Machine Learning Theory

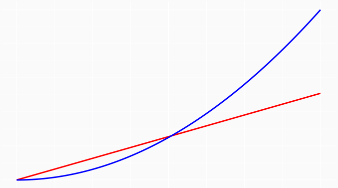
Sampling, Misspecification, and Non-Gaussian Noise

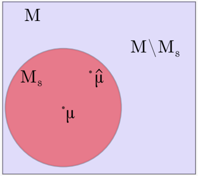
David A. Hirshberg

April 3, 2025

Emory University

Where We Left Off

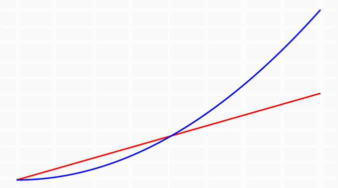


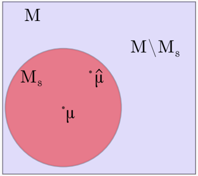


What do we know about the error of this least squares estimator $\hat{\mu}$?

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 \quad \text{for} \quad \text{convex } \mathcal{M}$$

Where We Left Off





What do we know about the error of this least squares estimator $\hat{\mu}$?

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 \quad \text{for} \quad \text{convex } \mathcal{M}$$

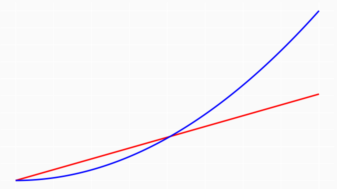
Here's what we proved in lecture.

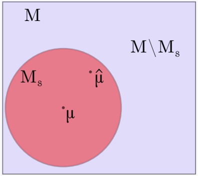
$$\lVert \hat{\mu} - \mu^\star \rVert_{L_2(\mathrm{P_n})} < s \quad \text{w.p. } 1-\delta \text{ for } \quad \frac{s^2}{2\sigma} \geq \mathrm{w}(\mathcal{M}_s^\circ) + s\sqrt{\frac{2\Sigma_n}{\delta n}}$$

where $\quad \mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g,\ v \rangle_{L_2(\mathrm{P_n})}$ and $\quad \Sigma_n = \sigma^2\{1 + 2\log(2n)\}$ for $g_i \overset{iid}{\sim} N(0,1)$

if $\quad Y_i = \mu(X_i) + \varepsilon_i \quad$ for $\quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad$ for $\quad \mu \in \mathcal{M}$

Where We Left Off





What do we know about the error of this least squares estimator $\hat{\mu}$?

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2 \quad \text{for} \quad \text{convex } \mathcal{M}$$

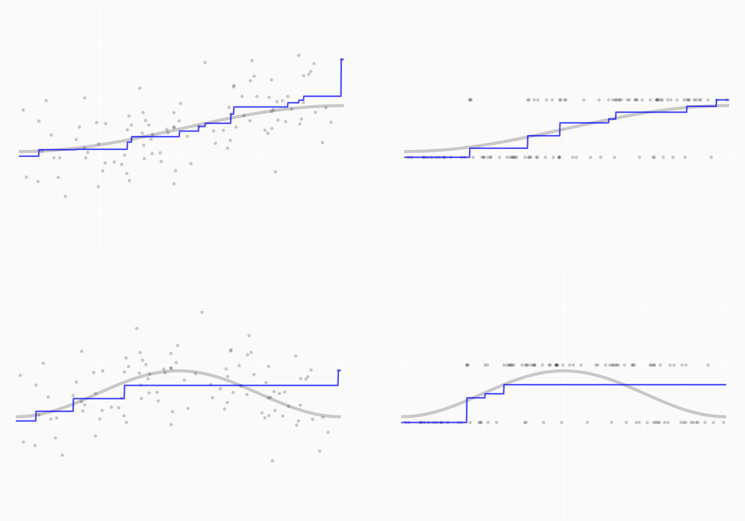
Here’s a simplified version of you’re proving for homework.

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

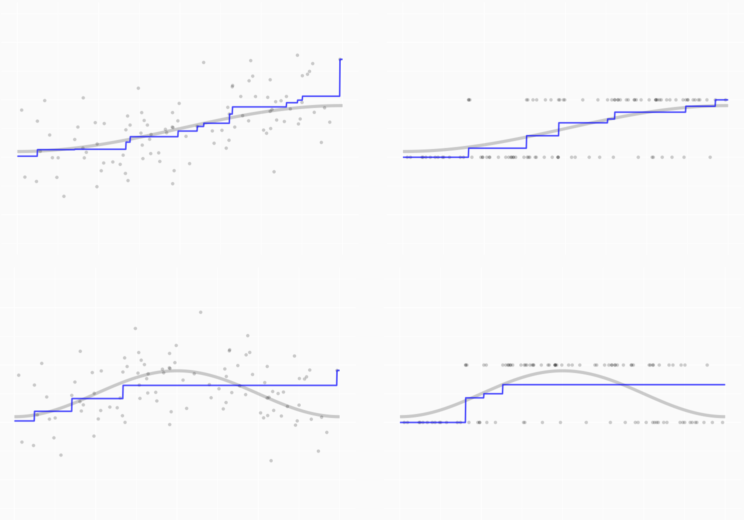
where $\mathrm{w}(\mathcal{V}) = \mathrm{E}\max_{v \in \mathcal{V}}\langle g,\ v\rangle_{L_2(\mathrm{P_n})}$ and $\Sigma_n = \sigma^2\{1 + 2\log(2n)\}$ for $g_i \overset{iid}{\sim} N(0,1)$

if $Y_i = \mu(X_i) + \varepsilon_i$ for $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ for $\mu \in \mathcal{M}$

When Does This Bound Apply?

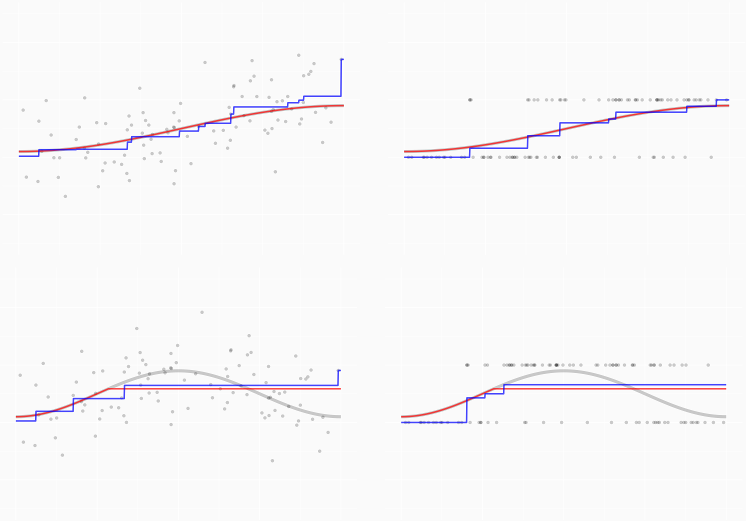


When Does This Bound Apply?

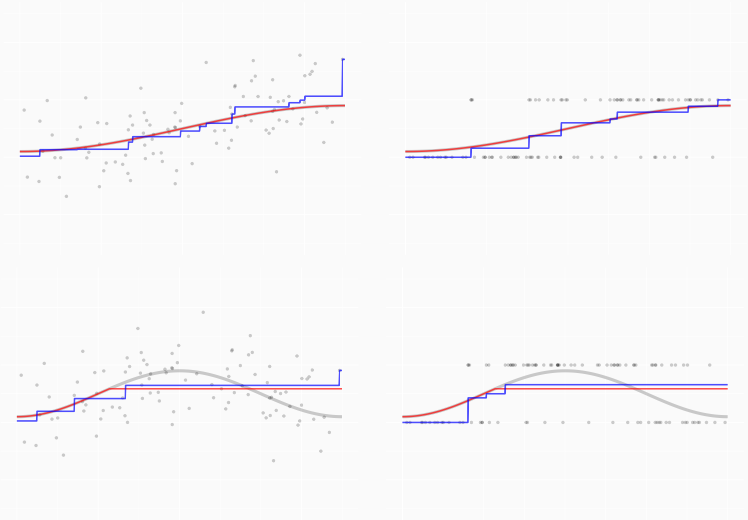


- The second column is out. We’ve assumed correct specfication.
- The second row is out. We’ve assumed normality.

Today, We Fix That



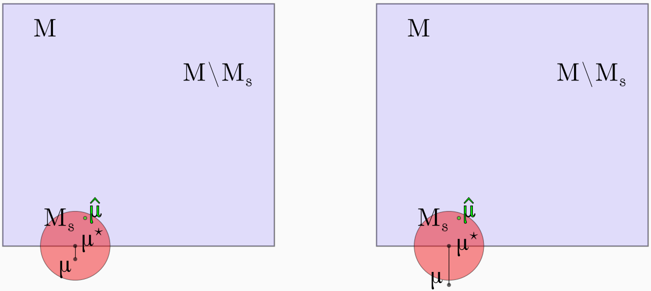
Today, We Fix That



- With misspecification, we estimate the model's best approximation to μ.
- Non-normality doesn't really matter much. We'll look at how it affects our bound.

Misspecification

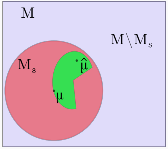
What happens when μ isn't in the model?

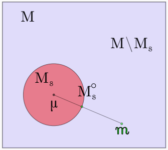


- Our error in estimating μ is bounded by a sum of two terms.
  - The critical radius s, i.e., the one satisfying $s^2/2\sigma \ge \mathrm{w}(\mathcal{M}_s^\circ) + s\sqrt{\frac{2\Sigma_n}{\delta n}}$.
  - The distance from μ to its best approximation in the model. Or really 3 times that.

  We showed this in the model selection lab using the Cauchy-Schwarz inequality.
- In convex models, we can say more.
  Our error in estimating $\mu^\star$ does not depend on its distance to μ.

Review





$$\hat{\mu} \quad \text{minimizes} \quad \ell(m) = \overset{\text{squared error loss}}{\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu(X_i)\}^2}$$

among curves m in a convex set $\mathcal{M}$.

- If μ is in the model, that tells us it's one of the curves with loss as small as μ's.

$$\text{i.e.} \quad m = \hat{\mu} \quad \text{satisfies} \quad \ell(m) \le \ell(\mu) \quad \text{if} \quad \mu \in \mathcal{M}.$$

- To prove $\hat{\mu}$ is in the neighborhood $\mathcal{M}_s$, we show that ...
- ...none of these curves is in the neighborhood's complement $M \setminus \mathcal{M}_s$.

$$\hat{\mu} \in \mathcal{M}_s \quad \text{if} \quad \ell(m) > \ell(\mu) \quad \text{for all} \quad m \in \mathcal{M} \setminus \mathcal{M}_s.$$

- i.e. we show the *loss difference* is strictly positive for curves in the complement.
- That's true if it's positive for curves on the neighborhood's boundary $\mathcal{M}_s^\circ$.

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

- And that boils down to the neighborhood's *squared radius* exceeding ...
- ...twice its boundary's *maximal inner product with noise* $\varepsilon = Y - m$.

