Machine Learning Theory

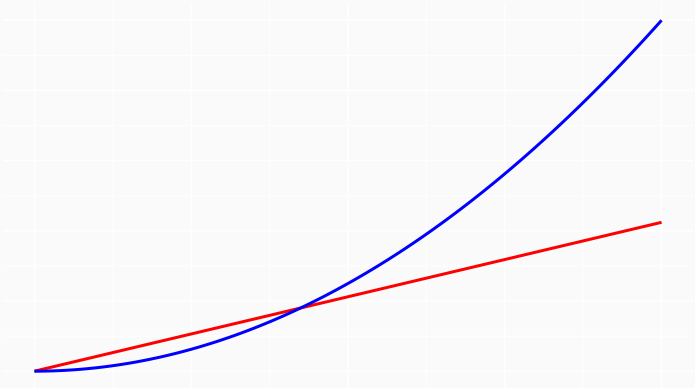
Misspecification

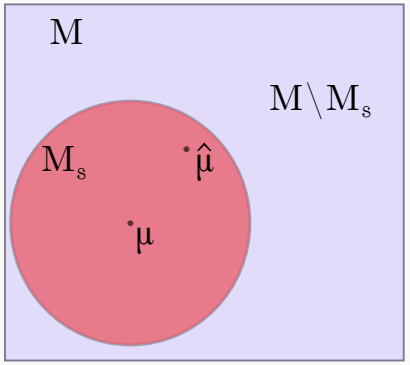
David A. Hirshberg

April 15, 2025

Emory University

Where We Left Off

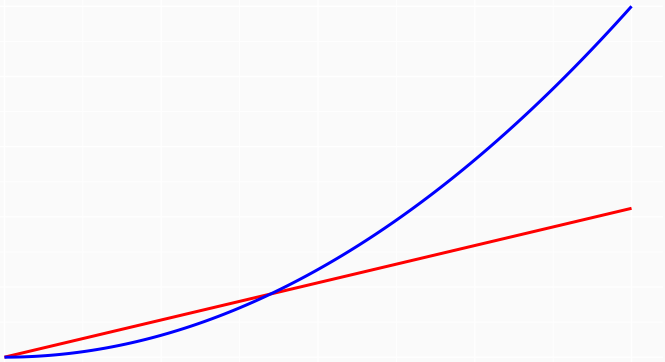


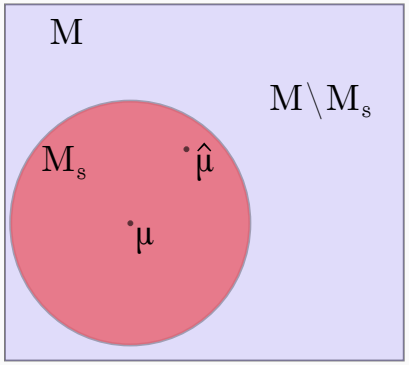


What do we know about the error of this least squares estimator $\hat{\mu}$?

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

Where We Left Off





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

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

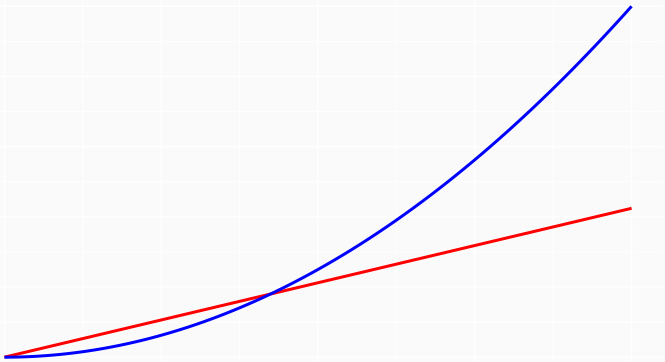
Here's what we've proven in lecture.

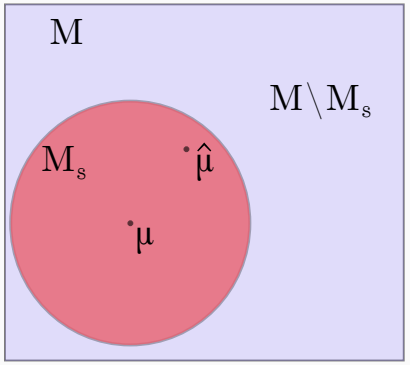
$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P_n})} < s \quad \text{w.p.} \ \ 1 - \delta \ \ \text{for} \quad \frac{s^2}{2\sigma} \geq \mathrm{w}(\mathcal{M}_s^\circ) + s\sqrt{\frac{2\Sigma_n}{\delta n}}$$

where $\Sigma_n = \sigma^2\{1 + 2\log(2n)\}$ and $\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g, \ v \rangle_{L_2(\mathrm{P_n})}$ 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 $\mu \in \mathcal{M}$

Where We Left Off





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

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

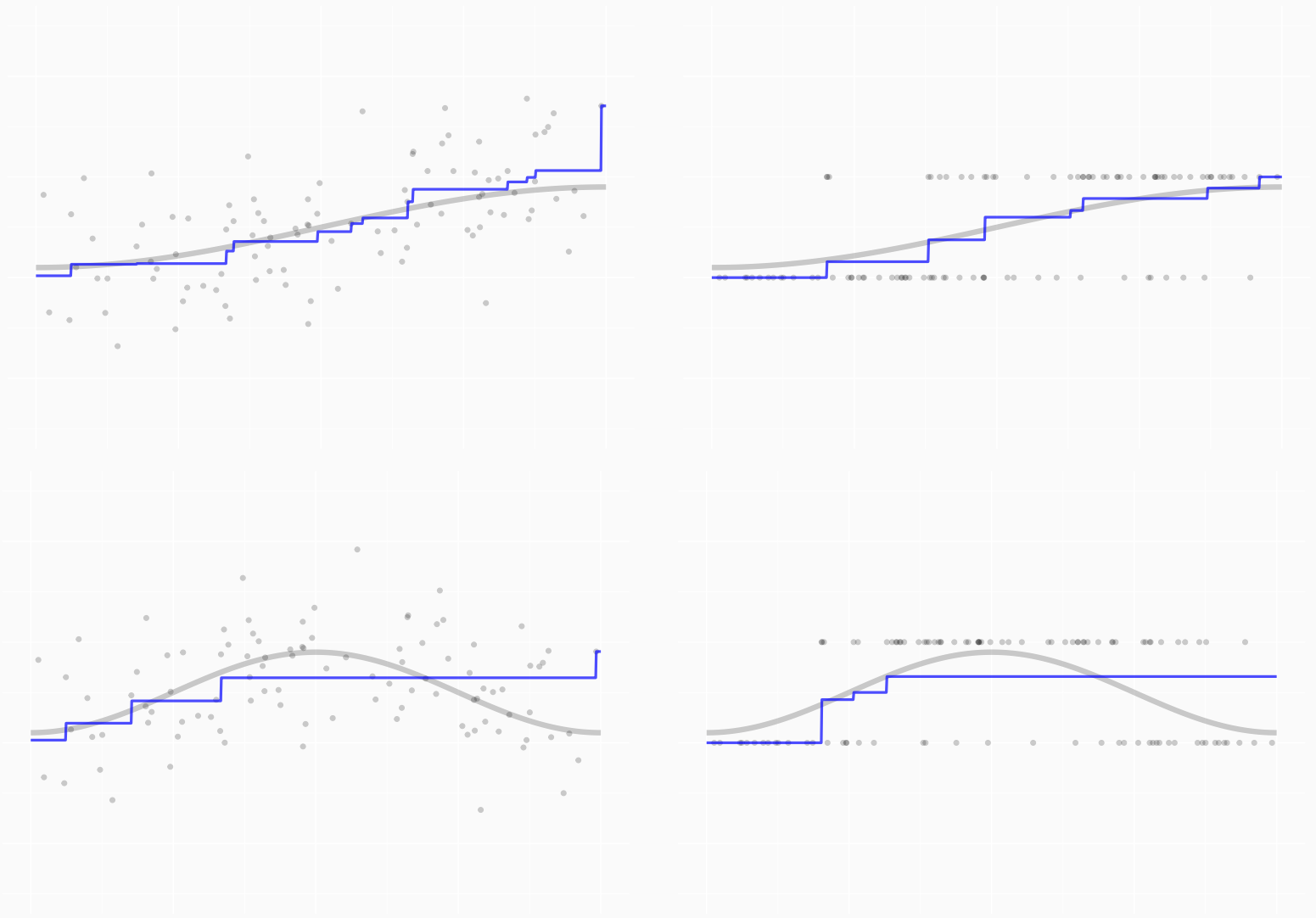
Here's a simplified version you're proving for homework.

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

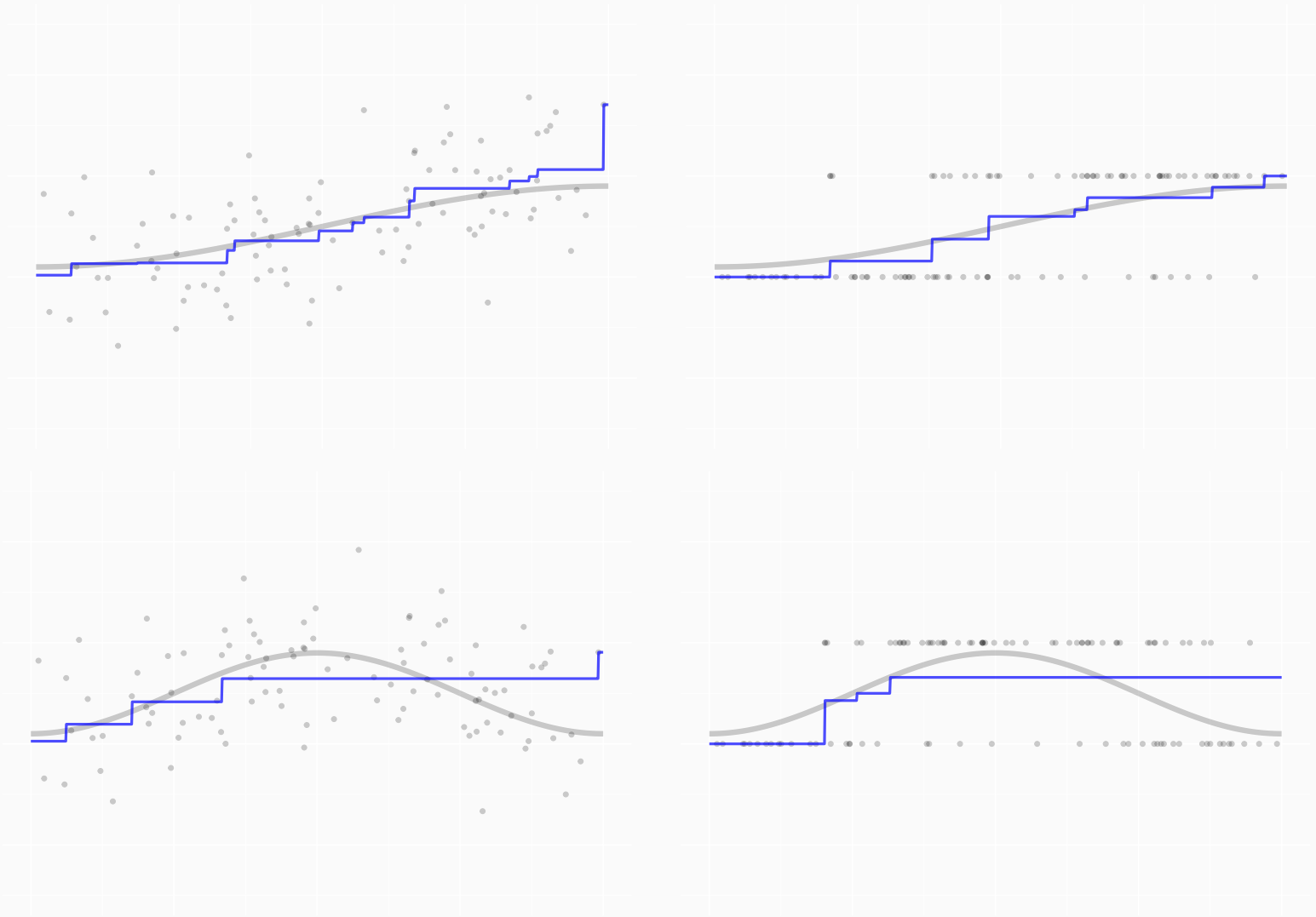
where $\Sigma_n = \sigma^2\{1 + 2\log(2n)\}$ and $\mathrm{w}(\mathcal{V}) = \mathrm{E} \max_{v \in \mathcal{V}} \langle g, \ v \rangle_{L_2(\mathrm{P_n})}$ for $g_i \overset{iid}{\sim} N(0,1)$

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?



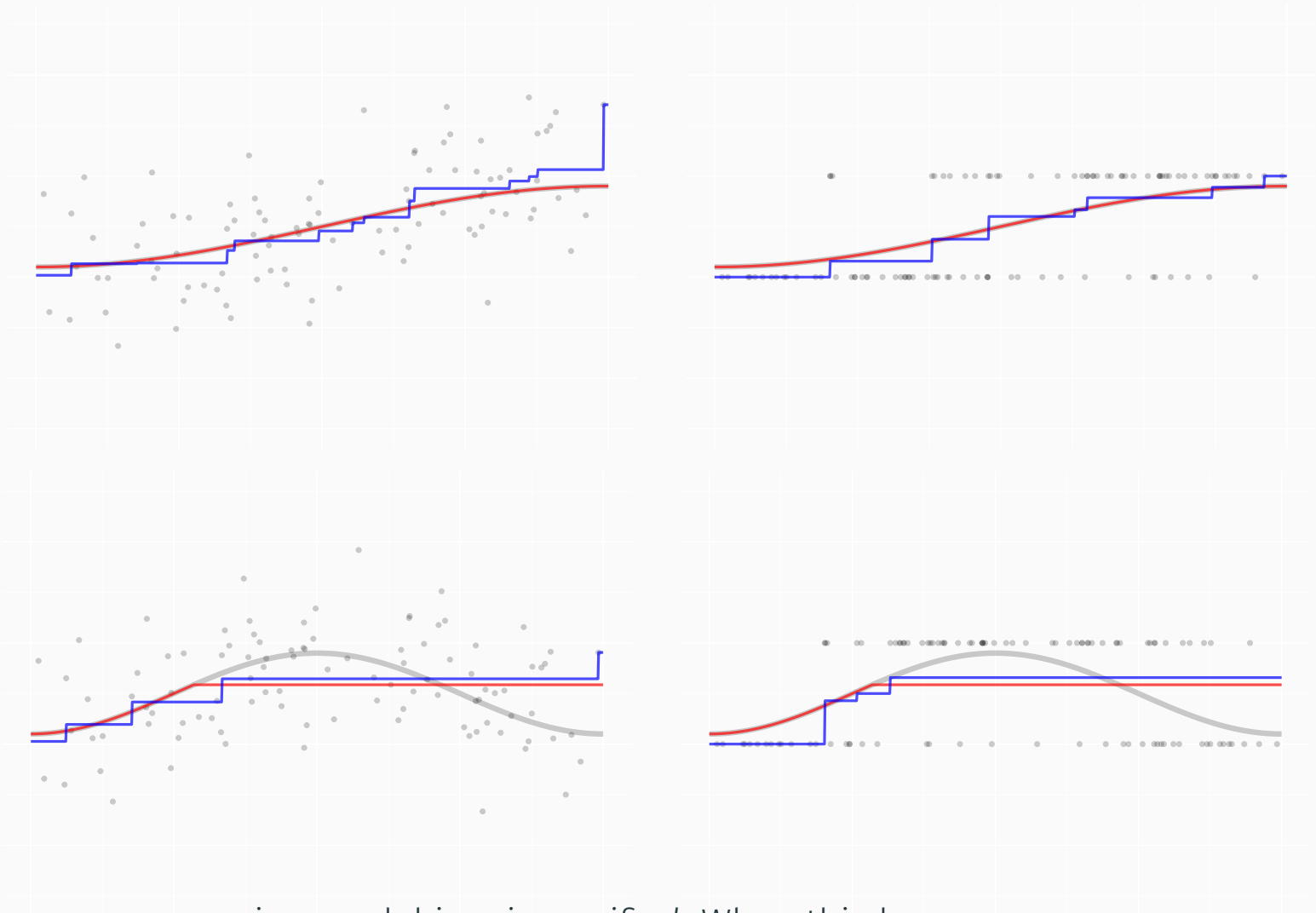
When Does This Bound Apply?



