Machine Learning Theory

Lecture 4: Least Squares in Finite Models

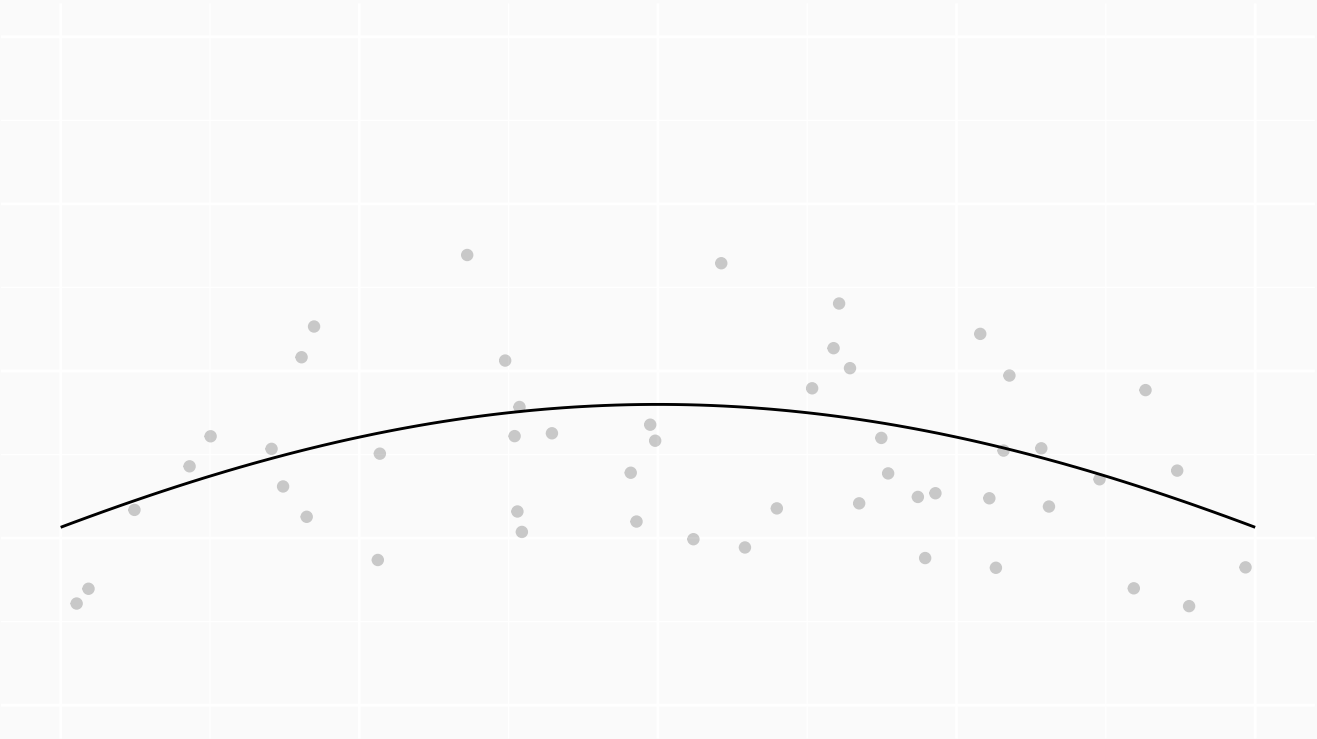
David A. Hirshberg

May 24, 2024

Emory University

Least squares with gaussian noise

We observe $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.



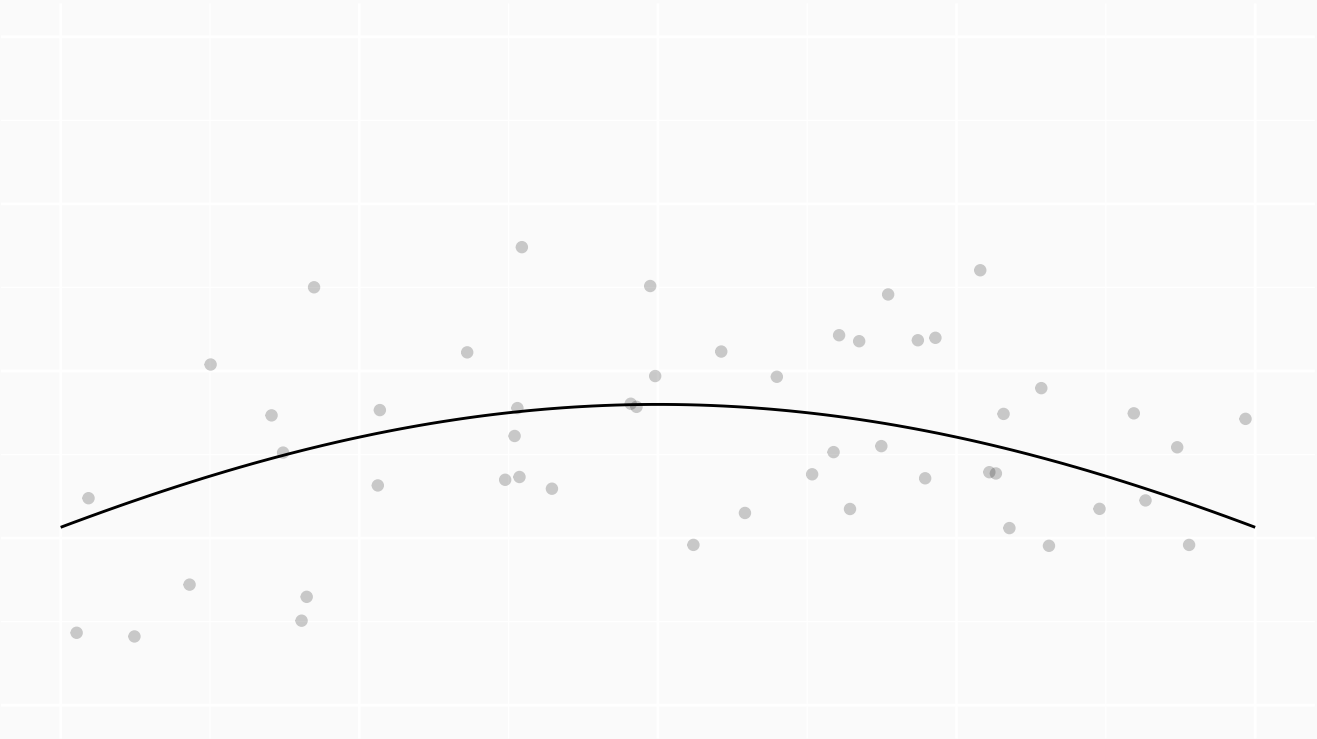
We're estimating the curve $\mu(x)$.
Our goal is get close in terms of *sample mean squared distance*.

This is the kind of statement we're after.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{with probability} \quad 1 - \delta$$

Least squares with gaussian noise

We observe $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.



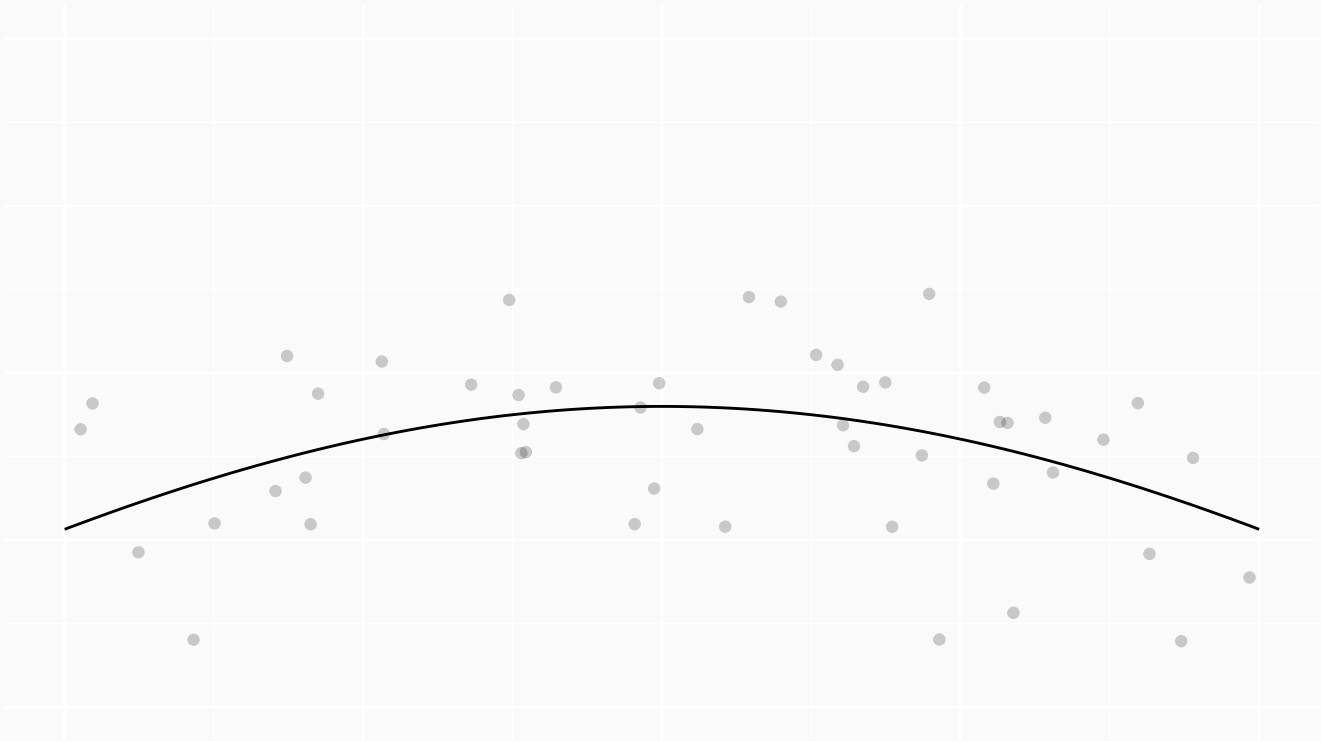
We're estimating the curve $\mu(x)$.
Our goal is get close in terms of *sample mean squared distance*.

This is the kind of statement we're after.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{with probability} \quad 1 - \delta$$

Least squares with gaussian noise

We observe $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

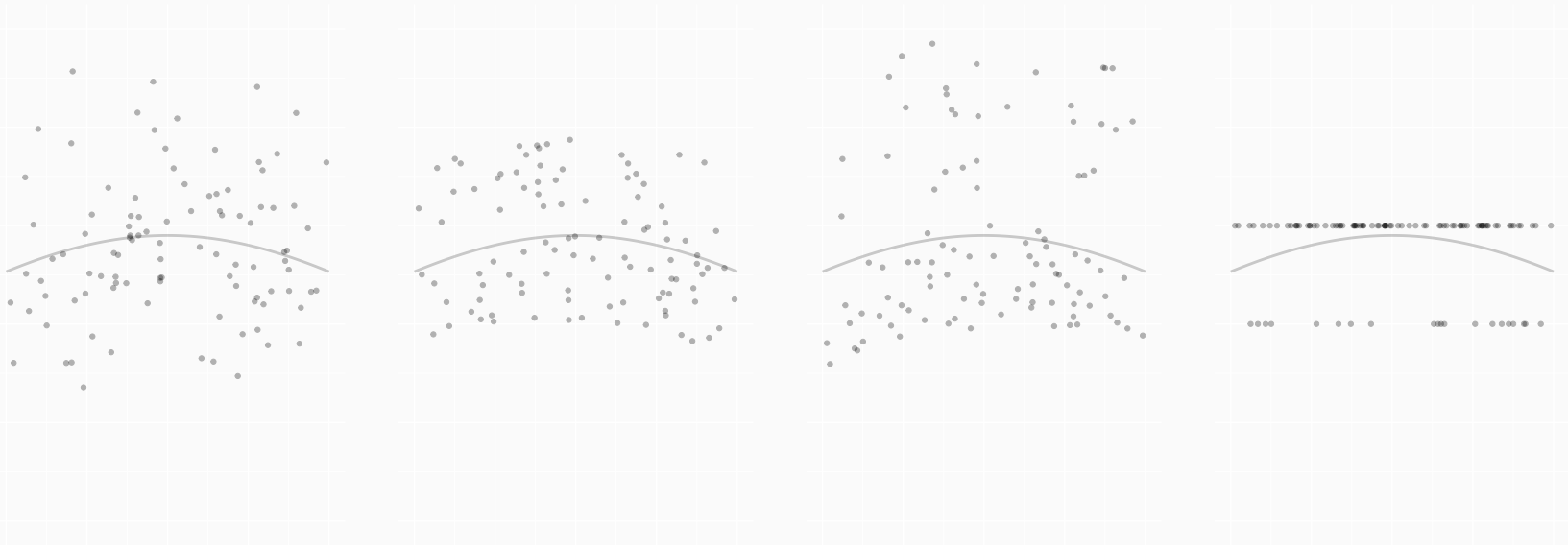


We're estimating the curve $\mu(x)$.
Our goal is get close in terms of *sample mean squared distance*.

This is the kind of statement we're after.

$$\|\hat{\mu} - \mu\|_{L_2(\mathrm{P_n})} < s \quad \text{with probability} \quad 1 - \delta$$

Why Gaussian?



- We'll focus on gaussian noise today because it's easy to think about.
- It makes the geometry simple.
- Once we've thought that case through, our results will generalize easily.
- None of the noise we see above will be a problem.

The two-observation case

To help with visualization, we'll look at the case with two observations.
This lets us plot our observation *vectors* Y as a point in the plane.

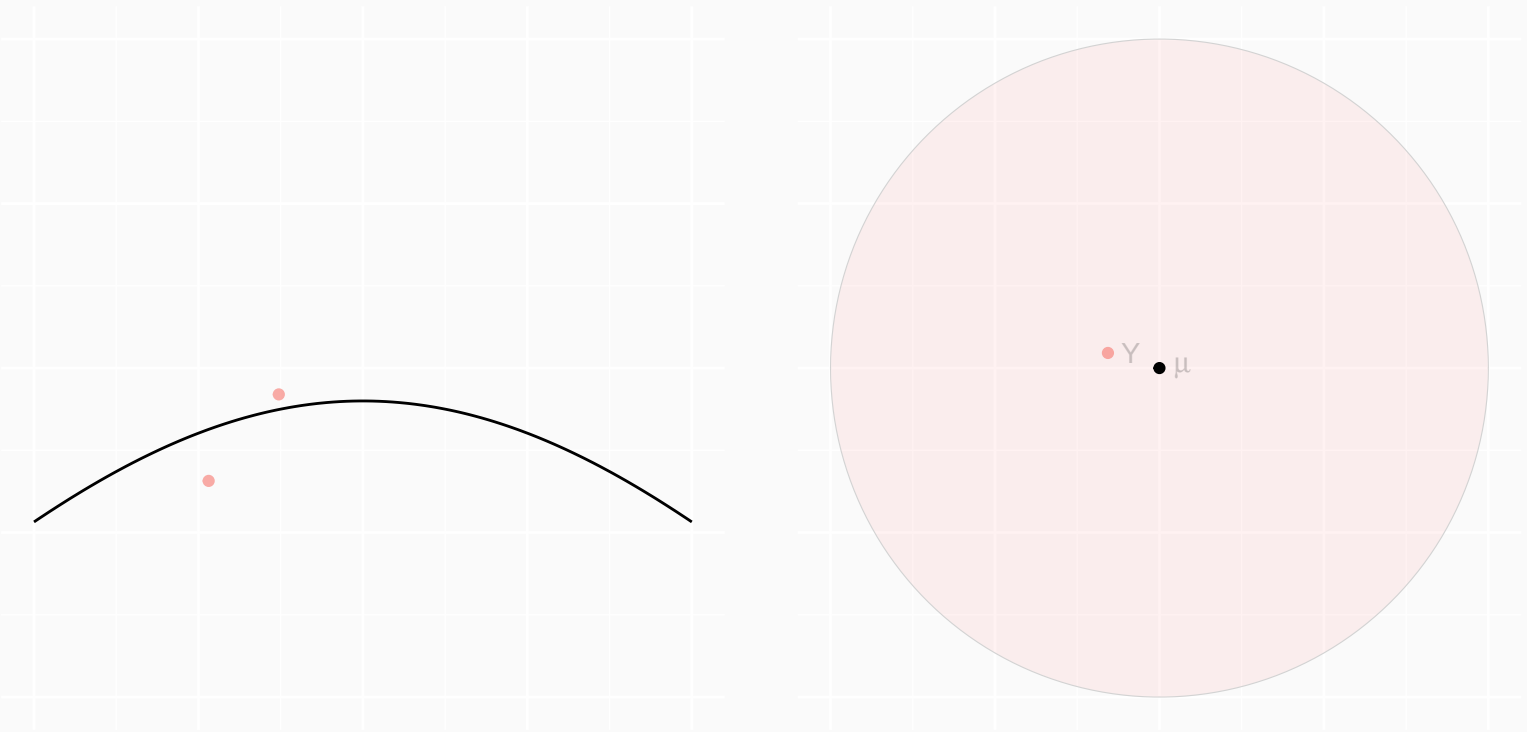


Figure 1: Right: $\mu(X_1)$ and Y_1 are x-coordinates; $\mu(X_2)$ and Y_2 are y-coordinates.

We'll use intuition we develop to understand what's going on in practical sample sizes.
We can plot a few random *replications* to get an idea of the distribution of Y.

The two-observation case

To help with visualization, we'll look at the case with two observations.
This lets us plot our observation *vectors* Y as a point in the plane.

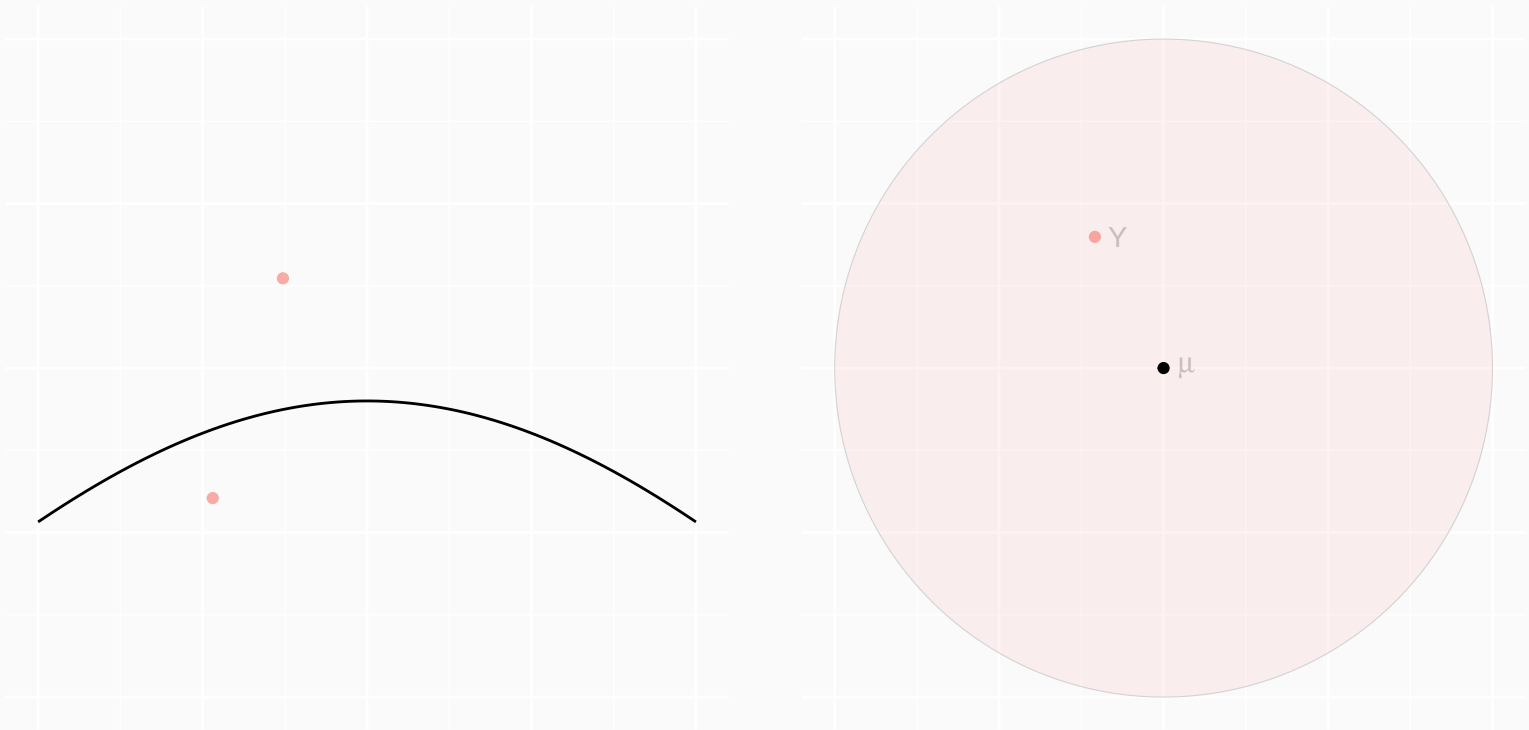


Figure 1: Right: $\mu(X_1)$ and Y_1 are x-coordinates; $\mu(X_2)$ and Y_2 are y-coordinates.