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

- Then we do a little probability and get our error bound.

The Argument with no if

For any $\mu^\star \in \mathcal{M}$, we can expand our mean squared error difference as before.

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

But our new 'noise' $\varepsilon_i^\star$ doesn't have mean zero. It's our old noise ε_i, minus something.

$$\varepsilon_i^\star = \{\underbrace{Y_i - \mu(X_i)}_{\varepsilon_i}\} - \{\underbrace{\mu^\star(X_i) - \mu(X_i)}_{\text{something}}\}.$$

So we can think of our mean squared error difference as having three terms:

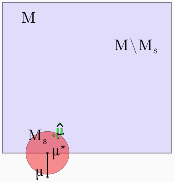
$$\begin{aligned} \ell(m) - \ell(\mu^\star) &= \|m-\mu^\star\|_{L_2(\mathrm{P_n})}^2 && \text{squared distance, like before;} \\ &\quad - \frac{2}{n}\sum_{i=1}^n \varepsilon_i\,\{m(X_i) - \mu^\star(X_i)\} && \text{a mean zero term, like before;} \\ &\quad + \frac{2}{n}\sum_{i=1}^n \{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\} && \text{and something else.} \end{aligned}$$

We can use our argument, ignoring the new term, if that term is always *non-negative*.

Why?

Why.

$$
\begin{aligned}
\ell(m) - \ell(\mu^\star) &= \|m - \mu^\star\|_{L_2(\mathrm{P_n})}^2 \\
&\quad - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \left\{m(X_i) - \mu^\star(X_i)\right\} \\
&\quad + \frac{2}{n}\sum_{i=1}^{n} \{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\}
\end{aligned}
$$



We want to show that if distance from m to $\mu^\star$ is big enough, it wins.

- In particular, it wins in the sense that the loss difference $\ell(m) - \ell(\mu^\star)$ is positive.
- That implies distance from $\hat{\mu}$ to $\mu^\star$ is smaller, as distance doesn't win in that case.

If this new term is non-negative, it helps distance win.

- If the loss difference is positive when we ignore a non-negative term ...
- ...then it's still positive when we don't.

$$
\ell(m) - \ell(\mu^\star) > 0 \quad \text{if} \quad \|m - \mu^\star\|_{L_2(\mathrm{P_n})}^2 - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \left\{m(X_i) - \mu^\star(X_i)\right\} > 0
$$

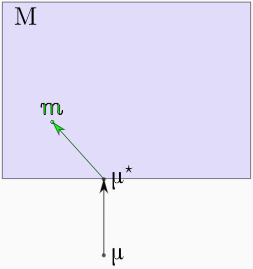
what we're used to

$$
\text{and} \quad \frac{2}{n}\sum_{i=1}^{n} \{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\} \geq 0
$$

new term

This only works if the new term is non-negative. Can we choose $\mu^\star \in \mathcal{M}$ so it is?

We can



The new term is always non-negative when we compare to the *best approximation* to μ in the model,

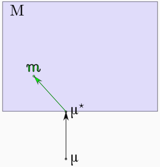
$$\mu^\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\|^2_{L_2(\mathrm{P_n})} \quad \text{satisfies} \quad \frac{2}{n}\sum_{i=1}^{n}\{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\}$$

$$\text{or in vector notation} \quad \frac{2}{n}\langle \mu^\star - \mu, m - \mu^\star \rangle_2 \ge 0 \quad \text{for all} \quad m \in \mathcal{M}.$$

It's proportional to the dot product between two vectors: $\mu \to \mu^\star$ and $\mu^\star \to m$.

- When the model $\mathcal{M}$ is convex, these vectors are always in the same direction.
- They both point 'in' to the model. That means the dot product is non-negative.

Proof



Claim. For any convex set $\mathcal{M}$ in an inner product space, [1]

$\mu^\star = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \|m - \mu\|$ satisfies

$\langle \mu^\star - \mu,\ m - \mu^\star \rangle \geq 0$ for all curves $m \in \mathcal{M}$.

Proof. Let $m_\lambda = \lambda(m - \mu^\star) + \mu^\star$.

$$\begin{aligned}\|m_\lambda - \mu\|^2 &= \langle \lambda(m - \mu^\star) + (\mu^\star - \mu),\ \lambda(m - \mu^\star) + (\mu^\star - \mu) \rangle \\ &= \lambda^2 \|m - \mu^\star\|^2 + \|\mu^\star - \mu\|^2 + 2\lambda \langle m - \mu^\star,\ \mu^\star - \mu \rangle.\end{aligned}$$

Because $m_\lambda \in \mathcal{M}$, it follows that this is at least as large as $\|\mu - \mu^\star\|^2$, so

$$0 \leq \lambda^2 \|m - \mu^\star\|^2 + 2\lambda \langle m - \mu^\star,\ \mu^\star - \mu \rangle$$

and therefore, dividing by $\lambda > 0$, that

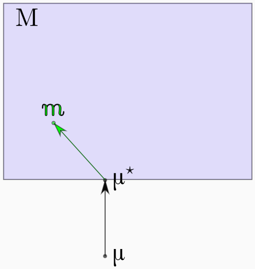
$$0 \leq \lambda \|m - \mu^\star\|^2 + 2 \langle m - \mu^\star,\ \mu^\star - \mu \rangle.$$

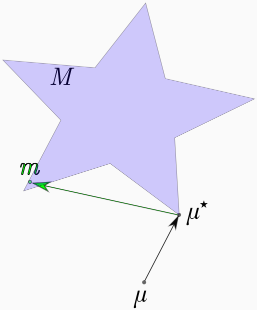
Because this holds for arbitrarily small $\lambda > 0$, it must also hold for $\lambda = 0$.

[1] An inner product space is a vector space with a norm $\|u\| = \sqrt{\langle u, u \rangle}$ induced by an inner product $\langle u, v \rangle$.

That’s not true for other choices

When $\mu^\star \in \mathcal{M}$ isn’t the closest point to μ,
these vectors can point in opposite directions.
That is, this dot product can be negative for some $m \in \mathcal{M}$.



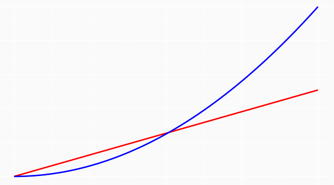


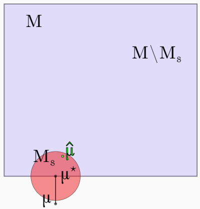
The same thing can happen *for the closest point* in a non-convex model.

Summary

When we use a convex model, the least squares estimator $\hat{\mu}$ converges to the model's closest point to μ. This generalizes our result without misspecification.

- If μ is in the model, that closest point is μ.
- Otherwise, it's something else.





We can bound our estimator's distance to that closest point $\mu^\star$ just like we've been bounding distance to μ when we assumed it was in the model.

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

$$\text{for} \quad \mathcal{M}_s = \left\{m \in \mathcal{M} : \|m - \mu^\star\|_{L_2(\mathrm{P_n})} \leq s\right\} \quad \text{and} \quad \Sigma_n = \sigma\{1 + 2\log(2n)\}$$

$$\textbf{if} \quad Y_i = \mu(X_i) + \varepsilon_i \quad \text{for} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad \textbf{for some function } \mu.$$

Misspecification

Examples

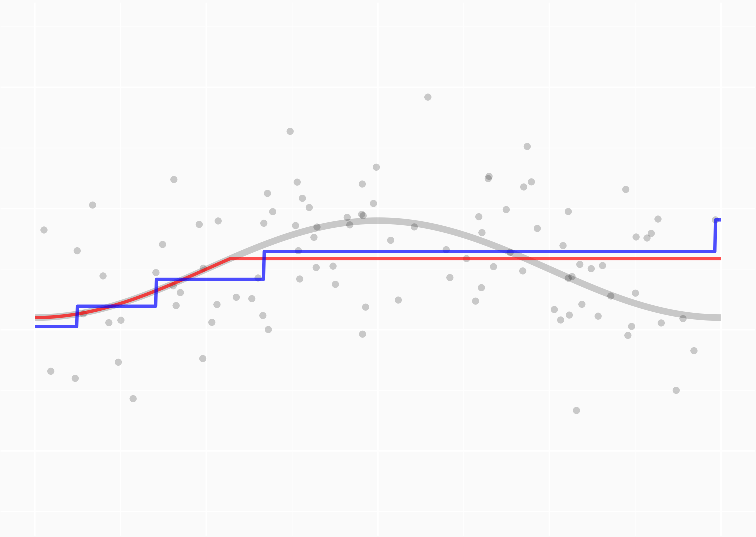


Figure 1: Increasing Curves ($n = 100$.)

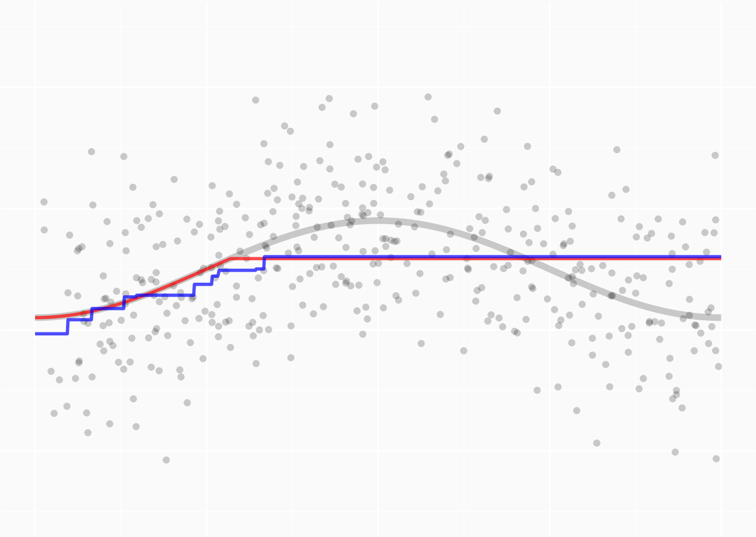


Figure 2: Increasing Curves ($n = 400$.)

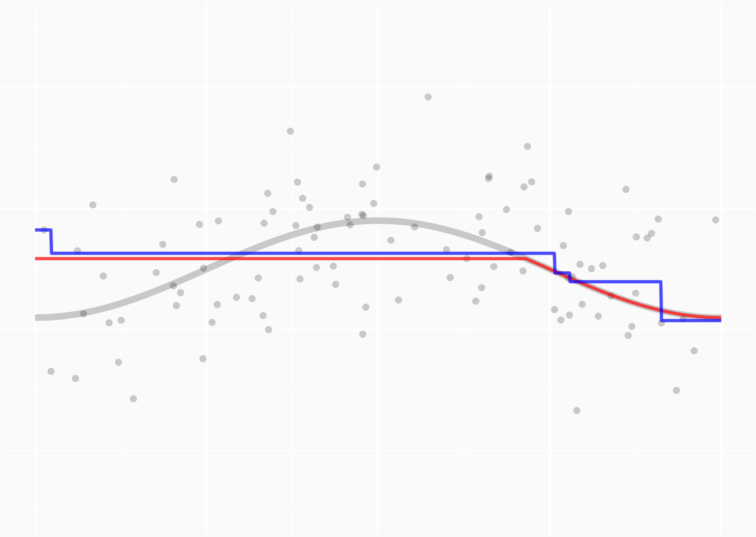


Figure 3: Decreasing Curves ($n = 100$.)