In the top-left only.

- The second column is out. We've assumed μ is in the model.
- The second row is out. We've assumed our noise is Gaussian.

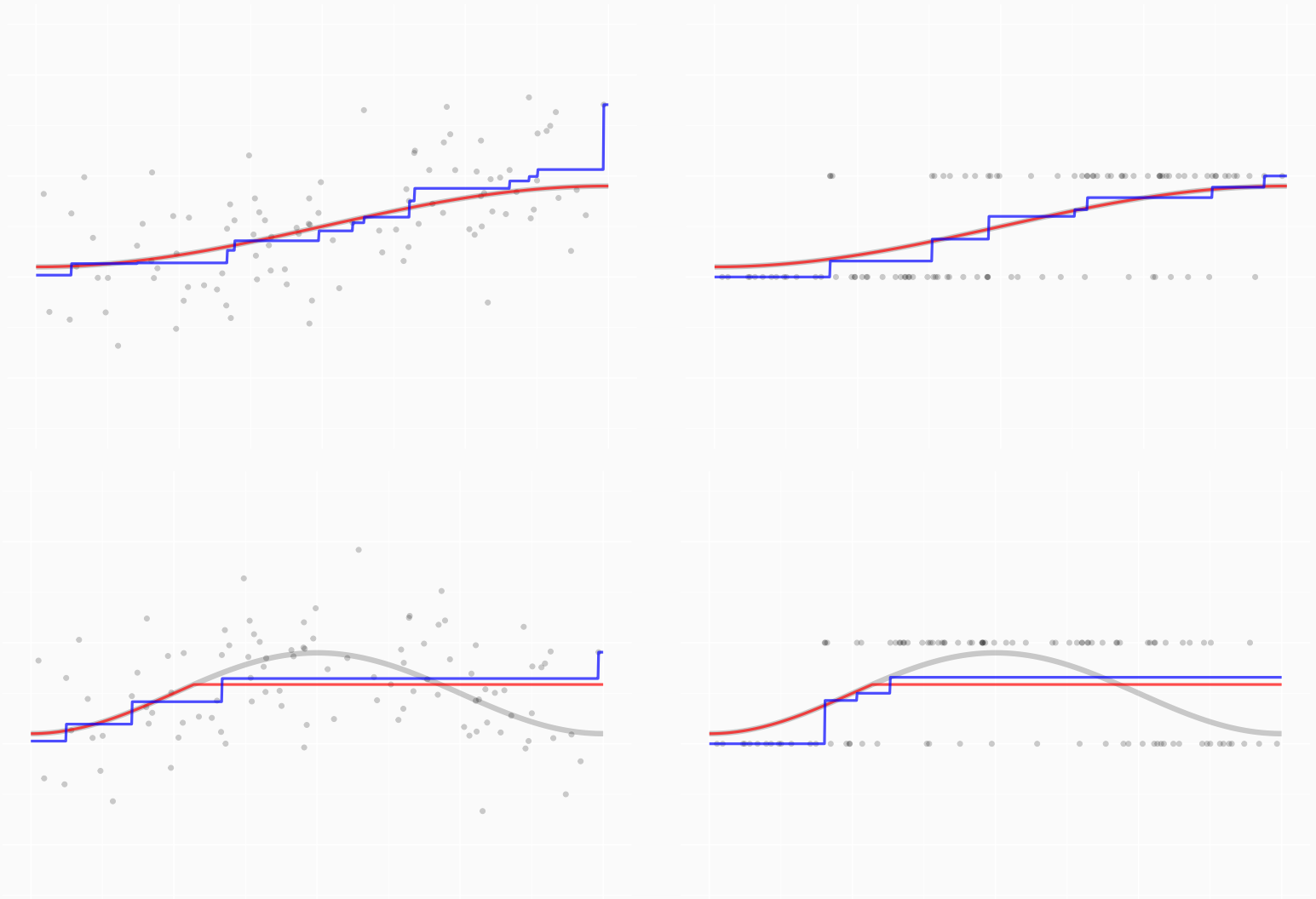
When μ isn't in the model



We say our regression model is *misspecified*. When this happens, ...

- we estimate the model's best approximation to μ. Otherwise, not much changes.
- We'll bound the distance between that and our estimator the same way we've been doing.

When our noise isn't Gaussian

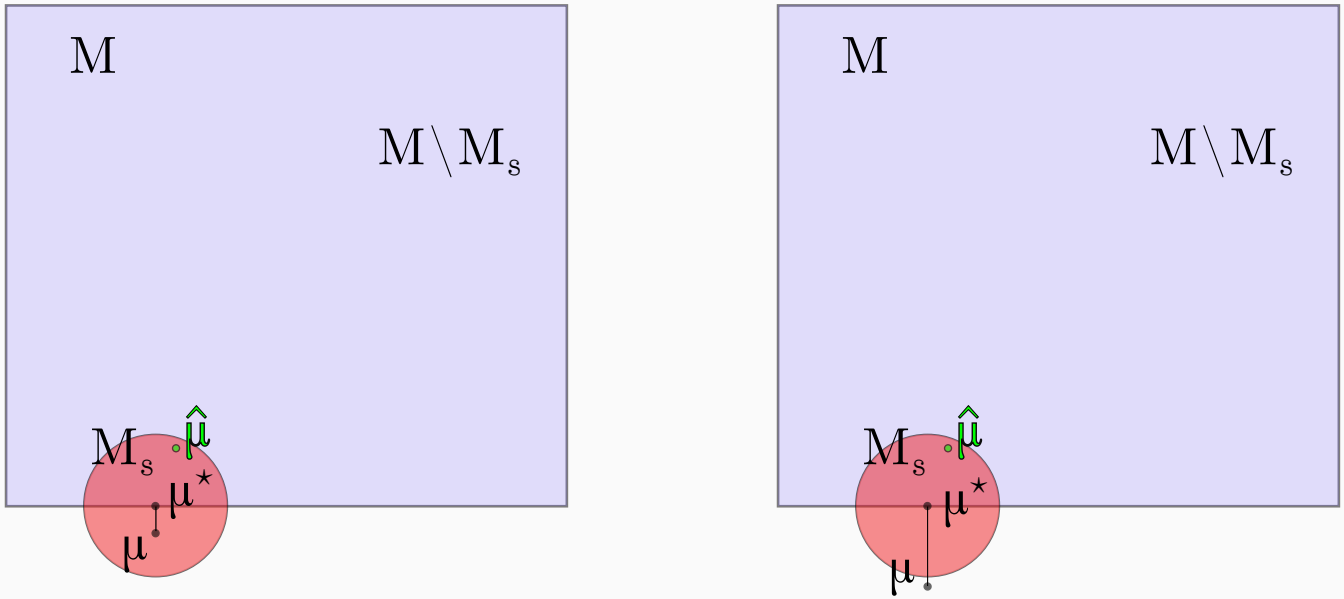


If you think of least squares as a gaussian noise thing, our noise is misspecified.

- We'll compare the difficulty of this problem to regression with gaussian noise.
- The *probabilistic classification* problem shown above is no harder than regression with gaussian noise with $\sigma = 1.25$.

Misspecification

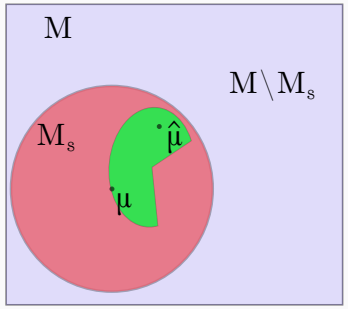
What happens when μ isn't in the model?

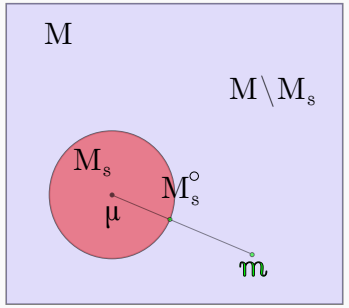


- Our error in estimating μ is bounded by a sum of two terms.
  - The critical radius s, i.e., the one satisfying $s^2/2\sigma \geq \mathrm{w}(\mathcal{M}_s^\circ) + s\sqrt{\frac{2\Sigma_n}{\delta n}}$.
  - The distance from μ to its best approximation in the model. Or really 3 times that.

  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





squared error loss

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

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.

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

- 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* $\varepsilon = Y - m$.

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

- 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.

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

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

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

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

$$\ell(m) - \ell(\mu^\star) = \|m - \mu^\star\|^2_{L_2(\mathrm{P_n})}$$ squared distance, like before;

$$- \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \{m(X_i) - \mu^\star(X_i)\}$$ a mean zero term, like before;

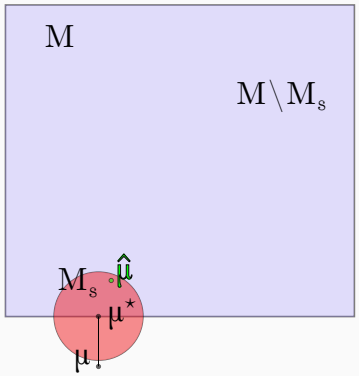
$$+ \frac{2}{n}\sum_{i=1}^{n} \{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\}$$ and something else.

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

Why?

Why.

$$\ell(m) - \ell(\mu^\star) = \|m - \mu^\star\|^2_{L_2(\mathrm{Pn})} - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \{m(X_i) - \mu^\star(X_i)\} + \frac{2}{n}\sum_{i=1}^{n}\{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\}$$



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.

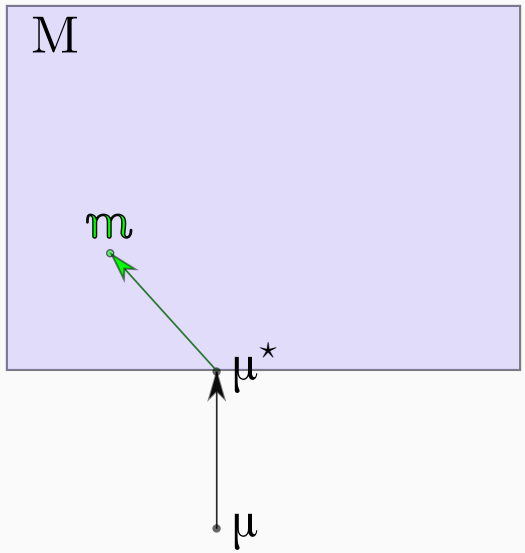
- If the loss difference is positive when we ignore a non-negative term ...
- ...then it's still positive when we don't.

$$\ell(m) - \ell(\mu^\star) > 0 \quad \text{if} \quad \|m - \mu^\star\|^2_{L_2(\mathrm{Pn})} - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \{m(X_i) - \mu^\star(X_i)\} > 0 \qquad \text{what we're used to}$$

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

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,

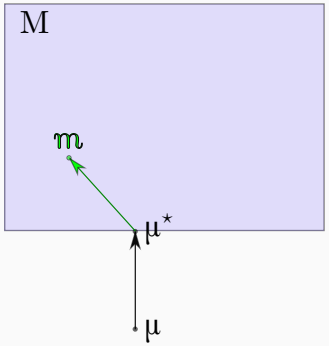
$$\mu^\star = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \|m - \mu\|^2_{L_2(\mathrm{P_n})} \quad \text{satisfies} \quad \frac{2}{n}\sum_{i=1}^{n}\{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\}$$

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

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]

$$\mu^\star = \operatorname*{argmin}_{m \in \mathcal{M}} \|m - \mu\| \quad \text{satisfies}$$

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

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

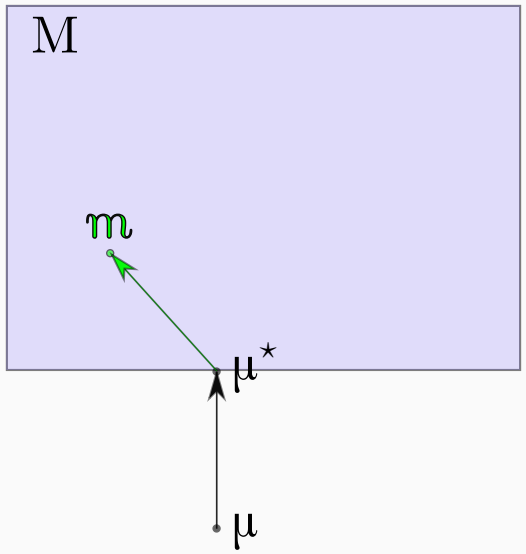
$$0 \leq \lambda \|m - \mu^\star\|^2 + 2 \langle m - \mu^\star,\ \mu^\star - \mu \rangle.$$

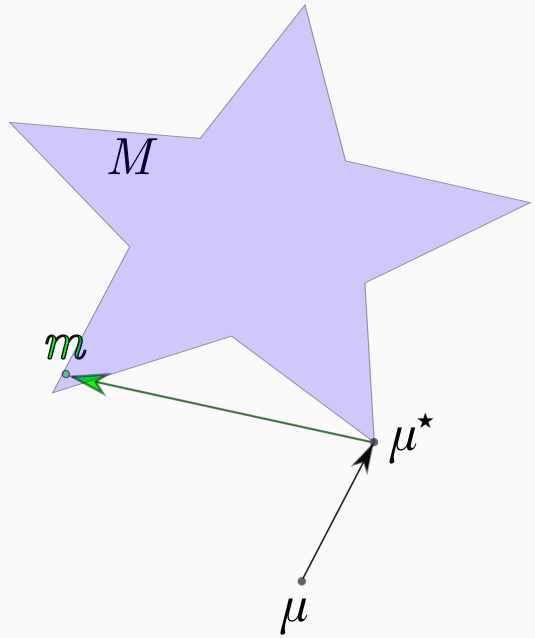
Because this holds for arbitrarily small $\lambda > 0$, it must also hold for $\lambda = 0$.

[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 $m \in \mathcal{M}$.



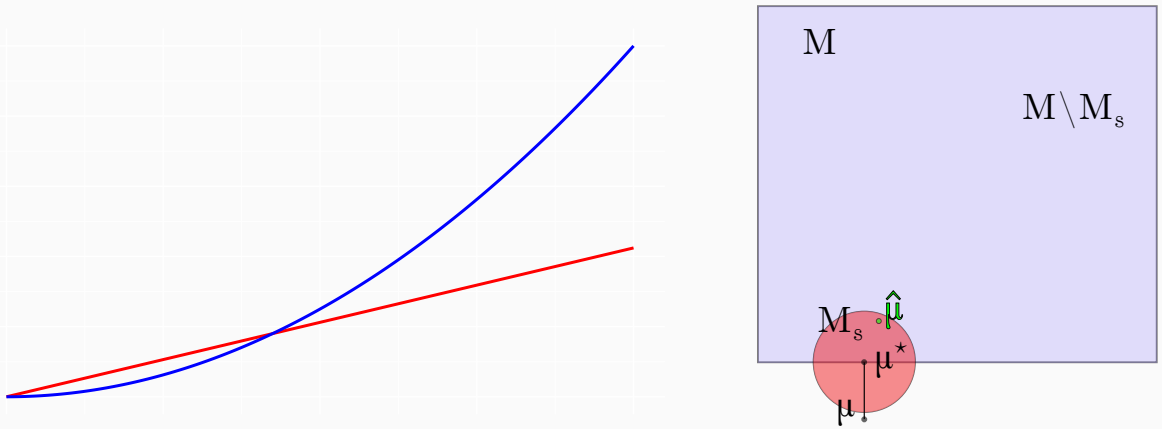


The same thing can happen *for the closest point* in a non-convex model.

