim c_\delta \, \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \frac{1}{n} \sum_{i=1}^{n} (\mathrm{E} - 1)\{m(X_i) - \mu(X_i)\}^2$$

Can we find a radius s satisfying this essentially the same way?

Yes. And we can understand why using *symmetrization*.

- Let's think about independent copy $X_1' \ldots X_n'$ of our observed points.
- They have the same distribution, so it's equivalent to center on the copy's mean.
- But they're independent, so we have a little more latitude in terms of arithmetic.

Symmetrization

Step 1

The mean (E_X) over $X_1 \dots X_n$ is equivalent to a mean ($\mathrm{E}_{X'}$) over our copy.

$$
\begin{aligned}
&\mathrm{E}_X \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} \big[(\mathrm{E}_X - 1) \{ m(X_i) - \mu(X_i) \}^2 \big] \\
&= \mathrm{E}_X \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} \big[\mathrm{E}_X \{ m(X_i) - \mu(X_i) \}^2 - \{ m(X_i) - \mu(X_i) \}^2 \big] \\
&= \mathrm{E}_X \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} \big[\mathrm{E}_{X'} \{ m(X_i') - \mu(X_i') \}^2 - \{ m(X_i) - \mu(X_i) \}^2 \big] \\
&= \mathrm{E}_X \max_{m \in \mathcal{M}_s^\circ} \mathrm{E}_{X'} \sum_{i=1}^{n} \big[\{ m(X_i') - \mu(X_i') \}^2 - \{ m(X_i) - \mu(X_i) \}^2 \big].
\end{aligned}
$$

Step 2

The mean of a maximum is larger than the maximum of a mean.

$$\mathrm{E}_X \max_{m \in \mathcal{M}_s^\circ} \mathrm{E}_{X'} \sum_{i=1}^{n} \left[\{m(X_i') - \mu(X_i')\}^2 - \{m(X_i) - \mu(X_i)\}^2 \right]$$

$$\leq \mathrm{E}_X \, \mathrm{E}_{X'} \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} \left[\{m(X_i') - \mu(X_i')\}^2 - \{m(X_i) - \mu(X_i)\}^2 \right].$$

Symmetrization

Step 3

Terms as symmetric, so multiplying by *signs* $s_i = \pm 1$ doesn't change anything.

$$\mathrm{E}_X \mathrm{E}_{X'} \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^n \left[\{m(X_i') - \mu(X_i')\}^2 - \{m(X_i) - \mu(X_i)\}^2 \right]$$

$$= \mathrm{E}_X \mathrm{E}_{X'} \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^n s_i \left[\{m(X_i') - \mu(X_i')\}^2 - \{m(X_i) - \mu(X_i)\}^2 \right].$$

where our signs $s_i = \pm 1$ each with probability $1/2$.

Symmetrization

Step 4

Each copy, multiplied by this random sign, is centered. Maximize each separately.

$$\mathrm{E}_X \, \mathrm{E}_{X'} \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^n s_i \left[\{m(X_i') - \mu(X_i')\}^2 - \{m(X_i) - \mu(X_i)\}^2 \right]$$

$$\leq \mathrm{E}_{X'} \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^n s_i \{m(X_i') - \mu(X_i')\}^2 + \mathrm{E}_X \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^n s_i \{m(X_i) - \mu(X_i)\}^2.$$

This leaves us with two copies of the same thing.

Symmetrization

Summary

We pay a factor of two to subsitute a random sign s_i for centering $(\mathrm{E} - 1)$.

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

$$\leq 2\,\mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \frac{1}{n} \sum_{i=1}^{n} s_i \, \{m(X_i) - \mu(X_i)\}^2.$$

What this Leaves Us With. In the Gaussian-noise Case.

What we're left with is like the gaussian width we used to bound the old term.

$$\mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \frac{2}{n} \sum_{i=1}^{n} s_i \left\{ m(X_i) - \mu(X_i) \right\}^2 \qquad \text{new term bound}$$

$$\mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \frac{2\sigma}{n} \sum_{i=1}^{n} g_i \left\{ m(X_i) - \mu(X_i) \right\} \qquad \text{old term bound.}$$

But there are a few differences.