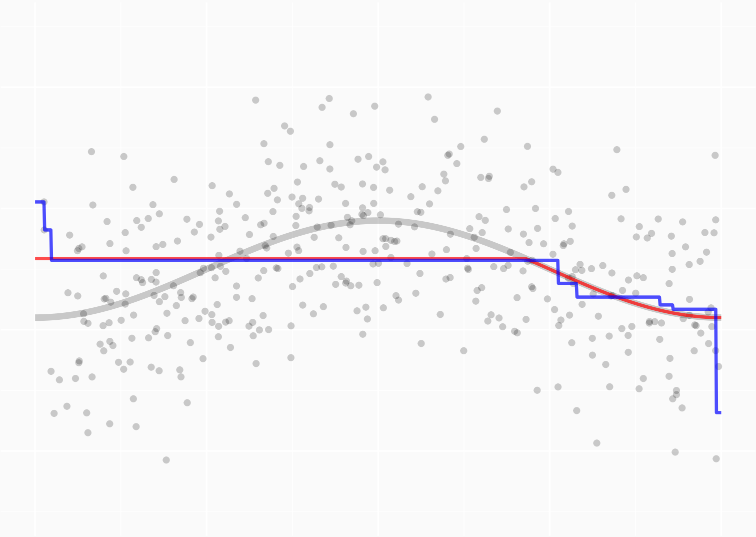


Figure 4: Decreasing Curves ($n = 400.$)

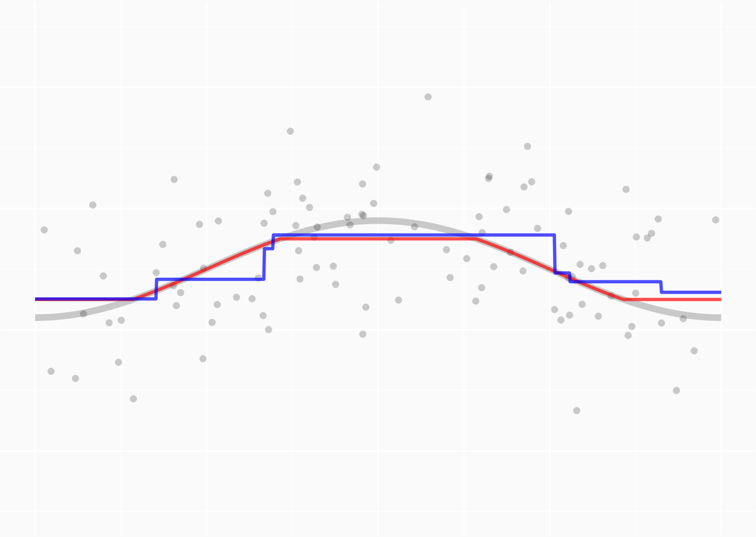


Figure 5: Bounded Variation Curves: $\rho_{\text{TV}} \leq 1$ $(n = 100.)$

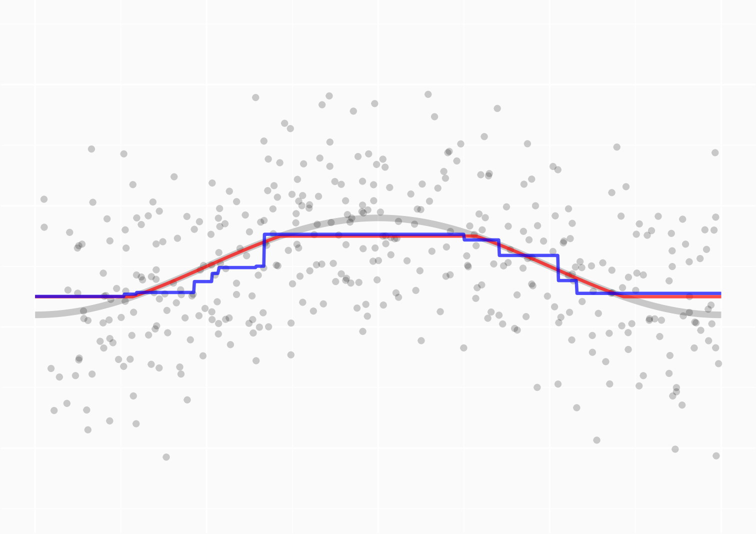


Figure 6: Bounded Variation Curves: $\rho_{\text{TV}} \leq 1$. ($n = 400$.)

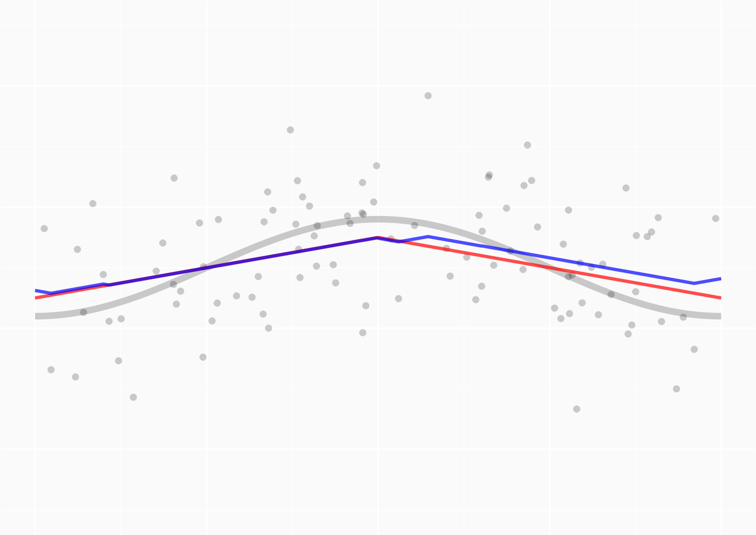


Figure 7: Lipschitz Curves: $\rho_{\text{Lip}} \leq 1$ $(n = 100.)$

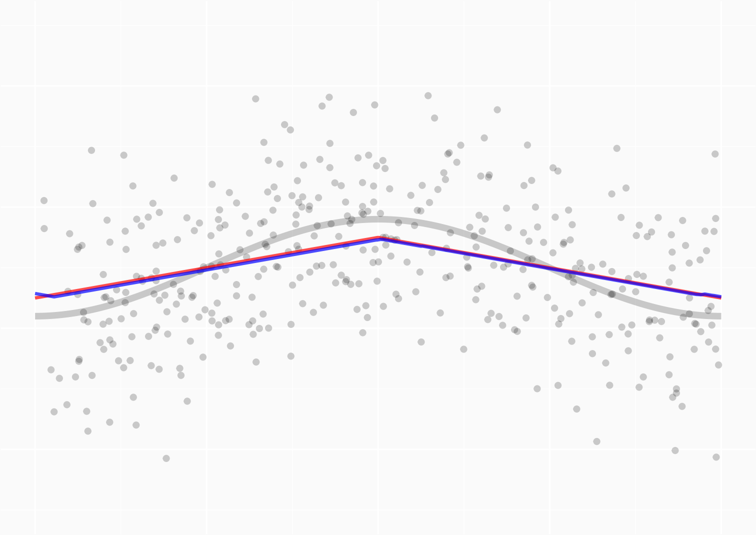


Figure 8: Lipschitz Curves: $\rho_{\mathsf{Lip}} \leq 1$ $(n = 400.)$

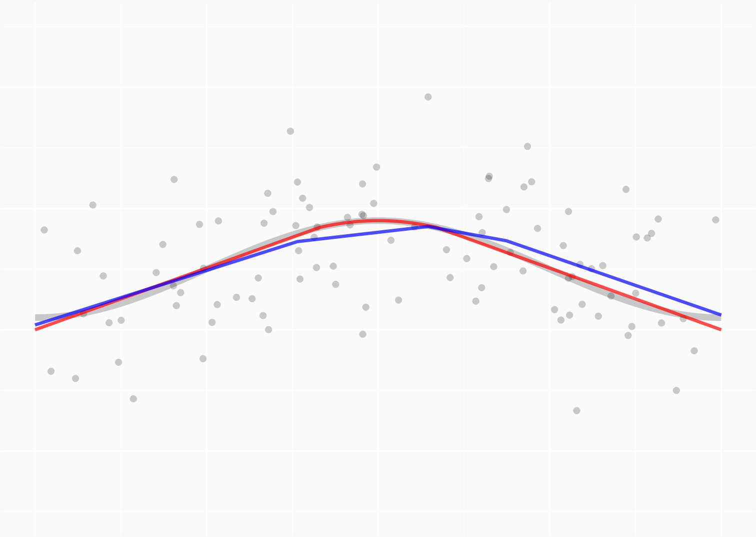


Figure 9: Concave Curves ($n = 100$.)

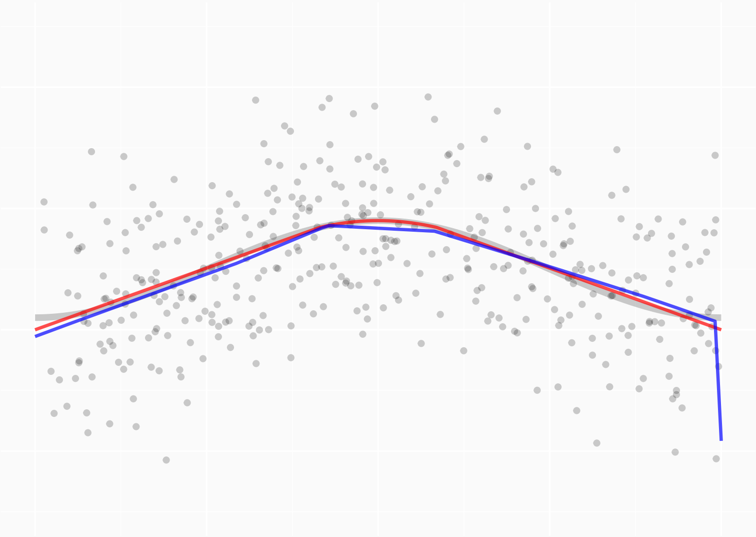
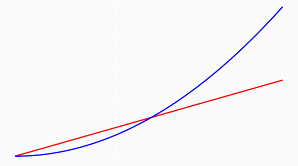


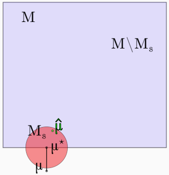
Figure 10: Concave Curves ($n = 400$.)

Non-Gaussian Noise

Starting Point

$$
\begin{aligned}
\ell(m) - \ell(\mu^\star) = \|m - \mu^\star\|^2_{L_2(\mathrm{P_n})} & \qquad \text{squared distance} \\
- \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \left\{ m(X_i) - \mu^\star(X_i) \right\} & \qquad \text{a mean zero term} \\
+ \frac{2}{n}\sum_{i=1}^{n} \{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\} & \qquad \text{a non-negative term}.
\end{aligned}
$$





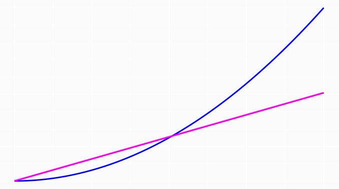
We can bound error using a corresponding *width*, no matter how noise is distributed.

