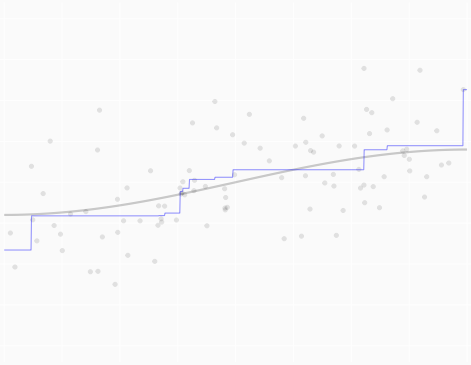
Non-Gaussian Noise

Review: Probabilistic Classification



Last time, we talked about *probabilistic classification*, i.e. regression with *classification noise*.

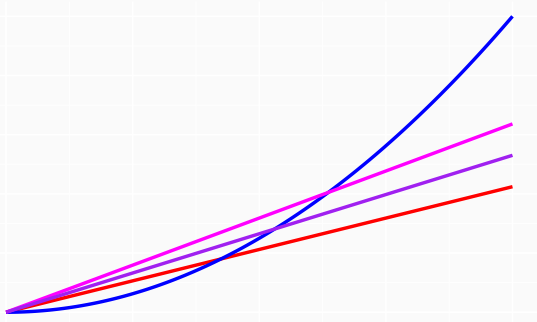
$$\hat{\mu} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2 \quad \text{where} \quad Y_i = \mu(X_i) + \varepsilon_i \text{ for } \varepsilon_i = \begin{cases} 1-\mu(X_i) & \text{w.p. } \mu(X_i) \\ -\mu(X_i) & \text{w.p. } 1-\mu(X_i) \end{cases}$$

By comparing widths, we showed that this is *easier* than regression with ...

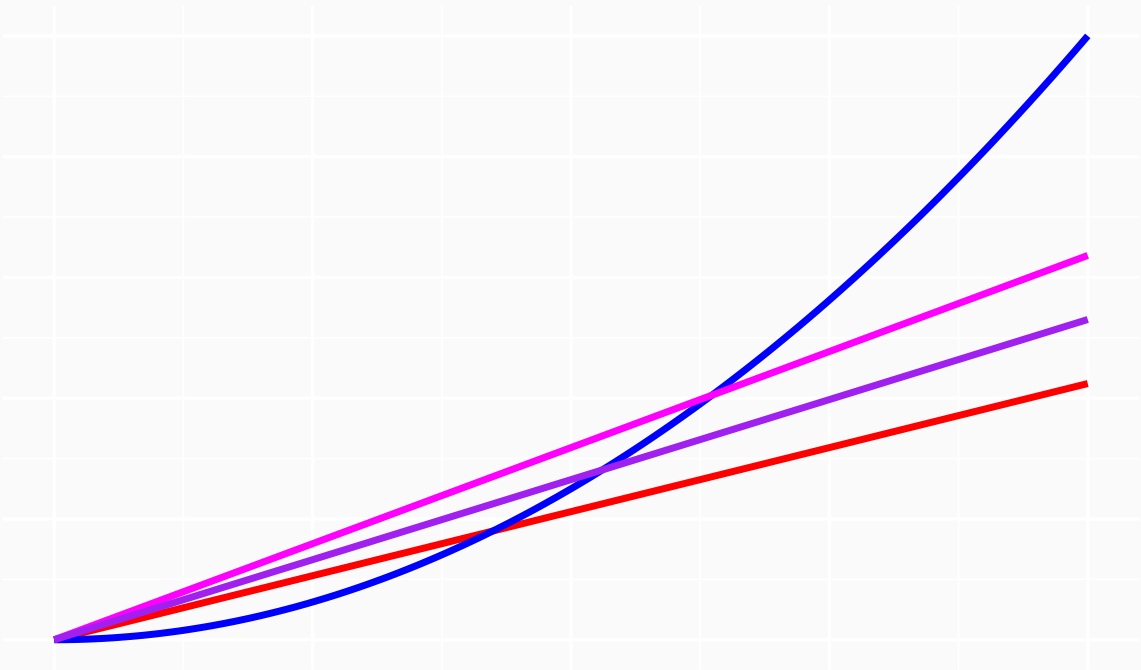
1. random sign noise, $s_i = \pm 1$ each w.p. $1/2$.
2. gaussian noise σg_i of standard deviation $\sigma = 1.25$.

Easier in the sense that our crossing-point argument gives us a better error bound.

$$\frac{s^2}{2} \geq 1.25\,\mathrm{w}(\mathcal{M}_s) \implies \mathrm{w}_s(\mathcal{M}_s) \implies \mathrm{w}_\varepsilon(\mathcal{M}_s)$$



Starting Point



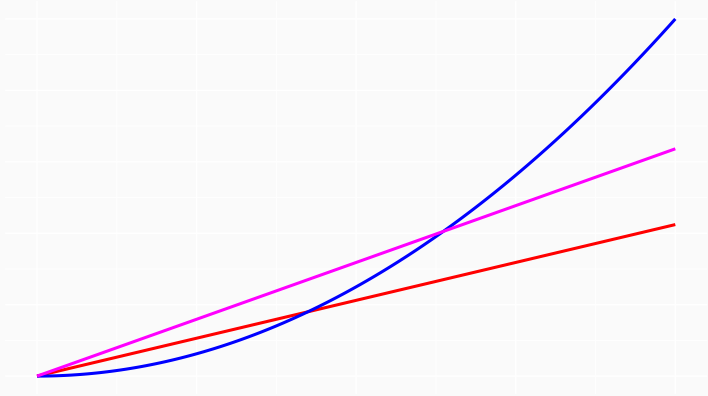
Today, we're going to generalize that result to regression with *any kind of noise*.
We'll start with the same abstract bound. It applies no matter how noise is distributed.

$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P_n})} < s + 2\sqrt{\frac{2\Sigma_n}{\delta n}} \quad \text{w.p. } 1-\delta \text{ for } \quad \frac{s^2}{2} \geq \mathrm{w}_\epsilon(\mathcal{M}_s)$$

$$\text{where} \quad \mathrm{w}_\epsilon(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle \epsilon, v \rangle_{L_2(\mathrm{P_n})} \quad \text{and} \quad \Sigma_n = \mathrm{E} \max_{i \in 1 \ldots n} \varepsilon_i^2 .$$

This bound depends on the model $\mathcal{M}$ and the distribution of the noise ε
in a complex, entangled way: through the width $\mathrm{w}_\varepsilon(\mathcal{M}_s)$.

Our Approach



To disentangle the impact of the model and noise distribution, we'll bound this width in terms of gaussian width.

$$\mathrm{w}_\epsilon(\mathcal{M}_s) \leq \alpha \mathrm{w}(\mathcal{M}_s)$$

for α depending on ε but not $\mathcal{M}$ or s.

At the heart of this comparison $\mathrm{w}_\epsilon(\cdot) \leq \alpha \mathrm{w}(\cdot)$ are two ideas.

1. **Symmetrization**. We'll substitute for ϵ_i a variant that's symmetric around zero.

   $$\epsilon_i \to \epsilon_i - \epsilon_i' \quad \text{where} \quad \epsilon_i' \text{ is an independent copy of } \epsilon_i$$

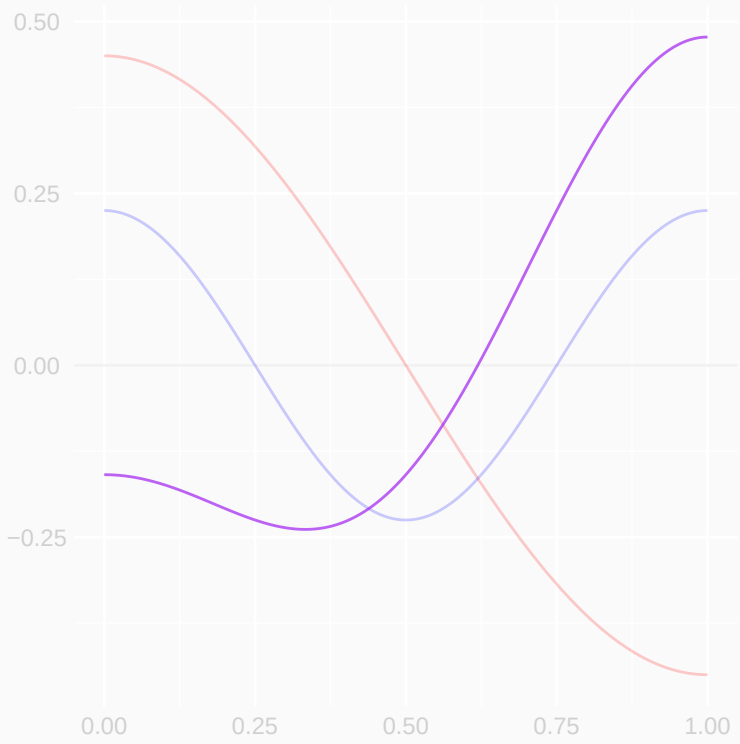
   This substitution *increases* width: $\mathrm{w}_\epsilon(\cdot) \leq \mathrm{w}_{\epsilon-\epsilon'}(\cdot)$.
2. **Contraction**. We'll substitute a gaussian vector[1] for our symmetrized noise $\epsilon - \epsilon'$. We can bound the impact of this substitution in a model-invariant way.