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.

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

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

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

Misspecification

Examples

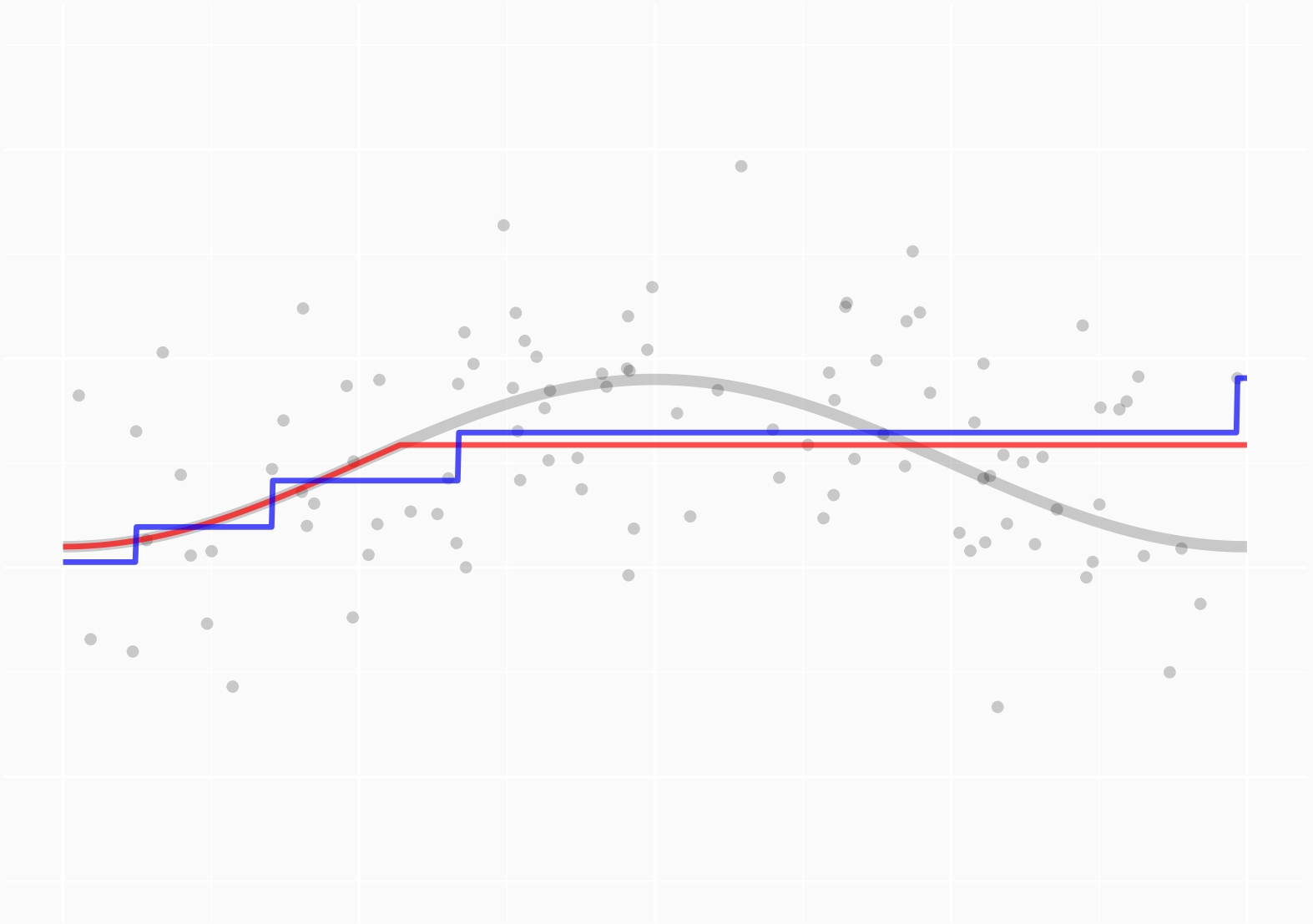


Figure 1: Increasing Curves ($n = 100$.)

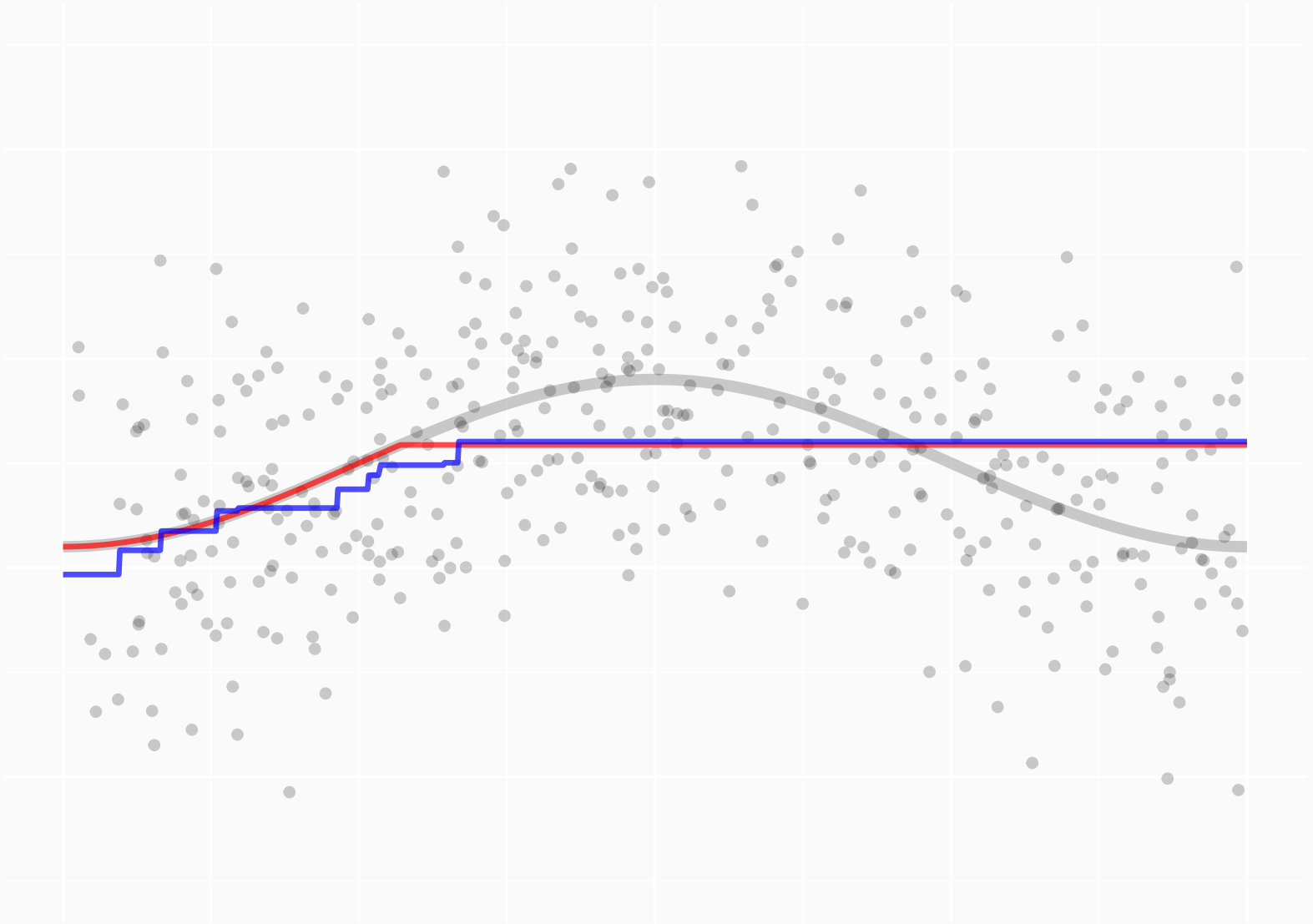


Figure 2: Increasing Curves ($n = 400$.)

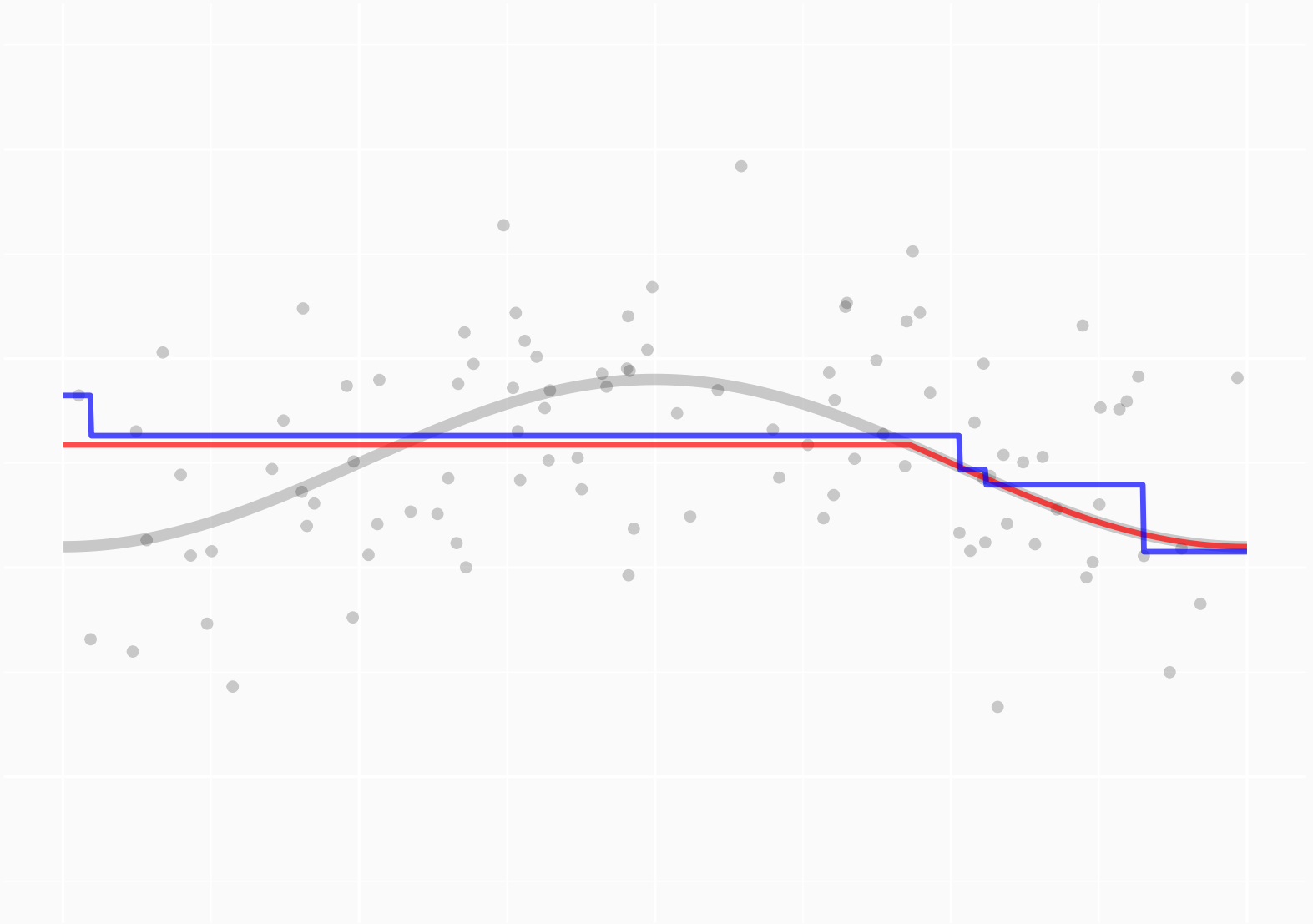


Figure 3: Decreasing Curves ($n = 100$.)

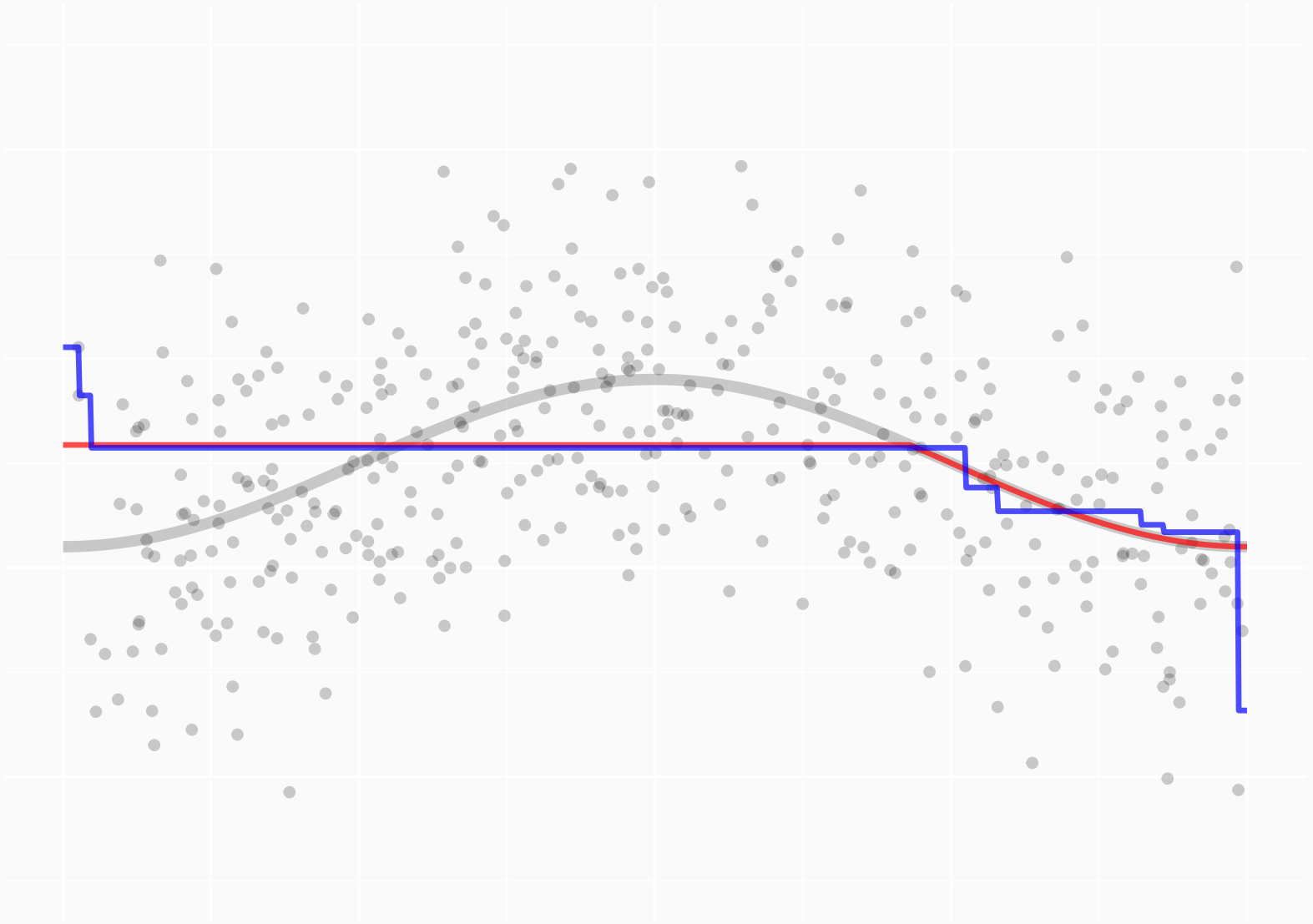


Figure 4: Decreasing Curves ($n = 400$.)