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

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

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

Plan for Today



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

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

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

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

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

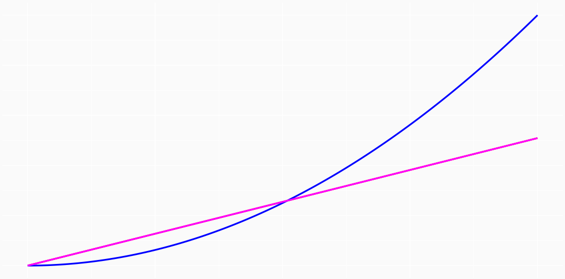
   $\epsilon_i \to \epsilon_i - \epsilon_i'$ where ϵ_i' is an independent copy of ϵ_i

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

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

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

A Simple Consequence: Width Comparison implies Radius Comparison



- If you have a width comparison $\mathrm{w}_\epsilon \le \alpha \mathrm{w}_\eta$ for some $\alpha \ge 1$.
- This implies a radius comparison $s_\epsilon \le \alpha s_\eta$ for all convex models $\mathcal{M}$.

$$s_\epsilon = \alpha s_\nu \quad \text{satisfies} \quad \frac{s_\epsilon^2}{2} \ge \mathrm{w}_\varepsilon(\mathcal{M}_{s_\epsilon}) \quad \text{if} \quad \frac{s_\eta^2}{2} \ge \mathrm{w}_\eta(\mathcal{M}_{s_\eta}) \quad \text{for} \quad \text{convex} \ \mathcal{M}$$

$$\text{and} \quad \mathrm{w}_\varepsilon \le \alpha \mathrm{w}_\eta \quad \text{for} \quad \alpha \ge 1.$$

- *Interpretation.*
  The noise ε makes regression at most 'α times harder' than the noise η.
- *This is simplistic and 'lossy'.*
  For most models, our width comparison implies a better radius comparison.

Proof: Width Comparisons imply Radius Comparisons

Claim. If $\mathrm{w}_\varepsilon \le \alpha\,\mathrm{w}_\eta$ for $\alpha \ge 1$, then for any convex model $\mathcal{M}$, the critical radius using noise ε is at most α times the critical radius using noise η, i.e.

$$\frac{(\alpha s)^2}{2} \ge \mathrm{w}_\varepsilon(\mathcal{M}_{\alpha s}) \quad \text{if} \quad \frac{s^2}{2} \ge \mathrm{w}_\eta(\mathcal{M}_s) \quad \text{and} \quad \mathrm{w}_\varepsilon \le \alpha\,\mathrm{w}_\eta \quad \text{for} \quad \alpha \ge 1.$$

Proof. If $s^2/2 \ge \mathrm{w}_\eta(\mathcal{M}_s)$, then

Proof: Width Comparisons imply Radius Comparisons

Claim. If $\mathrm{w}_\varepsilon \le \alpha\, \mathrm{w}_\eta$ for $\alpha \ge 1$, then for any convex model $\mathcal{M}$, the critical radius using noise ε is at most α times the critical radius using noise η, i.e.

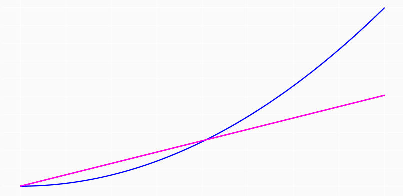
$$\frac{(\alpha s)^2}{2} \ge \mathrm{w}_\varepsilon(\mathcal{M}_{\alpha s}) \quad \text{if} \quad \frac{s^2}{2} \ge \mathrm{w}_\eta(\mathcal{M}_s) \quad \text{and} \quad \mathrm{w}_\varepsilon \le \alpha\, \mathrm{w}_\eta \quad \text{for} \quad \alpha \ge 1.$$

Proof. If $s^2/2 \ge \mathrm{w}_\eta(\mathcal{M}_s)$, then

$$\begin{aligned}
\alpha s/2 &\ge \alpha\, \mathrm{w}_\eta(\mathcal{M}_s)/s && \text{multiplying both sides by } \alpha/s \\
&\ge \alpha\, \mathrm{w}_\eta(\mathcal{M}_{\alpha s})/(\alpha s) && \text{using sublinearity of } f(s) = \mathrm{w}_\eta(\mathcal{M}_s) \\
&\ge \mathrm{w}_\varepsilon(\mathcal{M}_{\alpha s})/(\alpha s) && \text{using our premise } \alpha\, \mathrm{w}_\eta \ge \mathrm{w}_\varepsilon.
\end{aligned}$$

Multiplying both sides by αs, we get our claim.

Summary



Where we are. We have a bound that depends on the model $\mathcal{M}$ and the distribution of the noise ε in a complex and entangled way.

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

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

Where we're going. We'll derive a bound that depends on the model $\mathcal{M}$ and the distribution of the noise ε in simpler and disentangled way.

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

$$\text{where} \quad \mathrm{w}(\mathcal{V}) = \mathrm{E}\max_{v\in\mathcal{V}}\langle g, v\rangle_{L_2(\mathrm{P_n})} \quad \text{and} \quad M_n = \mathrm{E}\max_{i\in 1\ldots n}|\varepsilon_i|.$$

Better yet. We can simplify it into a bound that depends on only one measure of noise complexity.

$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P_n})} \le \sqrt{2\pi\Sigma_n}\left(s + \sqrt{\frac{2}{\delta n}}\right) \quad \text{because} \quad M_n \le \sqrt{\Sigma_n} \quad \text{and} \quad 2 \le \sqrt{2\pi}$$

Non-Gaussian Noise

Example: Probabilistic Classification

The Setting

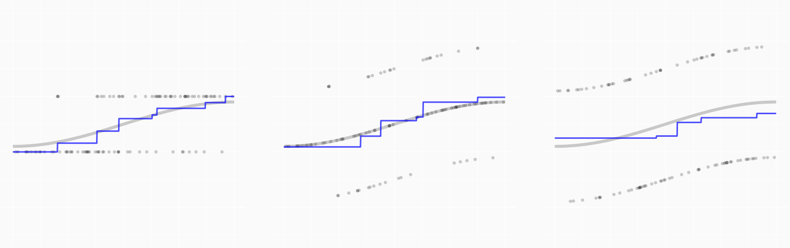


Figure 11: classification noise → symmetrized classification noise → random-sign noise

Suppose we have independent *binary observations*.

$$
\begin{aligned}
Y_i &= \begin{cases} 1 & \text{with conditional probability } \mu(X_i) \\ 0 & \text{otherwise} \end{cases} \\
&= \mu(X_i) + \varepsilon_i \quad \text{for} \quad \varepsilon_i = \begin{cases} 1-\mu(X_i) & \text{with conditional probability } \mu(X_i) \\ -\mu(X_i) & \text{with conditional probability } 1-\mu(X_i) \end{cases}.
\end{aligned}
$$

Note that this *classification noise* ε_i has conditional mean zero.

$$\mathrm{E}[\varepsilon_i \mid X_i] = \mu(X_i)\{1-\mu(X_i)\} + \{1-\mu(X_i)\}\{-\mu(X_i)\} = 0.$$

The Setting

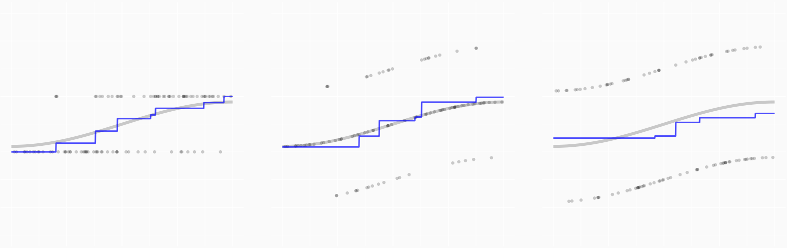


Figure 11: classification noise → symmetrized classification noise → random-sign noise

What we need to bound is *classification-noise width*

$$\mathrm{w}_\epsilon(\mathcal{V}) = \frac{1}{n}\,\mathrm{E}\max_{v\in\mathcal{V}}\sum_{i=1}^{n}\varepsilon_i v_i.$$

We'll show it's no bigger than a version with *symmetrized noise.*

$$\varepsilon_i - \varepsilon_i' = \begin{cases} +1 & \text{when } \ \varepsilon_i = 1-\mu(X_i),\ \varepsilon_i' = \mu(X_i) \\ -1 & \text{when } \ \varepsilon_i = \mu(X_i),\ \varepsilon_i' = 1-\mu(X_i) \\ \ \ 0 & \text{when } \ \varepsilon_i = \varepsilon_i' \end{cases}$$

The Setting

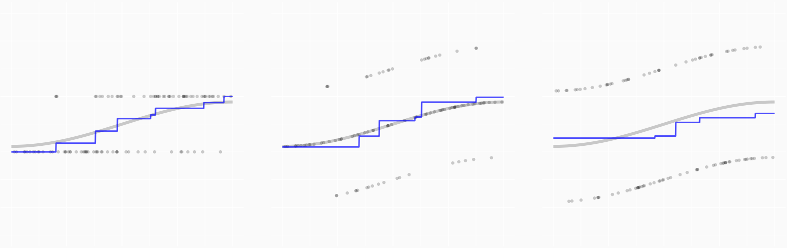


Figure 11: classification noise → symmetrized classification noise → random-sign noise

And we'll show that *this* is no bigger than a version with *random sign noise*

$$\mathrm{w}_{\epsilon}(\mathcal{V}) \le \mathrm{w}_{\epsilon-\epsilon'}(\mathcal{V}) \le \mathrm{w}_{s}(\mathcal{V}) \quad \text{where} \quad s_i = \pm 1 \ \text{ w.p. } \ 1/2.$$

The trick will be multiplying the symmetrized noise by a random sign.
It's already symmetric, so that doesn't change its distribution.

$$\varepsilon_i - \varepsilon_i' \overset{dist}{=} s_i(\varepsilon_i - \varepsilon_i')$$

Then we'll *contract out* the symmetrized noise, leaving the random sign. You'll see.

Step 1

We bound our maximum in terms of one involving symmetric noise.
We'll work with an *independent copy* ε' of our noise vector ε.

$$
\begin{aligned}
\mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} \varepsilon_i v_i &\overset{(a)}{=} \mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \mathrm{E}_{\varepsilon'} \varepsilon'_i) v_i \\
&\overset{(b)}{=} \mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \ \mathrm{E}_{\varepsilon'} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon'_i) v_i \\
&\overset{(c)}{\leq} \mathrm{E}_{\varepsilon} \, \mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon'_i) v_i .
\end{aligned}
$$

Why do these steps work?