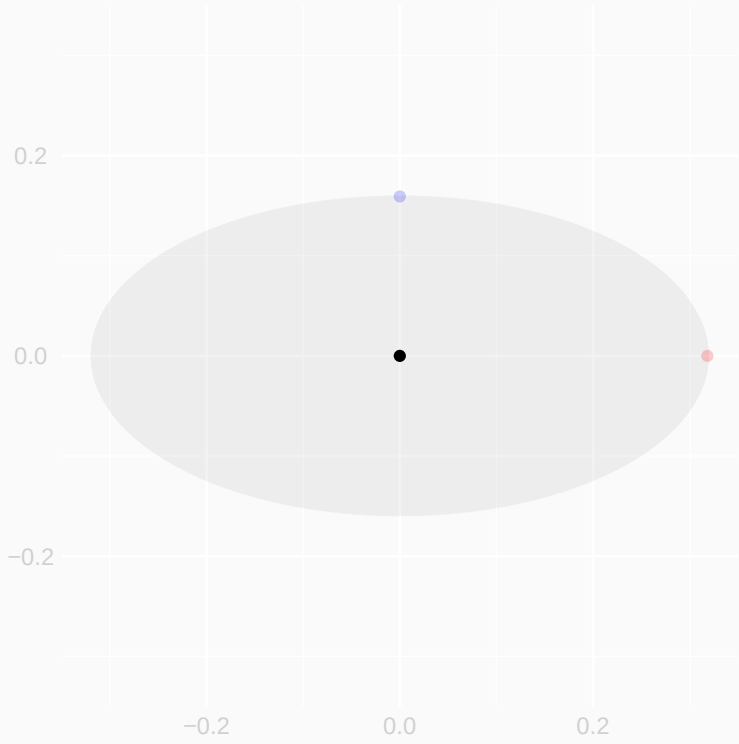
   $$\mathrm{w}_{\epsilon-\epsilon'}(\cdot) \leq 2M_n \, \mathrm{w}_s(\cdot) \leq \sqrt{2\pi} M_n \times \mathrm{w}(\cdot) \quad \text{for} \quad M_n = \mathrm{E} \max_{i \in 1 \ldots n} |\varepsilon_i|$$

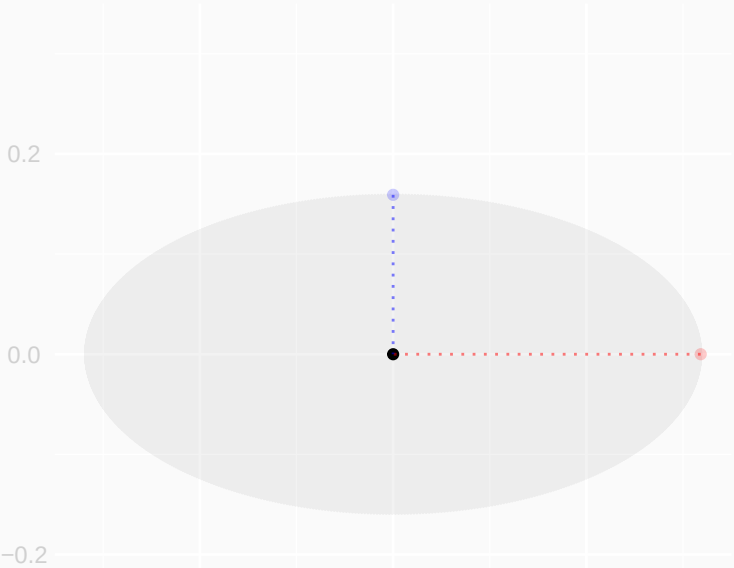
This lets us re-use our gaussian width calculations to analyze regression with any noise distribution.

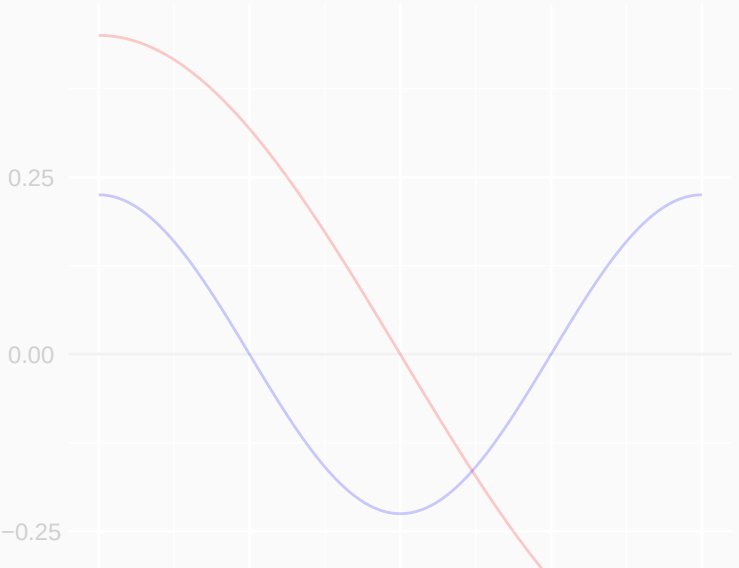
[1] or a random-sign vector

An Example


0.50
0.25
0.00
−0.25
0.00
0.25
0.50
0.75
1.00


0.2
0.0
−0.2
−0.2
0.0
0.2

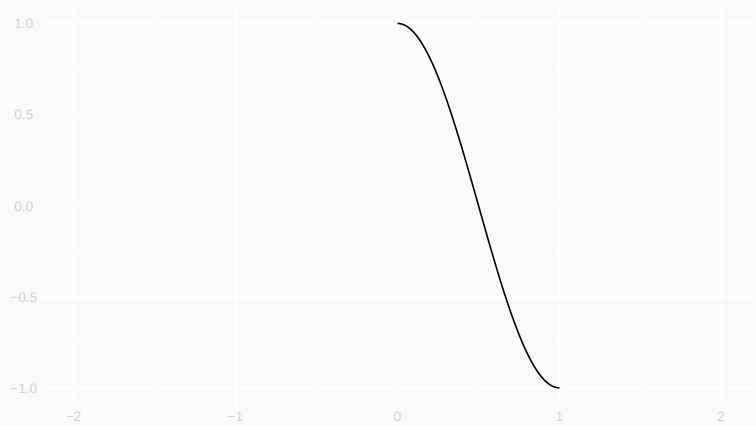

0.2
0.0
−0.2

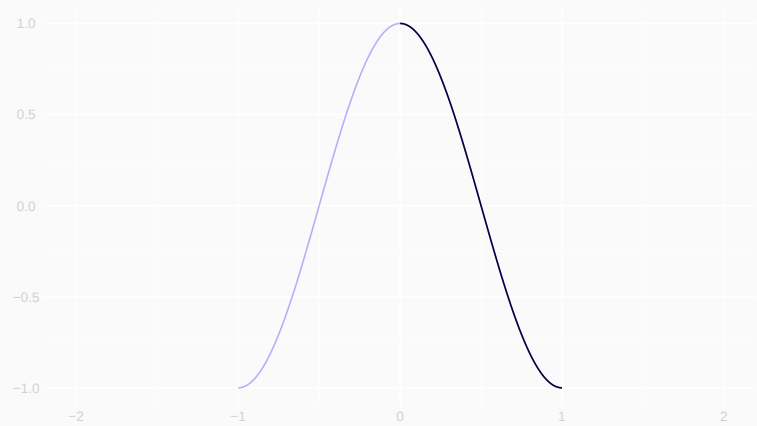

0.25
0.00
−0.25

Symmetrization

$$\mathrm{w}_{\varepsilon}(\mathcal{V}) \le \mathrm{w}_{s(\varepsilon-\varepsilon')}(\mathcal{V}) \le 2\,\mathrm{w}_{s\varepsilon}(\mathcal{V})$$

$$\begin{aligned}
\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} \varepsilon_i v_i &= \mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \mathrm{E}\,\varepsilon_i') v_i \\
&\overset{(a)}{\le} \mathrm{E}\,\mathrm{E}' \max_{v\in\mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon_i') v_i \\
&= \mathrm{E}_s\,\mathrm{E}\,\mathrm{E}' \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i(\varepsilon_i - \varepsilon_i') v_i \\
&\overset{(b)}{\le} \mathrm{E}_s\,\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i\varepsilon_i + \mathrm{E}_s\,\mathrm{E}' \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i\varepsilon_i' v_i = 2\,\mathrm{E}_s\,\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} \varepsilon_i s_i v_i.
\end{aligned}$$





Symmetrization

$$\mathrm{w}_{\varepsilon}(\mathcal{V}) \le \mathrm{w}_{s(\varepsilon-\varepsilon')}(\mathcal{V}) \le 2\,\mathrm{w}_{s\varepsilon}(\mathcal{V})$$

$$\begin{aligned}
\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} \varepsilon_i v_i &= \mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \mathrm{E}\,\varepsilon_i') v_i \\
&\overset{(a)}{\le} \mathrm{E}\,\mathrm{E}' \max_{v\in\mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon_i') v_i \\
&= \mathrm{E}_s\,\mathrm{E}\,\mathrm{E}' \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i(\varepsilon_i - \varepsilon_i') v_i \\
&\overset{(b)}{\le} \mathrm{E}_s\,\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i\varepsilon_i + \mathrm{E}_s\,\mathrm{E}' \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i\varepsilon_i' v_i = 2\,\mathrm{E}_s\,\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^{n} \varepsilon_i s_i v_i.
\end{aligned}$$

(a) Replacing ε_i with $s_i(\varepsilon_i - \varepsilon_i')$ is 'free'.

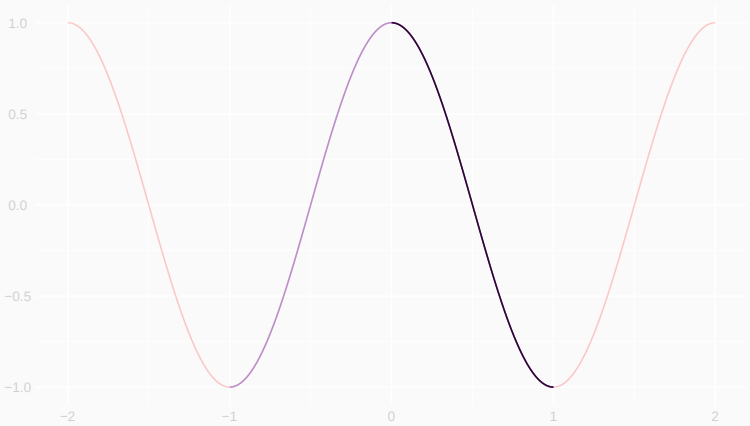
- We stopped here in our classification example because $\varepsilon_i - \varepsilon_i'$ was easy to bound.
- Generally, we take an extra step to express things in terms of ε_i again.