We'll use intuition we develop to understand what's going on in practical sample sizes.
We can plot a few random *replications* to get an idea of the distribution of Y.

The two-observation case

To help with visualization, we'll look at the case with two observations.
This lets us plot our observation *vectors* Y as a point in the plane.

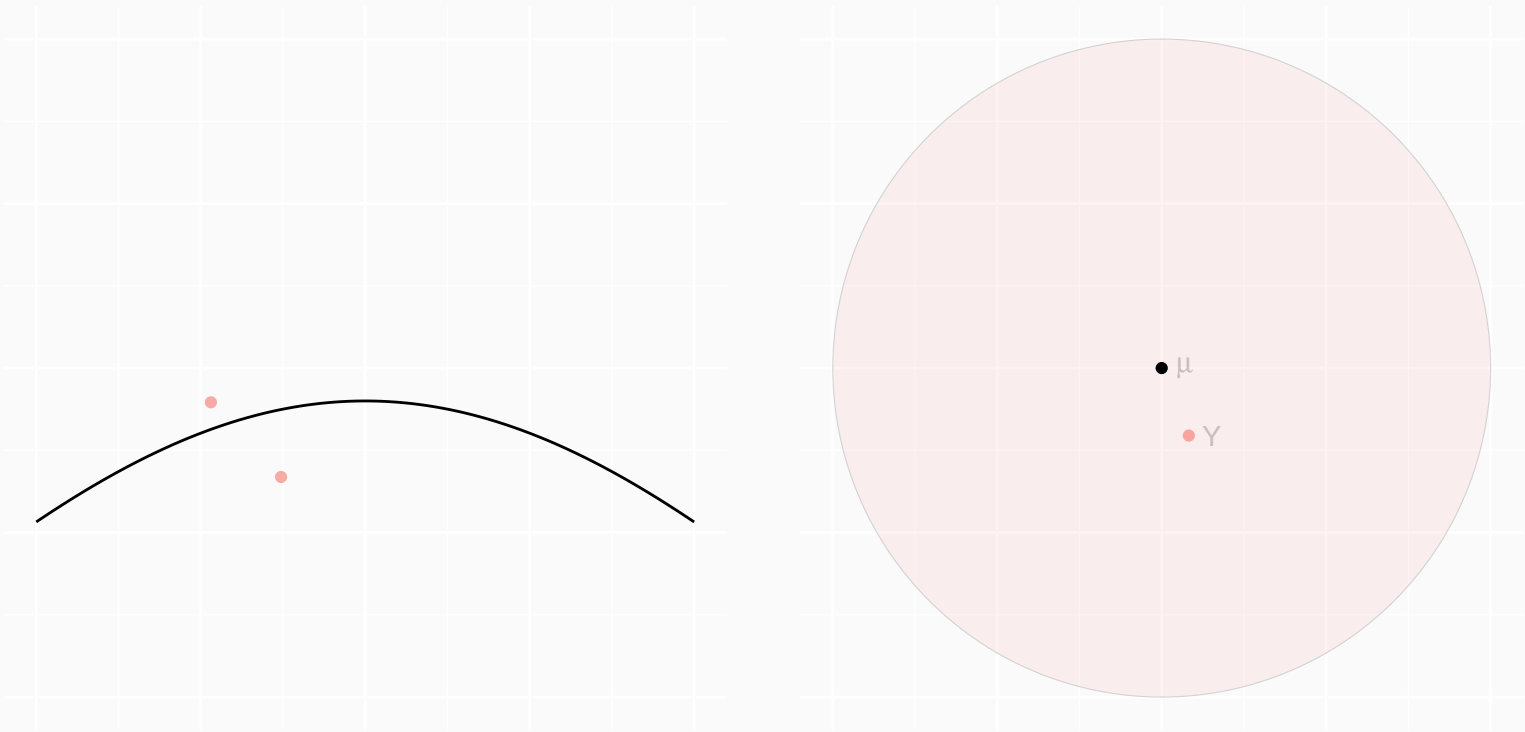


Figure 1: Right: $\mu(X_1)$ and Y_1 are x-coordinates; $\mu(X_2)$ and Y_2 are y-coordinates.

We'll use intuition we develop to understand what's going on in practical sample sizes.
We can plot a few random *replications* to get an idea of the distribution of Y.

The two-observation case

To help with visualization, we'll look at the case with two observations.
This lets us plot our observation *vectors* Y as a point in the plane.

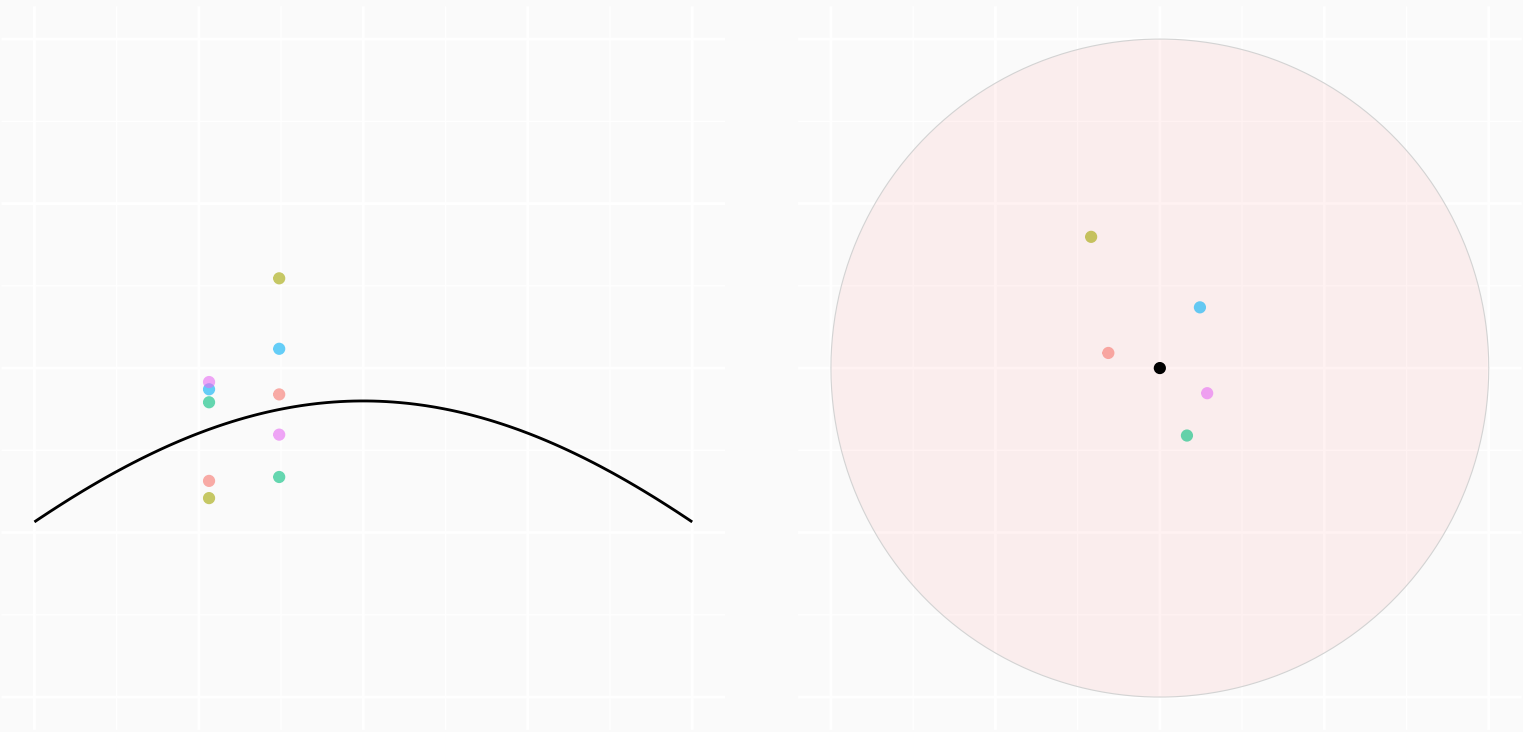


Figure 1: Right: $\mu(X_1)$ and Y_1 are x-coordinates; $\mu(X_2)$ and Y_2 are y-coordinates.

We can plot all these replications on top of each other, too.
Here's 5 replications.

The two-observation case

To help with visualization, we'll look at the case with two observations.
This lets us plot our observation *vectors* Y as a point in the plane.

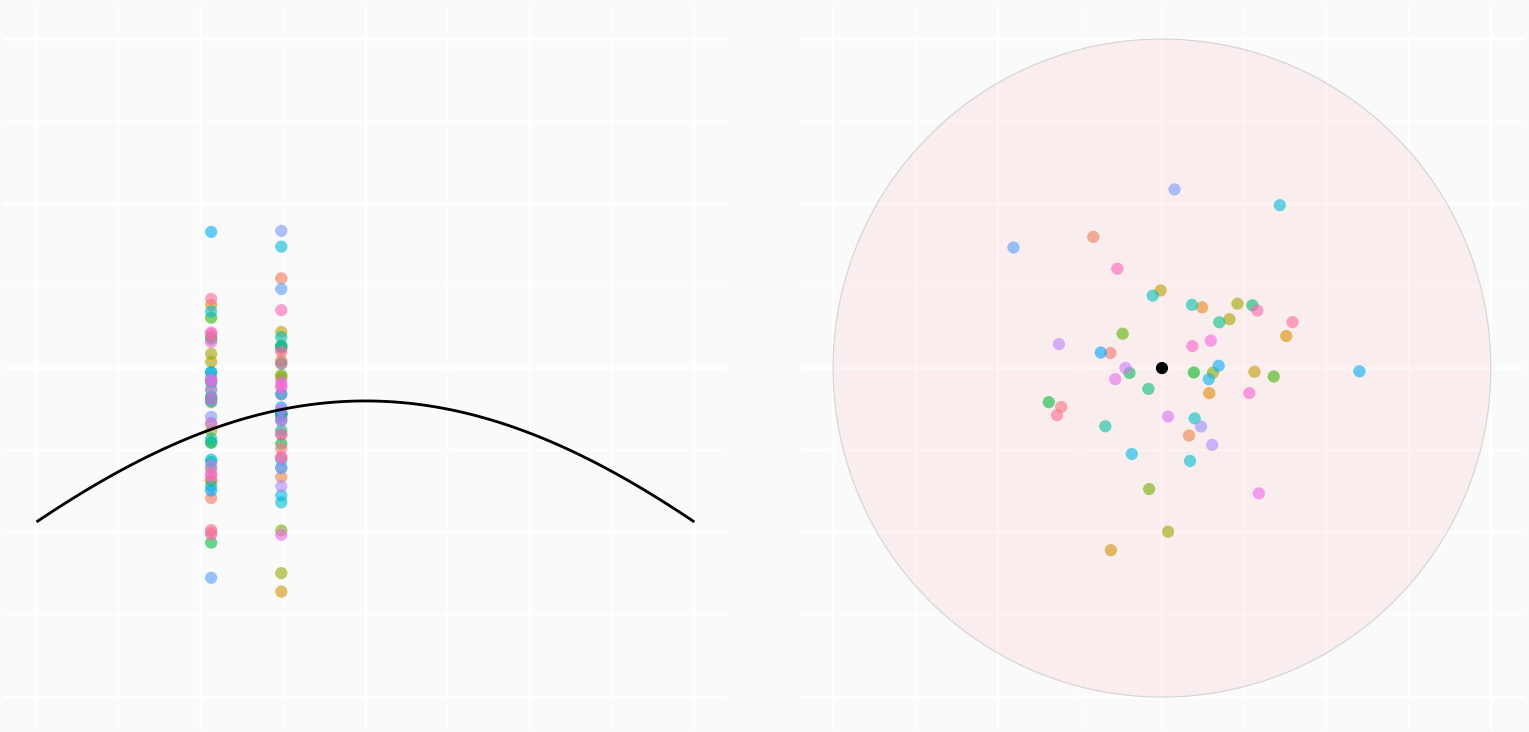


Figure 1: Right: $\mu(X_1)$ and Y_1 are x-coordinates; $\mu(X_2)$ and Y_2 are y-coordinates.

We can plot all these replications on top of each other, too.
Here's 50 replications.

The two-observation case

To help with visualization, we'll look at the case with two observations.
This lets us plot our observation *vectors* Y as a point in the plane.

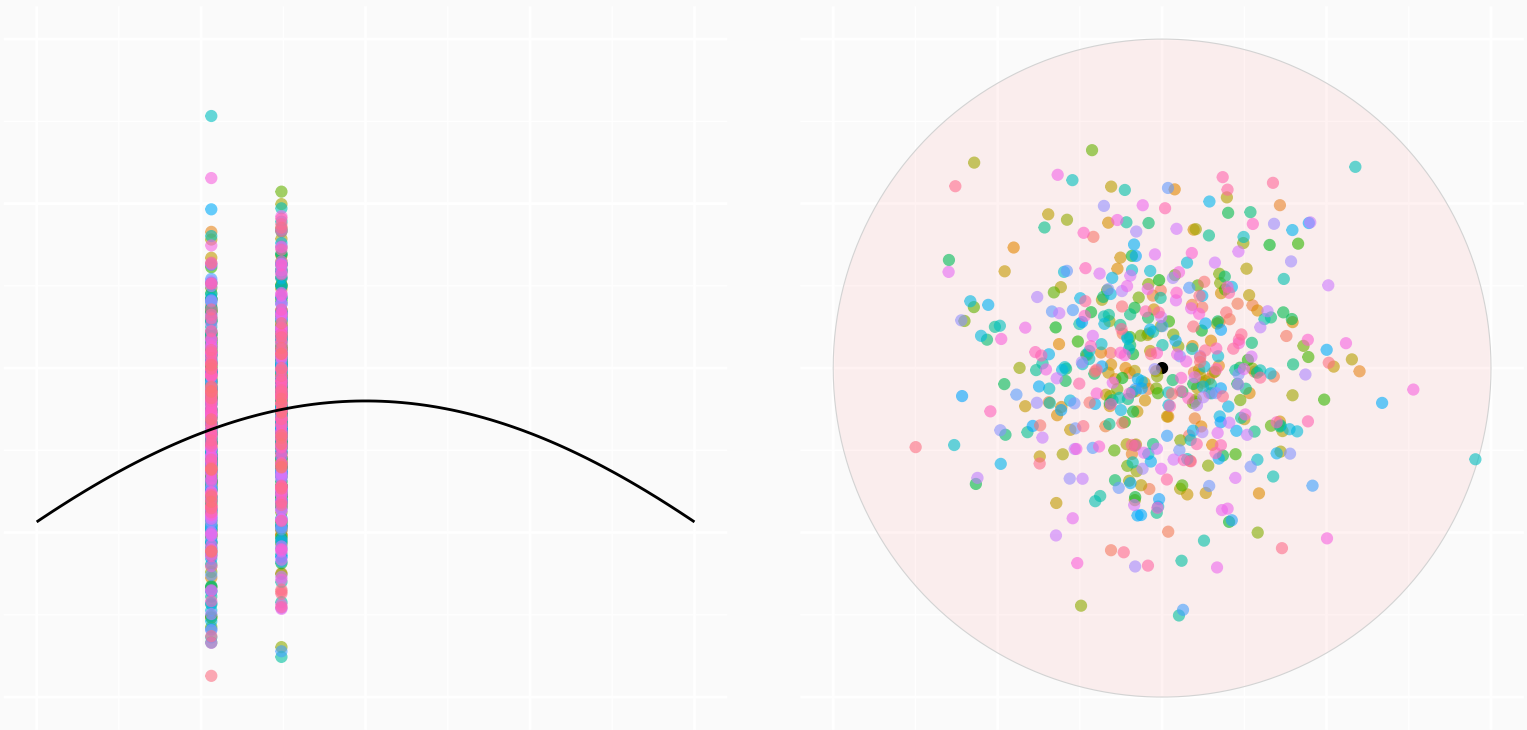
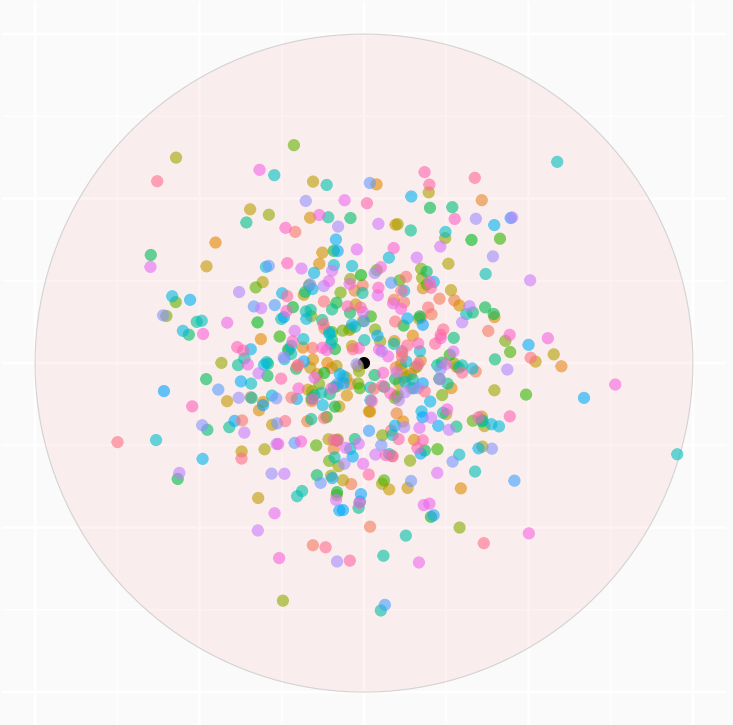


Figure 1: Right: $\mu(X_1)$ and Y_1 are x-coordinates; $\mu(X_2)$ and Y_2 are y-coordinates.

We can plot all these replications on top of each other, too.
Here's 500 replications.

The distribution of our observation vector

We see our observation vector Y is distributed in a kind of sphere around its mean μ.
That's because our *noise vector* $\varepsilon = Y - \mu$ is distributed in a sphere around zero.



The distribution of a vector of independent normals is *spherically symmetric*.
Its probability density depends on *length* but not *angle*.