(a) $\mathrm{E}_{\varepsilon'} \varepsilon'_i = 0.$

Step 1

We bound our maximum in terms of one involving symmetric noise.
We'll work with an *independent copy* ε' of our noise vector ε.

$$\begin{aligned}
\mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} \varepsilon_i v_i &\overset{(a)}{=} \mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \mathrm{E}_{\varepsilon'} \varepsilon'_i) v_i \\
&\overset{(b)}{=} \mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \ \mathrm{E}_{\varepsilon'} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon'_i) v_i \\
&\overset{(c)}{\leq} \mathrm{E}_{\varepsilon} \, \mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon'_i) v_i.
\end{aligned}$$

Why do these steps work?

(a) $\mathrm{E}_{\varepsilon'} \varepsilon'_i = 0.$

(b) Expectation is linear.

Step 1

We bound our maximum in terms of one involving symmetric noise.
We'll work with an *independent copy* ε' of our noise vector ε.

$$\begin{aligned}
\mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} \varepsilon_i v_i &\overset{(a)}{=} \mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \mathrm{E}_{\varepsilon'}\, \varepsilon'_i) v_i \\
&\overset{(b)}{=} \mathrm{E}_{\varepsilon} \max_{v \in \mathcal{V}}\ \mathrm{E}_{\varepsilon'} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon'_i) v_i \\
&\overset{(c)}{\leq} \mathrm{E}_{\varepsilon}\, \mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon'_i) v_i.
\end{aligned}$$

Why do these steps work?

(a) $\mathrm{E}_{\varepsilon'}\, \varepsilon'_i = 0.$

(b) Expectation is linear.

(c) Maximizing the average gives us something smaller than averaging the maxima.

Step 1

We bound our maximum in terms of one involving symmetric noise.
We'll work with an *independent copy* ε' of our noise vector ε.

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

Why do these steps work?

(a) $\mathrm{E}_{\varepsilon'}\, \varepsilon'_i = 0$.

(b) Expectation is linear.

(c) Maximizing the average gives us something smaller than averaging the maxima.

- In (c), we choose the maximizing $v \in \mathcal{V}$ *for each* ε'.

Step 1

We bound our maximum in terms of one involving symmetric noise.
We'll work with an *independent copy* ε' of our noise vector ε.

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

Why do these steps work?

(a) $\mathrm{E}_{\varepsilon'} \varepsilon'_i = 0.$

(b) Expectation is linear.

(c) Maximizing the average gives us something smaller than averaging the maxima.

- In (c), we choose the maximizing $v \in \mathcal{V}$ *for each* ε'.
- If we wanted to choose the same one each time, like we do in (b), we could.

Step 2

We introduce independent random signs $s_i = \pm 1$ w.p. $1/2$, changing nothing.

$$\mathrm{E}_{\varepsilon}\,\mathrm{E}_{\varepsilon'}\max_{v\in\mathcal{V}}\sum_{i=1}^{n}(\varepsilon_i-\varepsilon'_i)v_i=\mathrm{E}_{s}\,\mathrm{E}_{\varepsilon}\,\mathrm{E}_{\varepsilon'}\max_{v\in\mathcal{V}}\sum_{i=1}^{n}s_i(\varepsilon_i-\varepsilon'_i)v_i.$$

Why does this change nothing?

Step 2

We introduce independent random signs $s_i = \pm 1$ w.p. $1/2$, changing nothing.

$$\mathrm{E}_{\varepsilon}\,\mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} (\varepsilon_i - \varepsilon_i') v_i = \mathrm{E}_{s}\,\mathrm{E}_{\varepsilon}\,\mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^{n} s_i (\varepsilon_i - \varepsilon_i') v_i.$$

Why does this change nothing?

- Because the inner mean $(\mathrm{E}_{\varepsilon}\,\mathrm{E}_{\varepsilon'})$ doesn't depend on the signs s_i.
- That's because ε_i and ε_i' have the same distribution.
- And this implies $(\varepsilon_i - \varepsilon_i')$ and $(\varepsilon_i' - \varepsilon) = -(\varepsilon_i - \varepsilon_i')$ do, too.

Step 3

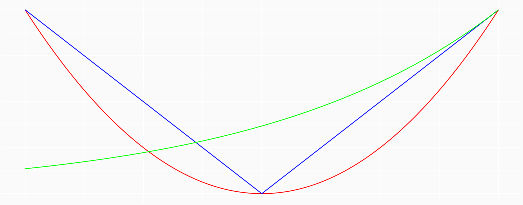
We swap the order of our averages and think about the inner average as a *function* of our vector of symmetric noise.

$$\mathrm{E}_s \mathrm{E}_\varepsilon \mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon_i')v_i = \mathrm{E}_\varepsilon \mathrm{E}_{\varepsilon'} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon_i')v_i$$

$$= \mathrm{E}_\varepsilon \mathrm{E}_{\varepsilon'} f(\varepsilon - \varepsilon') \quad \text{for} \quad f(u) = \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i u_i v_i.$$

This function f is convex.

What does that mean? These, for example, are all convex.



$$f\{(1-\lambda)a + \lambda b\} \le (1-\lambda)f(a) + \lambda f(b) \quad \text{for} \quad \lambda \in [0,1].$$ That's Convexity

Step 3

We swap the order of our averages and think about the inner average as a *function* of our vector of symmetric noise.

$$\mathrm{E}_s \mathrm{E}_\varepsilon \mathrm{E}_{\varepsilon'} \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon_i')v_i = \mathrm{E}_\varepsilon \mathrm{E}_{\varepsilon'} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon_i')v_i$$

$$= \mathrm{E}_\varepsilon \mathrm{E}_{\varepsilon'} f(\varepsilon - \varepsilon') \quad \text{for} \quad f(u) = \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i u_i v_i.$$

This function f is convex.

How do we know? Maximizing each term is better than maximizing their sum.

$$\begin{aligned} f\{(1-\lambda)a + \lambda b\} &= \mathrm{E}_s \max_{v \in \mathcal{V}} \left\{ (1-\lambda) \sum_{i=1}^n s_i a_i v_i + \lambda \sum_{i=1}^n s_i b_i v_i \right\} \\ &\le \mathrm{E}_s \left\{ \max_{v \in \mathcal{V}} (1-\lambda) \sum_{i=1}^n s_i a_i v_i + \max_{v \in \mathcal{V}} \lambda \sum_{i=1}^n s_i b_i v_i \right\} \\ &= (1-\lambda) \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i a_i v_i + \lambda \mathrm{E}_s \max_{v \in \mathcal{V}} \lambda \sum_{i=1}^n s_i b_i v_i \\ &= (1-\lambda) f(a) + \lambda f(b). \end{aligned}$$

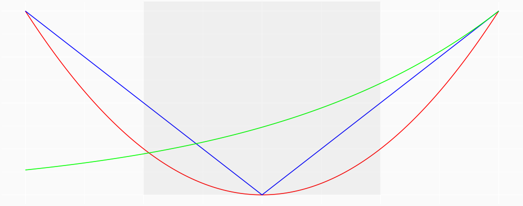
Step 3

We swap the order of our averages and think about the inner average as a *function* of our vector of symmetric noise.

$$
\begin{aligned}
\mathrm{E}_s\,\mathrm{E}_\varepsilon\,\mathrm{E}_{\varepsilon'} \max_{v\in\mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon'_i)v_i &= \mathrm{E}_\varepsilon\,\mathrm{E}_{\varepsilon'}\,\mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon'_i)v_i \\
&= \mathrm{E}_\varepsilon\,\mathrm{E}_{\varepsilon'}\, f(\varepsilon - \varepsilon') \quad \text{for} \quad f(u) = \mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^n s_i u_i v_i.
\end{aligned}
$$

This function f is convex.

Why does this matter? The max of a convex function over a cube occurs at a corner.



What cube?

Step 4

The vector of symmetric noise, $\varepsilon - \varepsilon'$, is in the *unit cube* $[-1,1]^n$.

$$\varepsilon_i - \varepsilon_i' = \begin{cases} 0 & \text{when } \varepsilon_i = \varepsilon_i' \\ +1 & \text{when } \varepsilon_i = 1 - \mu(X_i),\ \varepsilon_i' = \mu(X_i) \\ -1 & \text{when } \varepsilon_i = \mu(X_i),\ \varepsilon_i' = 1 - \mu(X_i). \end{cases}$$

The average over this random vector is bounded by the maximum over the cube it's in.

$$\begin{aligned} \mathrm{E}_\varepsilon \, \mathrm{E}_{\varepsilon'} \, \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i(\varepsilon_i - \varepsilon_i')v_i &\le \max_{u \in [-1,1]^n} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^n s_i u_i v_i \\ &= \max_{u \in [-1,1]^n} f(u) \qquad \text{max over the cube} \\ &= \max_{u \in \{-1,1\}^n} f(u) \qquad \text{max over its corners} \end{aligned}$$

Step 5

We characterize this maximum over corners. Remember what f is.

$$\max_{u\in\{-1,1\}^n} f(u) = \max_{u\in\{-1,1\}^n} \mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i u_i v_i$$

$$= \mathrm{E}_s \max_{v\in\mathcal{V}} \sum_{i=1}^{n} s_i v_i.$$

Why?

Hint. What's the distribution of s_i? And $s_i u_i$ for $u_i \in \{-1, 1\}$?

Step 5

We characterize this maximum over corners. Remember what f is.

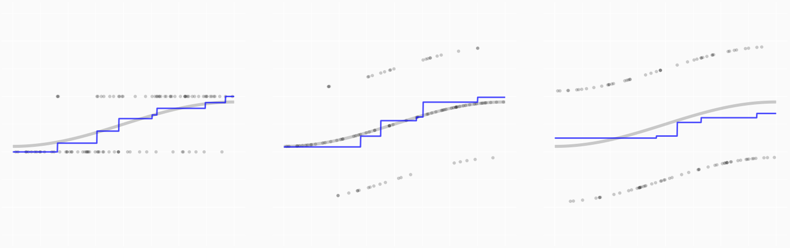
$$\begin{aligned}\max_{u \in \{-1,1\}^n} f(u) &= \max_{u \in \{-1,1\}^n} \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^{n} s_i u_i v_i \\ &= \mathrm{E}_s \max_{v \in \mathcal{V}} \sum_{i=1}^{n} s_i v_i.\end{aligned}$$

Why?

Hint. What's the distribution of s_i? And $s_i u_i$ for $u_i \in \{-1, 1\}$?

- For $u_i \in \{-1, 1\}$, the distributions of u_i and $s_i u_i$ are the same.
- So the distribution of the sum, and its maximum, are the same at every corner u.
- Including the vector of all ones $u = (1, 1, \ldots, 1)$.