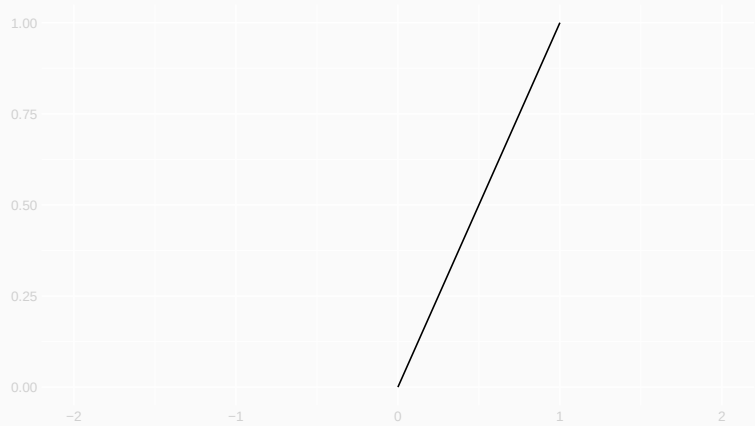
(b) Replacing ε_i with $s_i\varepsilon_i$ increases width by at most $2\times$.

Contraction

$$\mathrm{w}_\eta(\mathcal{V}) = \mathrm{w}_{s\eta}(\mathcal{V}) \leq \mathrm{E}\|\eta\|_\infty \, \mathrm{w}_\eta(\mathcal{V}) \quad \text{if} \quad \eta \overset{dist}{=} -\eta.$$

$$\begin{aligned}
\mathrm{E}_s \, \mathrm{E}_\eta \max_{v \in \mathcal{V}} \sum_{i=1}^n \eta_i s_i v_i &\leq \mathrm{E}_\eta \max_{\substack{u \in \mathbb{R}^n \\ |u_i| \leq \|\eta\|_\infty}} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n u_i s_i v_i \\
&= \mathrm{E}_\eta \|\eta\|_\infty \max_{u \in [-1,1]^n} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n u_i s_i v_i \\
&= \mathrm{E}_\eta \|\eta\|_\infty \times \max_{u \in \{-1,1\}^n} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n u_i s_i v_i \\
&= \mathrm{E}_\eta \|\eta\|_\infty \times \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i v_i
\end{aligned}$$





Contraction

$$\mathrm{w}_\eta(\mathcal{V}) = \mathrm{w}_{s\eta}(\mathcal{V}) \leq \mathrm{E}\|\eta\|_\infty \, \mathrm{w}_\eta(\mathcal{V}) \quad \text{if} \quad \eta \overset{dist}{=} -\eta.$$

$$\begin{aligned}
\mathrm{E}_s \, \mathrm{E}_\eta \max_{v \in \mathcal{V}} \sum_{i=1}^n \eta_i s_i v_i &\leq \mathrm{E}_\eta \max_{\substack{u \in \mathbb{R}^n \\ |u_i| \leq \|\eta\|_\infty}} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n u_i s_i v_i \\
&= \mathrm{E}_\eta \|\eta\|_\infty \max_{u \in [-1,1]^n} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n u_i s_i v_i \\
&= \mathrm{E}_\eta \|\eta\|_\infty \times \max_{u \in \{-1,1\}^n} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n u_i s_i v_i \\
&= \mathrm{E}_\eta \|\eta\|_\infty \times \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i v_i
\end{aligned}$$

- We can 'contract out' any symmetrically distributed noise vector η by ...
  1. multiplying in independent random signs s_i. Symmetry $\implies$ $s_i \eta_i \overset{dist}{=} \eta_i$.
  2. maximizing over a cube containing η.
- We just have to use a big enough cube.
  - In our classification example, $\eta = \varepsilon - \varepsilon'$ was in the unit cube $[-1, 1]^n$ deterministically.
  - Generally, we maximize over a random cube $[-\|\eta\|_\infty, \ \|\eta\|_\infty]^n$.
  - And we can pull out the cube's radius $\|\eta\|_\infty$ as a multiplicative factor.

Symmetrization, Contraction, and Gaussian Noise

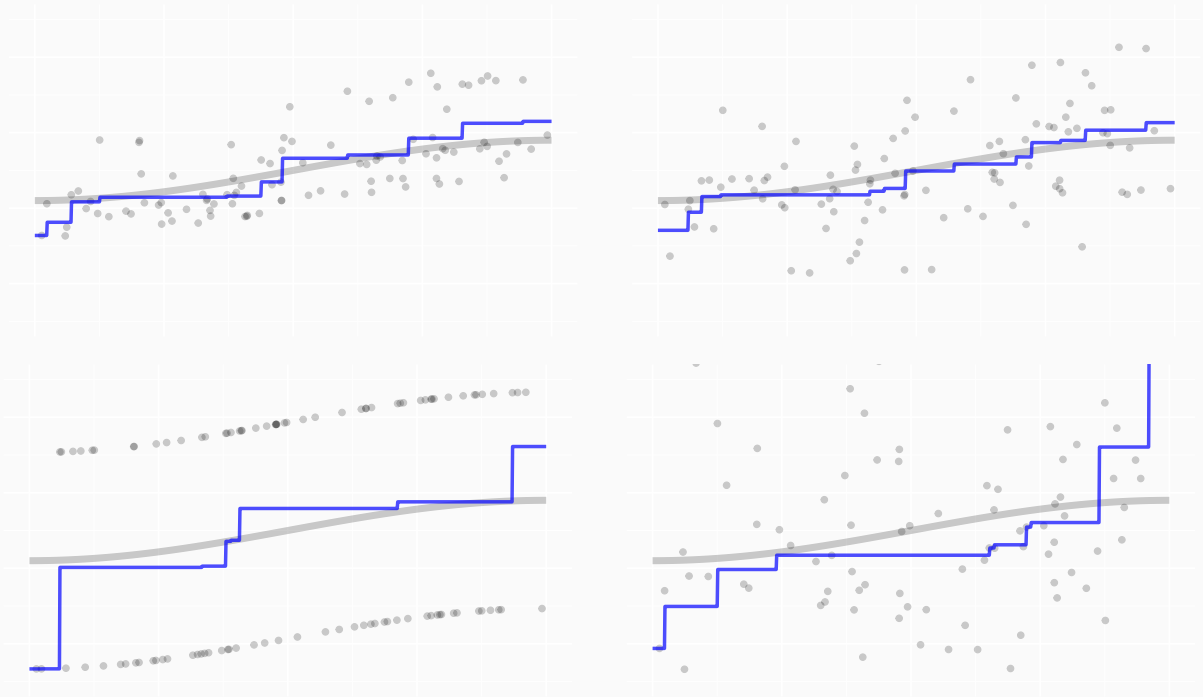


Figure 2: real noise $\rightarrow$ symmetrized noise $\downarrow$ scaled sign noise $\leftarrow$ scaled gaussian noise

After symmetrizing and introducing random signs, i.e. making the substitution

$$\varepsilon_i \rightarrow s_i(\varepsilon_i - \varepsilon_i'),$$

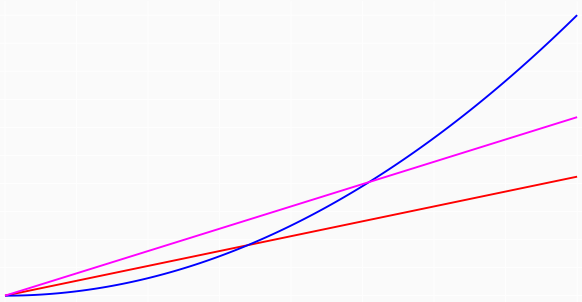
we 'contract out' the symmetrized noise $\varepsilon - \varepsilon'$ to get a bound in terms of random-sign width.

$$\mathrm{w}_\varepsilon(\mathcal{V}) \leq \mathrm{w}_{s(\varepsilon-\varepsilon')}(\mathcal{V}) \leq \|\varepsilon - \varepsilon'\|_\infty \mathrm{w}_s(\mathcal{V}) \leq \|\varepsilon - \varepsilon'\|_\infty \ 1.25 \ \mathrm{w}(\mathcal{V})$$

We can substitute 1.25 times gaussian width because that's at least as large as random sign width.

$$\mathrm{E} \max_{v\in\mathcal{V}} \sum_{i=1}^n g_i v_i = \mathrm{E}_s \mathrm{E}_g \max_{v\in\mathcal{V}} \sum_{i=1}^n |g_i| \ s_i v_i \geq \mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^n \underbrace{\mathrm{E}_g|g_i|}_{\approx 1/1.25} \ s_i v_i.$$

Implications for Regression



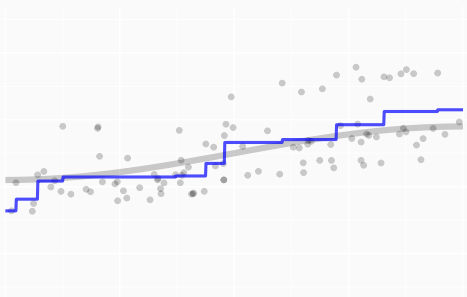
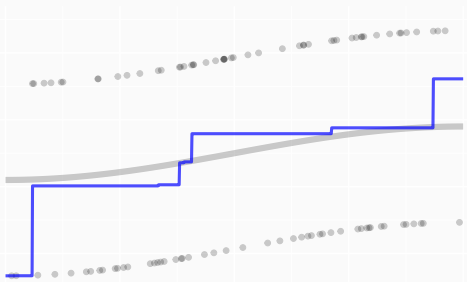
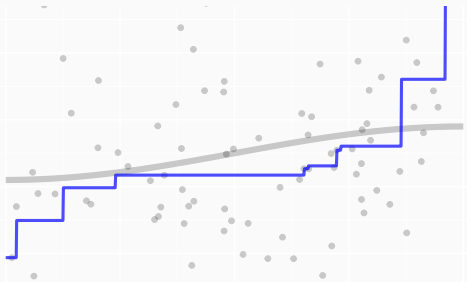
$$\mathrm{w}_\varepsilon(\mathcal{V}) \le M\,\mathrm{w}_s(\mathcal{V}) \le 1.25M\,\mathrm{w}(\mathcal{V})$$
$$\text{for} \quad \mathsf{M} = \mathrm{E}\|\varepsilon - \varepsilon'\|_\infty \le 2\,\mathrm{E}\|\varepsilon\|_\infty.$$

In terms of our crossing-point bounds, regression with arbitrary independent noise,

$$\text{i.e.} \quad Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \varepsilon_1 \ldots \varepsilon_n \text{ are independent ,}$$

is no harder than with scaled random sign noise or with gaussian noise

$$\text{i.e.} \quad Y_i = \mu(X_i) + Ms_i \quad \text{for} \quad s_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$$
$$\text{or} = \mu(X_i) + 1.25Mg_i \quad \text{for} \quad g_i \sim N(0,1)$$

The **scale factor** is $2\times$ the expected magnitude of our noise vector's largest element.