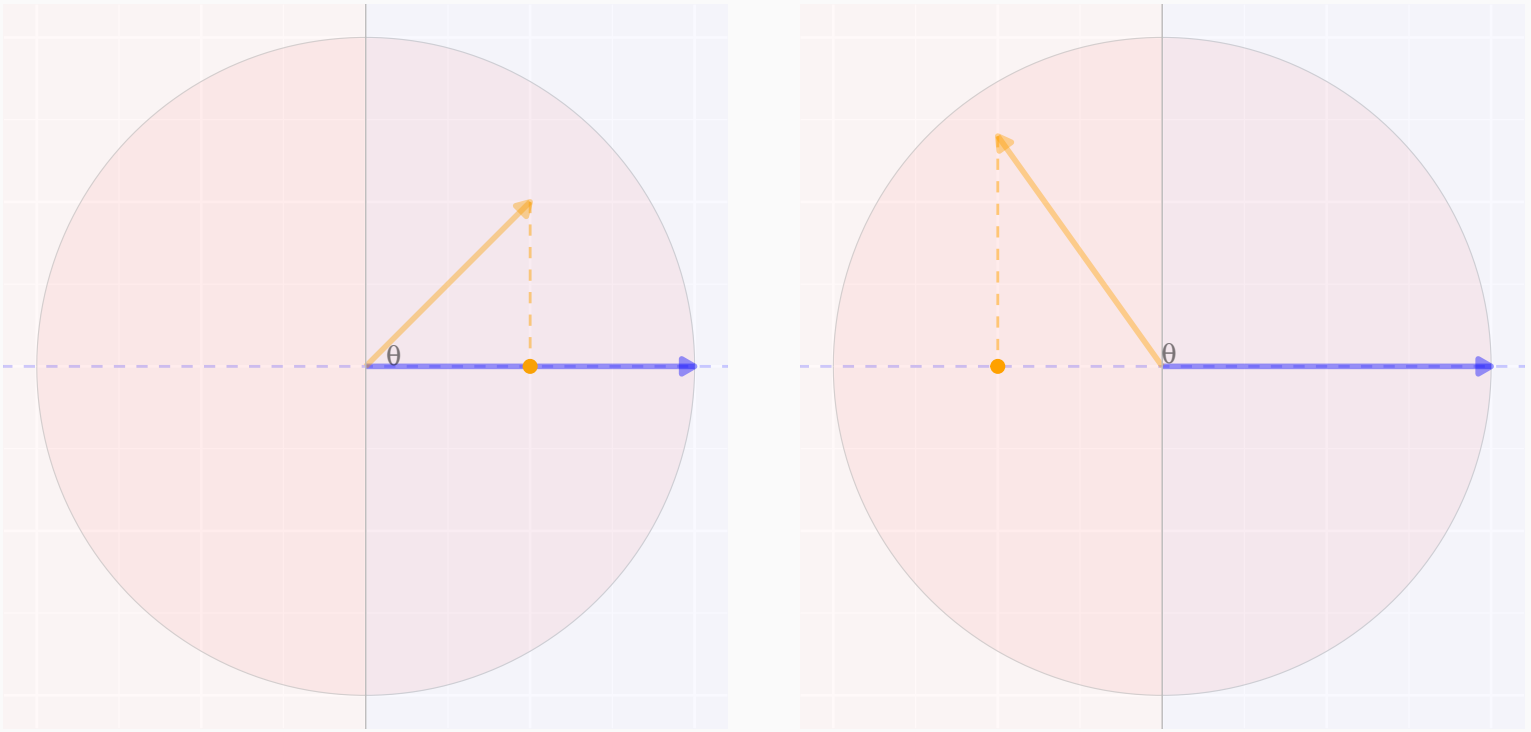
$$f_\varepsilon(x) = f_{\varepsilon_1}(x_1) \cdots f_{\varepsilon_n}(x_n) \propto e^{-\frac{x_1^2}{2\sigma^2}} \cdots e^{-\frac{x_n^2}{2\sigma^2}} = e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}} = e^{-\frac{\|x\|_2^2}{2\sigma^2}}$$

Projections

We'll be interested in *dot products*[1] of our noise vector ε with other vectors v.

$$\langle \varepsilon, v \rangle = \|\varepsilon\| \|v\| \cos(\theta)$$

When the other vector has length $\|v\| = 1$, that's the length of a *projection*.

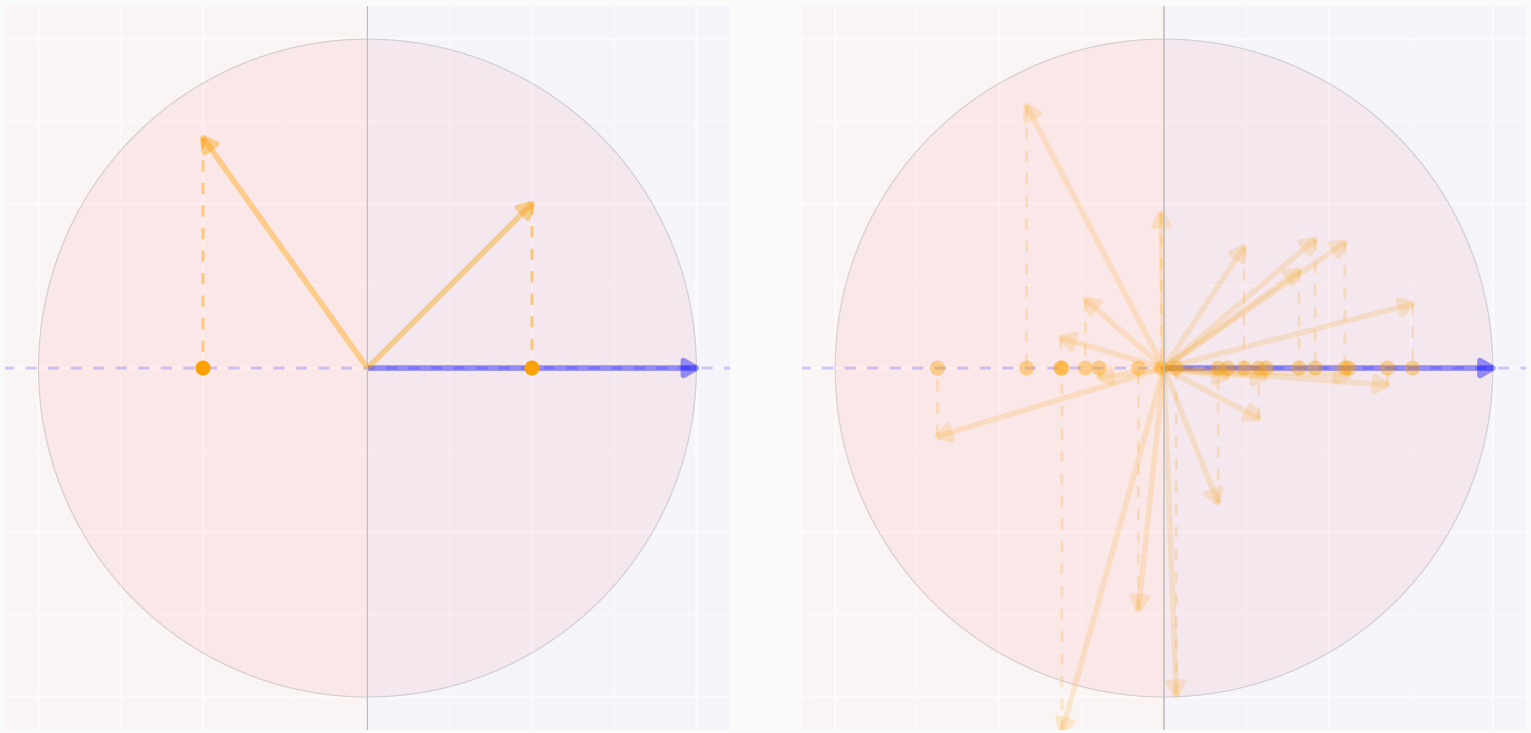


It's the length of the *component* of ε in the same direction as v.
Sometimes this will be positive; sometimes it'll be negative.

[1] Today, when we write $\langle u, v \rangle$ and $\|v\|$ we mean the dot product $\langle u, v \rangle_2$ and euclidean norm $\|v\|_2$ respectively.

The distribution of a projection

When we project onto one of the coordinate axes, we just get an element of our vector.

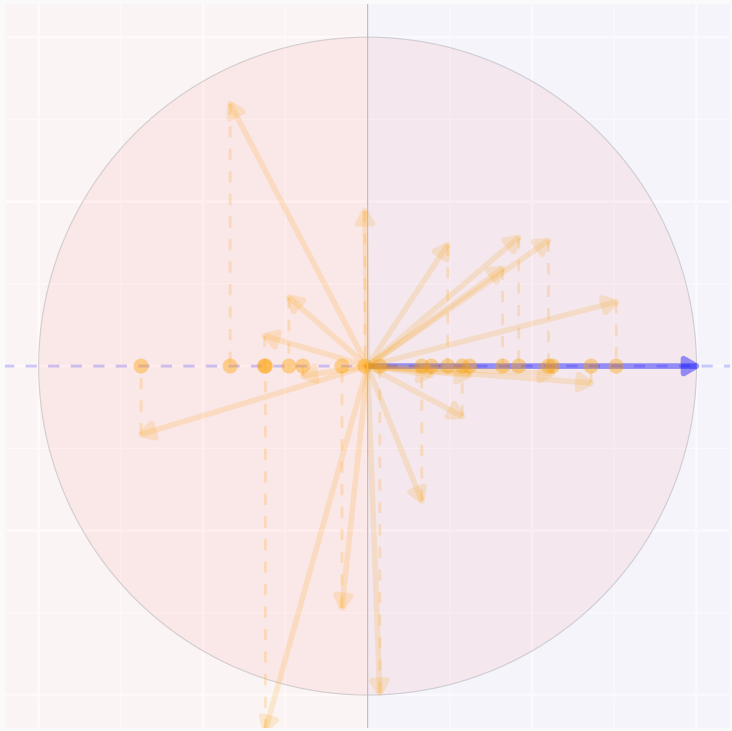


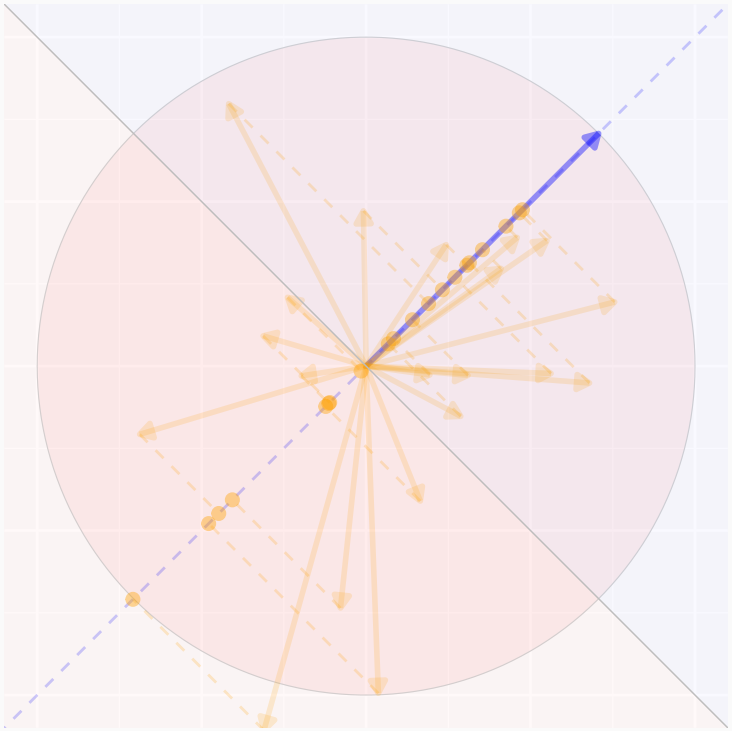
Here we project onto the first axis, so we get the first element of our noise vector ε. We know its distribution. On the right above, there's 20 samples from it.

$$\langle \varepsilon, v \rangle = \varepsilon_1 \sim N(0, \sigma^2)$$

The distribution of a projection

Our distribution is *spherically symmetric*, so there's nothing special about projection onto the axes.





The dot product still has the same distribution as our noise vector's elements.

$$\langle \varepsilon, v \rangle \sim N(0, \sigma^2) \quad \text{just like} \quad \varepsilon_i \sim N(0, \sigma^2)$$

High dimensional projections

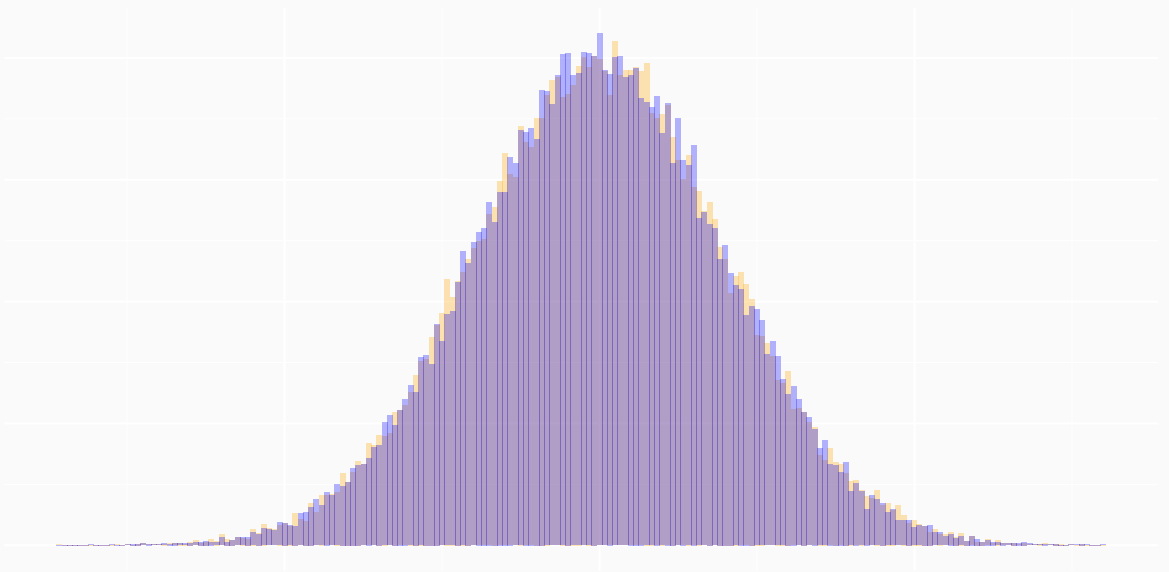


Figure 2: Overlaid histograms of 50k replications of ε_1 and $\langle \varepsilon, v/\|v\| \rangle$

We looked at this in 2D because that's what we can see, but it's true generally.

$$\left\langle \varepsilon, \frac{v}{\|v\|} \right\rangle \sim N(0, \sigma^2) \quad \text{when} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

If we have a vector of independent and identically distributed normals with mean zero, the length its projection in any one direction has that same distribution.

This is what makes it so pleasant to think about gaussian noise.

Warm Up

Least Squares in Models with Two Curves

Warm Up

Least Squares in Models with Two Curves

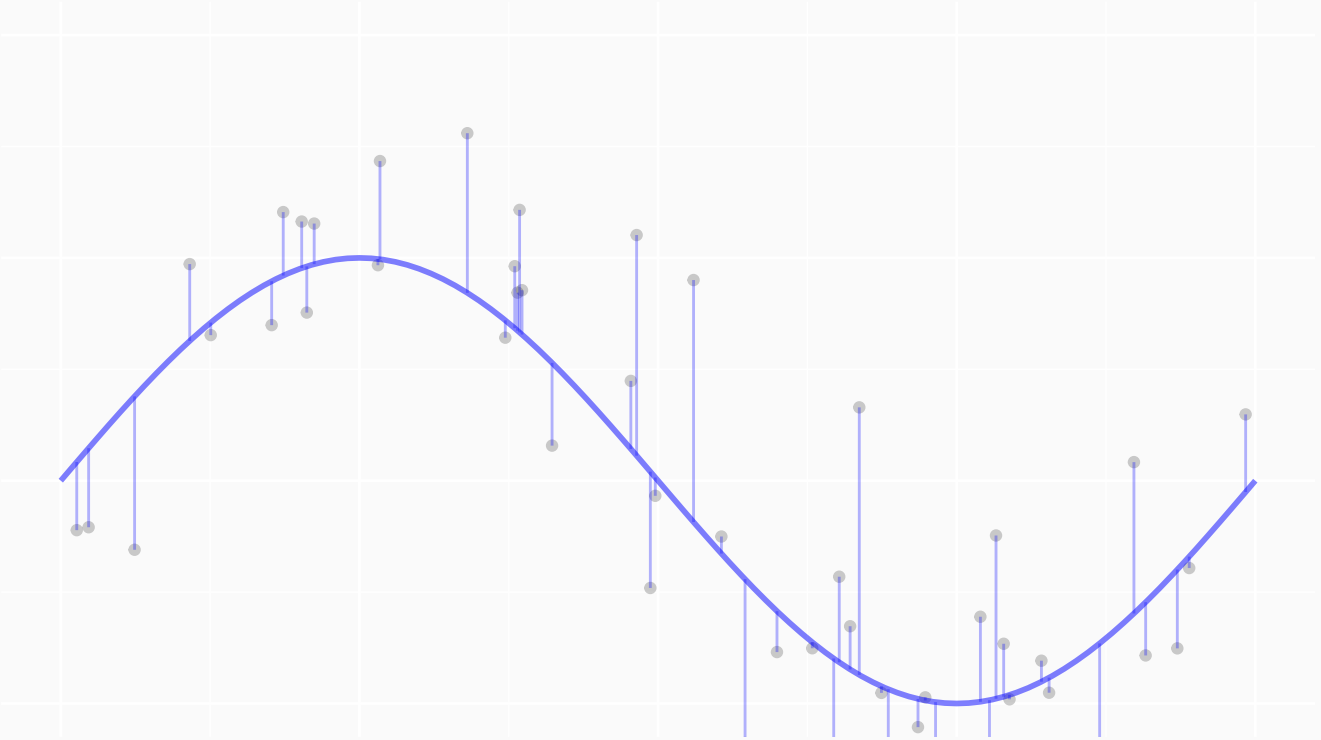
The Geometry of Least Squares

Visualization

This is what we minimize when we do least squares.

$$\ell(m) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2.$$

It's not *mean absolute error*, where we sum the heights of all these sticks.