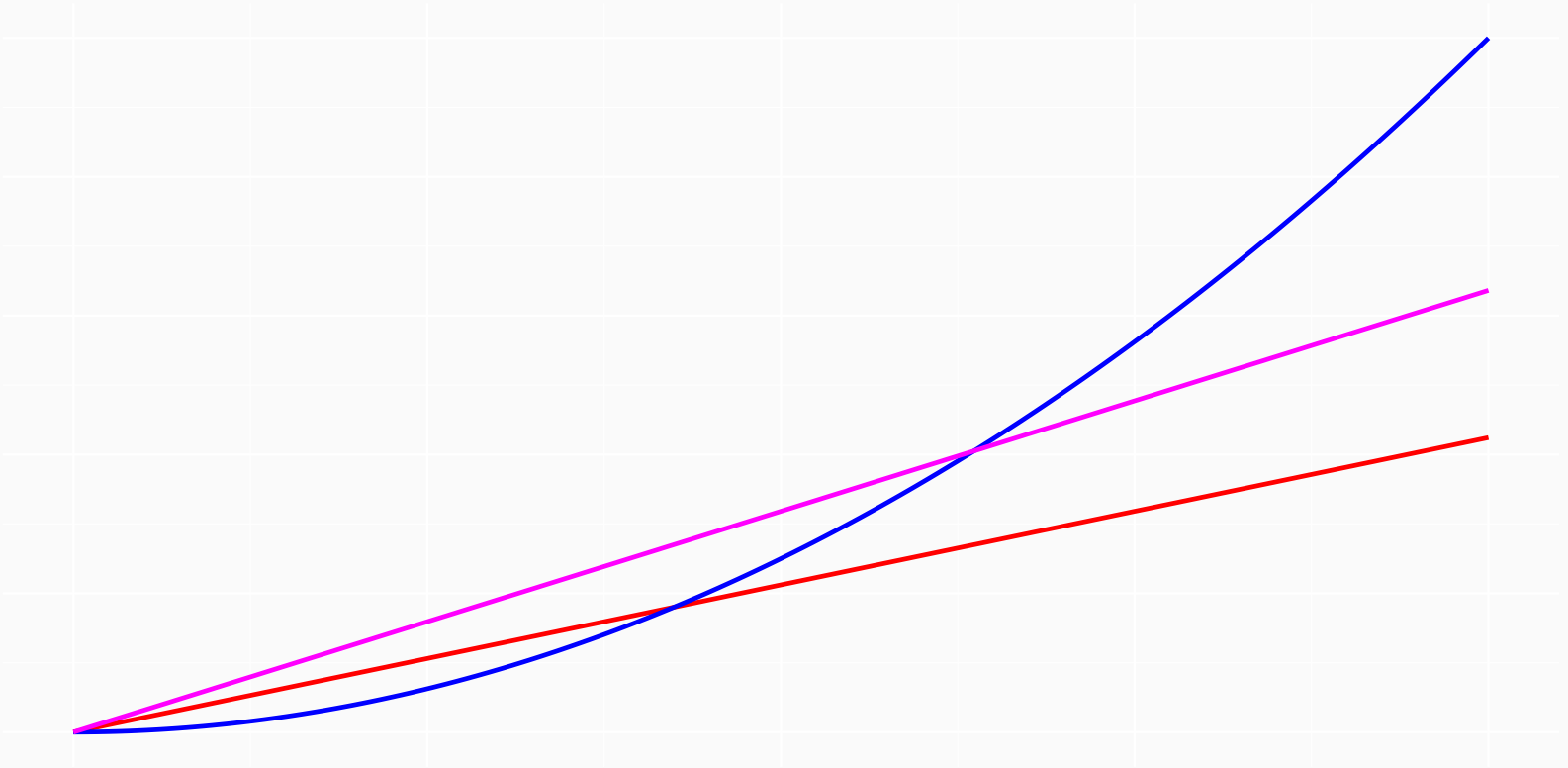


Figure 5: Bounded Variation Curves: $\rho_{\mathrm{TV}} \leq 1$ $(n = 100.)$

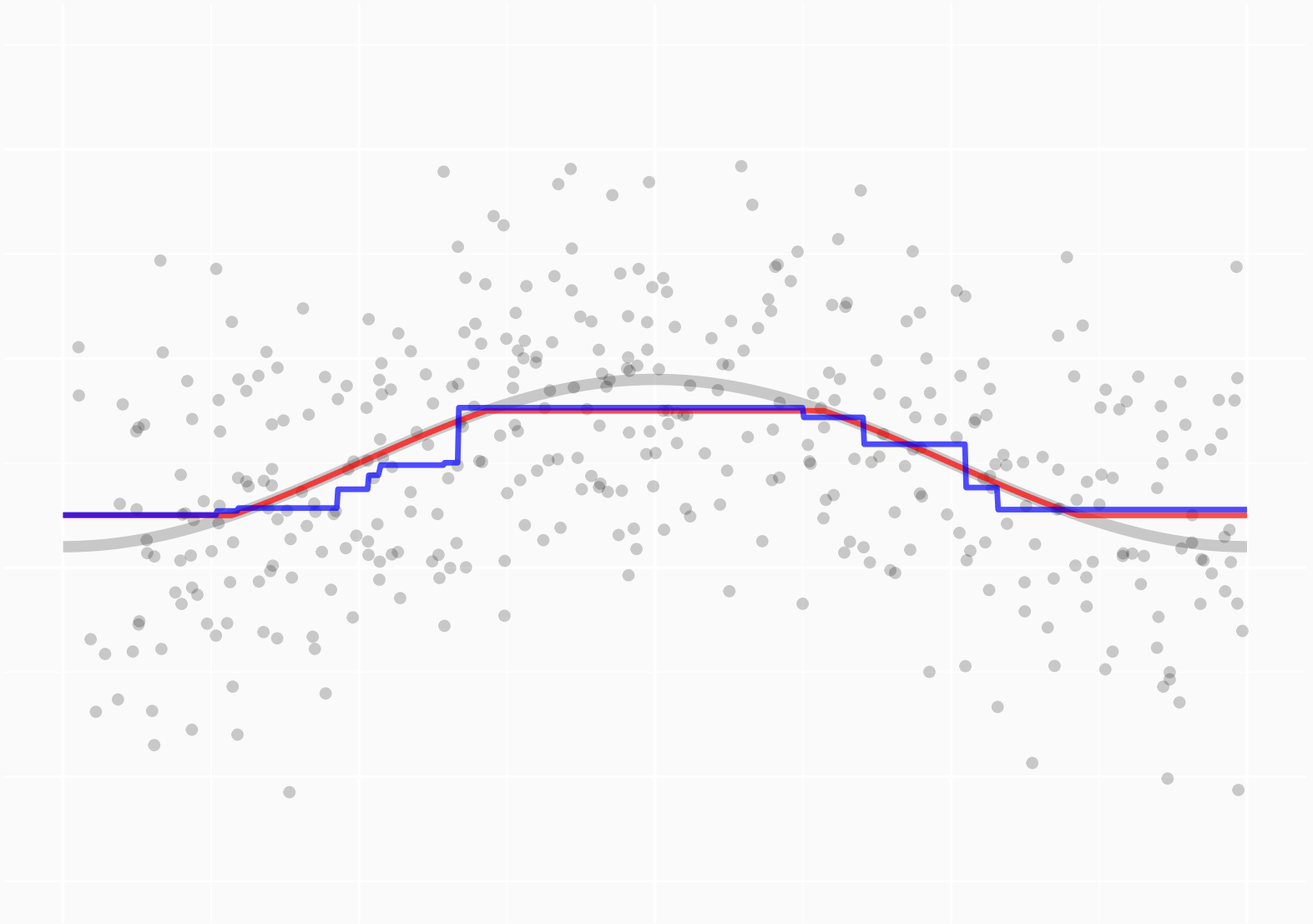


Figure 6: Bounded Variation Curves: $\rho_{\mathsf{TV}} \leq 1$. ($n = 400$.)