Lost Precision

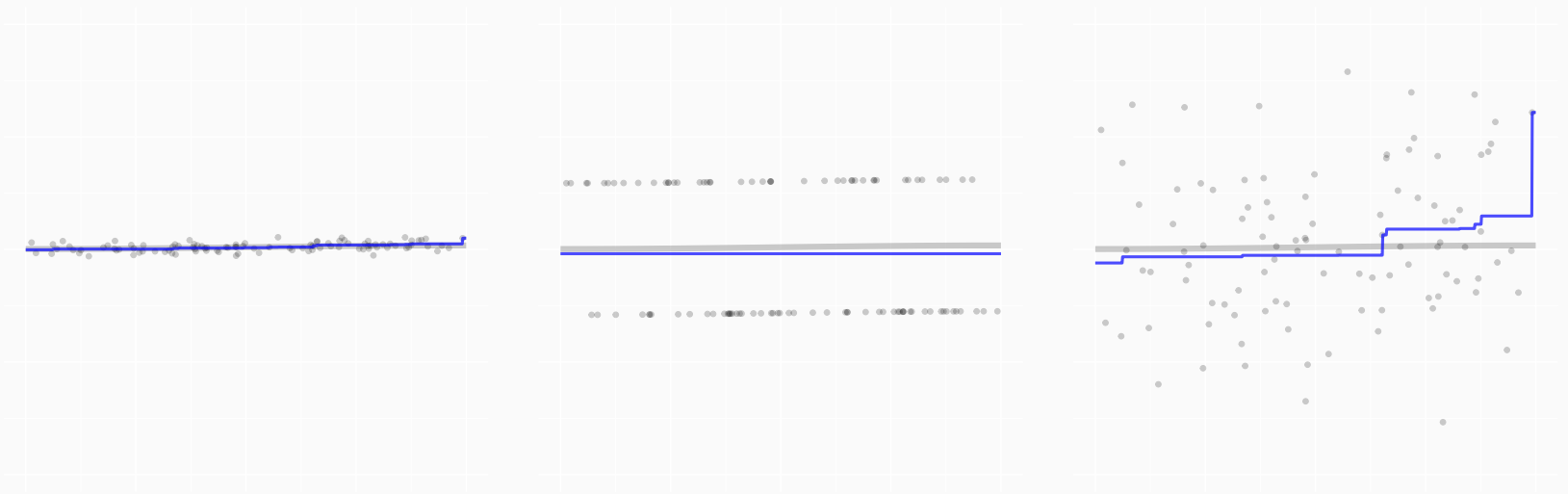


Figure 3: standard gaussian noise $\to$scaled random sign noise $\to$scaled gaussian noise

- This isn't the absolute best bound we can get.
- For example, if we start with standard gaussian noise, we lose ...
- ...a factor of roughly $7\sqrt{\log(2n)}$ going to random sign width and back.

$$\mathrm{w}_\varepsilon(\mathcal{V}) \le 2\,\mathrm{w}_{s\varepsilon}(\mathcal{V}) \le 2\times 2\underbrace{\sqrt{2\log(2n)}}_{\ge \mathrm{E}\|\varepsilon\|_\infty}\mathrm{w}_s(\mathcal{V}) \le 4\sqrt{2\log(n)}\times \underbrace{\sqrt{\frac{\pi}{2}}}_{=\mathrm{E}|g_1|}\ \mathrm{w}_\varepsilon(\mathcal{V}) \approx 7\sqrt{\log(2n)}\,\mathrm{w}_\varepsilon(\mathcal{V}).$$

(a) 'Symmetrization' cost us a factor of 2.

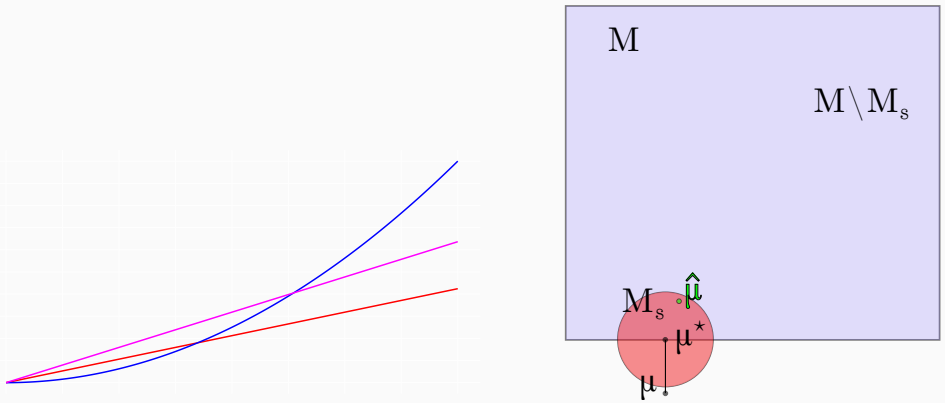
(b) Contraction costs us a factor of $\mathrm{E}\max_{i\le n}|\varepsilon_i| \le 2\sqrt{2\log(2n)}$. (See HW Appendix B)

(c) Converting random signs back to gaussians costs us a factor of $\sqrt{\frac{\pi}{2}} \approx 1.25$.

We're in the right ballpark. For sample sizes n between 50 and 50 million, that factor is between 15 and 30. But if we want a more precise error bound, we need to be a little more careful.

Sampling

What We've Done



We have a bound that's valid for any signal μ and any vector of independent noise ε.

$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P_n})} < 2\sqrt{\Sigma_n}\left(s + \sqrt{\frac{2}{\delta n}}\right) \quad \text{w.p. } 1-\delta \text{ for } \quad \frac{s^2}{2} \geq \mathrm{w}_s(\mathcal{M}_s)$$

- It depends on the model's size through the *critical radius* of random-sign width.

  s satisfying $s^2/2 \geq \mathrm{w}_s(\mathcal{M}_s)$ for $\mathcal{M}_s = \{m \in \mathcal{M} \,:\, \|m - \mu^\star\|_{L_2(\mathrm{P_n})} \leq s\}$

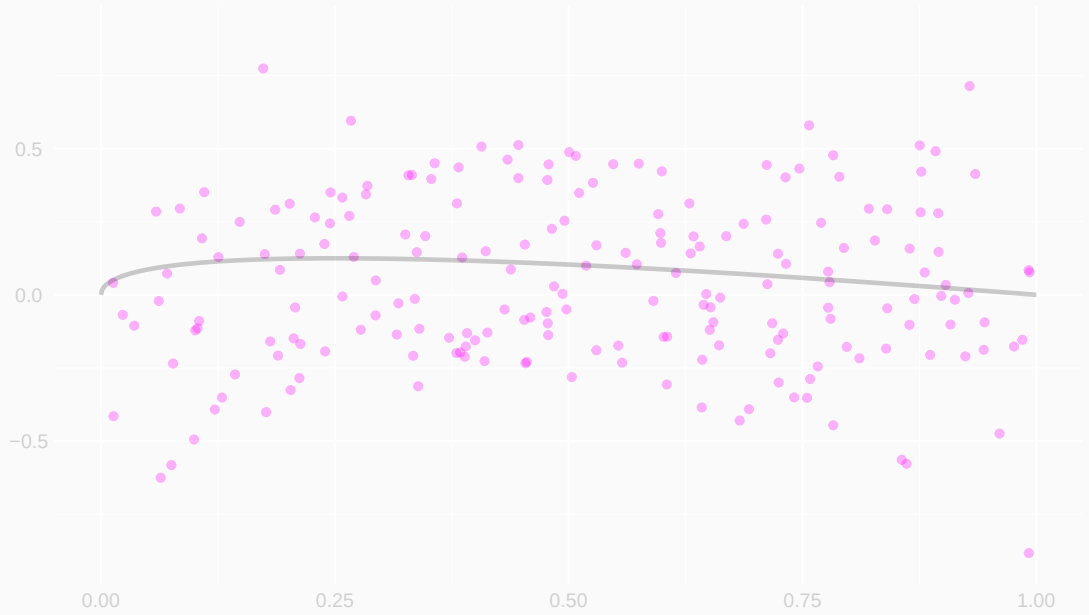
  - This is a one-number summary of the random-sign width of neighborhoods ...
  - ...of the model's best approximation to the signal. It's the summary that matters.
- It depends on the noise's size through the expected maximum square.

$$\Sigma_n = \mathrm{E} \max_{i \in 1 \ldots n} |\varepsilon_i|^2$$

What does this tell us?

Bounds like this say how close $\hat{\mu}$ and $\mu^\star$ are, on average, on our sample $X_1 \dots X_n$.

$$\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(X_i) - \mu^\star(X_i)\} < \dots$$



It doesn't tell us how close they are in the gaps between those points.