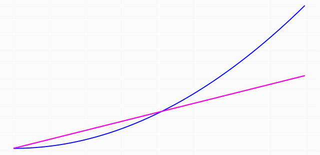
Summary



classification noise width $\leq$ symmetrized classification noise width $\leq$ random sign width

This means probabilistic classification is *easier* than regression
with random sign noise. Or, at least, that we get a better bound.

$$\frac{s^2}{2} \geq \mathrm{w}_s(\mathcal{M}_s) \quad \text{and} \quad \mathrm{w}_s(\mathcal{M}_s) \geq \mathrm{w}_\varepsilon(\mathcal{M}_s) \quad \Longrightarrow \quad \frac{s^2}{2} \geq \mathrm{w}_\varepsilon(\mathcal{M}_s)$$



Terminology

People call random sign width, or something like it, *Rademacher Complexity*.

$$\underset{\mathrm{w}_s(\mathcal{V})}{\text{Rademacher Complexity}(\mathcal{V})} = \mathrm{E}\max_{v \in \mathcal{V}} \langle s, v \rangle_{L_2(\mathrm{P_n})} \quad \text{for} \quad \text{i.i.d. } s_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$$

$$\text{or maybe } = \mathrm{E}\max_{v \in \mathcal{V}} \left| \langle s, v \rangle_{L_2(\mathrm{P_n})} \right|$$

- This second definition is the same if $\mathcal{V}$ is symmetric, i.e. $v \in \mathcal{V} \implies -v \in \mathcal{V}$.
- Otherwise, it can be a little bigger.
  - At most $2\times$ bigger. Prove it!
  - Use the bound $\max a, b \le a + b$ and the symmetry of s's distribution.

Non-Gaussian Noise

The General Case

Symmetrization and Contraction: Examples

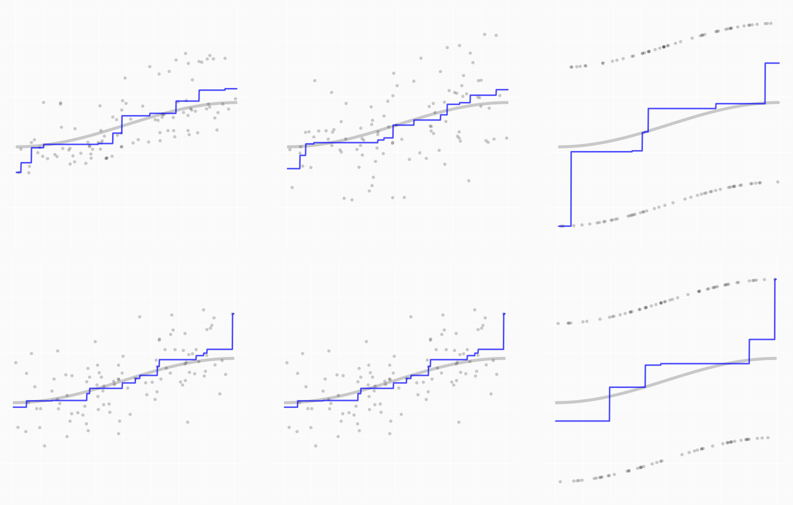


Figure 12: real noise → symmetrized noise → scaled sign noise

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 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 ×.

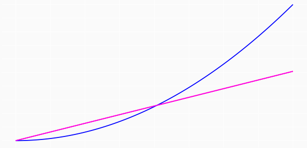
Contraction

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

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

- 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 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.

Implications for Regression



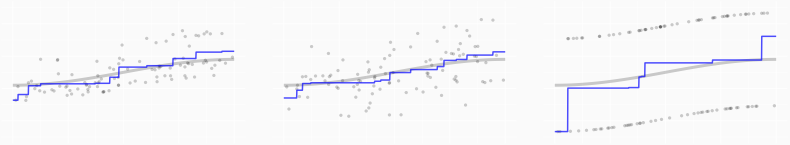
$$\mathrm{w}_\varepsilon(\mathcal{V}) \le \mathrm{E}\|\varepsilon_i - \varepsilon_i'\|_\infty\, \mathrm{w}_s(\mathcal{V}) \le 2\,\mathrm{E}\|\varepsilon_i\|_\infty\, \mathrm{w}_s(\mathcal{V})$$

Regression with arbitrary independent noise, i.e.

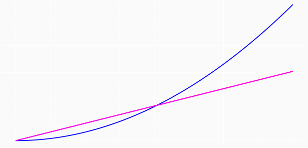
$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \varepsilon_1 \ldots \varepsilon_n \text{ are independent },$$

is no harder than with scaled-up random sign noise, i.e.

$$Y_i = \mu(X_i) + M s_i \quad \text{for} \quad M = \mathrm{E}\|\varepsilon_i - \varepsilon_i'\|_\infty \quad \text{and} \quad s_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}.$$



The Symmetric Case



$$\mathrm{w}_{\varepsilon}(\mathcal{V}) \le \mathrm{E}\|\varepsilon_i\|_{\infty}\, \mathrm{w}_s(\mathcal{V})$$

Regression with arbitrary independent *symmetric* noise, i.e.

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \varepsilon_1 \ldots \varepsilon_n \text{ are independent with } \varepsilon_i \overset{dist}{=} -\varepsilon_i,$$

is no harder than with scaled-up random sign noise, i.e.

$$Y_i = \mu(X_i) + Ms_i \quad \text{for}^2 \quad M = \mathrm{E}\|\varepsilon_i\|_{\infty} \quad \text{and} \quad s_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}.$$

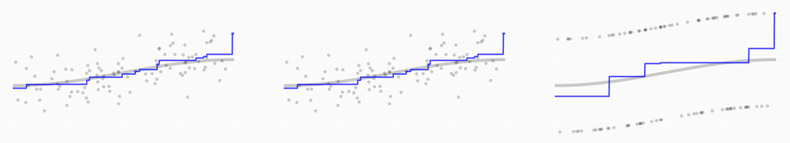
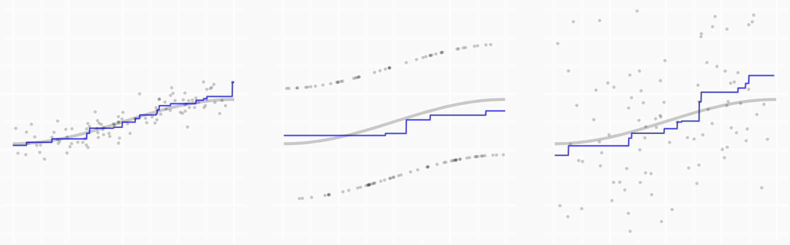


Figure 13: real noise → symmetrized noise → scaled sign noise

[2] $M = \mathrm{E}\|\varepsilon_i\|_{\infty} \le 2\sigma\sqrt{2\log(2n)}$ for $\varepsilon_i \sim N(0,\sigma^2)$. See Appendix B of the Gaussian Width Homework.

Non-Gaussian Noise

Comparison to the Gaussian Case



- So far, we've bounded arbitrary-noise width in terms of random-sign width.
- But often, it's easier to understand gaussian width. That's good enough.[3]

$$\underbrace{\frac{1}{2\sqrt{\log(2n)}}}_{\approx .2 \text{ for } n=100} \mathrm{w}_g(\mathcal{V}) \le \mathrm{w}_s(\mathcal{V}) \le \underbrace{\sqrt{\frac{\pi}{2}}}_{\approx 1.25} \mathrm{w}_g(\mathcal{V})$$

  - We just saw it can't be that much bigger than random-sign width.
  - And we can show it's at least $4/5$ as big.

$$\mathrm{E}\max_{v\in\mathcal{V}}\sum_{i=1}^{n} g_i v_i = \mathrm{E}_s\,\mathrm{E}_g \max_{v\in\mathcal{V}}\sum_{i=1}^{n}|g_i| s_i v_i \ge \mathrm{E}_s \max_{v\in\mathcal{V}}\sum_{i=1}^{n} \underbrace{\mathrm{E}_g|g_i|}_{=\sqrt{\frac{2}{\pi}}} s_i v_i.$$

[3] We can show $.125\,\mathrm{w}_g(\mathcal{V}) \le \mathrm{w}_s(\mathcal{V}) \le 1.25\,\mathrm{w}_g(\mathcal{V})$ for $n \le 10$ trillion by bounding $\mathrm{E}\|g\|_\infty$ more carefully.

Comparison in Steps

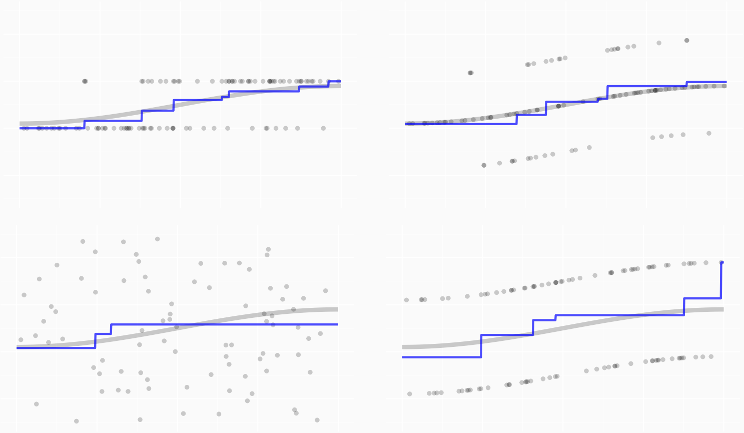


Figure 14: real noise → symmetrized noise ↓ scaled sign noise ← scaled gaussian noise

$$\mathrm{w}_{\varepsilon}(\mathcal{V}) \le \mathrm{w}_{\varepsilon-\varepsilon'}(\mathcal{V}) \le \underset{\le 2\,\mathrm{E}\|\varepsilon\|_\infty}{\mathrm{E}\|\varepsilon-\varepsilon'\|_\infty} \; \mathrm{w}_{s}(\mathcal{V}) \le \underset{\le\sqrt{2\pi}\approx 2.5\times\mathrm{E}\|\varepsilon\|_\infty}{\sqrt{\frac{\pi}{2}}\,\mathrm{E}\|\varepsilon-\varepsilon'\|_\infty} \; \mathrm{w}_{g}(\mathcal{V})$$

Implications for Regression

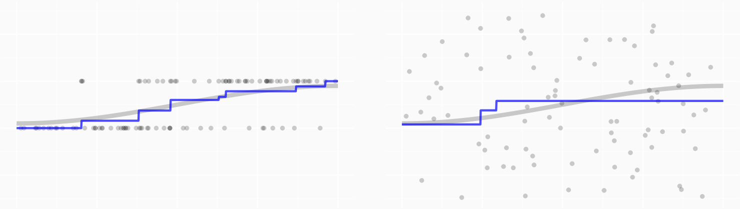


Figure 15: real noise → scaled gaussian noise

For any noise vector ε with independent components ε_i,

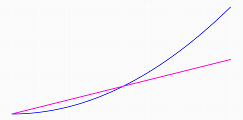
$$\mathrm{w}_\varepsilon(\mathcal{V}) \le 2\,\mathrm{E}\|\varepsilon\|_\infty \cdot \mathrm{w}_s(\mathcal{V}) \le \sqrt{2\pi}\,\mathrm{E}\|\varepsilon\|_\infty \cdot \mathrm{w}_g(\mathcal{V}).$$

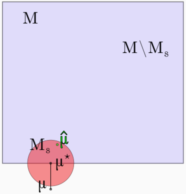
- We can bound the width w_ε in terms of
  1. random-sign width
  2. the maximum absolute value of ε's components.
- And we can bound random-sign width in terms of gaussian width.