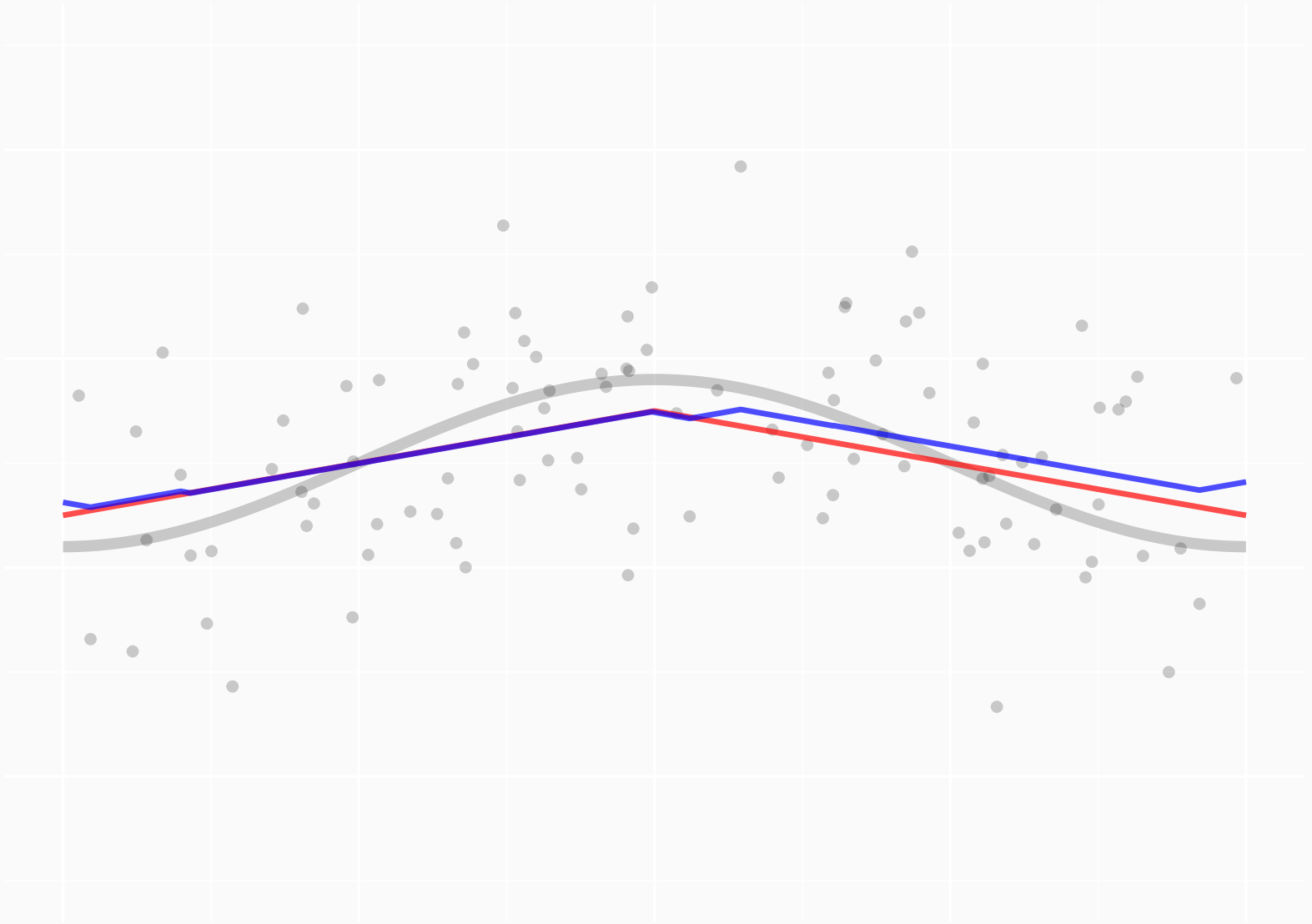


Figure 7: Lipschitz Curves: $\rho_{\text{Lip}} \leq 1$ $(n = 100.)$

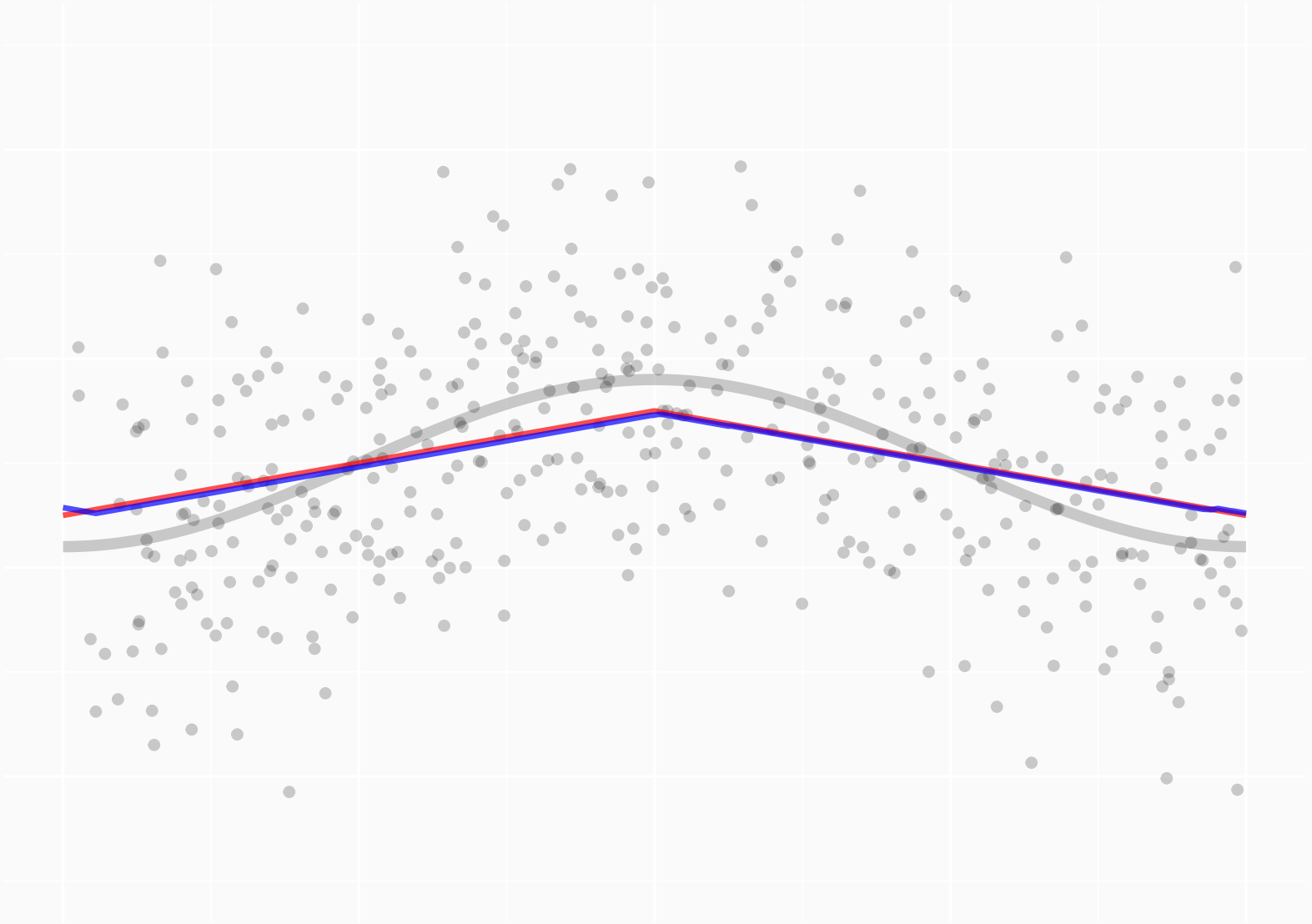
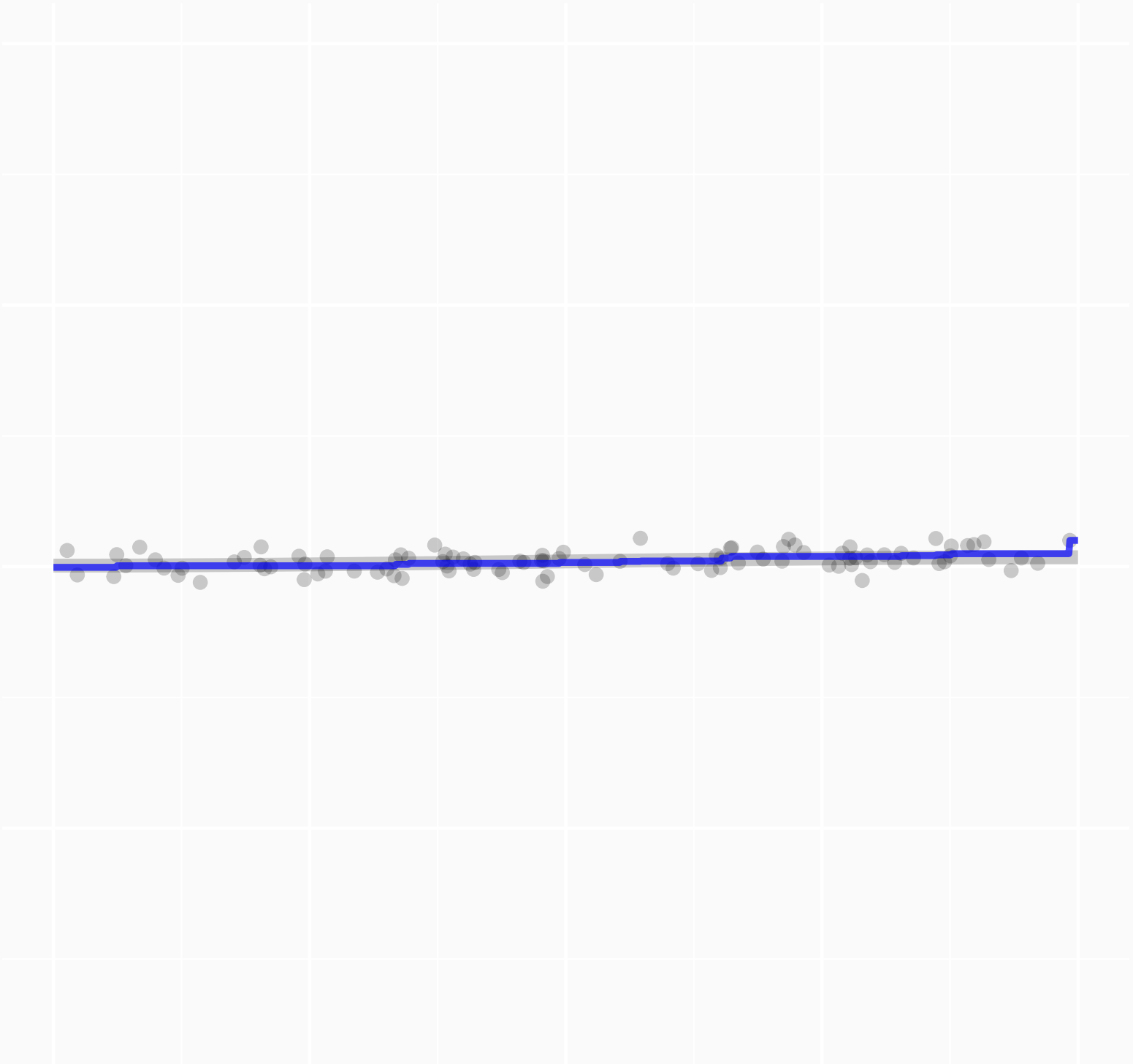
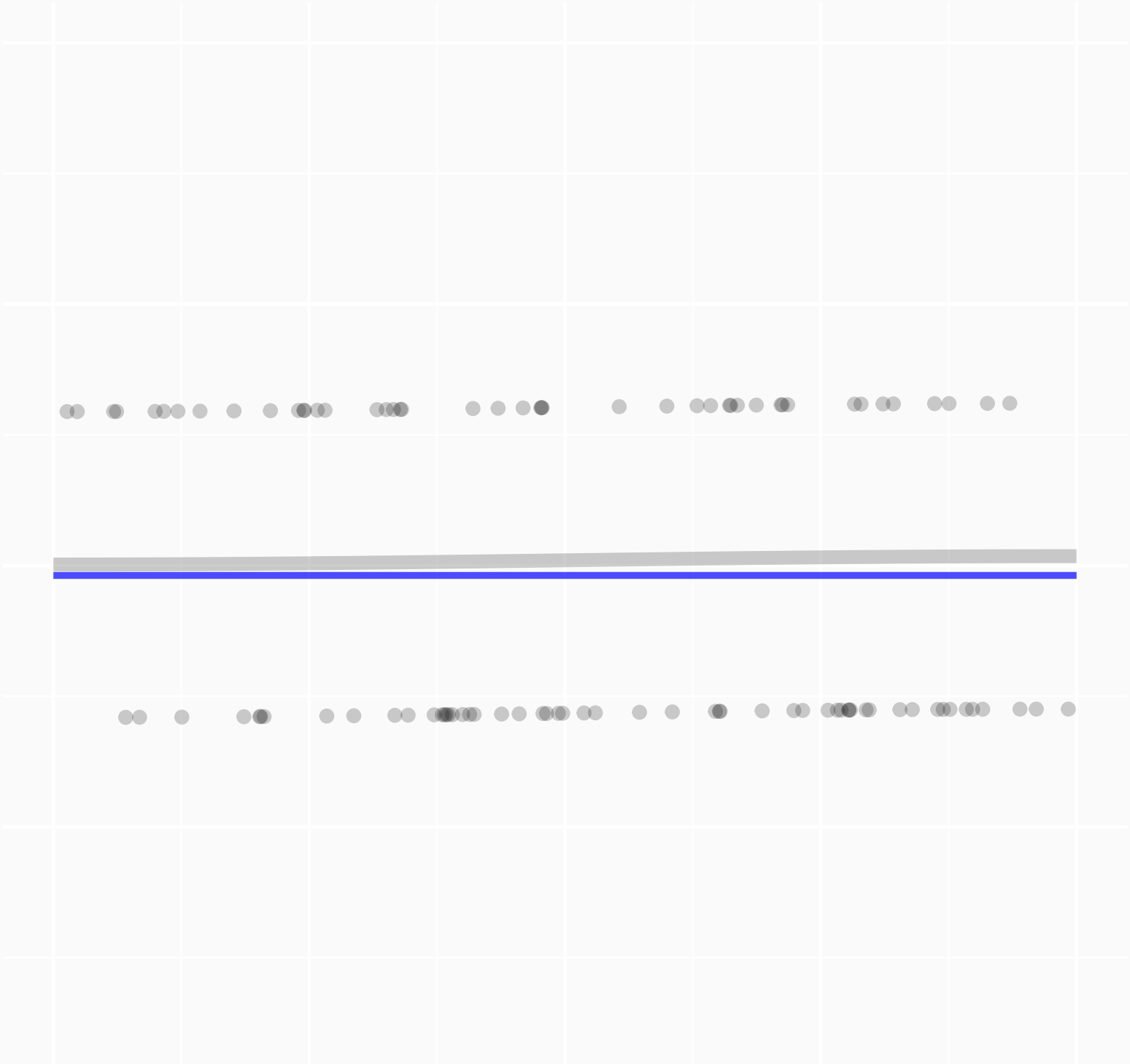


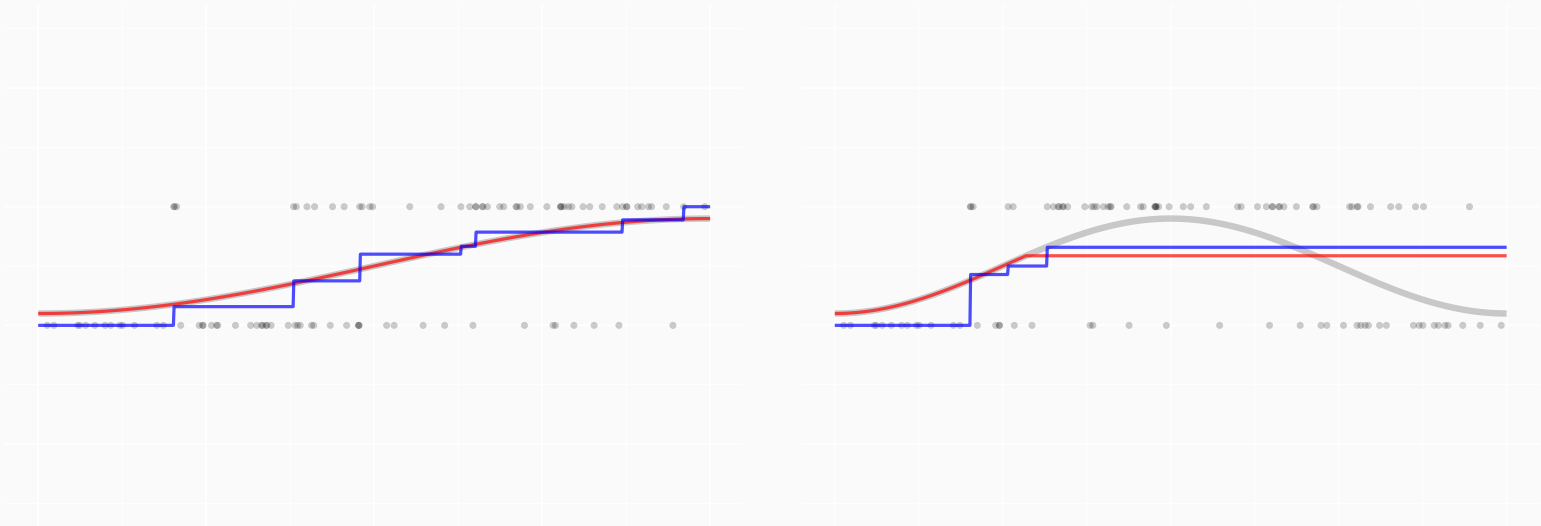
Figure 8: Lipschitz Curves: $\rho_{\text{Lip}} \leq 1$ $(n = 400.)$





Probabilistic Classification

What It Is



Suppose we have independent *binary observations*.

$$Y_i = \begin{cases} 1 & \text{w.p. } \mu(X_i) \\ 0 & \text{w.p. } 1 - \mu(X_i) \end{cases}$$

$$= \mu(X_i) + \varepsilon_i \quad \text{for} \quad \varepsilon_i = \begin{cases} 1 - \mu(X_i) & \text{w.p. } \mu(X_i) \\ -\mu(X_i) & \text{w.p. } 1 - \mu(X_i) \end{cases}.$$

We can think of this as regression with *classification noise* ε_i.
That's what's left after subtracting the mean $\mu(X_i) = \mathrm{E}[Y_i]$. It has mean zero.

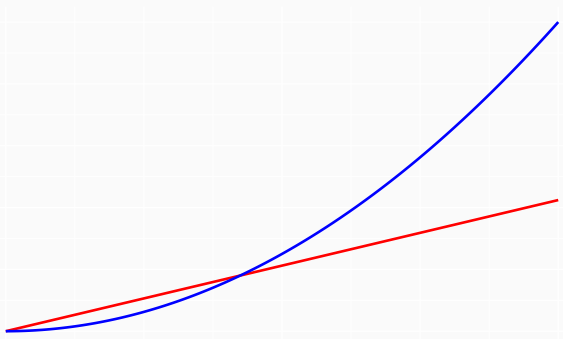
$$\mathrm{E}[\varepsilon_i] = \mu(X_i)\{1 - \mu(X_i)\} + \{1 - \mu(X_i)\}\{-\mu(X_i)\} = 0.$$

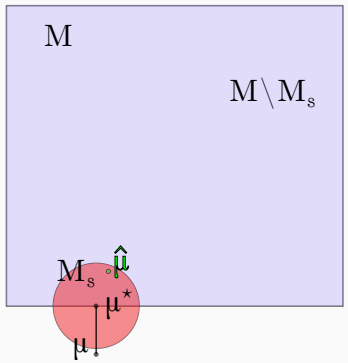
Starting Point: Our General Regression Error Bound

$$\ell(m) - \ell(\mu^\star) = \|m - \mu^\star\|^2_{L_2(\mathrm{P_n})} \qquad \text{squared distance}$$

$$- \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \{m(X_i) - \mu^\star(X_i)\} \qquad \text{a mean zero term}$$

$$+ \frac{2}{n}\sum_{i=1}^{n} \{\mu^\star(X_i) - \mu(X_i)\}\{m(X_i) - \mu^\star(X_i)\} \qquad \text{a non-negative term.}$$





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

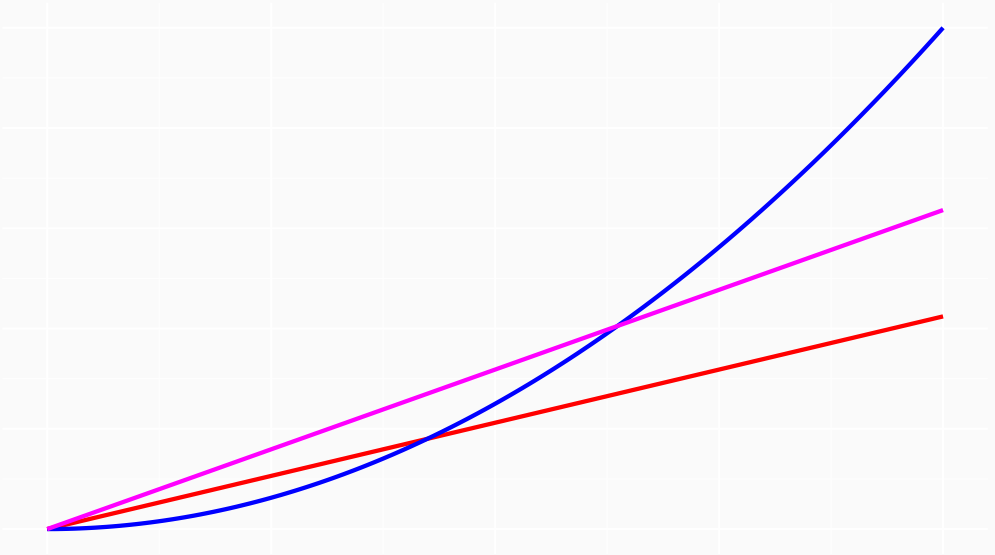
$$\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P_n})} < s + 2\sqrt{\frac{2\Sigma_n}{\delta n}} \quad \text{w.p. } 1-\delta \text{ for } \frac{s^2}{2} \geq \mathrm{w}_\varepsilon(\mathcal{M}_s)$$

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

We can take **s** to be the point where the red and blue curves cross.