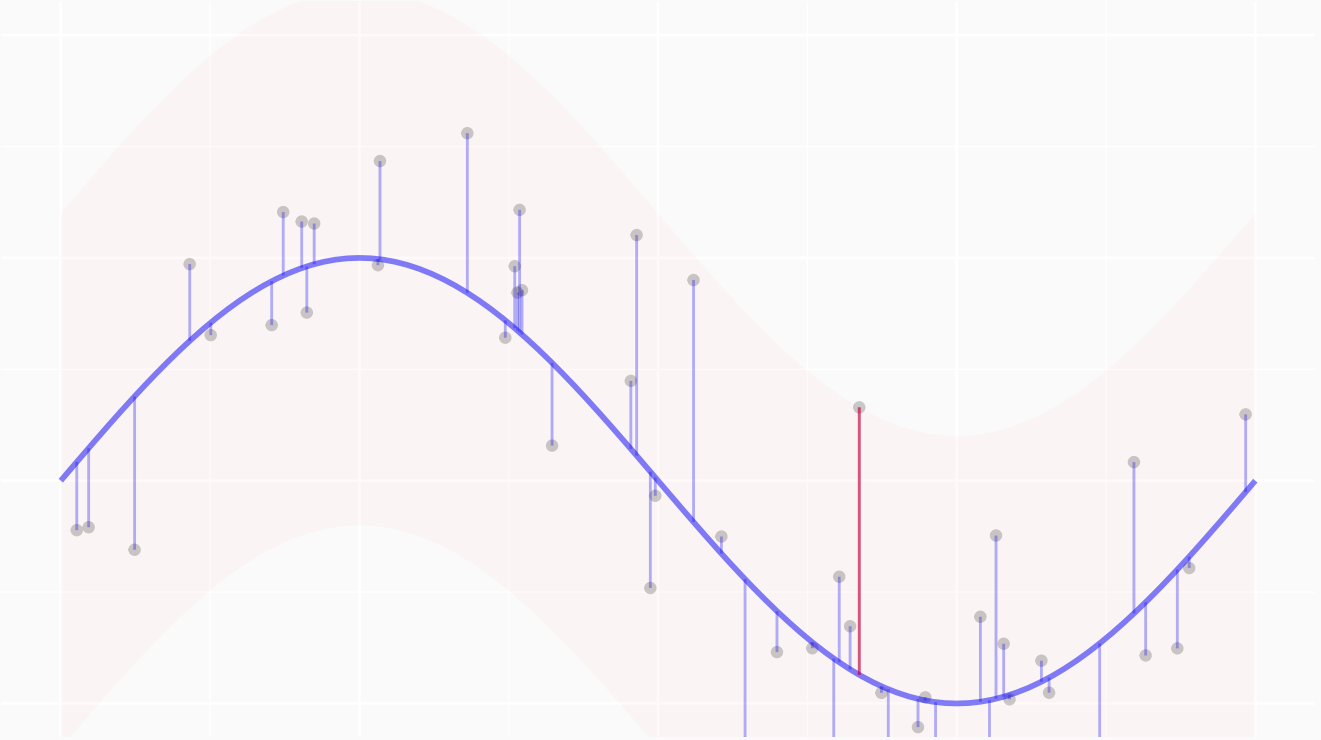


Visualization

This is what we minimize when we do least squares.

$$\ell(m) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2.$$

It's not *maximal absolute error*, where we take the biggest one.

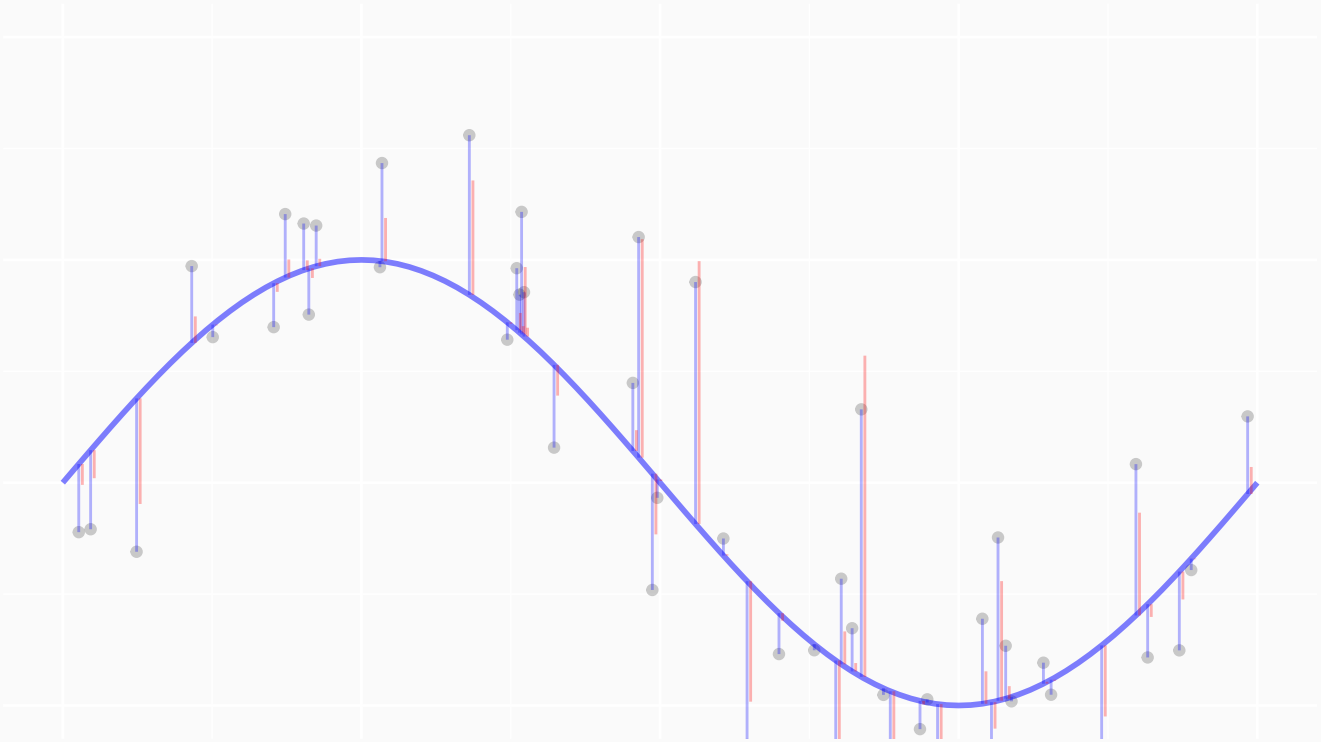


Visualization

This is what we minimize when we do least squares.

$$\ell(m) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2.$$

It's this, where we sum their squares.



We like squared error loss, in part, because it's easy to understand.

Comparing curves using squared error loss

If we compare the loss at any curve to the loss at μ, we get something very simple.

$$n \times \{\ell(m) - \ell(\mu)\} = \|m - \mu\|^2 - 2\langle \varepsilon,\ m - \mu \rangle$$

- It's the squared norm of the difference our our curve and μ.
- Plus a mean-zero term: the dot product of this difference with our noise vector.
  - Or, to be more precise, $-2\times$ this inner product.

Derivation.

Comparing curves using squared error loss

If we compare the loss at any curve to the loss at μ, we get something very simple.

$$n \times \{\ell(m) - \ell(\mu)\} = \|m - \mu\|^2 - 2\langle \varepsilon,\ m - \mu \rangle$$

- It's the squared norm of the difference our our curve and μ.
- Plus a mean-zero term: the dot product of this difference with our noise vector.
  - Or, to be more precise, $-2\times$ this inner product.

Derivation.

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

Least squares with two curves

Suppose our model contains two curves. And one is μ.

$$\mathcal{M} = \{m,\ \mu\}$$

We'll choose the wrong one, i.e. m, if it beats μ. If it has smaller loss.

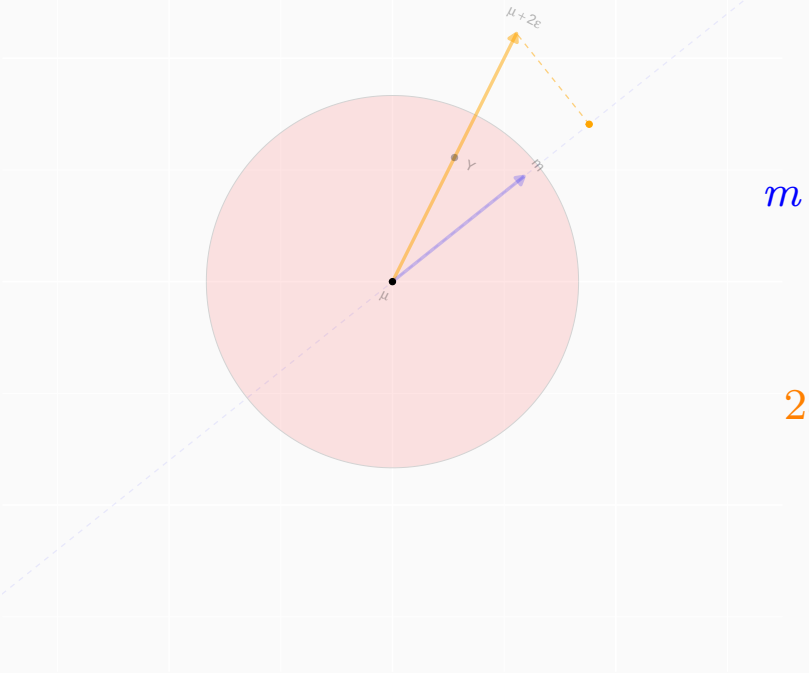
$$\hat{\mu} = m \quad \Longleftarrow \quad \ell(\mu) > \ell(m)$$

And equivalently if the loss difference we calculated is negative.

$$\begin{aligned} \ell(\mu) > \ell(m) \quad &\iff \quad 0 > \ell(m) - \ell(\mu) \\ \text{i.e.} \quad &\iff \quad 0 > \|m-\mu\|^2 - 2\langle \varepsilon, m-\mu \rangle \end{aligned}$$

And equivalently if this difference, divided by any positive number, is negative.
For example, divided by the distance between m and μ.

$$\begin{aligned} 0 > \ell(m) - \ell(\mu) \quad &\iff \quad 0 > \frac{\ell(m) - \ell(\mu)}{\|m-\mu\|} \\ \text{i.e.} \quad &\iff \quad 0 > \|m-\mu\| - 2\left\langle \varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle \end{aligned}$$



Featured.

μ. The center of the diagram.

$m - \mu$. The arrow from $\mu \to m$.

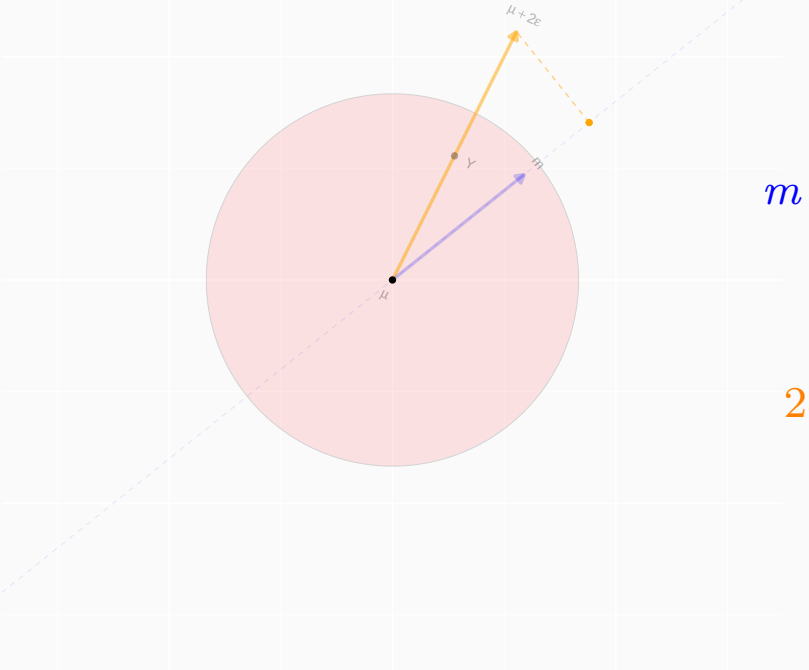
2ε. The arrow from $\mu \to \mu + 2\epsilon$ with midpoint $Y = \mu + \epsilon$.

$2\epsilon_{\text{proj}}$. The orange dot on the line through μ and m shows the projection of 2ϵ onto $m - \mu$.

We select the wrong curve iff the orange dot is past the tip of the blue arrow.

$$\hat{\mu} \neq \mu \iff \underbrace{\left\langle 2\varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle}_{=\|2\epsilon_{\text{proj}}\|} > \|m - \mu\|$$

In this case, that's what happens.



Featured.

μ. The center of the diagram.

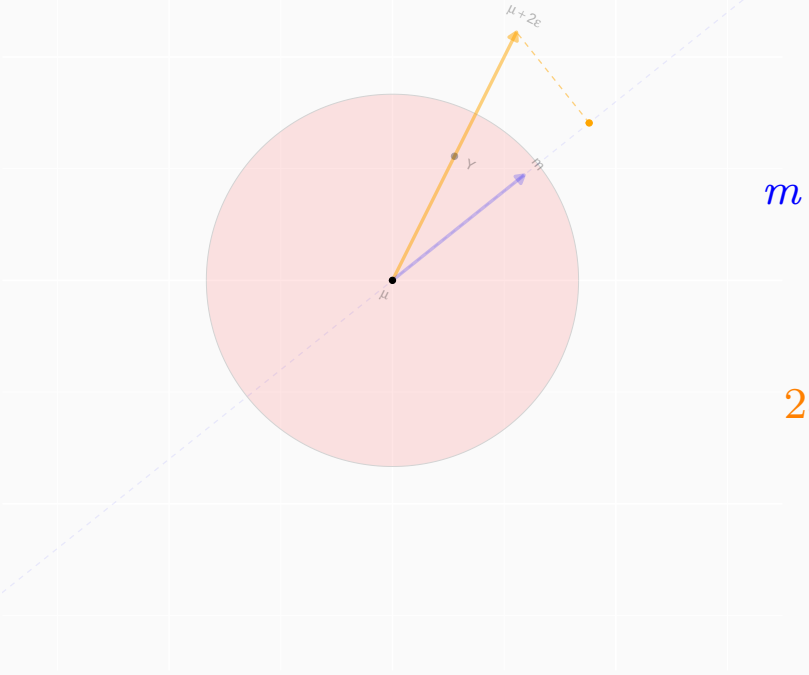
$m - \mu$. The arrow from $\mu \to m$.

2ε. The arrow from $\mu \to \mu + 2\epsilon$ with midpoint $Y = \mu + \epsilon$.

$2\epsilon_{\text{proj}}$. The orange dot on the line through μ and m shows the projection of 2ϵ onto $m - \mu$.

What's the point of all this?

- We've started with a statement that's visually obvious. Y is closer to m than to μ.
- And we've rephrased it as a statement that's easy to think about probabilistically.
- It's a statement about the size of a projection of a gaussian vector.
- i.e. a statement about the size of a gaussian random variable.



Featured.

μ. The center of the diagram.

$m - \mu$. The arrow from $\mu \to m$.

2ε. The arrow from $\mu \to \mu + 2\epsilon$ with midpoint $Y = \mu + \epsilon$.

$2\epsilon_{\text{proj}}$. The orange dot on the line through μ and m shows the projection of 2ϵ onto $m - \mu$.

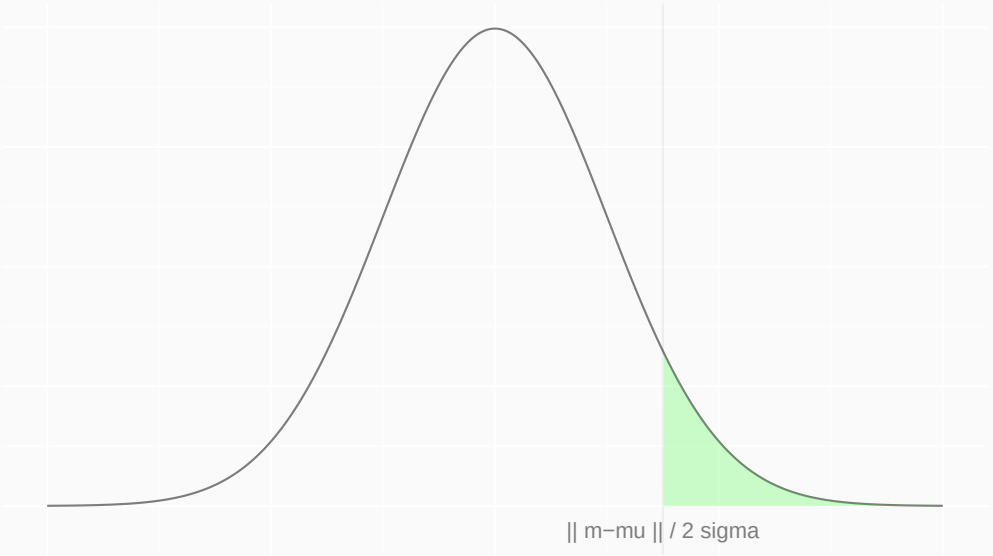
We select the wrong curve iff the orange dot is past the tip of the blue arrow.

$$\hat{\mu} \neq \mu \iff \underbrace{\left\langle 2\varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle}_{=\|2\epsilon_{\text{proj}}\|} > \|m - \mu\|$$

The **probability we choose wrong** is the probability that this happens.

$$P(\hat{\mu} \neq \mu) = P\left(\left\langle \varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle > \frac{\|m-\mu\|}{2} \right)$$

The Probability We Choose Wrong



That's easy to calculate in terms of the standard normal distribution.

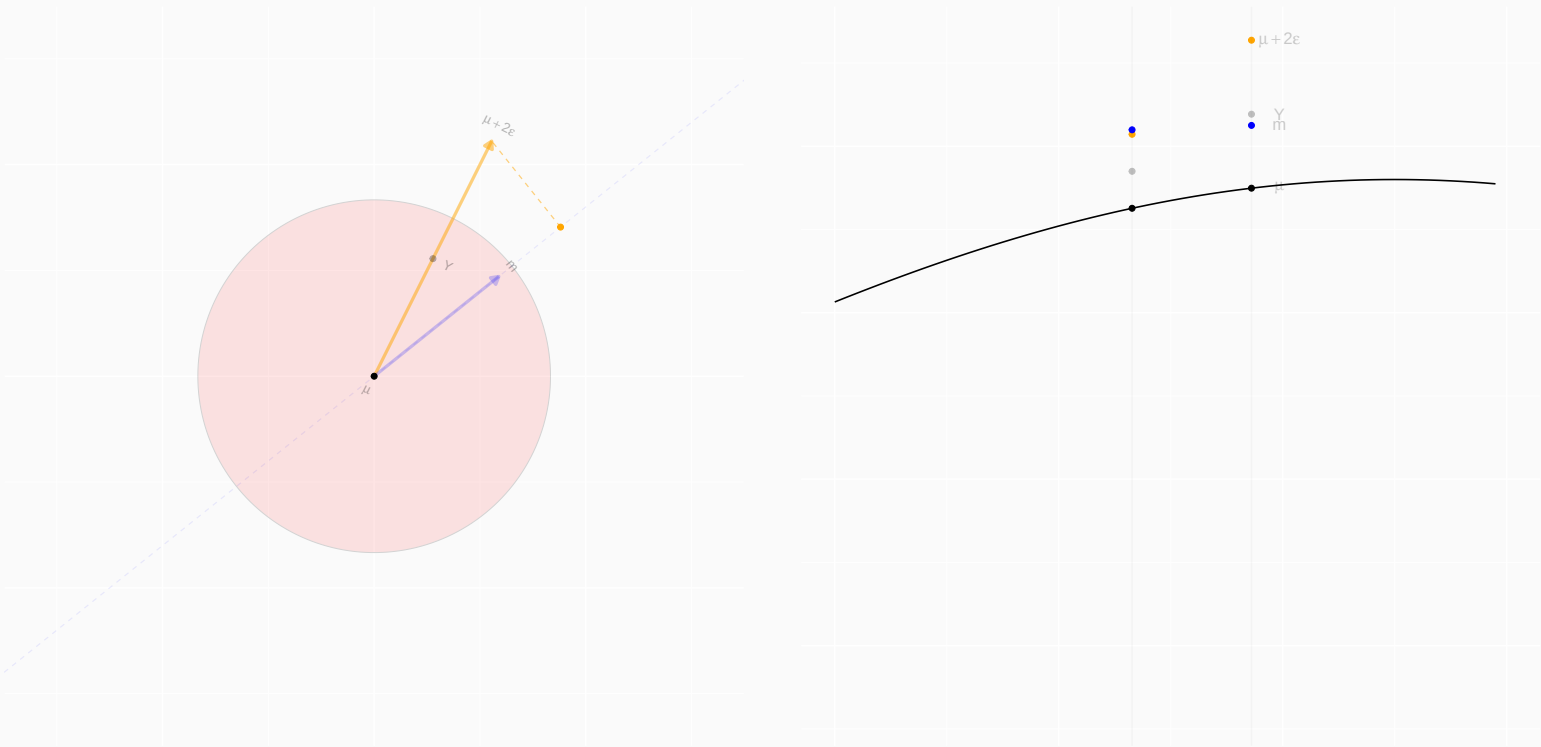
$$P\left(\hat{\mu} \neq \mu\right) = P\left(Z_m > \frac{\|m-\mu\|}{2\sigma}\right) \quad \text{where} \quad Z_m = \frac{\left\langle \varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle}{\sigma} \sim N(0,1)$$

Warm Up

Least Squares in Models with Two Curves

Visualizing What Happens

The Example We've Been Looking At



- We can see that m overshoots Y at X_1 and undershoots at X_2.
- But it's close. Closer, in aggregate than μ is. So that's what we select.
- That's what happens in this case. But the noise vector ε is random.
- Let's repeat this, drawing a new noise vector ε from the same distribution.