This means we don't have to bound a million different kinds of widths for each model. We can bound random-sign width or gaussian width. Whichever is easier.

Sampling

Starting Point





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

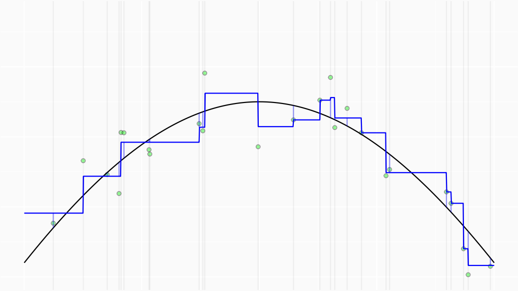
  - This is a one-number summary of the gaussian width of neighborhoods ...
  - ...of the model's best approximation to the signal. It's the summary that matters.
- It depends on the noise's size through the expected maximum square.

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

What does this tell us?

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

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



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

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

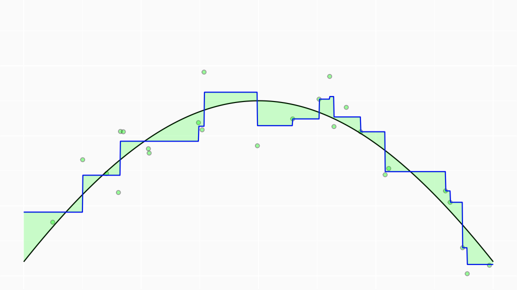
What Population Mean Squared Error Is

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

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

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

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

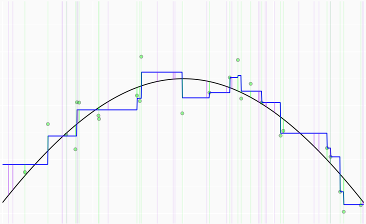


Why we care about Population Mean Squared Error: Generalization

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

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

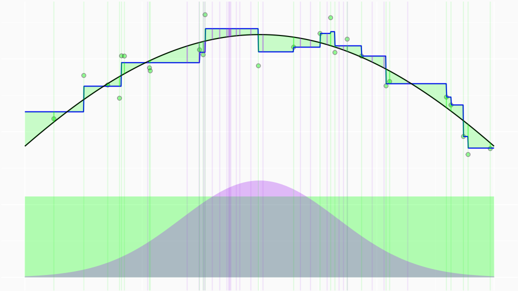
This can be a bit different from accuracy on our original sample $X_1 \ldots X_n$.



- 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.

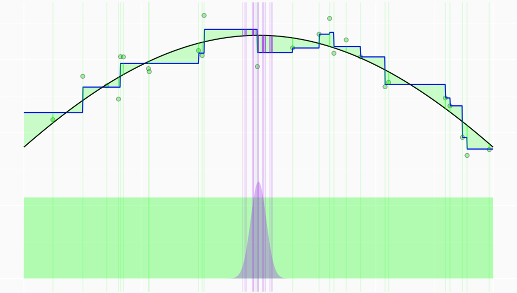


$$\frac{1}{n'}\sum_{i=1}^{n'}\big\{\hat\mu(X_i')-\mu(X_i')\big\}^2 \approx \|\hat\mu-\mu\|_{L_2(\mathrm{Q})}^2 = \int\{\hat\mu(x)-\mu(x)\}^2\,\frac{q(x)}{p(x)}\,p(x)\,dx$$

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

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'.



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

Sketch: What Our Sample MSE Bounds Tell Us About Population MSE.

Starting Point. Suppose $\hat{\mu}$ is in some neighborhood of $\mu^\star$.

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

We **bound the maximum** difference between population and sample MSE on that neighborhood.

$$\hat{\mu} \in \mathcal{M}_s \implies \|\hat{\mu} - \mu^\star\|^2_{L_2(\mathrm{P})} - \|\hat{\mu} - \mu^\star\|^2_{L_2(\mathrm{P_n})} \le Z := \max_{f \in \mathcal{M}_s - \mu^\star} \|f\|^2_{L_2(\mathrm{P})} - \|f\|^2_{L_2(\mathrm{P_n})}.$$

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

$$Z \text{ satisfies } Z \le \mathrm{E}\, Z + \sqrt{\frac{\mathrm{Var}(Z)}{\delta n}} \quad \text{w.p. } 1 - \delta$$

We use **symmetrization** to get a bound in terms of random-sign width.

(a) Write population MSE as an expectation over 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.

$$\begin{aligned}
n \times \mathrm{E}\, Z &\overset{(a)}{=} \mathrm{E}_X \max_{f \in \mathcal{M}_s - \mu^\star} \mathrm{E}_{X'} \sum_{i=1}^n f(X_i')^2 - \sum_{i=1}^n f(X_i)^2 \\
&\overset{(b)}{\le} \mathrm{E}_X\, \mathrm{E}_{X'}\, \mathrm{E}_s \max_{f \in \mathcal{M}_s - \mu^\star} \sum_{i=1}^n s_i \Big\{ f(X_i')^2 - f(X_i)^2 \Big\} \\
&\overset{(c)}{\le} \mathrm{E}_X\, \mathrm{E}_{X'}\, \mathrm{E}_s \max_{f', f \in \mathcal{M}_s - \mu^\star} \sum_{i=1}^n s_i f'(X_i')^2 + (-s_i) f(X_i) \\
&= 2\, \mathrm{E}_X\, \mathrm{E}_s \max_{f \in \mathcal{M}_s - \mu^\star} \sum_{i=1}^n s_i f(X_i)^2
\end{aligned}$$

Contracting Out Lipschitz Functions

This expectation bound is $2\times$ the expected random-sign width of the *squares* of the functions in our neighborhood.

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

$$\le 4\,\mathrm{E}_X \max_{f\in\mathcal{M}_s-\mu^\star}\|f\|_{L_\infty(\mathrm{P_n})}\ \underbrace{\mathrm{E}_s \max_{f\in\mathcal{M}_s-\mu^\star}\sum_{i=1}^n s_i f(X_i)}_{n\times \mathrm{w}_s(\mathcal{M}_s-\mu^\star)}$$

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

Lemma (The Lipschitz Contraction Lemma)

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

Application. Taking $\psi_i(x)=x^2$ for all i and $\mathcal{V}=\{f(X_1)\ldots f(X_n)\ :\ f\in\mathcal{M}_s-\mu^\star\}$,

$$\mathrm{E}_s \max_{f\in\mathcal{M}_s-\mu^\star}\sum_{i=1}^n s_i\ \underbrace{f(X_i)^2}_{\psi_i\{f(X_i)\}} \le L \max_{f\in\mathcal{M}_s-\mu^\star}\sum_{i=1}^n s_i f(X_i)$$

$$\text{for}\qquad L=\max_i \max_{f\in\mathcal{M}_s-\mu^\star}|\psi_i'\{f(X_i)\}| = \max_i \max_{f\in\mathcal{M}_s-\mu^\star}|2f(X_i)|.$$

Implications

If the curves in $\mathcal{M}_s - \mu^\star$ are all bounded by a constant B, the difference between population and sample MSE is at most …

$$\max_{m\in\mathcal{M}_s} \|m-\mu^\star\|^2_{L_2(\mathrm{P})} - \|m-\mu^\star\|^2_{L_2(\mathrm{P_n})}$$

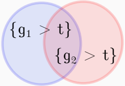
$$\le \mathrm{E}\,Z + \sqrt{\frac{\mathrm{Var}(Z)}{\delta n}} \quad \text{w.p. } 1-\delta$$

$$\le 4B\,\mathrm{E}_X\,\mathrm{w}_s(\mathcal{M}_s) + Bs \times \sqrt{\frac{2\Sigma_n}{\delta n}} \qquad \text{using Efron-Stein on } \mathrm{Var}(Z)$$

$$\le 4B\,\mathrm{w}_s(\mathcal{M}_s) + Bs \times \left(4+\sqrt{\Sigma_n}\right)\sqrt{\frac{2}{\delta n}} \qquad \text{using Efron-Stein on } \mathrm{Var}(\mathrm{w}_s(\mathcal{M}_s))$$

We will combine this with our bound on sample MSE from the previous section.

$$\|\hat\mu-\mu\| \in \mathcal{M}_{s_\delta} \quad \text{for} \quad s_\delta := 2\sqrt{\Sigma_n}\left(s+\sqrt{\frac{2}{\delta n}}\right) \quad \text{w.p. } \ 1-\delta \ \text{ if } \quad \frac{s^2}{2} \ge \mathrm{w}_s(\mathcal{M}_s)$$



By the union bound, these two bounds hold simultaneously w.p. $\ge 1-2\delta$. Consequently …

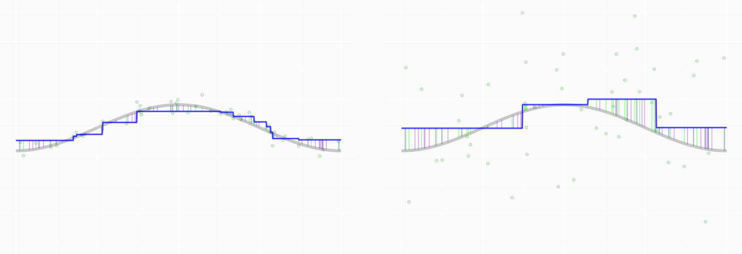
$$\|\hat\mu-\mu\|^2_{L_2(\mathrm{P})} = \|\hat\mu-\mu\|^2_{L_2(\mathrm{P_n})} + \left\{\|m-\mu\|^2_{L_2(\mathrm{P})} - \|m-\mu\|^2_{L_2(\mathrm{P_n})}\right\}$$

$$\le s_\delta^2 + \left\{4B\,\mathrm{w}_{s_\delta}(\mathcal{M}_{s_\delta}) + Bs_\delta \times \left(4+\sqrt{\Sigma_n}\right)\sqrt{\frac{2}{\delta n}}\right\}$$

$$\le s_\delta^2 + \left\{2Bs_\delta s + Bs_\delta \times \left(4+\sqrt{\Sigma_n}\right)\sqrt{\frac{2}{\delta n}}\right\} \approx \{s_\delta + Bs\}^2$$

because $\quad \mathrm{w}_{s_\delta}(\mathcal{M}_{s_\delta}) \le \frac{s_\delta}{s}\,\mathrm{w}_s(\mathcal{M}_s) \le \frac{s_\delta}{s}\,\frac{s^2}{2} = \frac{s_\delta\, s}{2}.$

Interpretation



$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P})} \le\approx 2\{B + \sqrt{\Sigma_n}\}\left(s + \sqrt{\frac{2}{\delta n}}\right) \quad \text{w.p. } \; 1 - \delta \; \text{ if } \quad \frac{s^2}{2} \ge \mathrm{w}_s(\mathcal{M}_s)$$

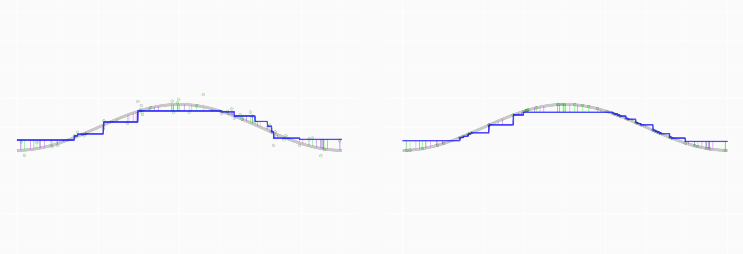
This is the bound we'd get on sample MSE with additional scaled random-sign noise.