Q. How does this error bound compare to the one we get with Gaussian noise?

Error Bounds and Width Comparison



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

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

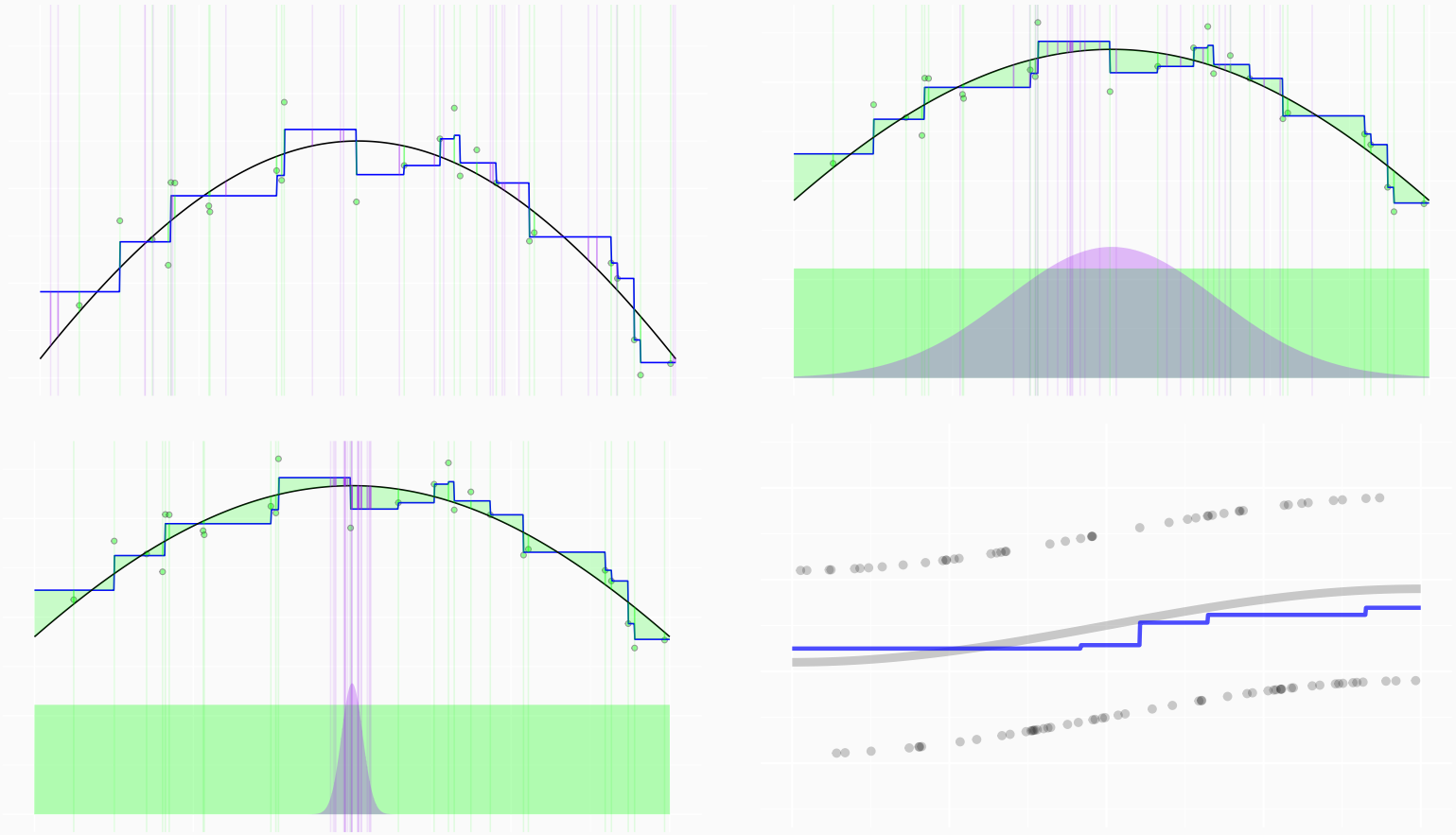
We'll show that 'classification-noise width' is no larger than $1.25 \times$ gaussian width.

$$1.25\, \mathrm{w}(\mathcal{V}) \geq \mathrm{w}_\varepsilon(\mathcal{V}) \quad \text{for any set} \quad \mathcal{V}$$

This implies an error bound for probabilistic classifiction that's no worse than the one we'd get with gaussian noise of standard deviation 1.25.

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

This is a Multi-Step Comparison



We compare versions of this maximum with ....

Top Left. Our original noise, ε_i.

$\rightarrow$ *Symmetrized noise*, $\varepsilon_i - \varepsilon'_i$ where ε' is an independent copy of ε.

$\downarrow$ *Random-sign noise* $s_i = \pm 1$ each w.p. $1/2$.

$\leftarrow$ *Gaussian noise* $g_i \sim N(0, \sigma^2)$ for $\sigma = \sqrt{\pi/2} \approx 1.25$.

Step 1. *Symmetrization*

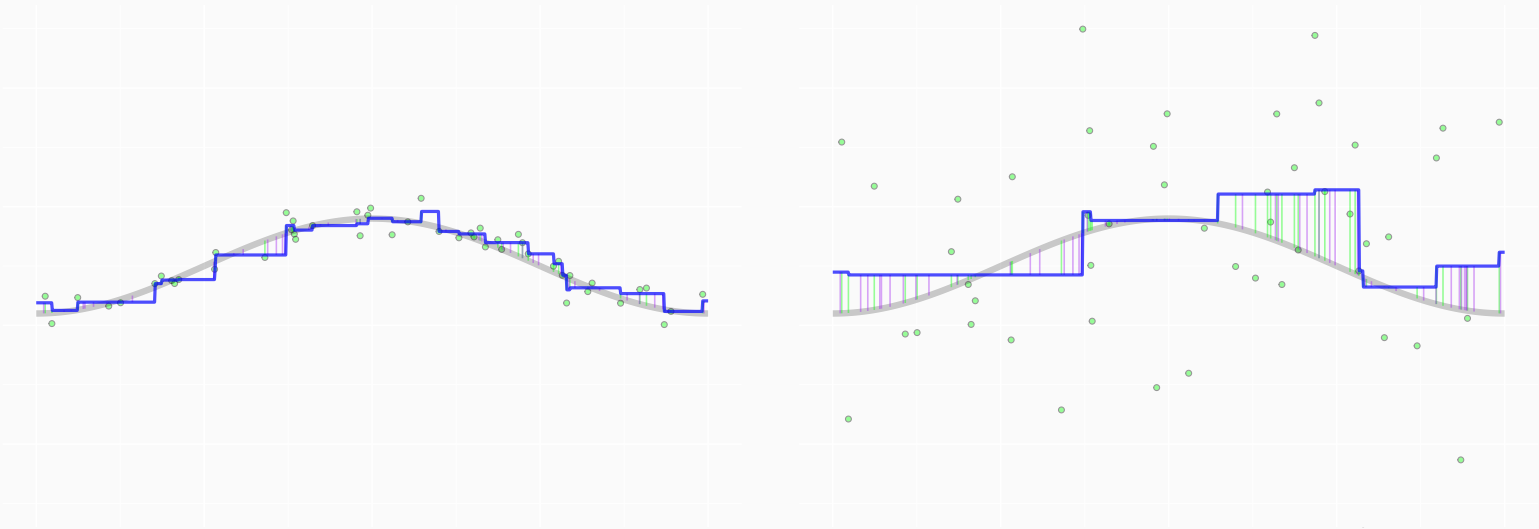


Figure 11: Our data with noise ε_i and the symmetrized version $\varepsilon_i - \varepsilon_i'$

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

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

Step 1. *Symmetrization*

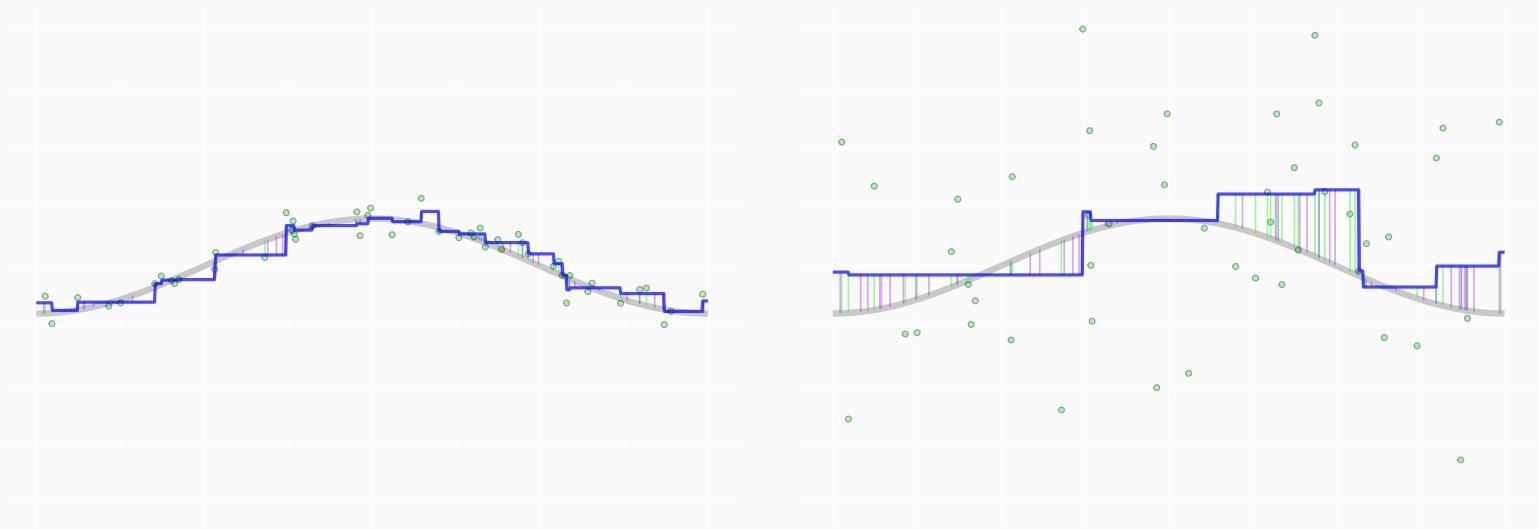


Figure 11: Our data with noise ε_i and the symmetrized version $\varepsilon_i - \varepsilon_i'$

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

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

Step 1. *Symmetrization*

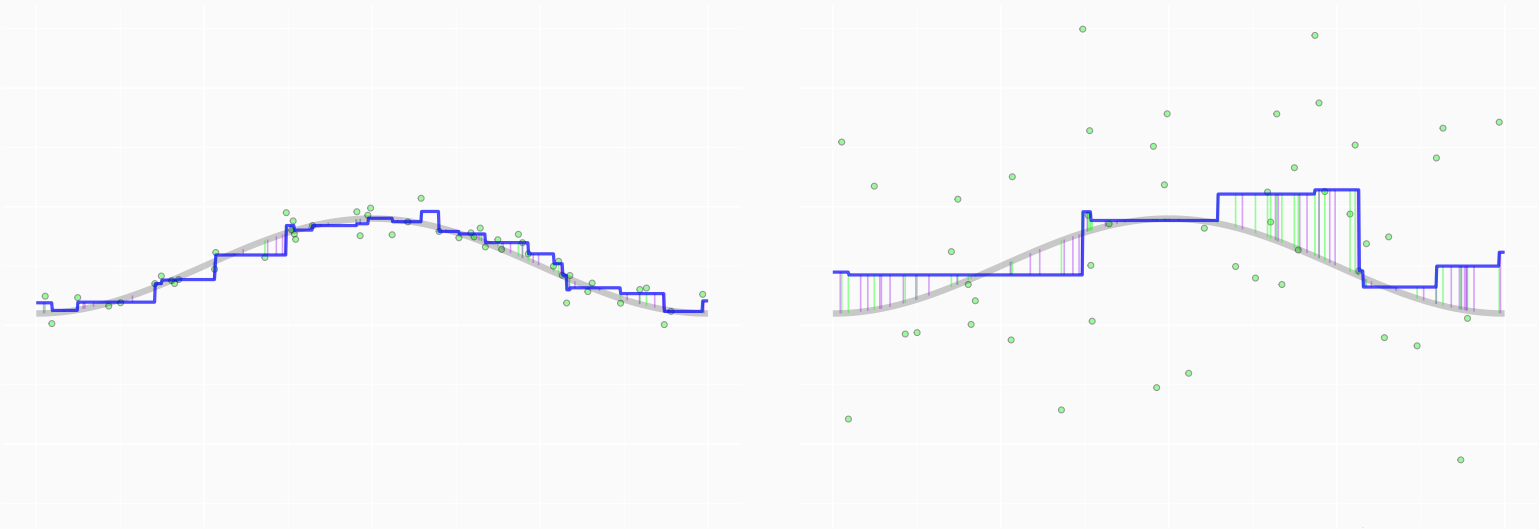


Figure 11: Our data with noise ε_i and the symmetrized version $\varepsilon_i - \varepsilon_i'$

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

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

Step 1. *Symmetrization*

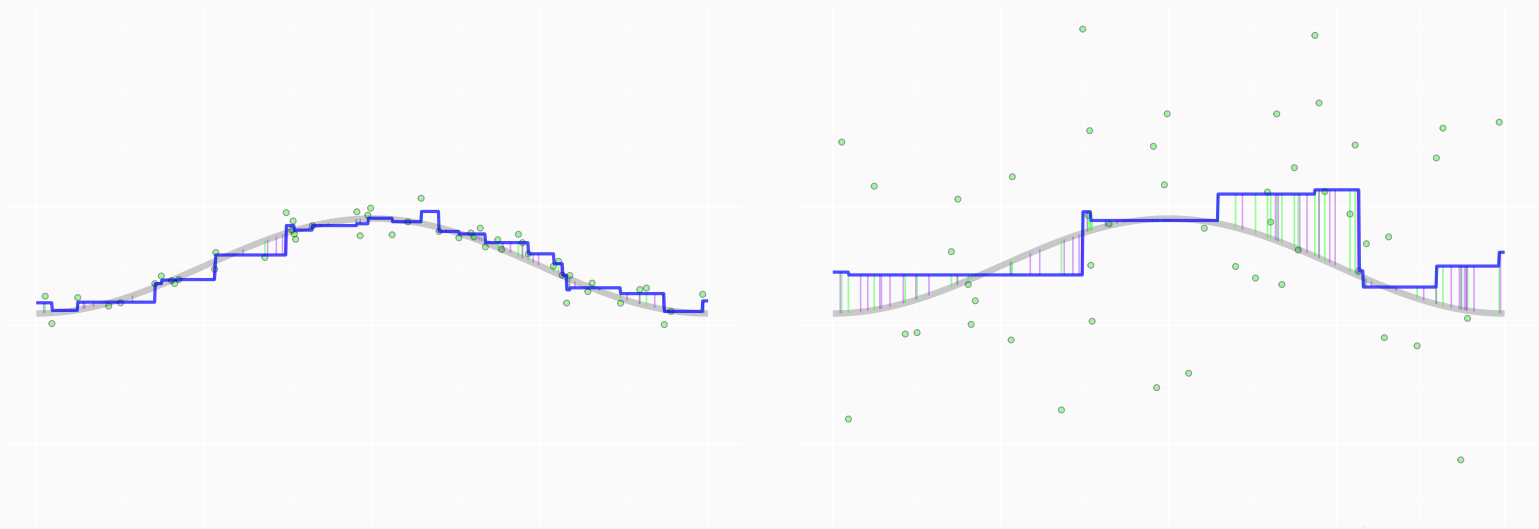


Figure 11: Our data with noise ε_i and the symmetrized version $\varepsilon_i - \varepsilon_i'$