- Let's think about what happens when $X_1 \dots X_n$ is are drawn independently from some distribution P. Think sampling with replacement from a population.
- We'll bound the *population root mean squared error* $\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P})}$.

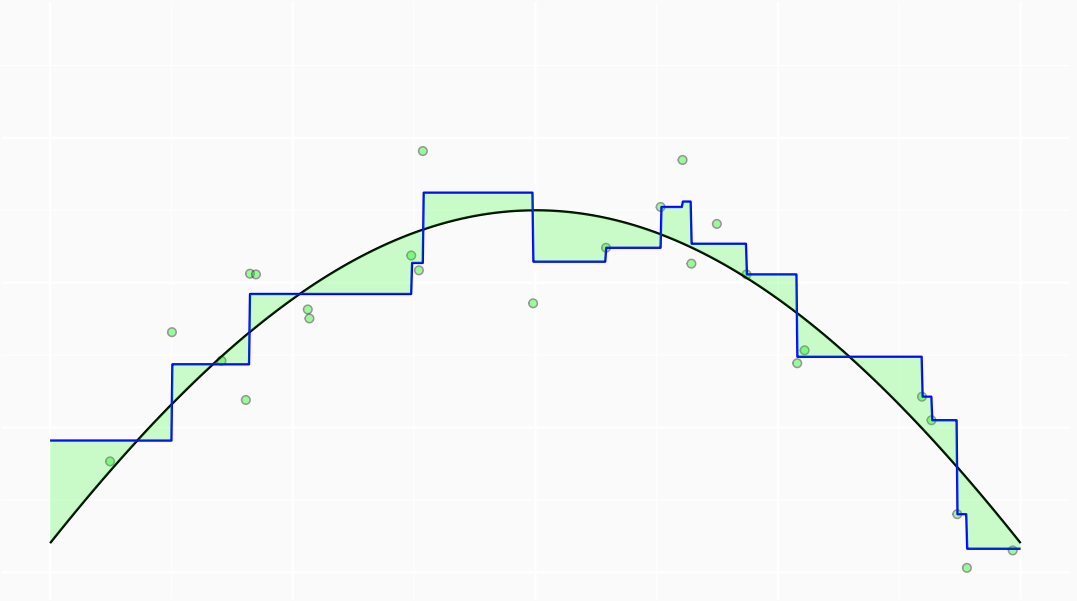
What Population Mean Squared Error Is

It's the mean squared error we make at random point X' distributed like $X_1 \dots X_n$.

$$\|\hat{\mu} - \mu^\star\|^2_{L_2(\mathrm{P})} = \mathrm{E}_{X'}\left[\{\hat{\mu}(X') - \mu^\star(X')\}^2\right]$$

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

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

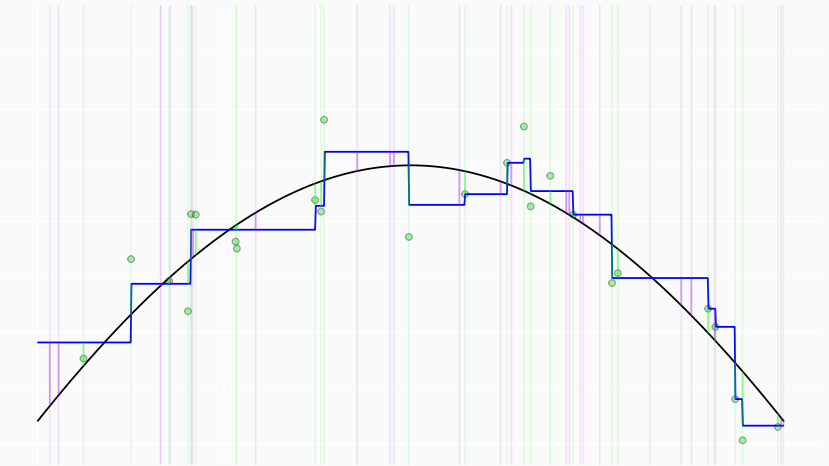


Why we care about Population Mean Squared Error: Generalization

If we're interested in average accuracy for a bunch of new points $X'_1 \dots X'_{n'}$ distributed like $X_1 \dots X_n$, that's more or less exactly what it is.

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

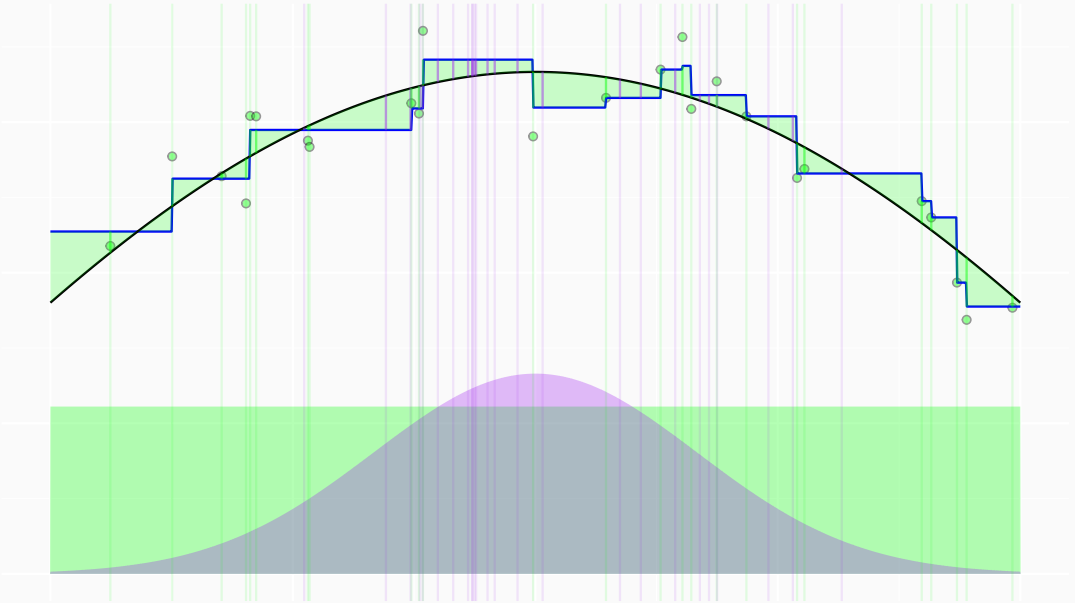
This can be a bit different from accuracy on our original sample $X_1 \dots X_n$.



- BV regression spends its 'variation budget' jumping to fit on the original sample.
- Between those points, it doesn't know whether it should jump or not.
  - So we can 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.

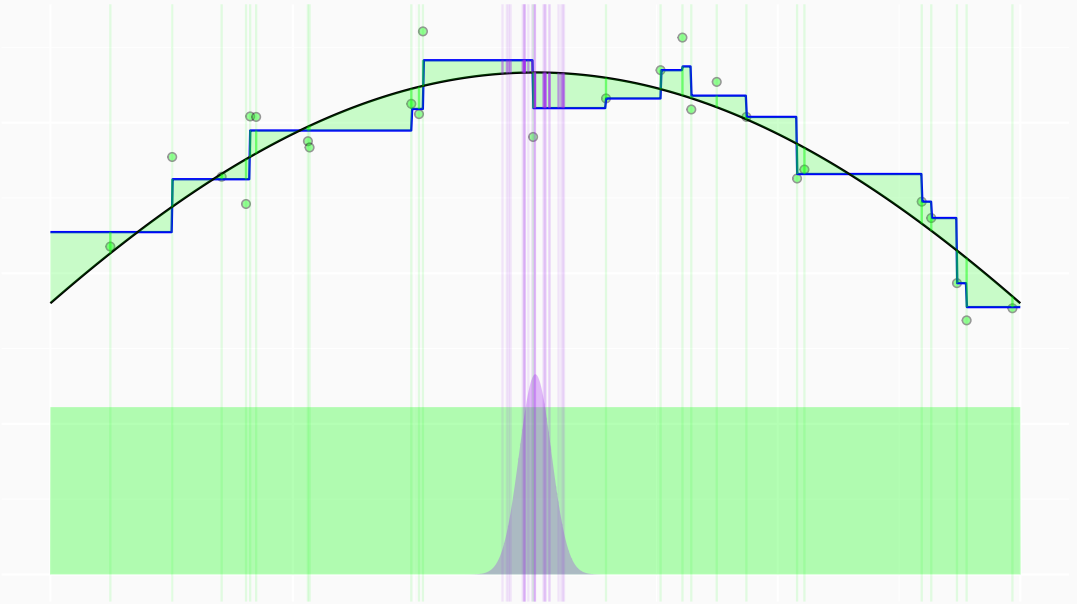


$$\frac{1}{n'}\sum_{i=1}^{n'}\{\hat{\mu}(X_i') - \mu(X_i')\}^2 \approx \|\hat{\mu} - \mu\|^2_{L_2(\mathrm{Q})} = \int \{\hat{\mu}(x) - \mu(x)\}^2 \frac{q(x)}{p(x)} p(x)\, dx$$