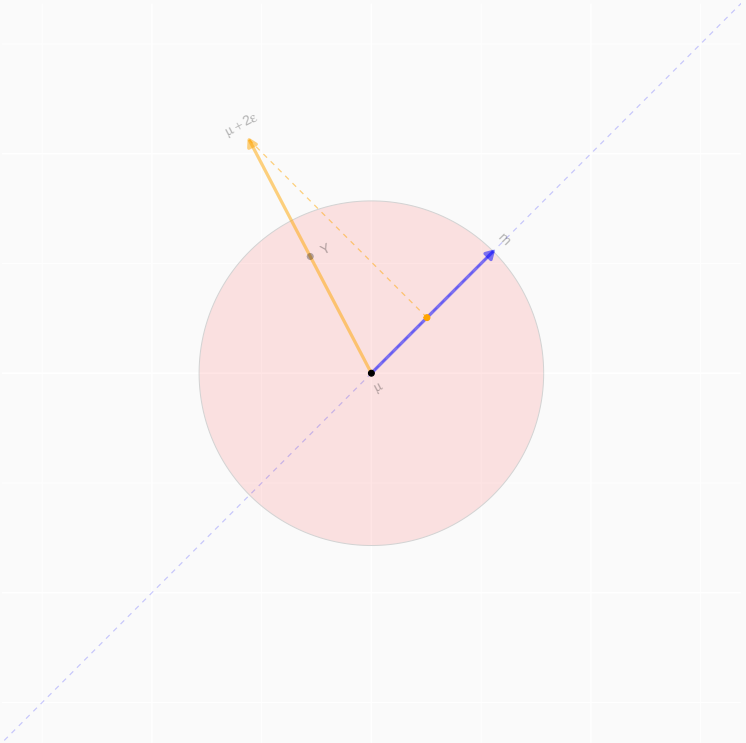
Exercise.

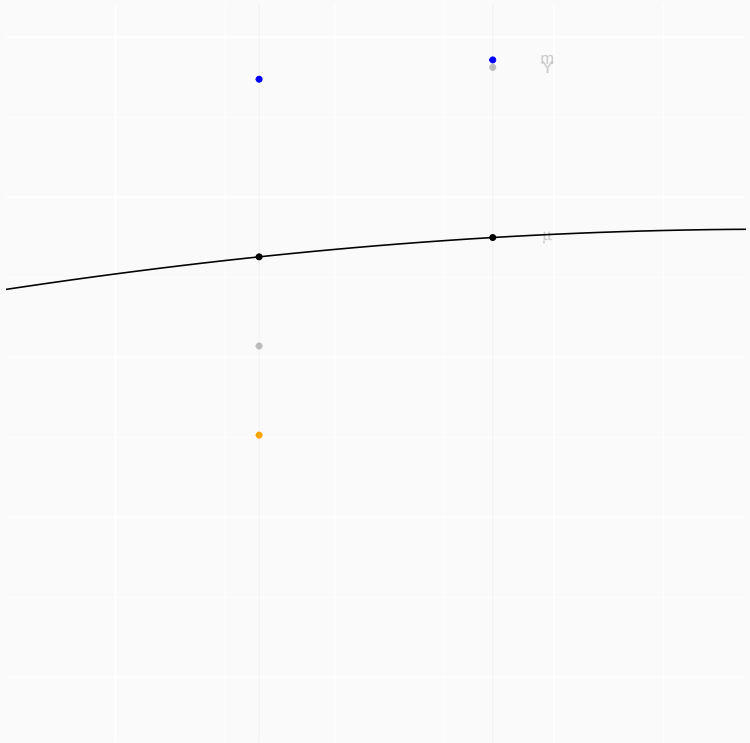
ε is *random*, and therefore so is our choice $\hat{\mu}$.

Here's what we see for a ten random draws of ε.

How often does our least squares estimator $\hat{\mu}$ choose correctly?

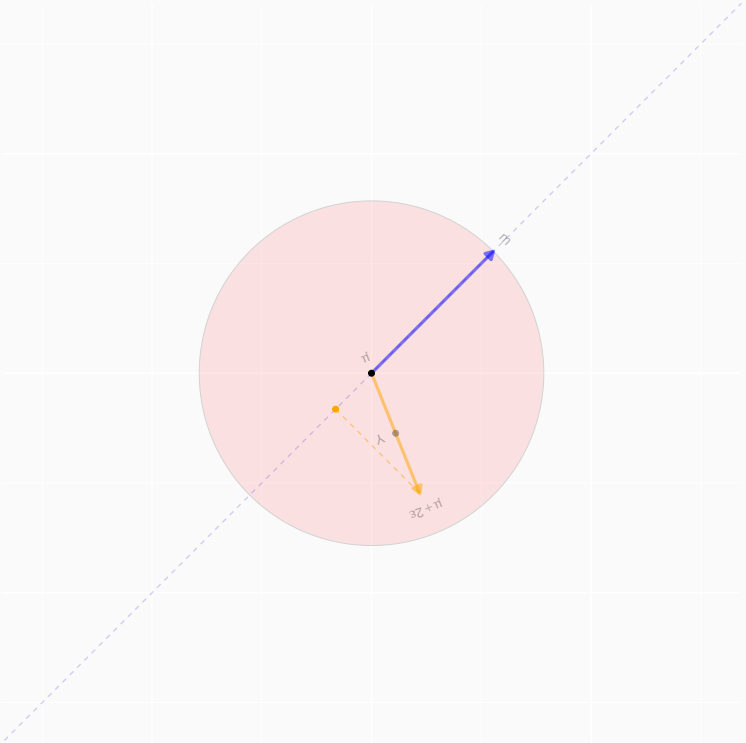
Sample 1

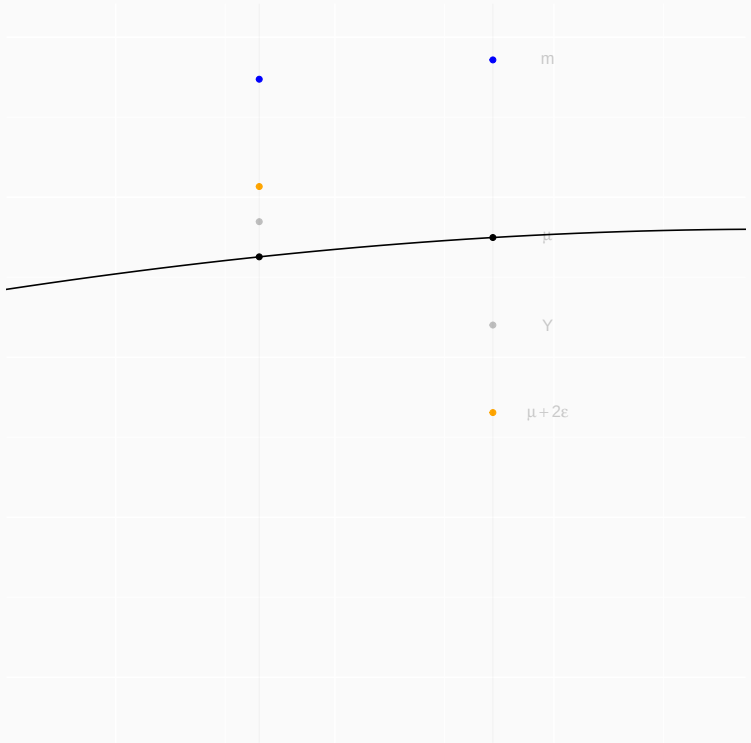




$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - \left\langle 2\varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \begin{array}{lll} > 0 & \Longrightarrow & \hat{\mu} = \mu. \\ < 0 & \Longrightarrow & \hat{\mu} = m. \end{array}$$

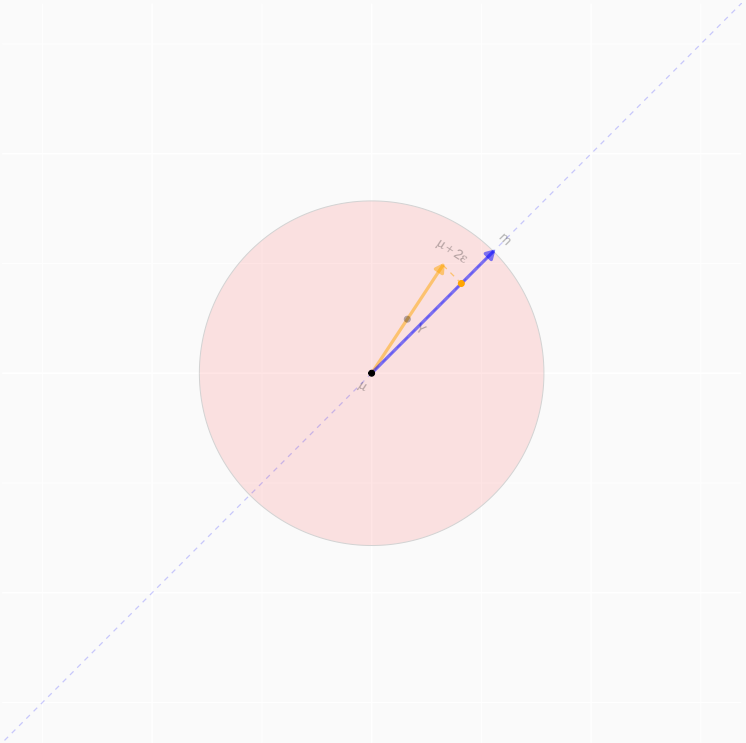
Sample 2

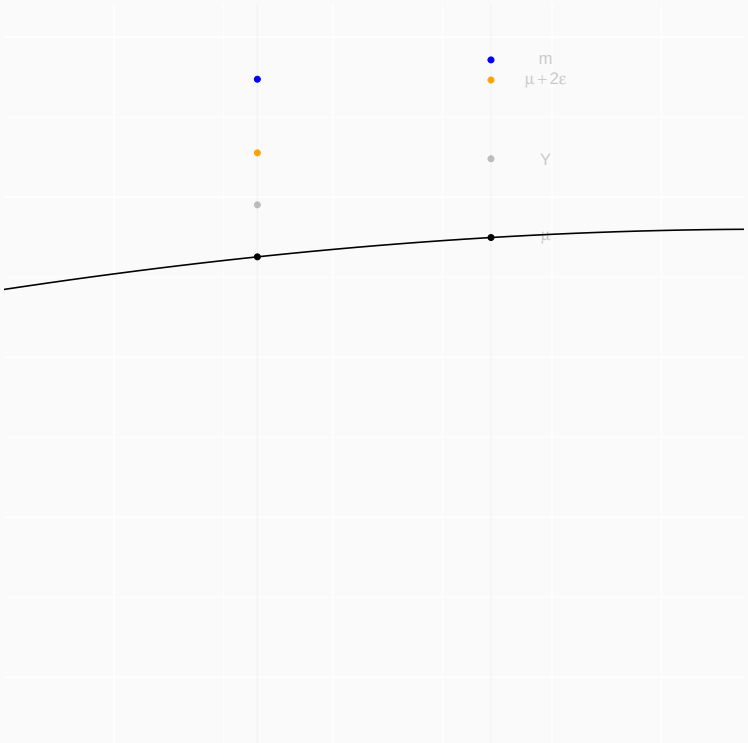




$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - \left\langle 2\varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \begin{array}{lll} > 0 & \Longrightarrow & \hat{\mu} = \mu. \\ < 0 & \Longrightarrow & \hat{\mu} = m. \end{array}$$

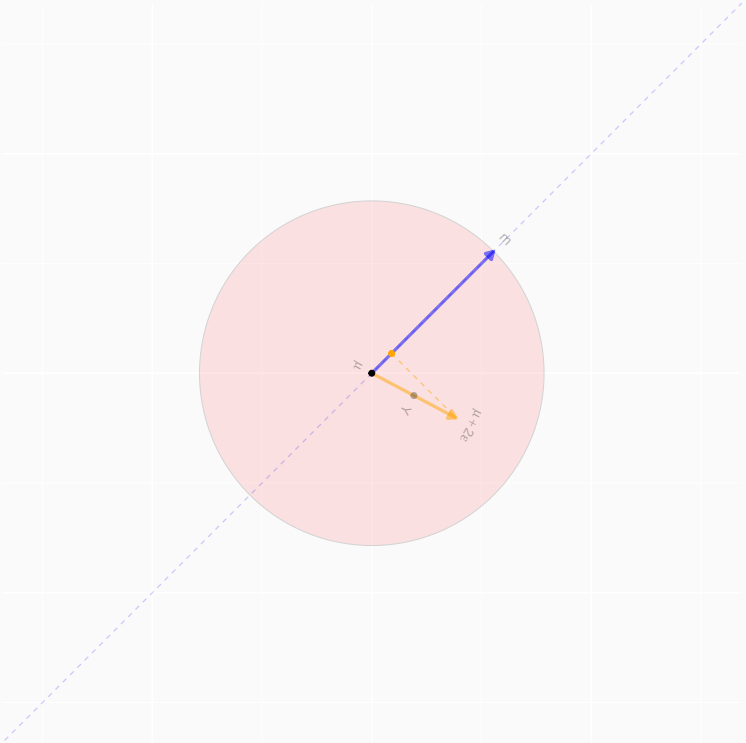
Sample 3





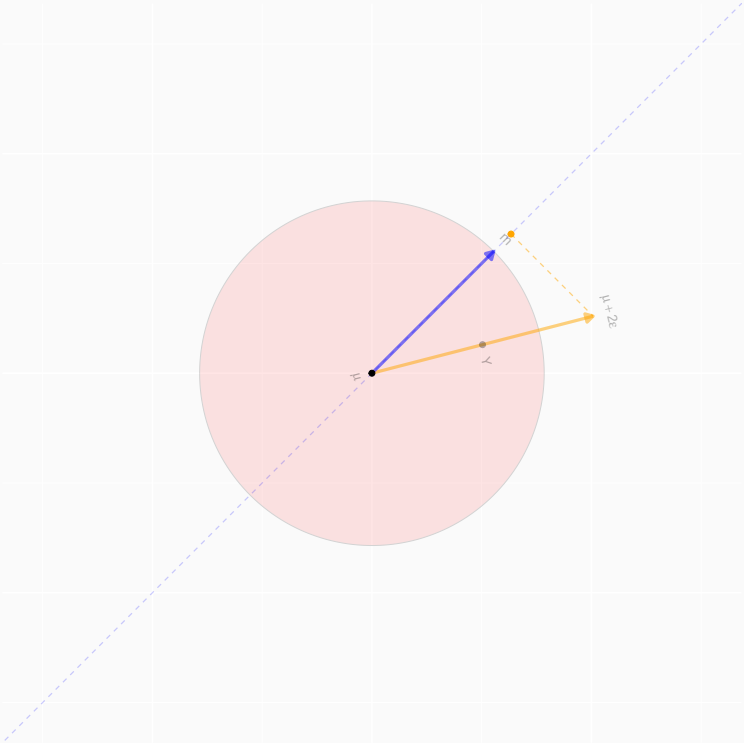
$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - \left\langle 2\varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \begin{array}{lll} > 0 & \Longrightarrow & \hat{\mu} = \mu. \\ < 0 & \Longrightarrow & \hat{\mu} = m. \end{array}$$

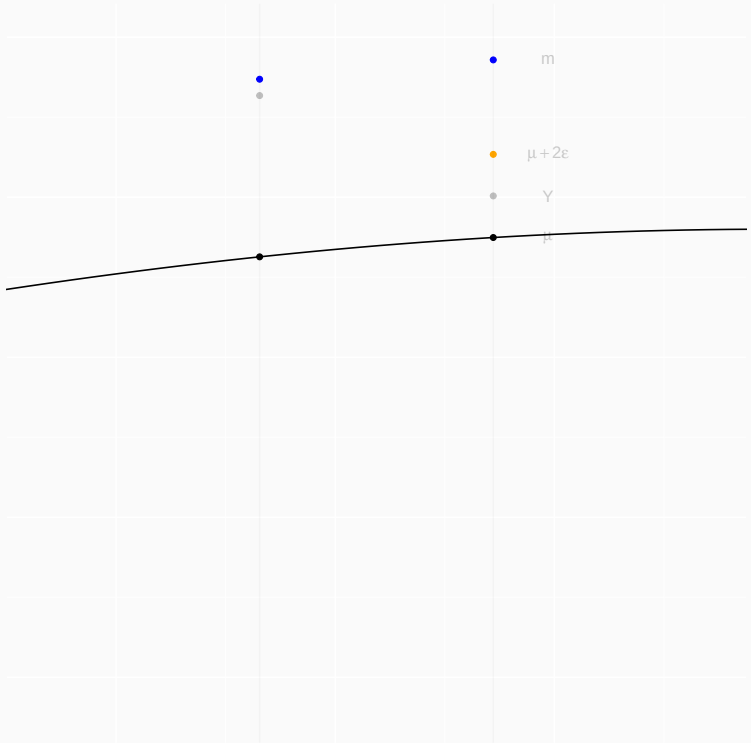
Sample 4



$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - \left\langle 2\varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \begin{array}{lll} > 0 & \Longrightarrow & \hat{\mu} = \mu. \\ < 0 & \Longrightarrow & \hat{\mu} = m. \end{array}$$