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

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

Step 1. *Symmetrization*

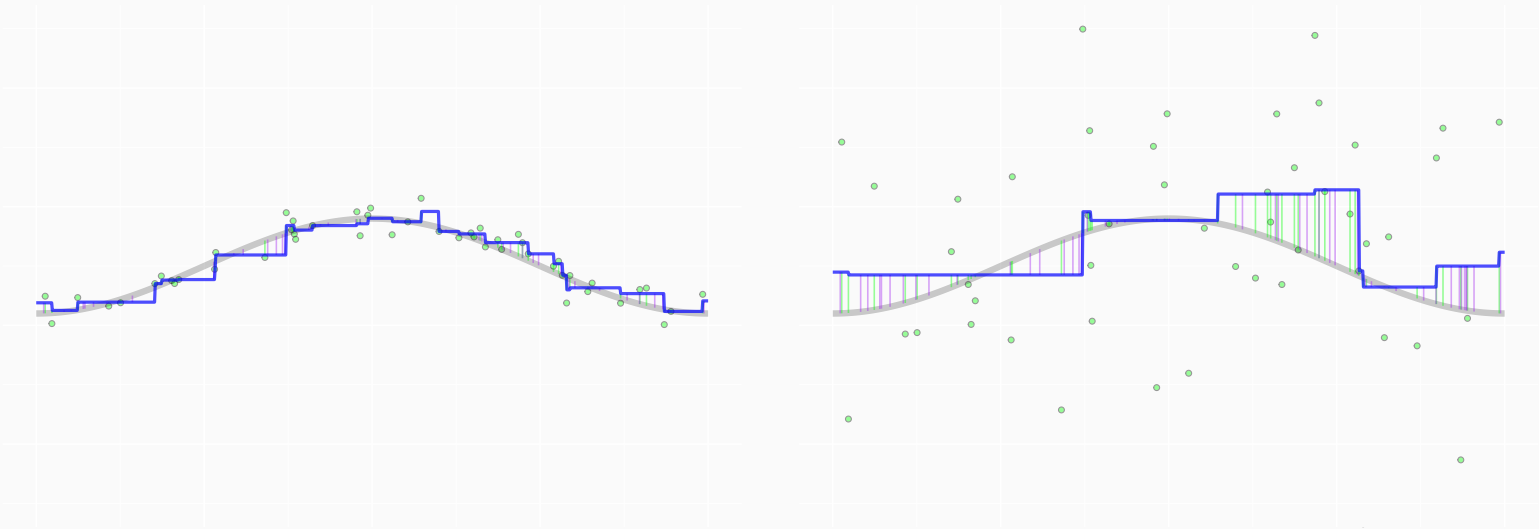


Figure 11: Our data with noise ε_i and the symmetrized version $\varepsilon_i - \varepsilon_i'$

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

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

Step 1′. *Symmetrization* with *Random Signs*

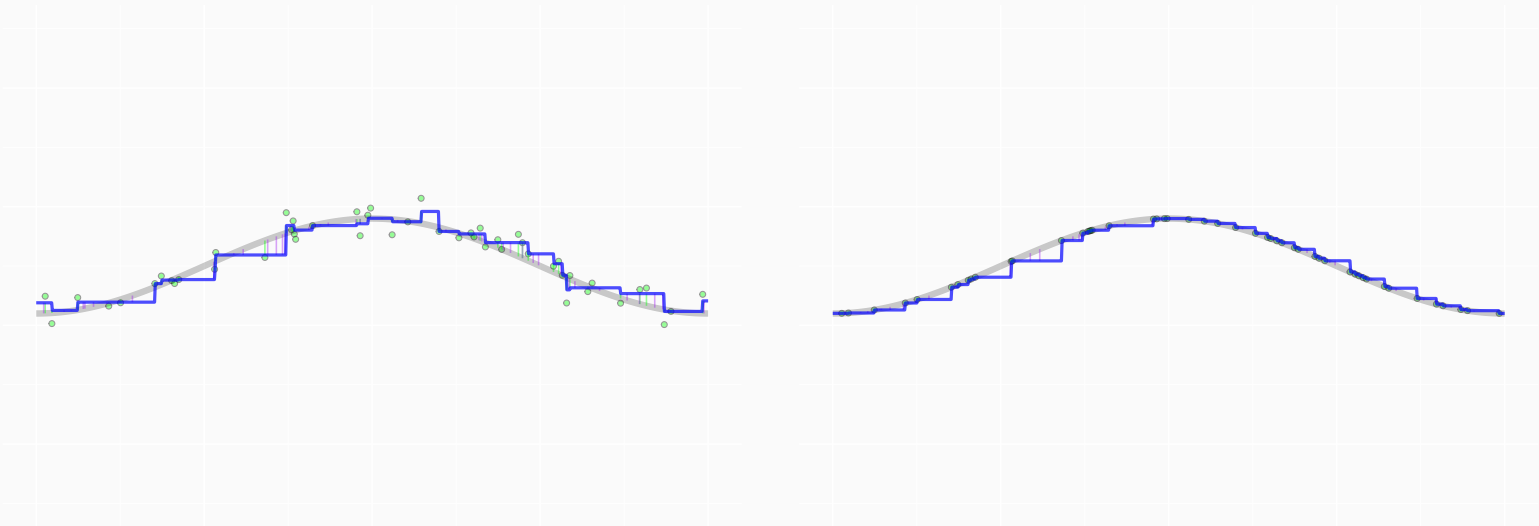


Figure 12: Our data with symmetrized noise $\varepsilon_i - \varepsilon_i'$ and $s_i(\varepsilon_i - \varepsilon_i')$.

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 1′. *Symmetrization* with *Random Signs*

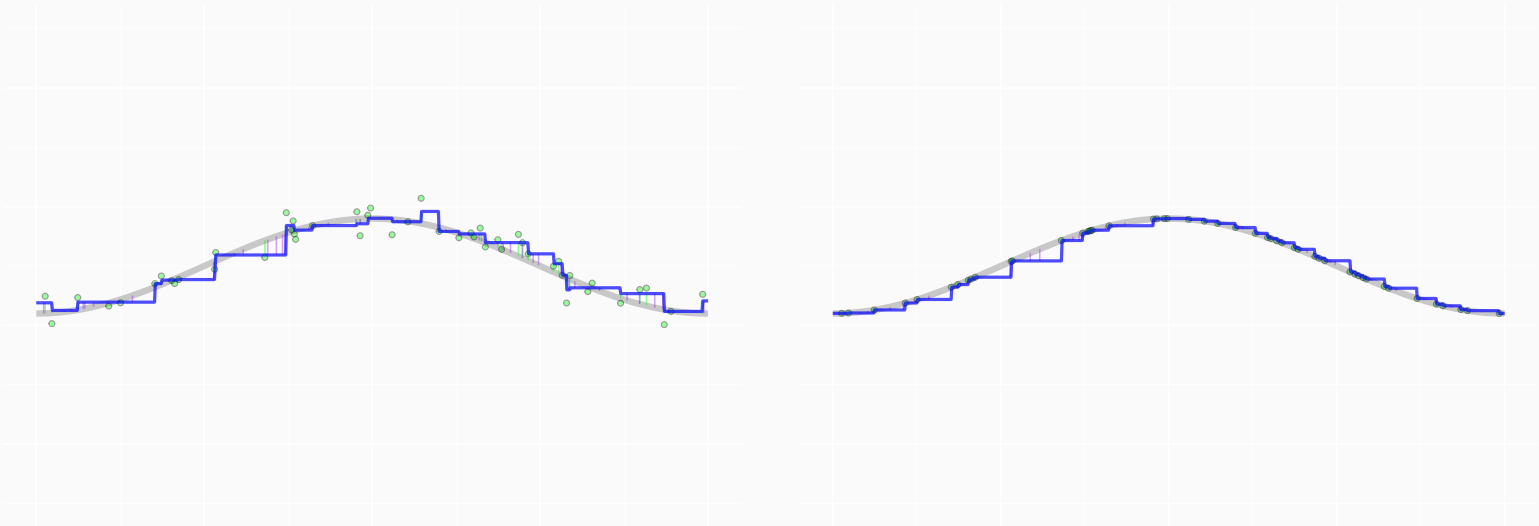


Figure 12: Our data with symmetrized noise $\varepsilon_i - \varepsilon_i'$ and $s_i(\varepsilon_i - \varepsilon_i')$.

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 $(\varepsilon_i' - \varepsilon_i) = -(\varepsilon_i - \varepsilon_i')$ do, too.

Step 1. About the Symmetrized Noise

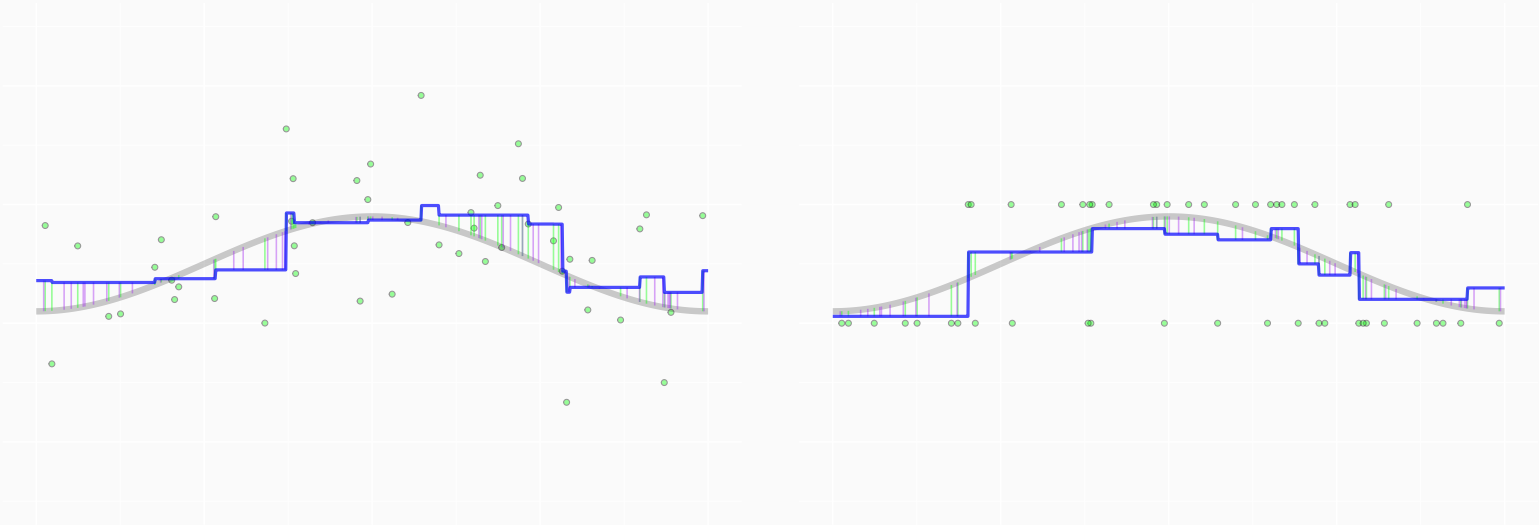


Figure 13: Our data with noise ε_i and the symmetrized version $\varepsilon_i - \varepsilon_i'$

Our symmetrized noise, $\varepsilon_i - \varepsilon_i'$ takes on 3 values: $0, +1, -1$.

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

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

Step 2. *Contraction*

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

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

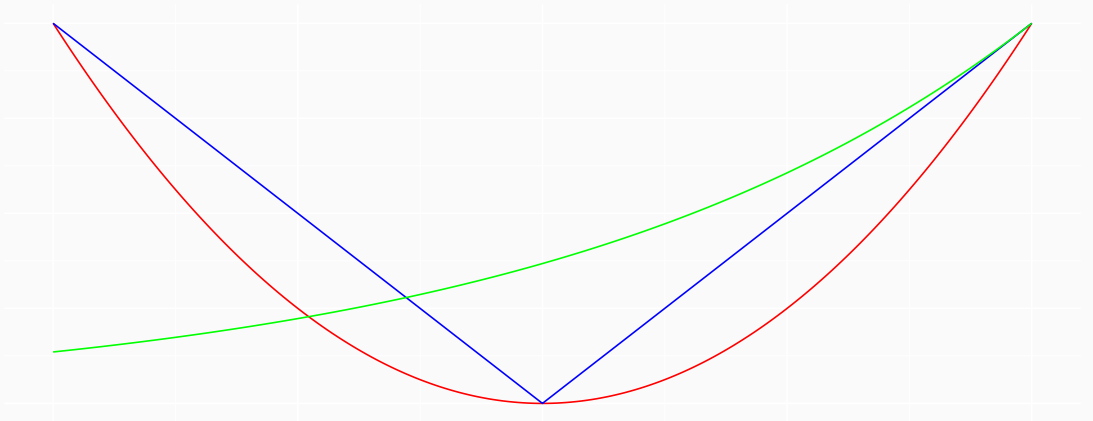
That noise is in the unit cube $[-1,1]^n$, so we can bound that function's *average* over the noise by its *maximum* over the cube.

Step 2. *Contraction* and Convexity

$$\mathrm{E}_s\,\mathrm{E}_\varepsilon\,\mathrm{E}_{\varepsilon'}\max_{v\in\mathcal{V}}\sum_{i=1}^{n}s_i(\varepsilon_i-\varepsilon_i')v_i\le\max_{u\in[-1,1]^n}f(u)\quad\text{for}\quad f(u)=\mathrm{E}_s\max_{v\in\mathcal{V}}\sum_{i=1}^{n}s_iu_iv_i$$

That function is *convex*.

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



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

Step 2. *Contraction* and Convexity

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

That function is *convex*.

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

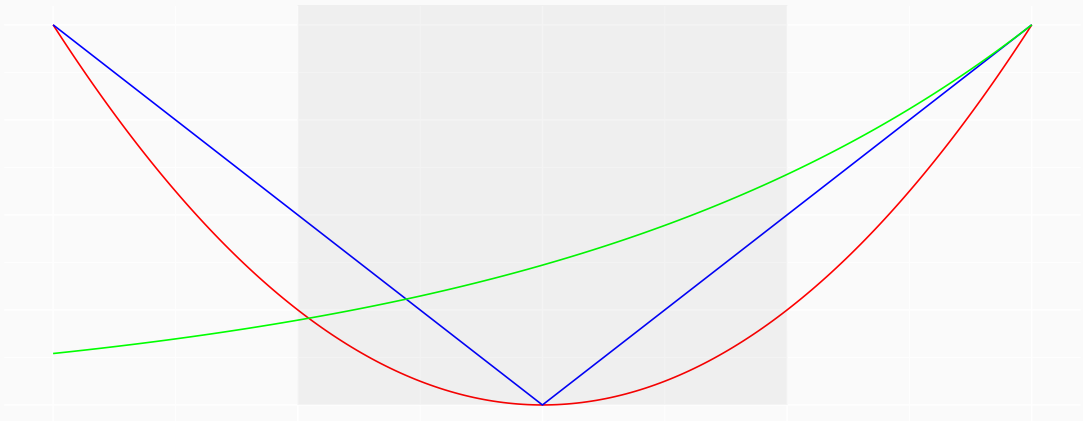
$$\begin{aligned}
f\{(1-\lambda)a + \lambda b\} &= \mathrm{E}_s \max_{v \in \mathcal{V}} \left\{ (1-\lambda) \sum_{i=1}^{n} s_i a_i v_i + \lambda \sum_{i=1}^{n} s_i b_i v_i \right\} \\
&\leq \mathrm{E}_s \left\{ \max_{v \in \mathcal{V}} (1-\lambda) \sum_{i=1}^{n} s_i a_i v_i + \max_{v \in \mathcal{V}}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

Step 2. *Contraction* and Convexity

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

That function is *convex*.

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



And it's easy to characterize this maximum over corners. It's just random-sign width.

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

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

Why this maximum is just random-sign width.