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

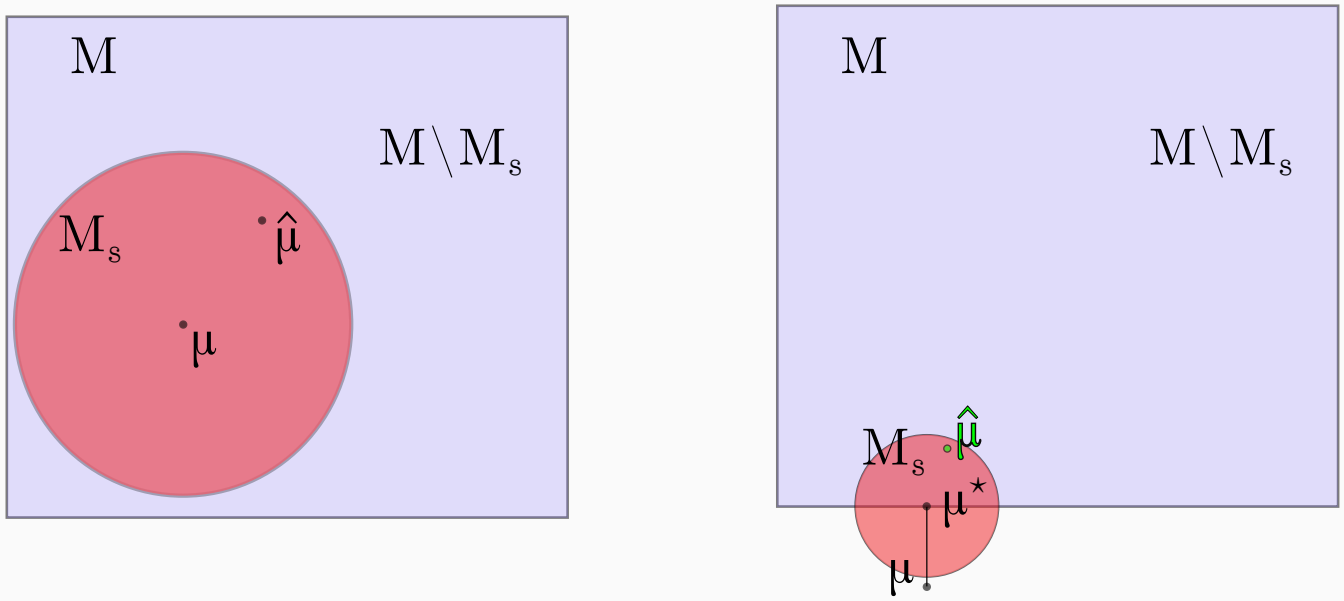
Why we care about Population Mean Squared Error: Generalization

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'.



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

Same Argument, Different Neighborhood



- We want to show that $\hat{\mu}$ is in a *population-distance neighborhood* of μ.
- Or, if we've chosen the model wrong, at least its best population-distance approximation.

$$\hat{\mu} \in \mathcal{M}_s \quad \text{for} \quad \mathcal{M}_s = \{m \in \mathcal{M} \,:\, \|m - \mu^\star\|_{L_2(\mathrm{P})} \leq s\} \quad \text{for} \quad \mu^\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\|_{L_2(\mathrm{P})}$$

- We'll do this using essentially the same argument we used to bound sample MSE.
  1. We know that the $\hat{\mu}$'s squared error loss is at least as good as $\mu^\star$'s.
  2. We find a radius s for which every curve with this property is in the neighborhood $\mathcal{M}_s$.
- It amounts to showing the loss difference $\ell(m) - \ell(\mu^\star)$ is positive outside this neighborhood.

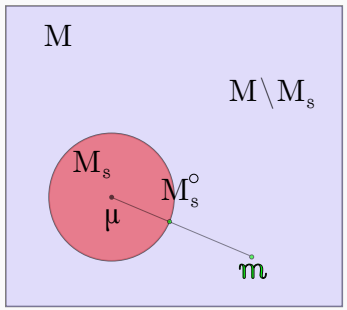
$$m \in \mathcal{M}_s \quad \text{if} \quad m \in \mathcal{M}_s \quad \text{and} \quad \ell(m) - \ell(\mu^\star) > 0 \text{ for all } m \in \mathcal{M} \setminus \mathcal{M}_s$$

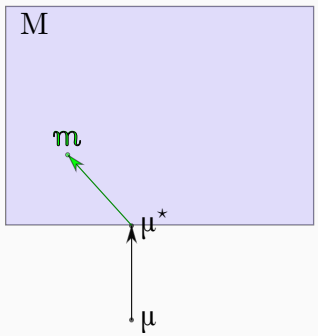
Reduction to a Maximal Inequality

$$\ell(m) - \ell(\mu^\star) = \frac{1}{n}\sum_{i=1}^{n} Z_i(m) := \{m(X_i) - \mu^\star(X_i)\}^2 - 2\ \{Y_i - \mu^\star(X_i)\}\{m(X_i) - \mu^\star(X_i)\}$$

$$= \mathrm{E}\, Z_i(m) + \frac{1}{n}\sum_{i=1}^{n} Z_i(m) - \mathrm{E}\, Z_i(m).$$

Convexity Helps as Usual.





1. The loss difference is positive outside the neighborhood if it's positive on its boundary.

$$m \in \mathcal{M}_s \quad \text{if} \quad m \in \mathcal{M}_s \quad \text{and} \quad \ell(m) - \ell(\mu^\star) > 0 \text{ for all } m \in \cancel{\mathcal{M} \setminus \mathcal{M}_s}\ \mathcal{M}_s^\circ$$

2. The projection theorem tells us an unwanted term in $\mathrm{E}\, Z_i(m)$ is non-negative.

$$-\mathrm{E}\left[\{Y_i - \mu^\star(X_i)\}\{m(X_i) - \mu^\star(X_i)\}\right] = -\mathrm{E}\left[\left\{\overset{=\mu(X_i)}{\mathrm{E}[Y_i \mid X_i]} - \mu^\star(X_i)\right\}\{m(X_i) - \mu^\star(X_i)\}\right]$$

$$= \langle \mu^\star - \mu, m - \mu^\star \rangle_{L_2(P)} \geq 0 \quad \text{for all} \quad m \in \mathcal{M}$$

It follows that $m \in \mathcal{M}_s$ if $m \in \mathcal{M}_s$ and $s^2 > \max_{m \in \mathcal{M}_s^\circ} \frac{1}{n}\sum_{i=1}^{n} Z_i(m) - \mathrm{E}\, Z_i(m)$

Bounding the New Maximum

We show this maximum is **approximately constant**, i.e. close to its expectation.

$$\bar{Z} := \max_{m \in \mathcal{M}_s^\circ} \frac{1}{n} \sum_{i=1}^{n} Z_i(m) - \mathrm{E}\, Z_i(m) \quad \text{satisfies} \quad \bar{Z} \le \mathrm{E}\, \bar{Z} + \sqrt{\frac{\mathrm{Var}(\bar{Z})}{\delta n}} \quad \text{w.p. } 1 - \delta$$

We use **symmetrization** to bound its expectation in terms of random-sign width.

(a) Write the centers $\mathrm{E}\, Z_i(v)$ in terms of an independent copy of our sample.
(b) Compare the result to a maximum of an average of symmetric random variables.
(c) Introduce random signs and compare to two copies of a simpler maximum.

$$\begin{aligned} n \times \mathrm{E}\, \bar{Z} &\overset{(a)}{=} \mathrm{E}_Z \max_{m \in \mathcal{M}_s^\circ} \mathrm{E}_{Z'} \sum_{i=1}^{n} \big\{ Z_i(m) - Z_i'(m) \big\} \\ &\overset{(b)}{\le} \mathrm{E}_Z\, \mathrm{E}_{Z'}\, \mathrm{E}_s \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} s_i \big\{ Z_i(m) - Z_i'(m) \big\} \\ &\overset{(c)}{\le} \mathrm{E}_Z\, \mathrm{E}_{Z'}\, \mathrm{E}_s \max_{m, m' \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} s_i Z_i(m) + (-s_i) Z_i(m') \\ &= 2\, \mathrm{E}_Z\, \mathrm{E}_s \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} s_i Z_i(m) \end{aligned}$$

We can use the **Efron-Stein inequality** to bound the variance. Come back and try it later!

$$\mathrm{Var}(\bar{Z}) \overset{\text{why?}}{\le} \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{E} \big\{ Z_i(\hat{m}) - Z_i'(\hat{m}) \big\}_+^2 \quad \text{for} \quad \hat{m} = \operatorname*{argmax}_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^{n} Z_i(m) - \mathrm{E}\, Z_i(m)$$

$$\le \dots$$

Contracting Out Lipschitz Functions

What we get is $2\times$ the expected random-sign width of some set of vectors,
but it's not just the set of the vectors in our neighborhood $\mathcal{M}_s - \mu^\star$.

$$n\times \mathrm{E}\, Z \le 2\,\mathrm{E}\,\mathrm{E}_s \max_{m\in\mathcal{M}_s^\circ} \sum_{i=1}^n s_i Z_i(m) = \{m(X_i)-\mu^\star(X_i)\}^2 - 2\ \{Y_i-\mu^\star(X_i)\}\{m(X_i)-\mu^\star(X_i)\}$$

$$\le 4\,\mathrm{E}\left\{\max_{m\in\mathcal{M}_s^\circ} \|m-\mu\|_{L_\infty(\mathrm{P_n})} + \|\varepsilon\|_{L_\infty(\mathrm{P_n})}\right\} \overbrace{\mathrm{E}_s \max_{m\in\mathcal{M}_s^\circ} \sum_{i=1}^n s_i\{m(X_i)-\mu^\star(X_i)\}}^{n\times \mathrm{w}_s(\mathcal{M}_s^\circ-\mu^\star)}$$

We've compared that to the width of the neighborhood itself using ...