1. We've got a random sign s_i instead of standard normals g_i.
2. We've got squared differences $(m - \mu)^2$ instead of $m - \mu$.
3. It doesn't scale with the noise level σ.

For the most part, the third difference is the only one that matters.

Why?

- A mean like this with random signs is never much bigger than one with gaussians.
- The squaring only matters if there's some X_i where $m - \mu$ is big.

Both of these are implied by a *contraction inequality* for random-sign width.
A new one. We'll investigate this, using simulation as well as theory, in this week's lab.

Preview: Consequences in the Gaussian-noise Case.

In effect, when we're interested in population MSE

- random samping from the population acts like gaussian noise
- that noise's standard deviation σ_{samp} is no larger than $2\times$ the extreme values of the functions in our centered neighborhood.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P})} \leq s \quad \text{with probability } 1-\delta \text{ if } \quad s^2 \geq 2c_\delta \left(\sigma_{samp} + \sigma_{noise}\right) \mathrm{w}(\mathcal{M}_s^\circ)$$

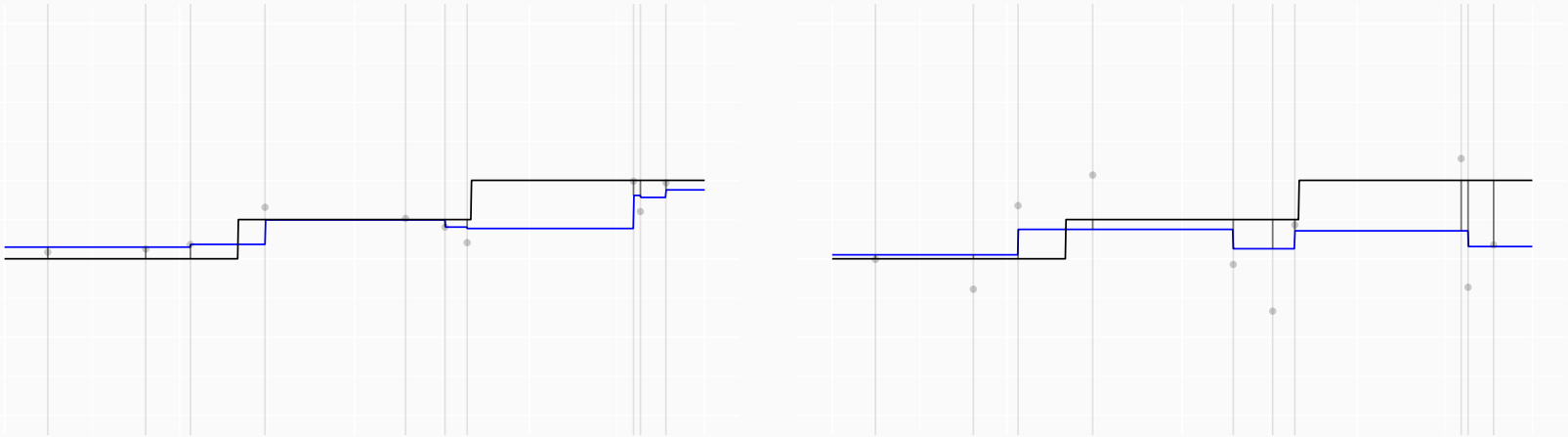
$$\text{where} \quad \sigma_{samp} := 2 \max_{m \in \mathcal{M}_s^\circ} \|m - \mu\|_\infty$$

$$\text{and} \quad \sigma_{noise} := \max_i \sqrt{\mathrm{E}\{ Y_i - \mu(X_i)\}^2}$$

When we're interested in sample MSE, we don't have that. Therefore

- sample MSE $\approx$ population MSE unless there's relatively little noise.
- sample MSE $\lesssim$ population MSE generally.

Low Noise vs High Noise



Left. With little noise, our estimator $\hat{\mu}$ fits substantially better at the sample points X_i.

Right. With more, it doesn't. The observations are far enough from μ that we can't estimate it all that precisely even where we have some data.