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

- 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.
  - i.e. this function $f(u)$ takes on the same value at every corner.
  - including the vector of all ones $u = (1, 1, \ldots, 1)$.

Step 2. *Contraction* Visualized.

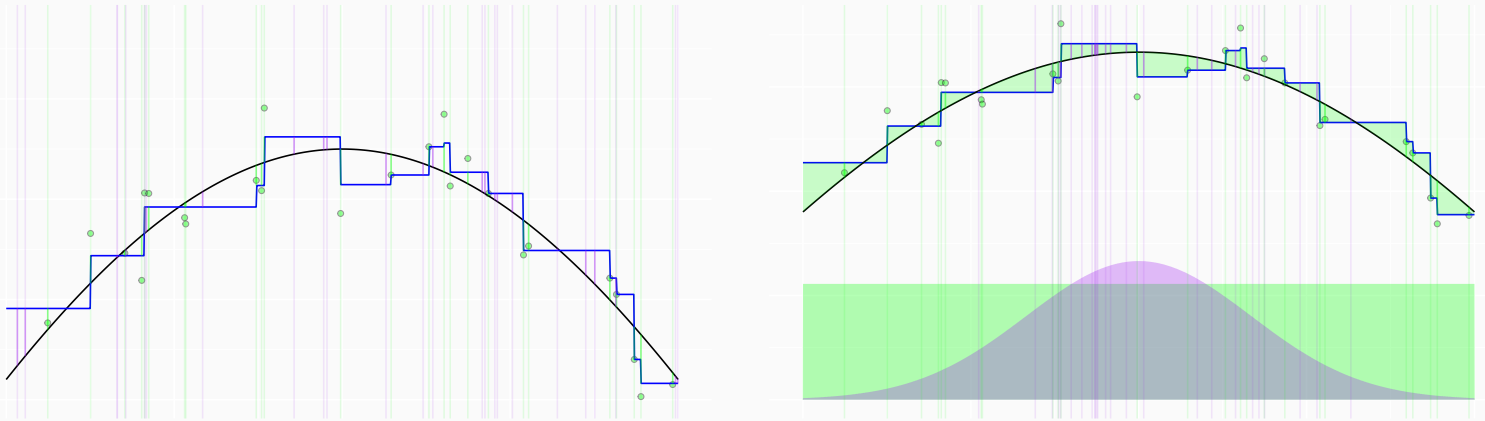


Figure 14: Our data with symmetrized noise $s_i(\varepsilon_i - \varepsilon_i')$ and random-sign noise s_i.

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

It makes sense that we'd get a bigger average with (only) random signs.
In effect, we've replaced zero-noise observations ($\varepsilon_i = \varepsilon_i'$) with noisy ones.

Step 3. Comparison to Gaussian Width

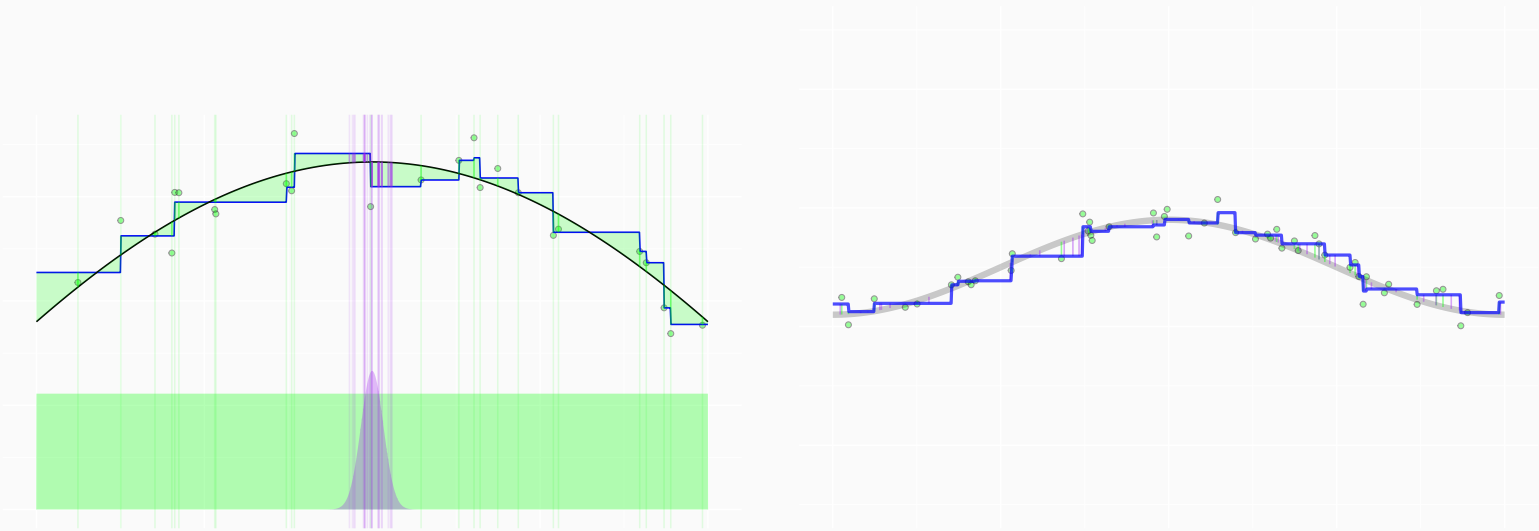


Figure 15: Our data with random-sign noise s_i and gaussian noise σg_i for $\sigma = 1.25$

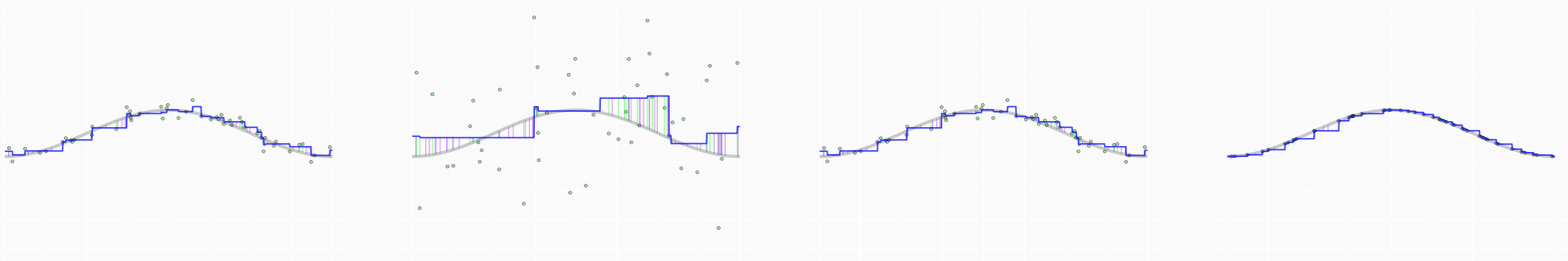
We can compare random-sign width to gaussian width. It's at most 1.25 times as big.

$$\mathrm{w}_s(\mathcal{V}) \le \sigma\, \mathrm{w}(\mathcal{V}) \quad \text{for any set} \quad \mathcal{V} \quad \text{where} \quad \sigma = \frac{1}{\mathrm{E}|g_i|} = \sqrt{\frac{\pi}{2}} \approx 1.25.$$

To show that, we use our 'two+ maxes are better than one' bound in reverse.

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

Summary



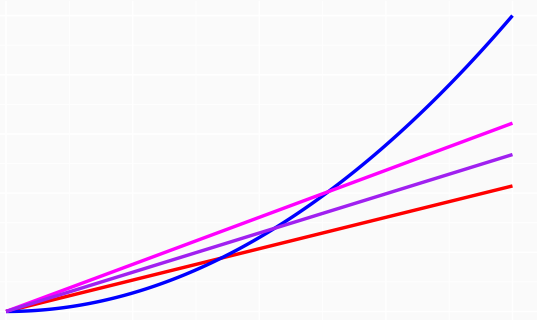
With each step, width gets bigger.

That means probabilistic classification 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, at least, in the sense that our argument gives us a better error bound.

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



Terminology

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

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

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

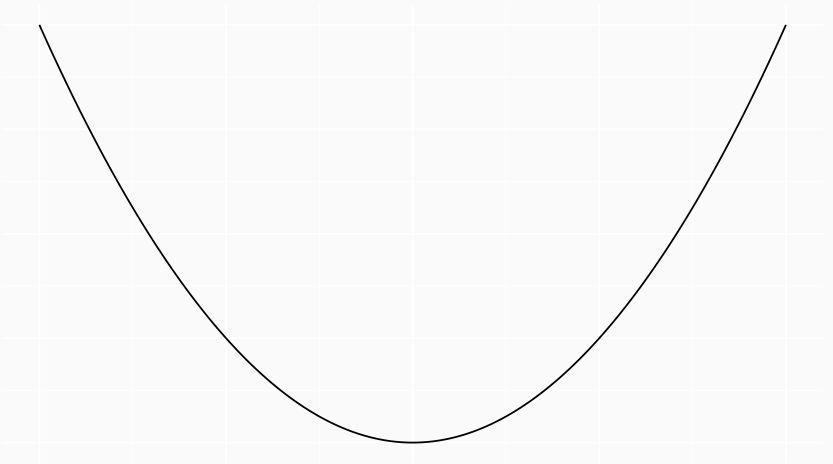
- This second definition is the same if $\mathcal{V}$ is symmetric, i.e. $v \in \mathcal{V} \implies -v \in \mathcal{V}$.
- Otherwise, it can be a little bigger.
  - At most $2\times$ bigger. Prove it!
  - Use the bound $\max a, b \leq a + b$ and the symmetry of s's distribution.

Convex Functions Are Maximized At Extreme Points

Definition

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

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



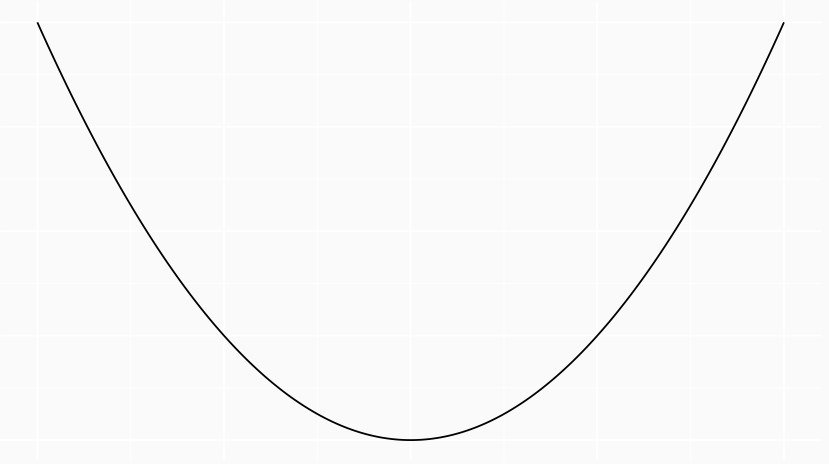
We can give this a *probabilistic interpretation* for a random variable Z_λ.

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

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.

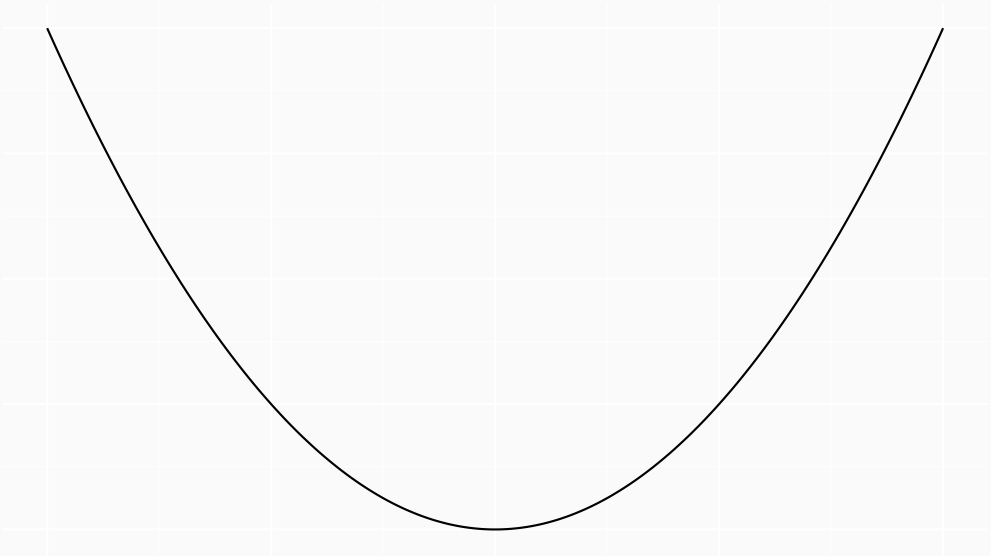
$$f(\mathrm{E}\, Z_\lambda) \leq \mathrm{E}\, f(Z_\lambda) \quad \text{where} \quad Z_\lambda = \begin{cases} a & \text{w.p. } 1-\lambda \\ b & \text{w.p. } \lambda \end{cases}$$

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.

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

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.

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

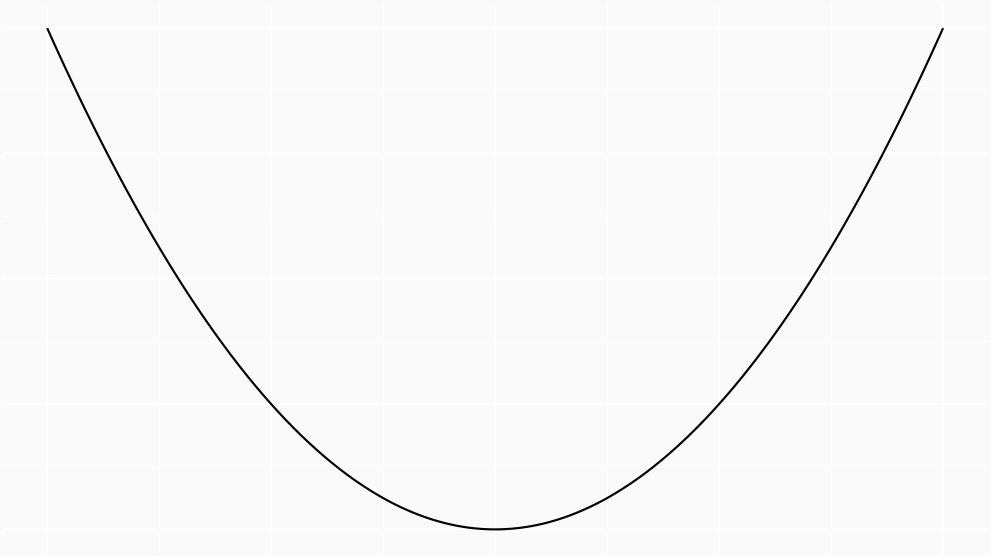
Inductive Step.

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.

$$\max_{x\in[a,b]} f(x) = \max_{\lambda\in[0,1]} \underbrace{f\{(1-\lambda)a+\lambda b\}}_{f(\mathrm{E}\, Z_\lambda)} \leq \max_{\lambda\in[0,1]} \underbrace{(1-\lambda)f(a)+\lambda f(b)}_{\mathrm{E}\, f(Z_\lambda)} = \max\{f(a), f(b)\}$$

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 $u_1 \dots u_K$.

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

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.

$$\max_{u \in \mathcal{U}} f(u) = \max_{\lambda \in \Lambda} f\left(\sum_i \lambda_i u_i\right) \leq \max_{\lambda \in \Lambda} \sum_i \lambda_i f(u_i) \leq \max_i f(u_i)$$

$$f(\mathrm{E}\, Z_\lambda) \qquad\qquad \mathrm{E}\, f(Z_\lambda)$$

where

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