Lemma (Lipschitz Comparison)

$$\mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^n s_i\psi_i(v_i) \le L\,\mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^n s_i v_i \ \text{ if } \ |\psi_i(u_i)-\psi_i(v_i)| \le L|u_i-v_i| \ \text{ for all } \ u,v\in\mathcal{V}.$$

For $\psi_i(v) = v_i^2 - 2\{Y_i-\mu^\star(X_i)\}v_i$ and $V = \{m(X_1)-\mu^\star(X_1)\ldots m(X_n)-\mu^\star(X_n) \,:\, m\in\mathcal{M}_s^\circ\}$,

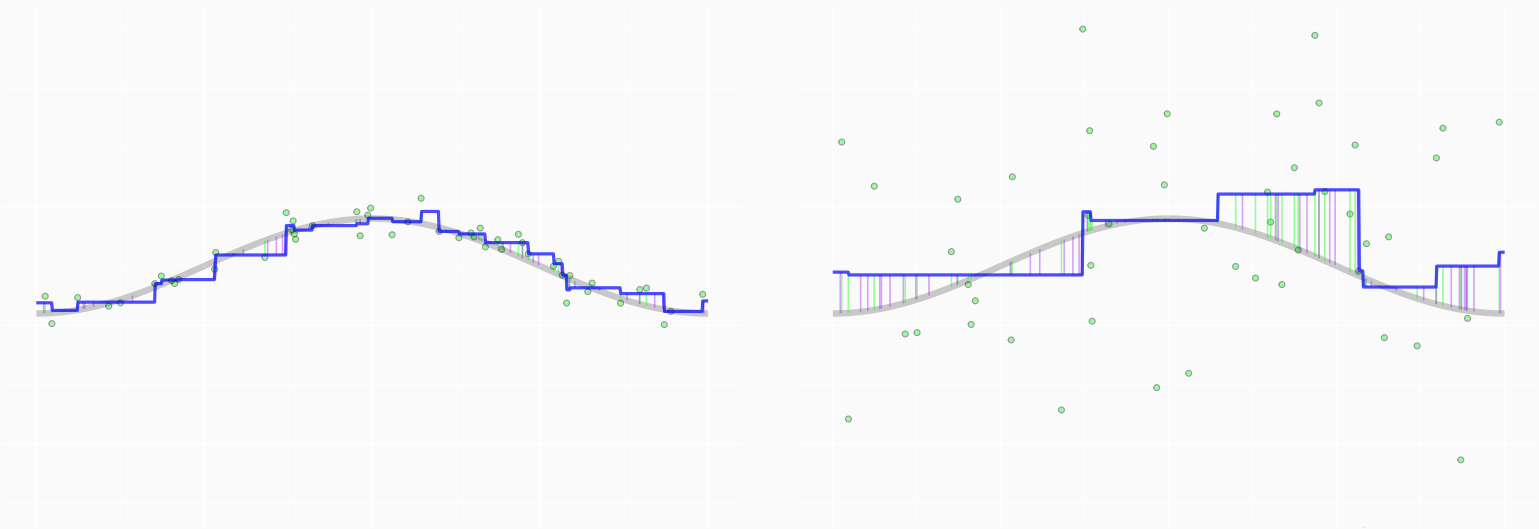
that's $\mathrm{E}_s \max_{m\in\mathcal{M}_s^\circ} \sum_{i=1}^n s_i\psi_i\{m(X_i)-\mu^\star(X_i)\} \le L \max_{m\in\mathcal{M}_s^\circ} \sum_{i=1}^n s_i\{m(X_i)-\mu^\star(X_i)\}$ (with $s_iZ_i(m)$ over the left sum)

where $L = \max_i \max_{m\in\mathcal{M}_s^\circ} |\psi_i'\{m(X_i)-\mu^\star(X_i)\}|$

$$= \max_i \max_{m\in\mathcal{M}_s^\circ} |2\{m(X_i)-\mu^\star(X_i)\} - 2\{Y_i-\mu^\star(X_i)\}|$$

$$\le 2 \max_{m\in\mathcal{M}_s^\circ} \|m-\mu\|_{L_\infty(\mathrm{P_n})} + 2\|\varepsilon\|_{L_\infty(\mathrm{P_n})}.$$

Interpretation



$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P})} \leq s \times 2\Big\{\sqrt{\Sigma_n} + B\Big\} + \sqrt{\frac{\mathrm{Var}(\bar{Z})}{\delta}} \quad \text{w.p. } 1-\delta$$

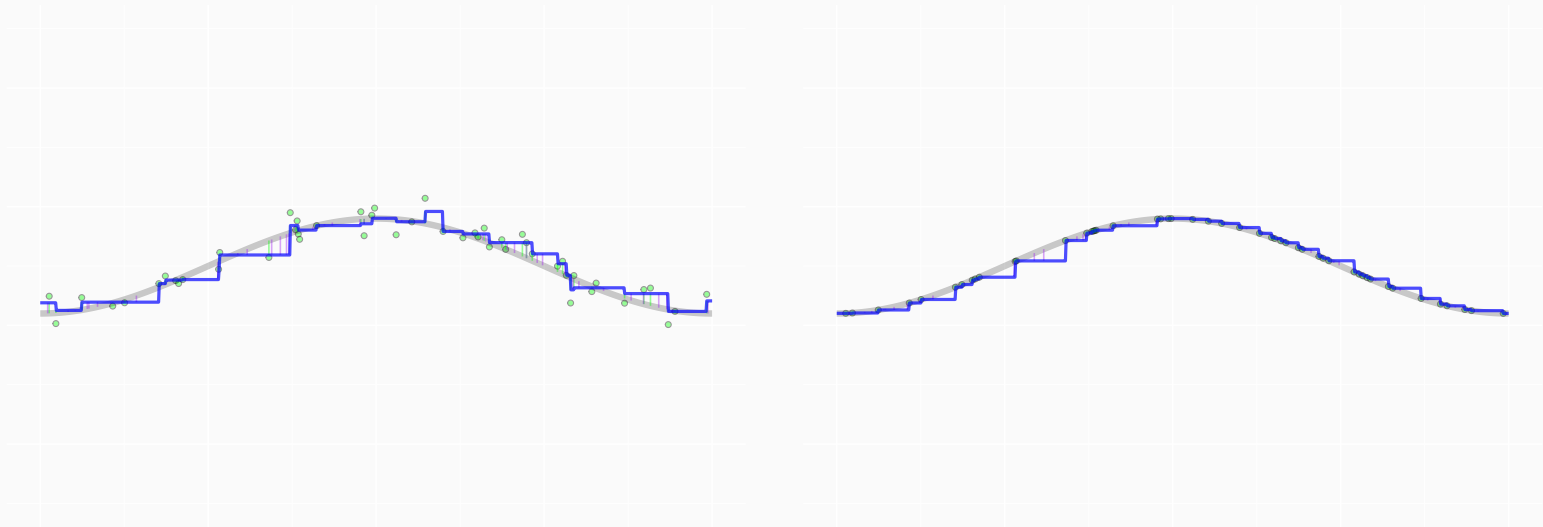
$$\text{if} \quad \frac{s^2}{2} \geq \mathrm{E}\, \mathrm{w}_s(\mathcal{M}_s) \quad \text{and} \quad \|m - \mu\|_\infty \leq B$$

This is the bound we'd get on sample MSE with additional scaled random-sign noise,

i.e. if we'd observed $\quad Y_i = \mu(X_i) + \varepsilon_i + B\, s_i$

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.

Signal Recovery



Signal Recovery is regression without any noise at all. In that case ($\Sigma_n = 0$),

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P})} \leq s \times 2\Big\{\sqrt{\Sigma_n} + B\Big\} + \sqrt{\frac{\mathrm{Var}(\bar{Z})}{\delta}} \quad \text{w.p. } 1-\delta$$

$$\text{if} \quad \frac{s^2}{2} \geq \mathrm{E}\, \mathrm{w}_s(\mathcal{M}_s) \quad \text{and} \quad \|m - \mu\|_\infty \leq B$$

This is the bound we'd get on sample MSE with *only* scaled random-sign noise.

$$\text{i.e. if we'd observed} \quad Y_i = \mu(X_i) + \varepsilon_i + Bs_i$$

- This is an extreme case of the low-noise regime. And it's still hard.
- When you want to estimate μ between the sample points $X_1 \ldots X_n$, ...
- ...what you want to see obscured by bounded 'sampling noise' $\in [-B, B]$.

References

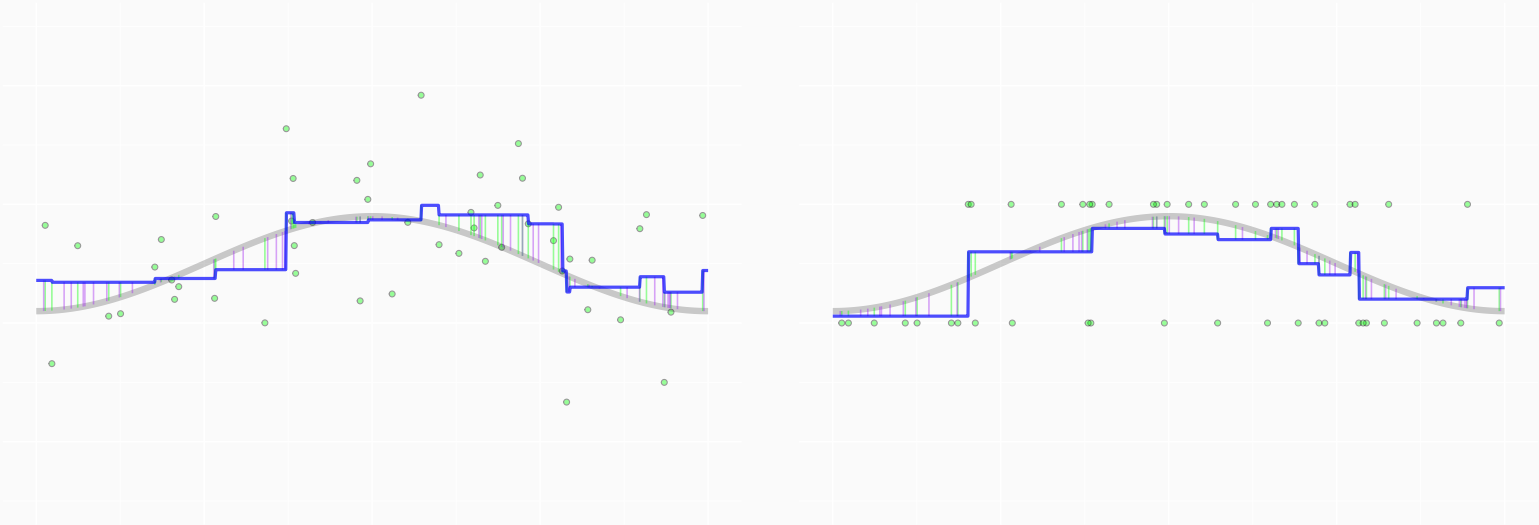
Chapter 6 of Talagrand's Upper and Lower Bounds for Stochastic Processes.