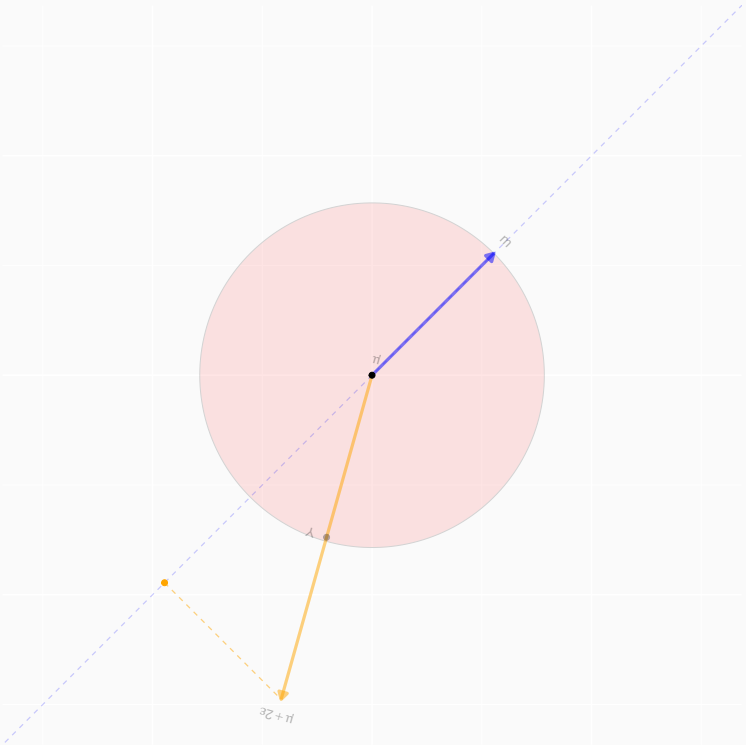
Sample 5

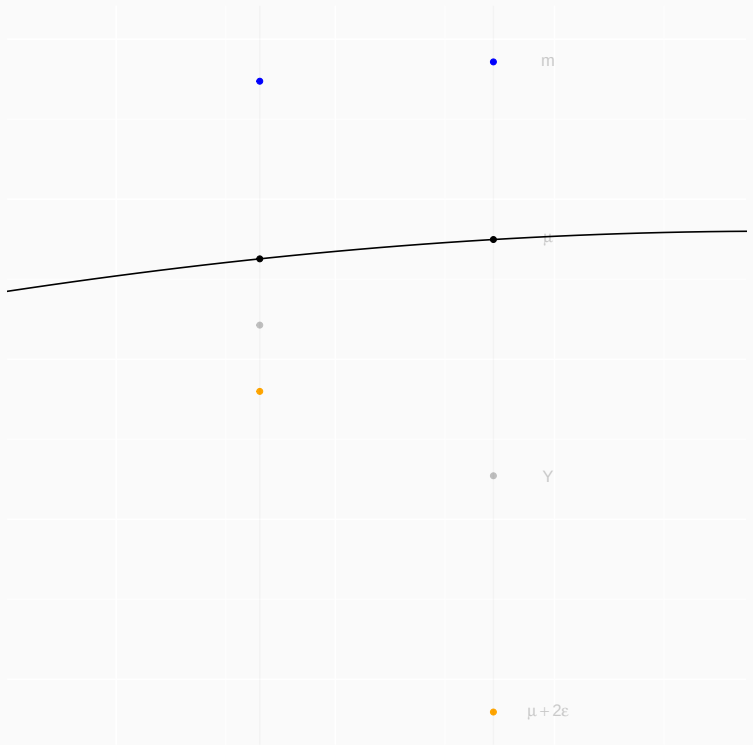




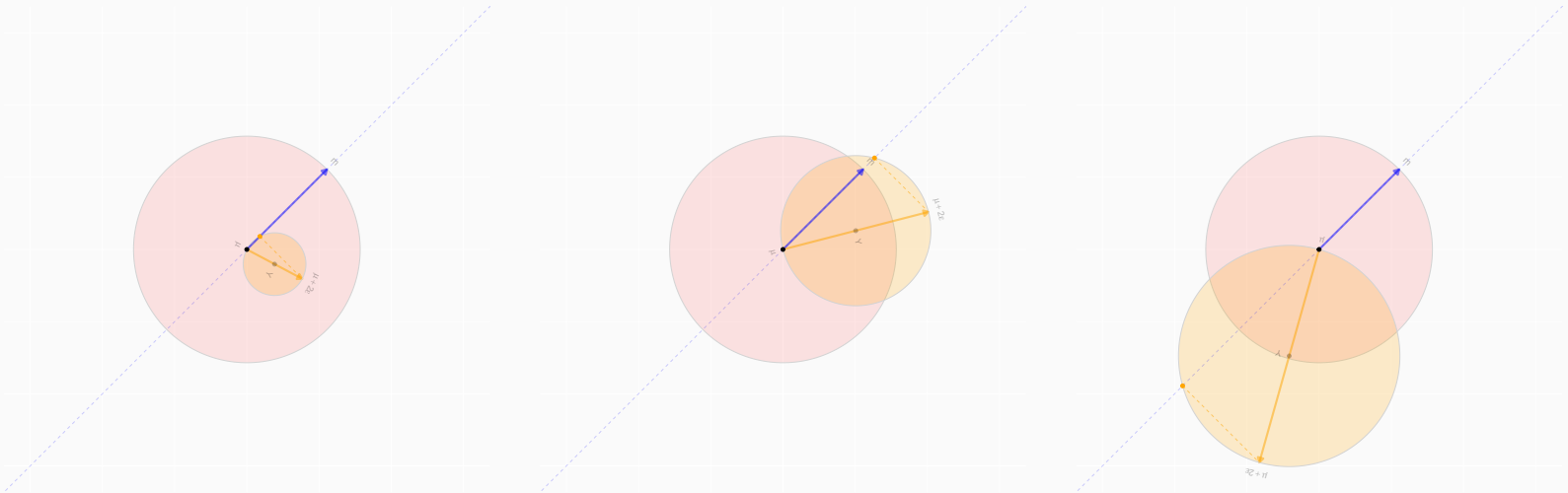
$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - \left\langle 2\varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \begin{array}{lll} > 0 & \Longrightarrow & \hat{\mu} = \mu. \\ < 0 & \Longrightarrow & \hat{\mu} = m. \end{array}$$

Sample 6

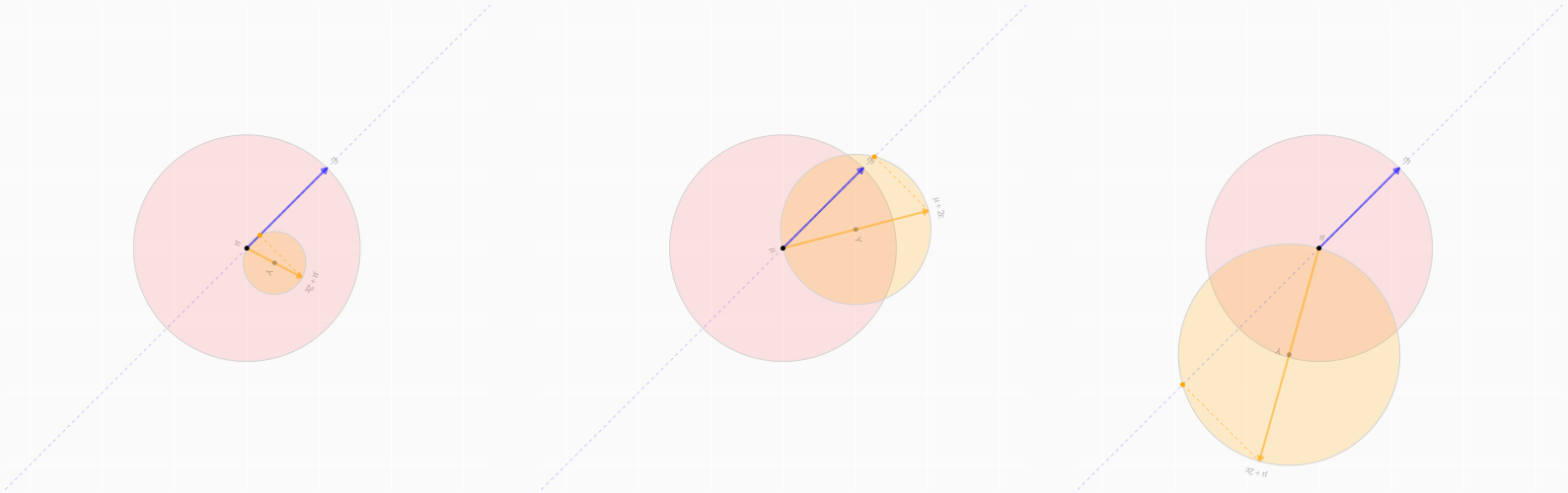




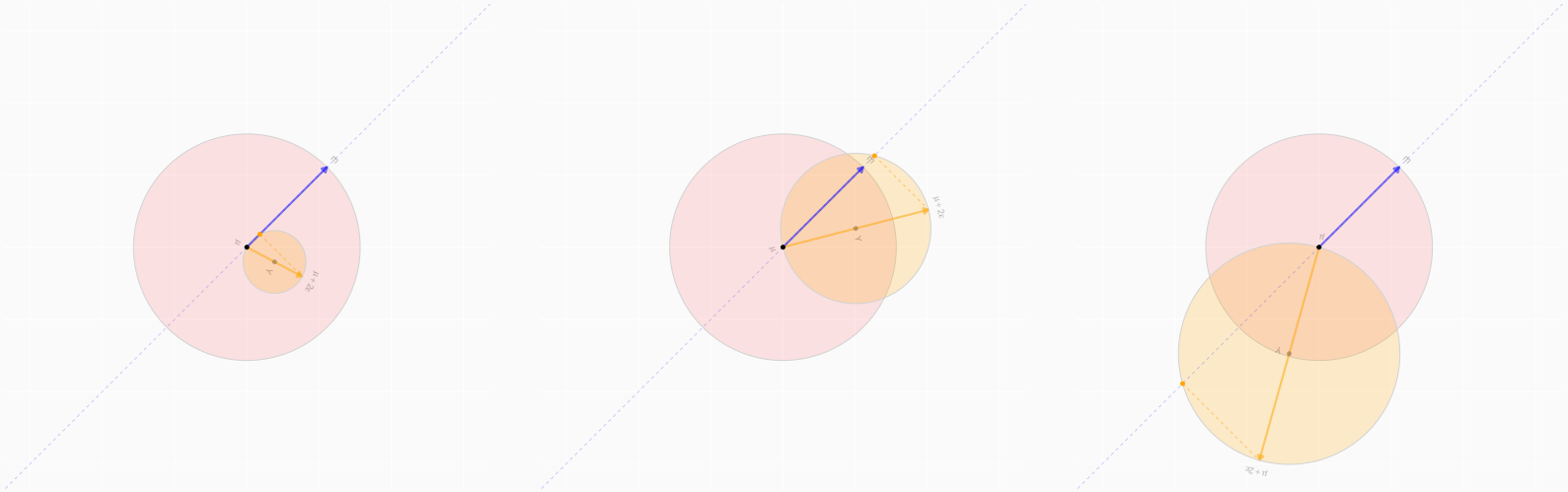
$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - \left\langle 2\varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \begin{array}{lll} > 0 & \Longrightarrow & \hat{\mu} = \mu. \\ < 0 & \Longrightarrow & \hat{\mu} = m. \end{array}$$



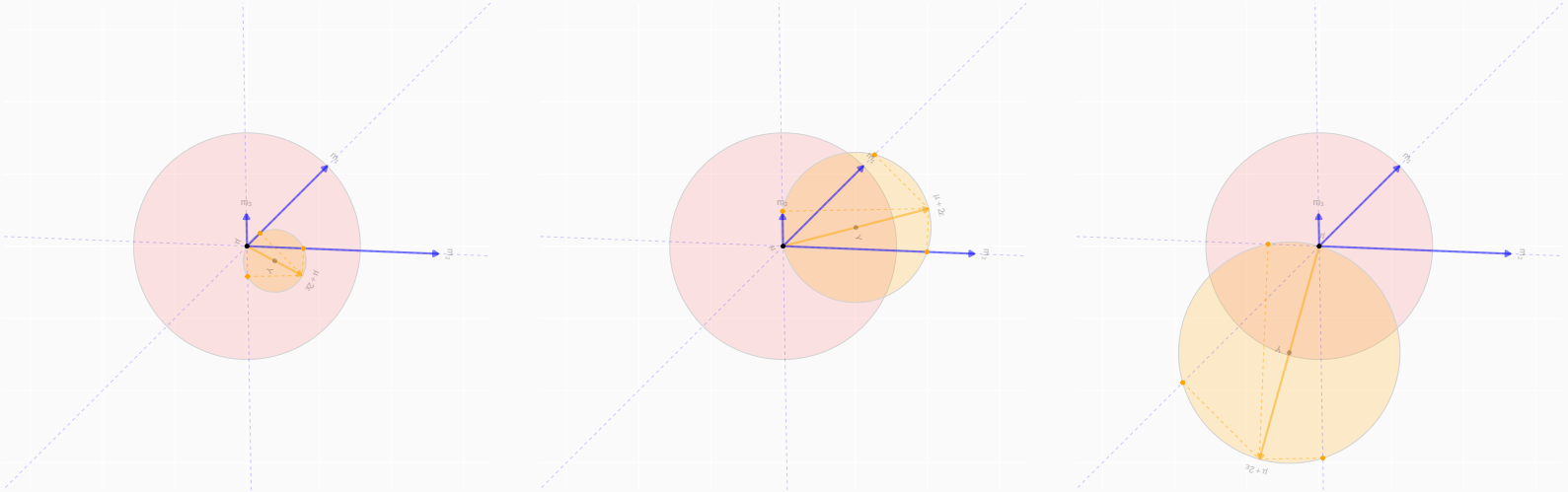
- We've been talking as if ...
  - the 'wrong curve' m were fixed. We've always drawn the same error vector $m - \mu$.
  - the noise vector ε varies. We've drawn a bunch of different ones.
- We've talked about which noise vectors ε that make us choose m.
- And the probability that we'll draw one of those.



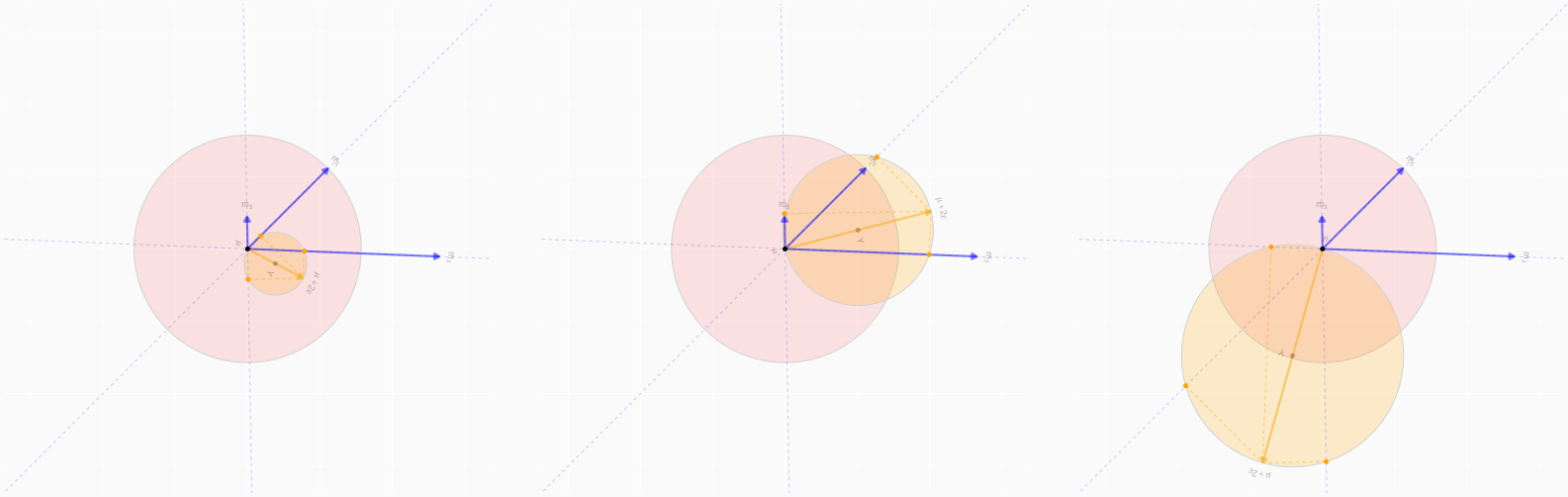
- To talk about realistic models, it helps to change perspective.
  - We'll think about the noise vector as fixed and the 'wrong curve' m varying.
  - Because in realistic models, we'll have a bunch of wrong curves.
- To help with this, I've tweaked the diagram.
- I've drawn in an orange circle around Y.
- **Q.** What is the significance of the orange circle
  Hint. Where is its center? And how far out does it go?



- **Q.** What is the significance of the orange circle?
- **A.** It's the 'overfitting zone'. The set of points closer to Y than μ is.
  - If m is in there, it beats μ. It fits the data better.
  - If it isn't, it doesn't. It fits worse.
- The better aligned our error $m - \mu$ is with the noise ε, the further out it gets to be.
- If it's perfectly aligned, i.e. if $m(X_i) \propto \varepsilon_i$, it gets to overshoot by a factor of 2.



- When we've more curves in our model, there's a better chance one is in the zone.
- Sometimes that's ok.
  - If the curve we choose is close to μ, it's no big deal.
  - e.g. if one had fallen in the zone in the left diagram.
- But it can be a problem.
  - If the curve we choose is far from μ, we're unhappy.
  - e.g. like we do in the middle diagram.



- It looks like a big part of this is the size of the zone.
  - i.e. the length $\|\varepsilon\|$ of our noise vector.
- That's where our 2D intuition fails us. In higher dimensions, things are different.
- The radius of the zone barely changes. $\|\varepsilon\|$ is approximately constant.

$$\frac{\|\varepsilon\|^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \approx \mathrm{E}[\varepsilon_i^2] = \sigma^2 \qquad \text{by the large of large numbers}$$

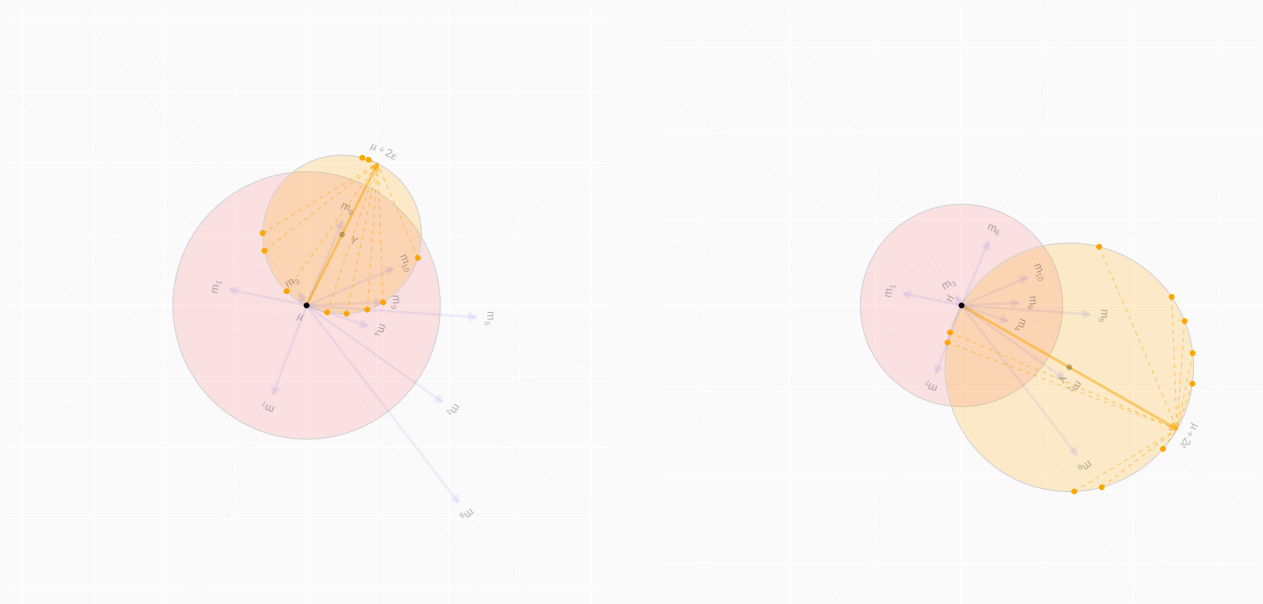
- But there are *tons* of directions. The chance we'll 'miss' is bigger than you'd think.
- Even if you take a lot of shots, i.e. even if your model $\mathcal{M}$ includes a lot of curves.

Least Squares in Finite Models

Least Squares in Finite Models

The Idea

When we've got a lot of curves in our model, it's likely that the overfitting zone will land on one or more of them.