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



$$\|\hat{\mu}-\mu\|_{L_2(\mathrm{P})} \lessapprox 2\{\sqrt{\Sigma_n}+B\}\left(s+\sqrt{\frac{2}{\delta n}}\right) \quad \text{w.p. } 1-\delta \text{ if } \quad \frac{s^2}{2} \ge \mathrm{w}_s(\mathcal{M}_s) \quad \text{and} \quad \max_{m\in\mathcal{M}_s} \|m$$

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

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

Signal Recovery is regression without any noise at all.

- 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 is still obscured by ‘sampling noise’.

References

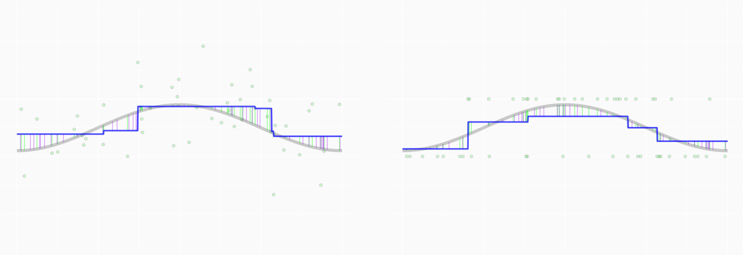
Chapter 6 of Talagrand's Upper and Lower Bounds for Stochastic Processes.

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

Appendices

Appendices

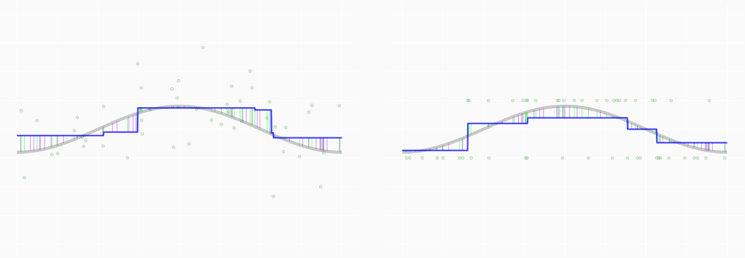
Boundedness



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

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P})} \le\approx 2\{B + \sqrt{\Sigma_n}\}\left(s + \sqrt{\frac{2}{\delta n}}\right) \quad \text{w.p. } 1 - \delta \text{ if } \quad \frac{s^2}{2} \ge \mathrm{w}_s(\mathcal{M}_s) \quad \text{and} \quad \max_{m \in \mathcal{M}_s} \|m$$

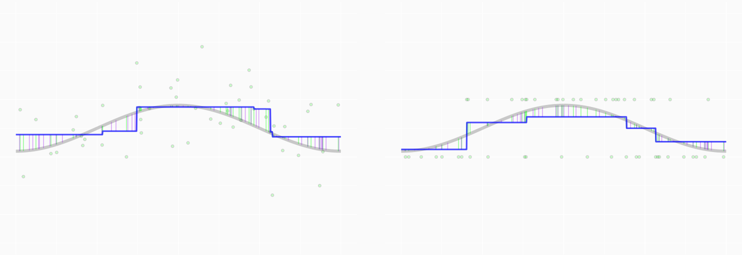
Getting a bound B can take a bit of work. There are options.



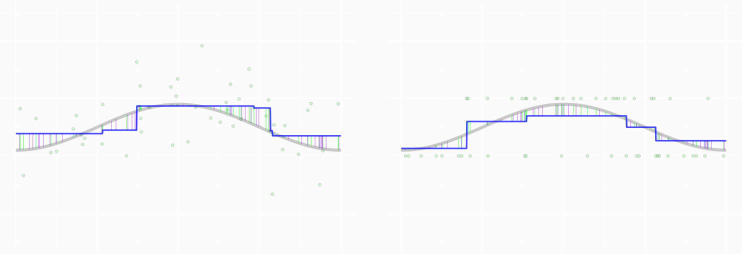
Option 1. Baking it into the Model.

$$\mathcal{M} = \{m \;:\; \|m\|_\infty \le B \quad \text{and} \quad \rho_{TV}(m) \le B\} \implies \|m - \mu^\star\|_\infty \le \|m\|_\infty + \|\mu^\star\|_\infty \le 2B$$

$$\mathcal{M} = \{m \;:\; |m(0)| = 0 \quad \text{and} \quad \rho_{TV}(m) \le B\} \implies \ldots$$



Option 2. Arguing Based on Bounded Data.



Others?

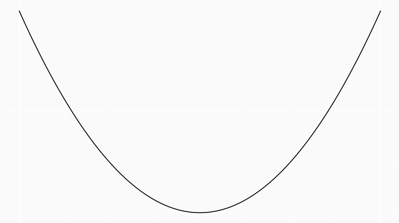
Appendices

Convex Functions Are Maximized At Extreme Points

Definition

A function f is convex if *secants* lie above the curve.

$$f\{(1-\lambda)a + \lambda b\} \leq (1-\lambda)f(a) + \lambda f(b) \quad \text{for} \quad \lambda \in [0,1]$$



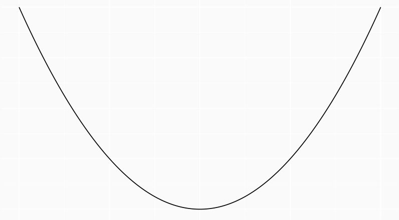
We can give this a *probabilistic interpretation* for a random variable Z_λ.

$$f(\mathrm{E}\, Z_\lambda) \leq \mathrm{E}\, f(Z_\lambda) \quad \text{where} \quad Z_\lambda =$$

Definition

A function f is convex if *secants* lie above the curve.

$$f\{(1-\lambda)a+\lambda b\} \leq (1-\lambda)f(a)+\lambda f(b) \quad \text{for} \quad \lambda \in [0,1]$$



We can give this a *probabilistic interpretation* for a random variable Z_λ.

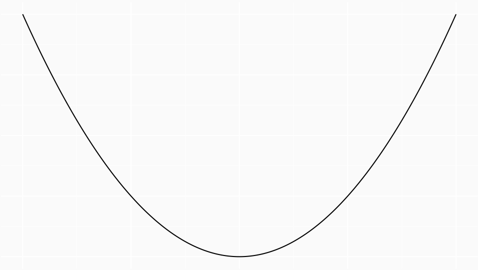
$$f(\operatorname{E} Z_\lambda) \leq \operatorname{E} f(Z_\lambda) \quad \text{where} \quad Z_\lambda = \begin{cases} a & \text{w.p. } \; 1-\lambda \\ b & \text{w.p. } \; \lambda \end{cases}$$

Jensen’s Inequality

In fact, this is true all random variables Z.
If f is convex, its mean value exceeds its value at the mean.

$$f(\mathrm{E}\,Z) \le \mathrm{E}\,f(Z)$$

That’s called Jensen’s Inequality.



You can prove it for discrete random variables via induction.

Jensen's Inequality Proof

Base case.

It's true for random variables taking on 2 values.

$$f(\lambda_1 z_1 + \lambda_2 z_2) \leq \lambda_1 f(z_1) + \lambda_2 f(z_2) \quad \text{if} \quad \lambda_1, \lambda_2 \geq 0 \quad \text{satisfy} \quad \lambda_1 + \lambda_2 = 1$$

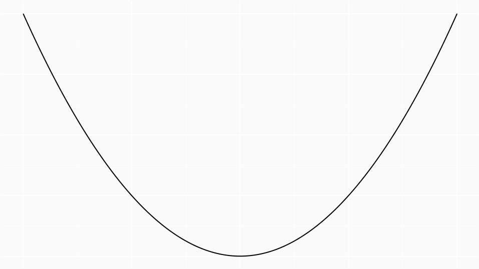
Inductive Step.

We'll show that if it's true for random variables taking on $n-1$ values, then it's also true for ones taking on n values.

$$\begin{aligned}
f\left\{\sum_{i=1}^{n} \lambda_i z_i\right\} &= f\left\{(1-\lambda_n)\left(\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} z_i\right) + \lambda_n z_n\right\} \\
&\leq (1-\lambda_n) f\left(\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} z_i\right) + \lambda_n f(z_n) \\
&\leq (1-\lambda_n) \sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} f(z_i) + \lambda_n f(z_n) \\
&= \sum_{i=1}^{n-1} \lambda_i f(z_i) + \lambda_n f(z_n)
\end{aligned}$$

Maxima of Convex Functions

Convex functions have no local maxima.



That means the maximum of a convex function over an interval occurs at an endpoint.

Proof.

$$\max_{x\in[a,b]} f(x) = \max_{\lambda\in[0,1]} \underset{f(\operatorname{E} Z_\lambda)}{f\{(1-\lambda)a+\lambda b\}} \le \max_{\lambda\in[0,1]} \underset{\operatorname{E} f(Z_\lambda)}{(1-\lambda)f(a)+\lambda f(b)} = \max\{f(a), f(b)\}$$

This is essentially true in higher dimensions as well.
We just need the right generalizations of *interval* and its *endpoints*.

Convex Polytopes

The natural generalizations a *convex polytope* and its *extreme points*.

Definitions.

A **convex polytope** is the set of all weighted averages of some set of vectors $u_1 \ldots u_K$.

$$\mathcal{U} = \left\{ \sum_i \lambda_i u_i \ : \ \lambda \in \Lambda \right\} \quad \text{where} \quad \Lambda = \left\{ \lambda \ : \ \lambda_i \geq 0 \ \text{ for all } i \ \text{ and } \ \sum_i \lambda_i = 1 \right\}$$

Its **extreme points** are the subset of these vectors that are not redundant.
That is, they're the ones we cannot write as weighted averages of the others.

Examples.

- A triangle is the set of weighted averages of its three vertices, its extreme points.
- A square is the set of weighted averages of its four vertices, its extreme points.
- A cube in $\mathbb{R}^n$ is the set of weighted averages of its 2^n vertices, its extreme points.

Maxima of Convex Functions over Polytopes

The maximum of a convex function over a convex polytope occurs at an extreme point.

Proof.

It’s more-or-less the same as the one-dimensional case.

We apply Jensen’s inequality to a *random extreme point* Z_λ.

$$\max_{u\in\mathcal{U}} f(u) = \max_{\lambda\in\Lambda} \underset{f(\mathrm{E}\, Z_\lambda)}{f\left(\sum_i \lambda_i u_i\right)} \le \max_{\lambda\in\Lambda} \underset{\mathrm{E}\, f(Z_\lambda)}{\sum_i \lambda_i f(u_i)} \le \max_i f(u_i)$$

where

$$Z_\lambda = \begin{cases} u_1 & \text{w.p. } \lambda_1 \\ \vdots & \vdots \\ u_K & \text{w.p. } \lambda_K \end{cases}$$