- Random Signs vs. Gaussians: Proposition 6.22
- Contraction: Lemma 6.4.5
- Lipschitz Contraction: Theorem 6.5.1

Appendices

Appendices

Boundedness

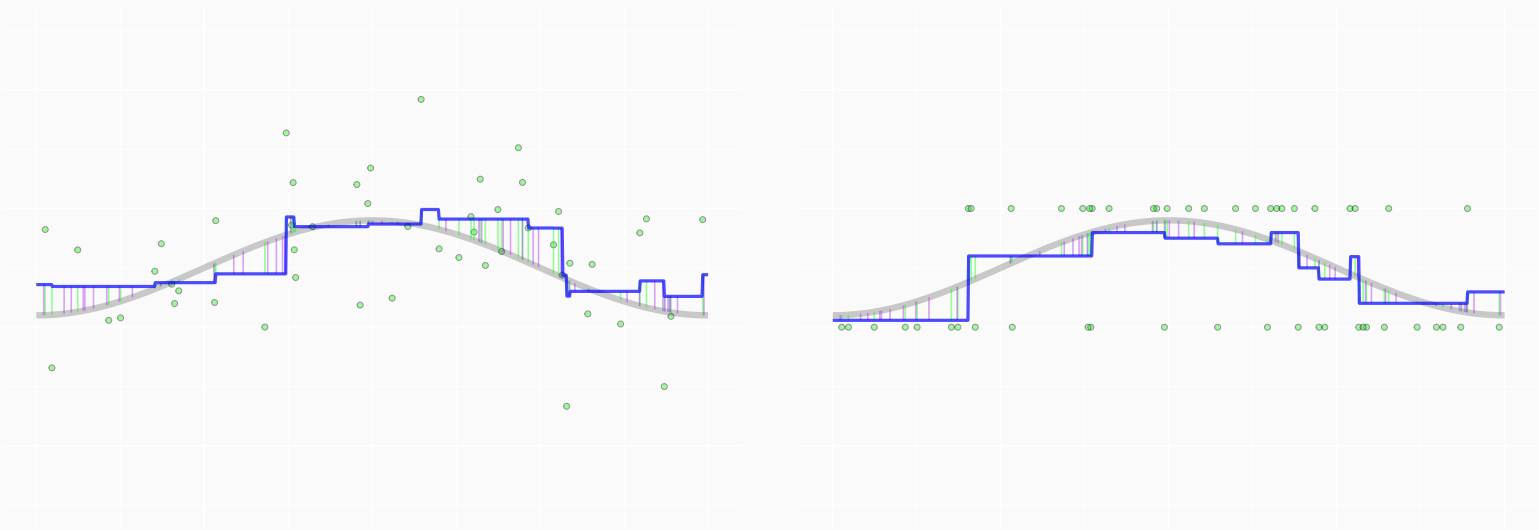


Our Population MSE bound introduces a new consideration: boundedness of $\|m - \mu\|_\infty$ in neighborhoods of $\mu^\star$.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P})} \leq s \times 2\Big\{\sqrt{\Sigma_n} + B\Big\} + \sqrt{\frac{\mathrm{Var}(\bar{Z})}{\delta}} \quad \text{w.p. } 1 - \delta$$

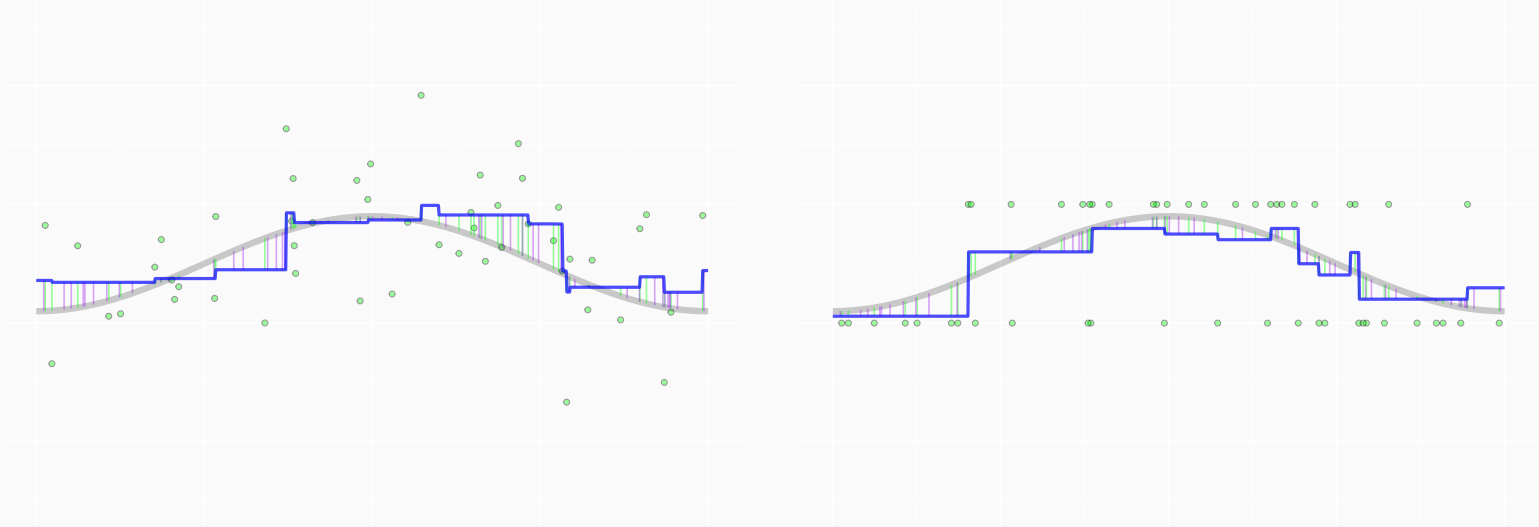
$$\text{if} \quad \frac{s^2}{2} \geq \mathrm{E}\, \mathrm{w}_s(\mathcal{M}_s) \quad \text{and} \quad \|m - \mu\|_\infty \leq B$$

Getting a bound B can take a bit of work. There are options.



Option 1. Baking it into the Model.

$$\mathcal{M} = \{m \ : \ \|m\|_\infty \leq B \quad \text{and} \quad \rho_{TV}(m) \leq B\}$$
$$\implies \|m - \mu\|_\infty \leq \|m\|_\infty + \|\mu\|_\infty \leq B + \|\mu\|_\infty$$
$$\mathcal{M} = \{m \ : \ m(0) = 0 \quad \text{and} \quad \rho_{TV}(m) \leq B\}$$
$$\implies$$

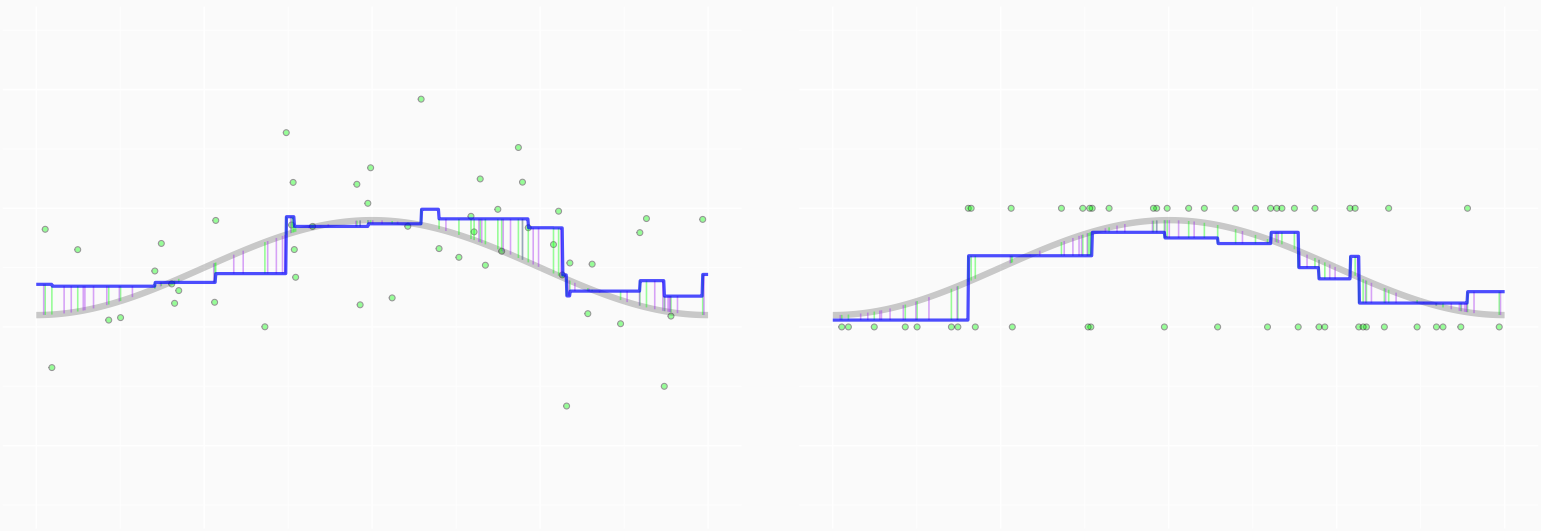


Option 2. Arguing Based on Bounded Data.
In many models, you can show that $\hat{\mu}$ is will be within the range of the data.

$$\text{i.e.} \quad \min_{i \leq n} Y_i \leq \hat{\mu}(x) \leq \max_{i \leq n} Y_i$$

This is true, in particular, for Monotone and Bounded Variation Regression.
We can add this constraint to our model when doing our analysis.

$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P})} < s \quad \text{if} \quad \ell(m) - \ell(\mu^\star) > 0 \quad \text{for all} \quad m \in \mathcal{M} \ldots$$
$$\ldots \text{ with } \|m\|_\infty \leq B \text{ and } \|m - \mu^\star\|_{L_2(\mathrm{P})} \geq s$$



The are other options.