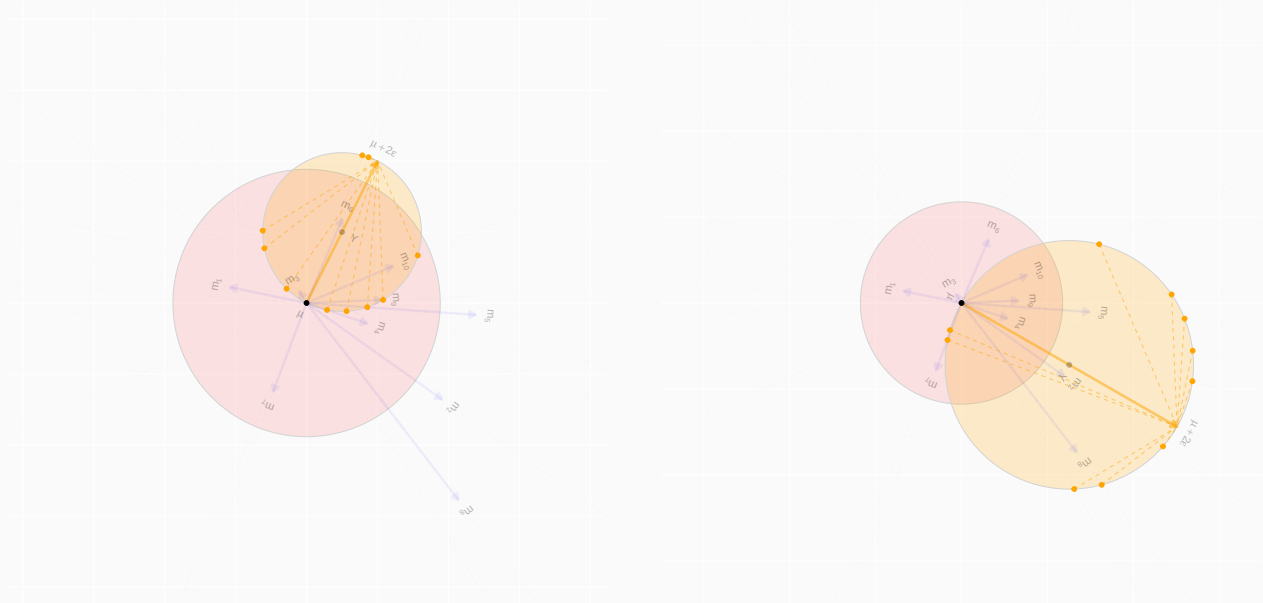
To understand our estimator $\hat{\mu}$, we use the property that it's *always* one of these curves in the zone. If μ is in the model, anyway.

$$\hat{\mu} \text{ minimizes } \ell(m) \text{ among } m \in \mathcal{M} \iff \ell(\hat{\mu}) - \ell(m) \leq 0 \quad \text{for all} \quad m \in \mathcal{M}$$

$$\implies \ell(\hat{\mu}) - \ell(\mu) \leq 0 \quad \text{if} \quad \mu \in \mathcal{M}$$

i.e. $\hat{\mu}$ is in the zone

When we've got a lot of curves in our model, it's likely that the overfitting zone will land on one or more of them.



We can be confident that $\hat{\mu}$ is close to μ when every curve m in the zone is. On the left, the claim that $\hat{\mu}$ is within our circle around μ is correct. On the right, it may not be.

Least Squares in Finite Models

Formal Argument

The characters

Old Friends

- (X_i, Y_i) for $i = 1 \ldots n$. The data.
- $\mu(x)$, the estimation target. A curve.
- $\mathcal{M}$, the model. A set of curves we hope contains μ.
- $\hat{\mu}$, our estimate. Some curve in the model, chosen because it fits the data.
- m, an anonymous curve. Whatever curve we're thinking about at the moment.

New Ones

- $\mathcal{M}_s$, a *neighborhood* of the target.
  - It's the subset of curves in our model that are close to μ.
  - We're trying to show that $\hat{\mu}$ is in it.
- $\mathcal{M} \setminus \mathcal{M}_s$, its complement.
  - It's the subset of curves in our model that aren't close to μ.
  - It's equivalent to show that $\hat{\mu}$ *is not* one of the curves in it.

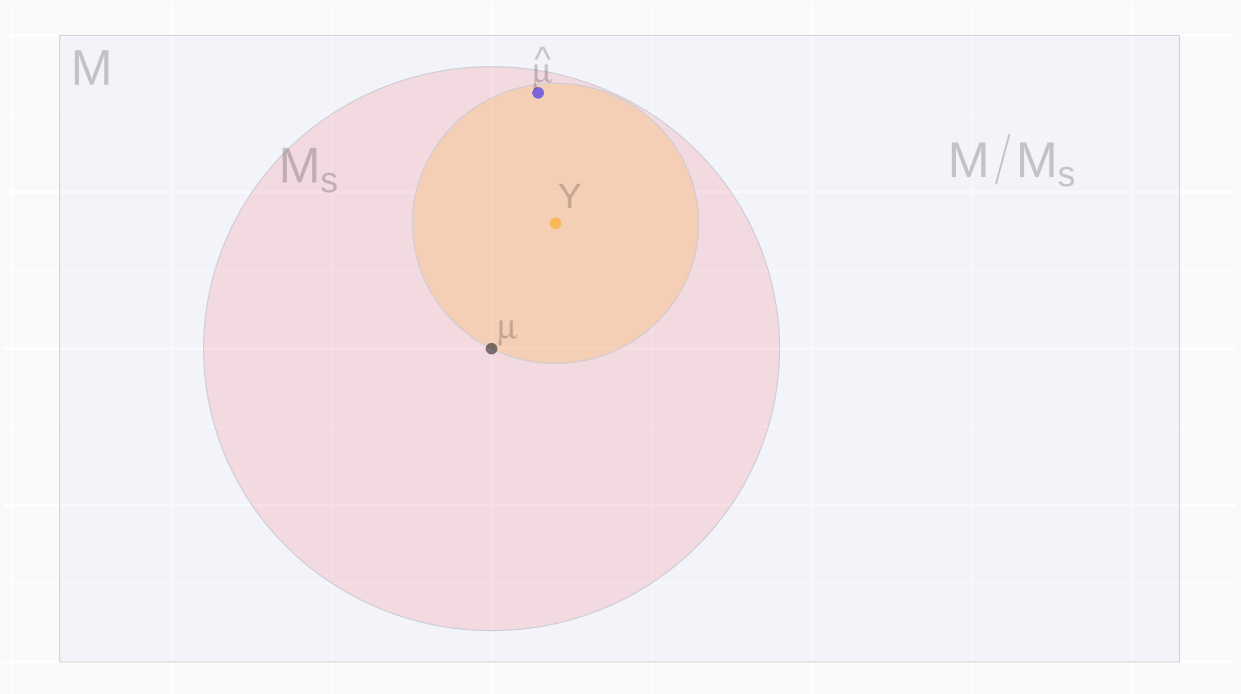
For now, we'll think of $X_1 \ldots X_n$ as deterministic.
If they are random, we *condition* on them.

What we're doing

This is the kind of statement we're after.

$$\text{With probability } 1-\delta, \quad \|\hat{\mu}-\mu\|_{L_2(\mathrm{P_n})} < s$$
$$\text{i.e.} \quad \|\hat{\mu}-\mu\| < s\sqrt{n}.$$

All we're doing is choosing the *neighborhood radius* s so this is true.
Everything else in the picture stays the same.



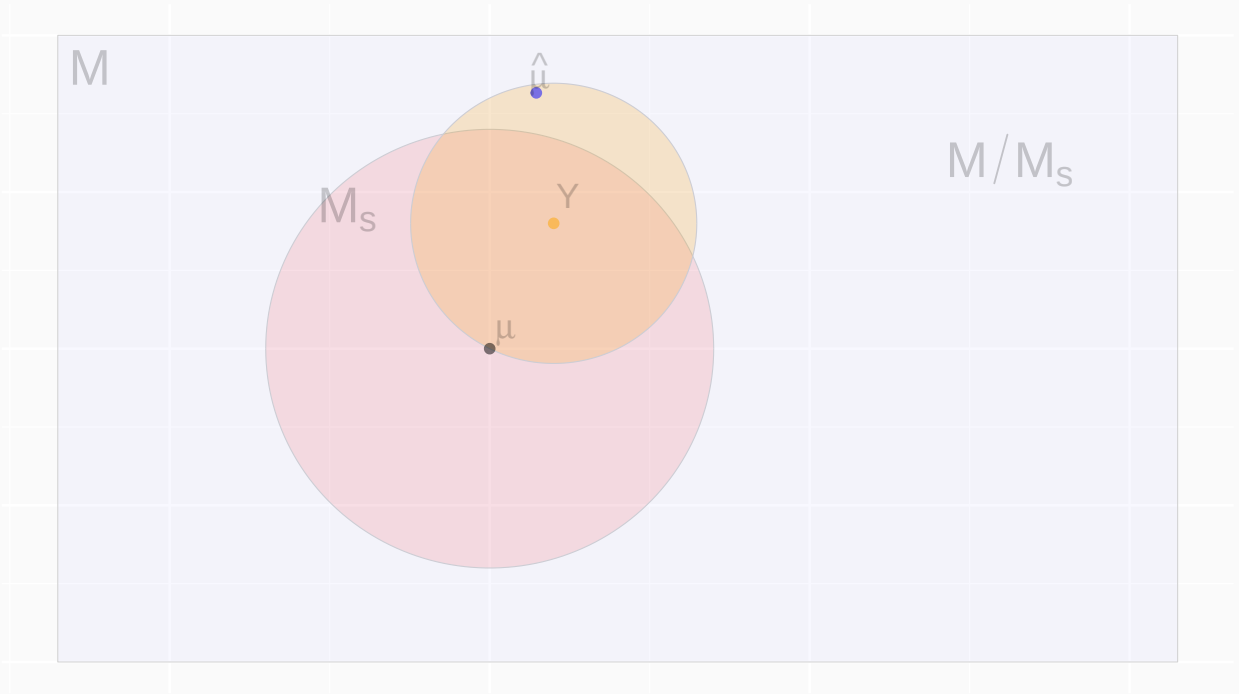
This is what things look like when we get it right.

What we're doing

This is the kind of statement we're after.

$$\text{With probability } 1-\delta, \quad \|\hat{\mu}-\mu\|_{L_2(\mathrm{P_n})} < s$$
$$\text{i.e.} \quad \|\hat{\mu}-\mu\| < s\sqrt{n}.$$

All we're doing is choosing the *neighborhood radius* s so this is true.
Everything else in the picture stays the same.

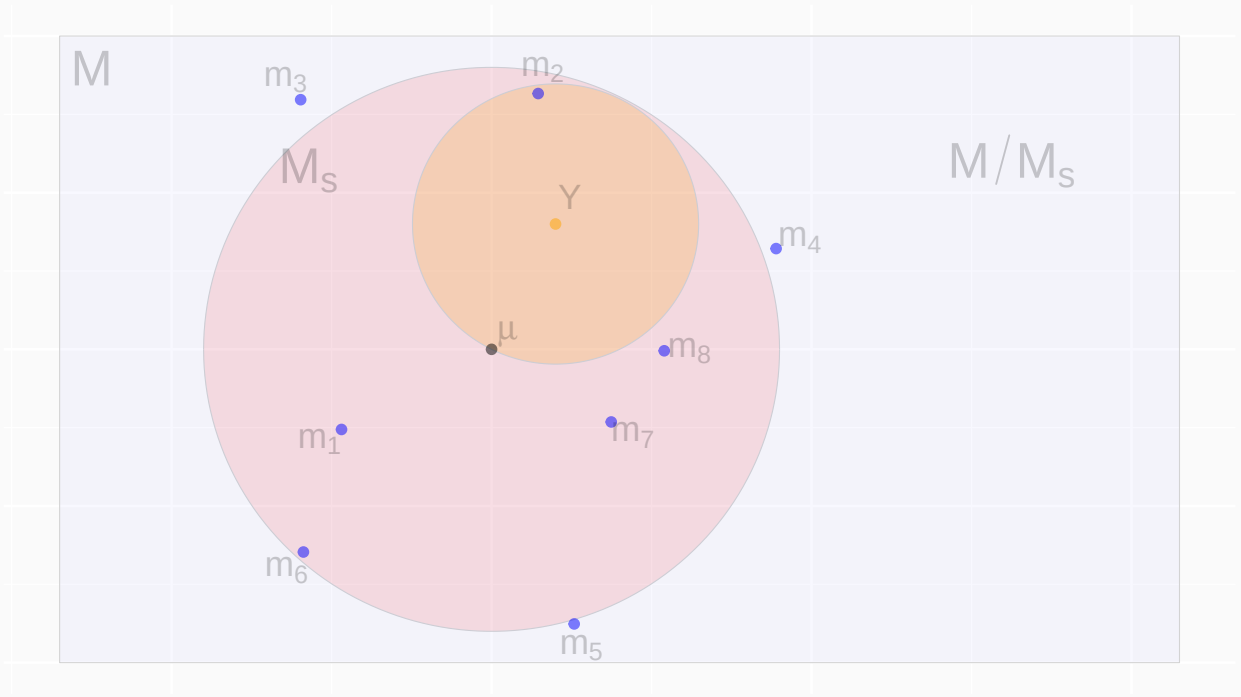


This is what things look like when we don't.

How we'll do it

We'll choose our radius to include *every curve* in the overfitting zone.

$$\ell(m) < \ell(\mu) \implies \|m\| < s\sqrt{n}$$

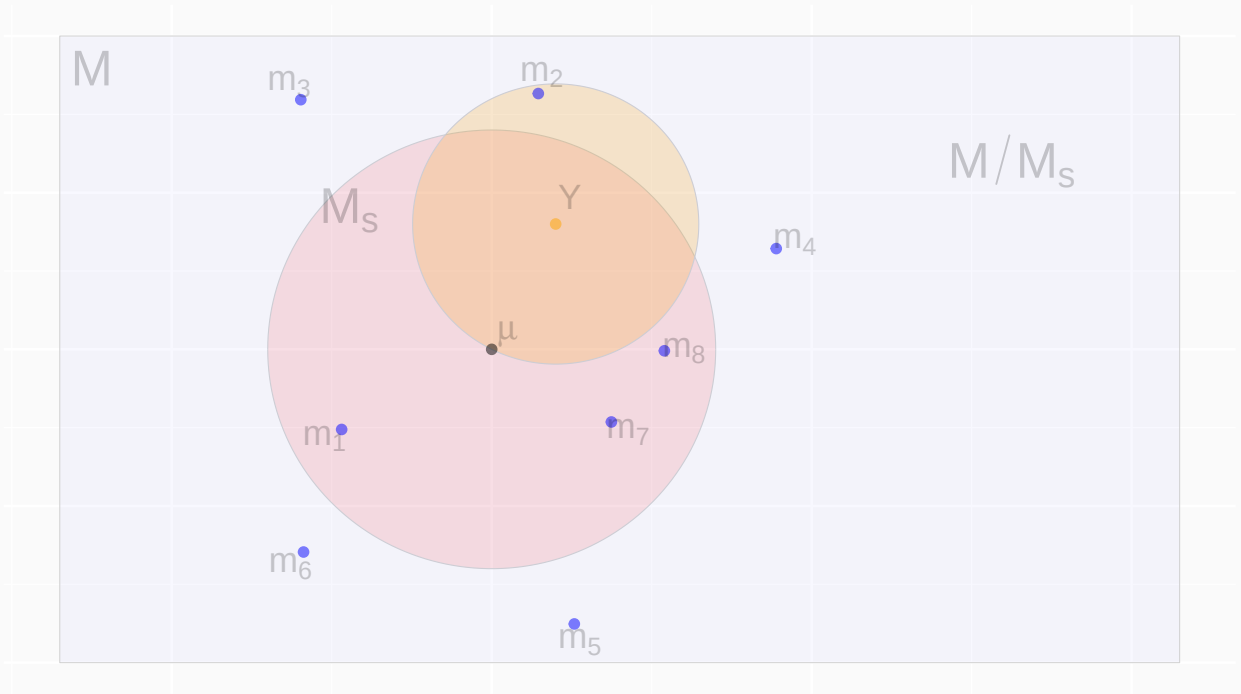


This is what things look like when we get it right.

How we'll do it

We'll choose our radius to include *every curve* in the overfitting zone.

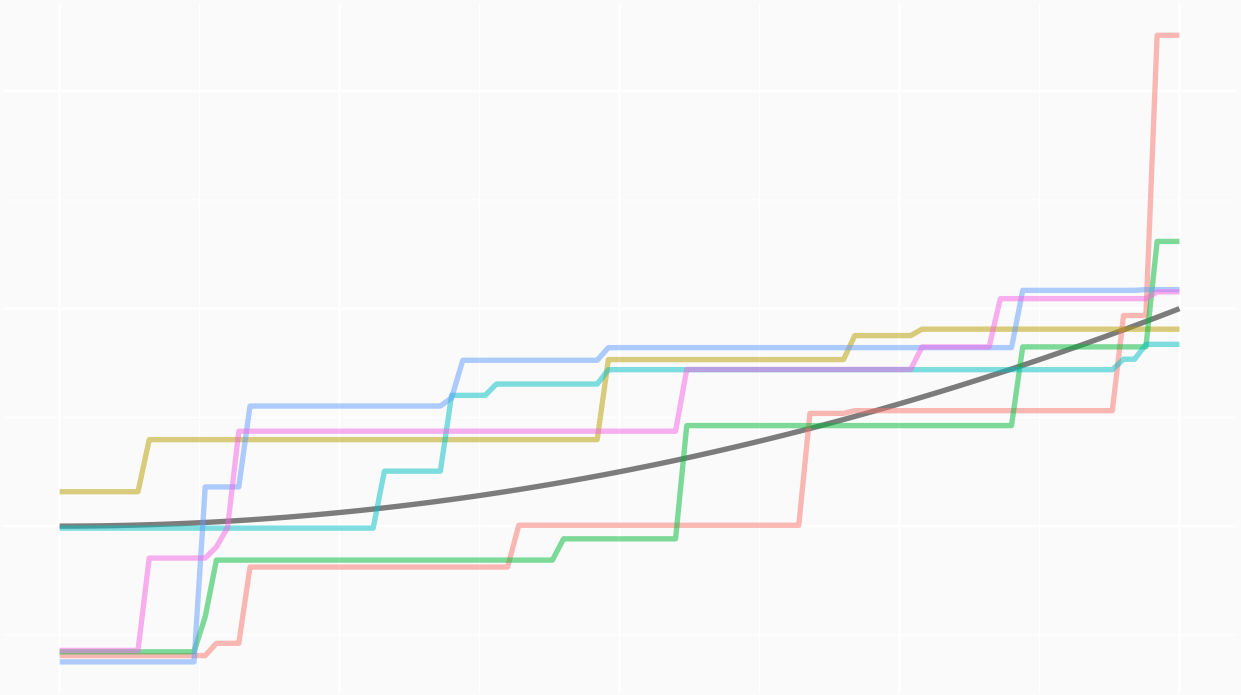
$$\ell(m) < \ell(\mu) \implies \|m\| < s\sqrt{n}$$



This is what things look like when we don't.

What a neighborhood really looks like

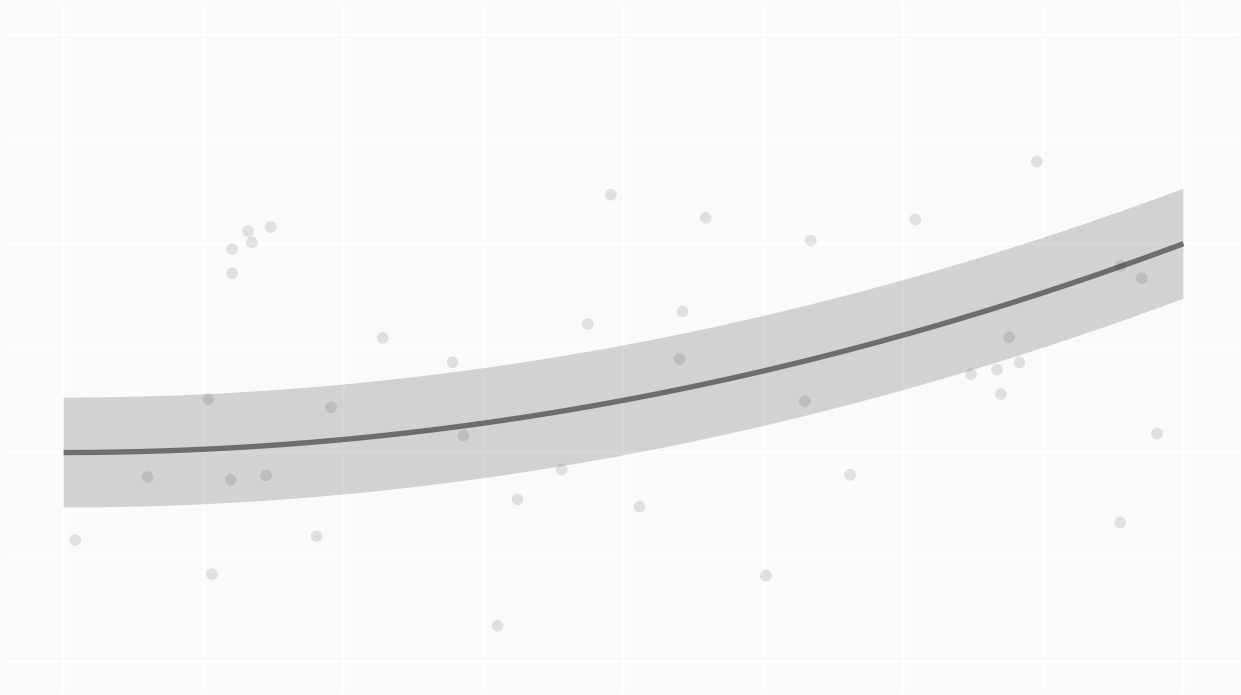
Today, our model is a finite set of curves.



Like these.

What a neighborhood really looks like

Today, our model is a finite set of curves.



A neighborhood is the subset of these curves that's close enough to μ.
Say within the gray tube.

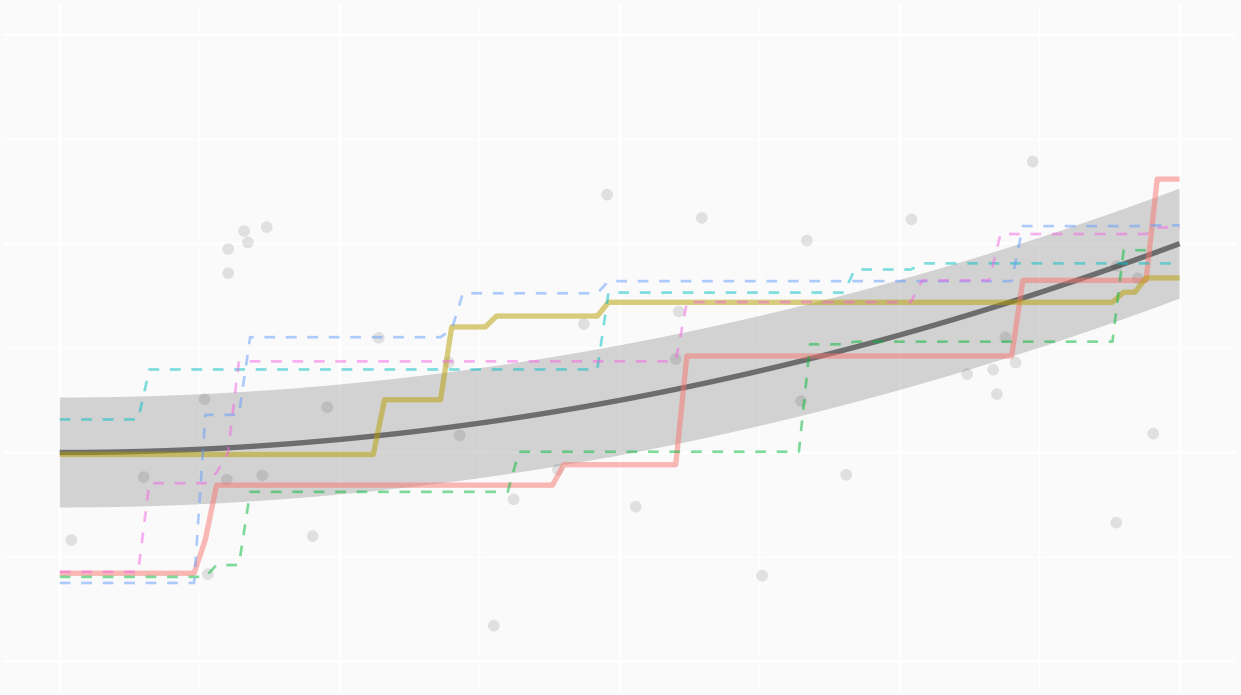
Caveat.

The gray tube is the set of curves that are close in terms of the infinity norm.

$$\mathcal{M}_s^\infty = \{m \in \mathcal{M} : \|m - \mu\|_\infty < s\}$$

What a neighborhood really looks like

Today, our model is a finite set of curves.



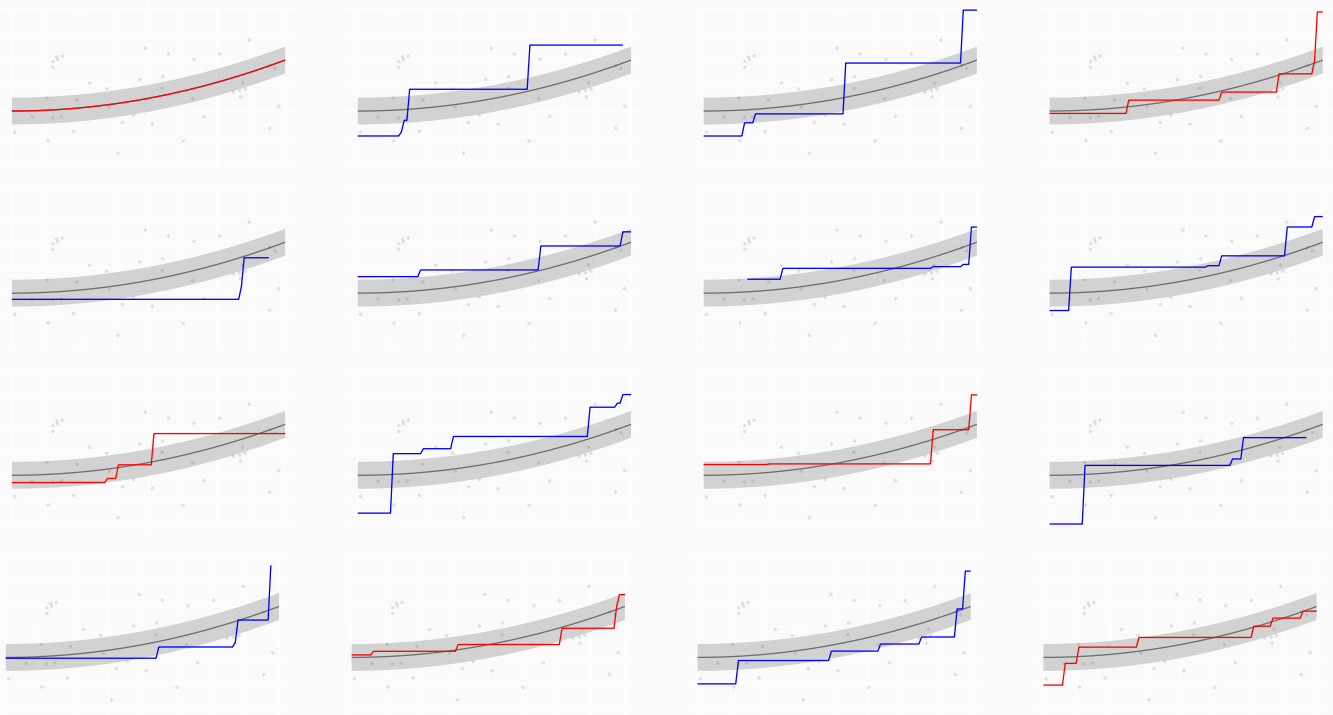
We're talking about the set of curves that are close in terms of the sample two-norm.

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

Think of these as curves that are mostly, but not necessarily always, in the tube.
These are plotted as solid lines above. Those in the complement are dashed.

Neighborhoods in model selection

When we do model selection, we'll tend to have many fairly similar curves.
And it's hard to look at them all drawn on top of each other.



I've drawn curves in our neighborhood in red and curves in its complement in blue.

The argument in words

What we know is that $\hat{\mu}$ beats or ties every other curve in the model.
That's what a minimizer (argmin) does.

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

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

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

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

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

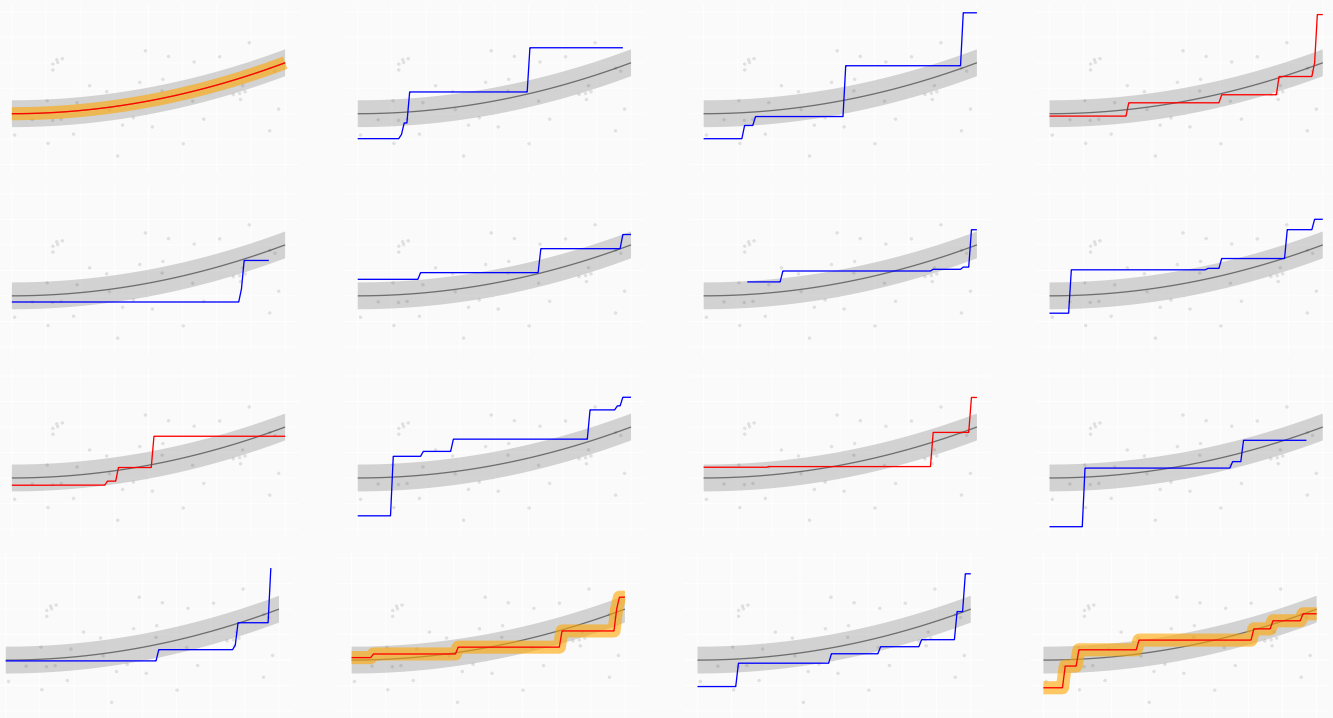
And because $\hat{\mu}$ *is* in the model, that means $\hat{\mu}$ is in the neighborhood.

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

When our two if clauses are true, this argument implies $\hat{\mu}$ is in our neighborhood.
So if they're true with some probability, $\hat{\mu}$ is in the neighborhood with that probability.

Conclusion

Suppose μ is in our model.
Then we can be confident $\hat{\mu}$ is one of the curves in our neighborhood



if we're confident none of the curves in its complement
are curves that can beat or tie μ, i.e., curves in the overfitting zone.

In Mathematical Notation

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

The right radius s

To see how big we need to go, let's revisit a characterization from our warm-up.

$$\frac{\ell(m) - \ell(\mu)}{\|m - \mu\|} = \|m - \mu\| - 2\left\langle \varepsilon, \frac{m - \mu}{\|m - \mu\|} \right\rangle \qquad \text{where} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$= \|m - \mu\| - 2\sigma Z_m \qquad \text{where} \quad Z_m = \frac{\left\langle \varepsilon, \frac{m-\mu}{\|m-\mu\|} \right\rangle}{\sigma} \sim N(0, 1).$$

This means our loss difference exceeds zero for a curve m when the distance from m to μ is large relative to the mean-zero normal random variable $2\sigma Z_m$.

$$\ell(m) - \ell(\mu) > 0 \quad \text{if} \quad \|m - \mu\| > 2\sigma Z_m$$

And it exceeds zero for all curves in our neighborhood's complement if the *minimum* of the left side is larger than the *maximum* of the right.

$$s\sqrt{n} = \min_{m \in \mathcal{M} \setminus \mathcal{M}_s} \|m - \mu\| > 2\sigma \max_{m \in \mathcal{M} \setminus \mathcal{M}_s} Z_m.$$

i.e. if our neighborhood's radius exceeds a maximum of normals.

Finding the right radius

We're after some radius satisfying this lower bound with probability $1 - \delta$.

$$s\sqrt{n} > 2\sigma \max_{m \in \mathcal{M} \setminus \mathcal{M}_s} Z_m.$$

This is implied if s satisfies a slightly larger lower bound.

$$s\sqrt{n} > 2\sigma \max_{m \in \mathcal{M}} Z_m.$$

If there are K functions $m \in \mathcal{M}$, we need a bound on the maximum of K standard normals like Z_m. Like this one.[2]

$$2\sqrt{\log(K)} > \max\{g_1 \dots g_K\} \qquad \text{with probability} \quad 1 - 1/K.$$

Making this substitution, the radius s is good with probability $1 - 1/K$ when

$$s\sqrt{n} = 2\sigma \times 2\sqrt{\log(K)} = 4\sigma\sqrt{\log(K)}.$$

Summary. If $\mu \in \mathcal{M}$, then

$$\min_{\hat{\mu} \in \mathcal{M}_s} \|\hat{\mu} - \mu\| < s\sqrt{n} \quad \text{with probability} \quad 1 - 1/K \quad \text{for} \quad s = 4\sigma\sqrt{\log(K)/n}.$$

Sometimes we drop constants and say $\hat{\mu}$ converges to μ at the rate $\sqrt{\log(K)/n}$.

[2] Throughout the semester, $\log$ will mean the natural log: $\log(e^x) = x$.

What's left

All that's left is proving our bound on the maximum of normals. We use two tools.

- The union bound
- The gaussian tail bound

The union bound

- The maximum of K things exceeds a threshold t if and only if *at least one of them does.*

$$\max g_1 \ldots g_K \geq t \quad \text{if and only if} \quad g_1 \geq t \text{ or } \ldots \text{ or } g_K \geq t$$

- This means the probability that it happens is the probability that *at least one thing happens.*

$$\mathrm{P}(\max g_1 \ldots g_K \geq t) = \mathrm{P}(g_1 \geq t \text{ or } \ldots \text{ or } g_K \geq t).$$

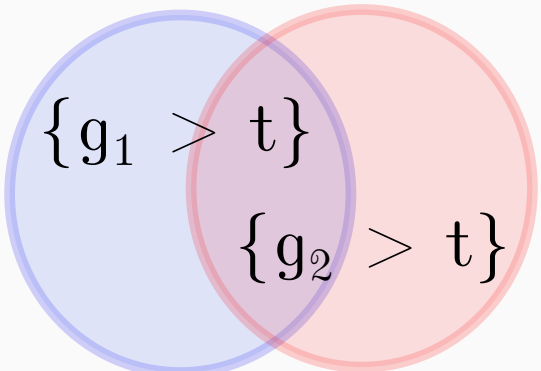
- And this is no larger than the *sum of probabilities* that one happens*.

$$\mathrm{P}(g_1 \geq t \text{ or } \ldots \text{ or } g_K \geq t) \leq \mathrm{P}(g_1 \geq t) + \ldots + \mathrm{P}(g_K \geq t).$$

- All of our probabilities are the same, so this sum is K times the probability of one.

$$\mathrm{P}(\max g_1 \ldots g_K \geq t) \leq K\,\mathrm{P}(g_1 \geq t).$$

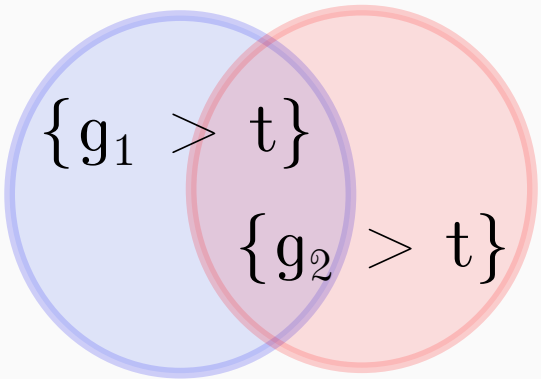
* Why?



* Why?

- Think about if we had two things, A and B, e.g. $\{g_1 \geq t\}$ and $\{g_2 \geq t\}$.
- And lets break down how A **or** B can happen into *disjoint events*, i.e., things that cannot happen simultaneously.
  - Event 1: A happens.
  - Event 2: B happens and A *doesn't*.
- The probability that A **or** B happens is the sum of the probabilities of these disjoint events. And Event 2 happens *less often* than B.

$$\mathrm{P}(A \text{ or } B) = P(A) + P(B \text{ and not } A) \leq P(A) + P(B)$$



The gaussian tail bound

Here's where we are.

$$\mathrm{P}(\max g_1 \dots g_K \geq t) \leq KP(g_1 \geq t).$$

And g_1 is standard normal, so* ...

$$\mathrm{P}(g_1 \geq t) \leq e^{-t^2/2} \qquad \text{for} \qquad t > \frac{1}{\sqrt{2\pi}}.$$

And therefore ...

$$\begin{aligned} K\,\mathrm{P}(g_1 \geq t) &\leq Ke^{-t^2/2} = e^{\log(K)-t^2/2} \\ &= e^{-\log(K)} = 1/K \qquad \text{for} \qquad t = 2\sqrt{\log(K)}. \end{aligned}$$

This means that

$$\mathrm{P}\left\{ \max g_1 \dots g_K \geq 2\sqrt{\log(K)} \right\} \leq 1/K \qquad \text{or equivalently}$$

$$\max g_1 \dots g_K < 2\sqrt{\log(K)} \qquad \text{with probability} \quad 1 - 1/K.$$

* Why?

$$
\begin{aligned}
\mathrm{P}(g_1 \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \\
&\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2} && \text{because } \frac{x}{t} \geq 1 \\
&= \frac{1}{t\sqrt{2\pi}} \int_t^\infty x e^{-x^2/2} \\
&= \frac{1}{t\sqrt{2\pi}} \left\{ -e^{-x^2/2} \right\} \Big|_{x=t}^{\infty} && \text{because } x e^{-x^2/2} = \frac{d}{dx}\left\{ -e^{-x^2/2} \right\} \\
&= \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}
\end{aligned}
$$