POLS 510
CSSS 510

Maximum Likelihood Methods
for the Social Sciences

Introduction to Maximum Likelihood Estimation

Christopher Adolph

Political Science *and* CSSS
University of Washington, Seattle

Onwards, from probability to modelling

We've worked up from sets and sample spaces

to the idea of random variables distributed according to functions

chosen for the plausibility of their assumptions

Now, we'll see how this culminates in practical models of social phenomena

The goal henceforth is inference:
using the data we know to uncover things we don't know yet

Key requirement:
quantifying the uncertainty of our inferences (that's why we need probability)

Outline

Notation for models of random variables

Basic concepts for likelihood inference

Interpreting profile likelihoods

Deriving maximum likelihood estimates (MLEs)

An MLE for heteroskedastic data

Numerical methods of finding MLEs

Statistical properties of MLEs

Summarizing the uncertainty of MLEs

From random variables to models

Suppose we're studying outcome $\mathbf{y}$ (which could be votes, disease incidence, or unemployment), and we decide $\mathbf{y}$ is distributed f

Stochastic component: $\mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha})$

From random variables to models

Suppose we're studying outcome $\mathbf{y}$ (which could be votes, disease incidence, or unemployment), and we decide $\mathbf{y}$ is distributed f

Stochastic component: $\mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha})$

Systematic component: $\boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta})$

This formulation encompasses all the models in this class.

From random variables to models

Suppose we're studying outcome $\mathbf{y}$ (which could be votes, disease incidence, or unemployment), and we decide $\mathbf{y}$ is distributed f

Stochastic component: $\mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha})$

Systematic component: $\boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta})$

This formulation encompasses all the models in this class.

You are used to seeing linear regression written this way:

$$\begin{aligned} y_i &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \\ \varepsilon_i &\sim f_{\mathcal{N}}(0, \sigma^2) \end{aligned}$$

From random variables to models

Suppose we're studying outcome $\mathbf{y}$ (which could be votes, disease incidence, or unemployment), and we decide $\mathbf{y}$ is distributed f

Stochastic component: $\mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha})$

Systematic component: $\boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta})$

This formulation encompasses all the models in this class.

You are used to seeing linear regression written this way:

$$\begin{aligned} y_i &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \\ \varepsilon_i &\sim f_{\mathcal{N}}(0, \sigma^2) \end{aligned}$$

In our notation, this is equivalent to assuming $f(\cdot)$ is the Normal distribution:

$$\begin{aligned} y_i &\sim f_{\mathcal{N}}(\mu_i, \sigma^2) \\ \mu_i &= \mathbf{x}_i\boldsymbol{\beta} \end{aligned}$$

From random variables to models

By choosing a different distribution to be $f(\cdot)$, we get a new model

Our notation allows this, but the error term format doesn't (only works for Normal)

From random variables to models

By choosing a different distribution to be $f(\cdot)$, we get a new model

Our notation allows this, but the error term format doesn't (only works for Normal)

For example, the Bernoulli leads to *binary choice logit*:

$$\begin{aligned} y_i &\sim f_{\text{Bern}}(\pi_i) \\ \pi_i &= \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\beta})} \end{aligned}$$

From random variables to models

By choosing a different distribution to be $f(\cdot)$, we get a new model

Our notation allows this, but the error term format doesn't (only works for Normal)

For example, the Bernoulli leads to *binary choice logit*:

$$\begin{aligned} y_i &\sim f_{\text{Bern}}(\pi_i) \\ \pi_i &= \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\beta})} \end{aligned}$$

And the Poisson leads to *Poisson event count regression*:

$$\begin{aligned} y_i &\sim f_{\text{Pois}}(\lambda_i) \\ \lambda_i &= \exp(\mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

and so on. . .

From random variables to models

Notes on stochastic components

Number of parameters in the stochastic component varies by distribution

"Extra" parameters often called nuisance parameters. (Too dismissive?)

Number of stochastic "layers" is variable – we'll see nested distributions

From random variables to models

Notes on stochastic components

Number of parameters in the stochastic component varies by distribution

"Extra" parameters often called nuisance parameters. (Too dismissive?)

Number of stochastic "layers" is variable – we'll see nested distributions

Notes on systematic components

Systematic component is not always linear

Often a transformation from unbounded $\mathbf{x}_i\boldsymbol{\beta}$ to some range, especially

the positive real numbers, $\mathbb{R}^+$

a real interval such as $[0, 1]$

From random variables to models

Learning from a model often relies on two quantities

Expected values often equal the systematic component: $\mathbb{E}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$

Predicted values are *draws* from the stochastic component: $\tilde{\mathbf{y}}|\mathbf{X}, \boldsymbol{\beta}$

Computing quantities like these is usually the final goal of inference

From random variables to models

Learning from a model often relies on two quantities

Expected values often equal the systematic component: $\mathbb{E}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$

Predicted values are *draws* from the stochastic component: $\tilde{\mathbf{y}}|\mathbf{X}, \boldsymbol{\beta}$

Computing quantities like these is usually the final goal of inference

To compute these quantities of interest, we need to estimate unknowns like $\boldsymbol{\beta}$

The intermediate goal of inference is estimating unknown parameters

We will often denote the set of all parameters (e.g., $\boldsymbol{\beta}$ and σ^2) as $\boldsymbol{\theta}$

In applications, we typically don't know the values of $\boldsymbol{\theta}$,
so we attempt to infer them from the data, $\mathbf{y}$

In probabilistic terms, we want to learn about $\boldsymbol{\theta}$ given $\mathbf{y}$, so we need to find $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

From models to inference

A catch: it's not possible to infer $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ from $\mathbf{y}$ alone *Why not?*

From models to inference

A catch: it's not possible to infer $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ from $\mathbf{y}$ alone *Why not?*

If we assume a distribution for $\mathbf{y}$, we can solve for $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$. . .

$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

From models to inference

A catch: it's not possible to infer $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ from $\mathbf{y}$ alone *Why not?*

If we assume a distribution for $\mathbf{y}$, we can solve for $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$. . .

$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})}{\mathrm{P}(\mathbf{y})} \qquad \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})}{\mathrm{P}(\boldsymbol{\theta})}$$

$$\mathrm{P}(\mathbf{y})\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y}) \qquad \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})$$

From models to inference

A catch: it's not possible to infer $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ from $\mathbf{y}$ alone *Why not?*

If we assume a distribution for $\mathbf{y}$, we can solve for $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$. . .

$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})}{\mathrm{P}(\mathbf{y})} \qquad \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})}{\mathrm{P}(\boldsymbol{\theta})}$$

$$\mathrm{P}(\mathbf{y})\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y}) \qquad \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})$$

$$\mathrm{P}(\mathbf{y})\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$$

From models to inference

A catch: it's not possible to infer $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ from $\mathbf{y}$ alone *Why not?*

If we assume a distribution for $\mathbf{y}$, we can solve for $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$. . .

$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})}{\mathrm{P}(\mathbf{y})} \qquad \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})}{\mathrm{P}(\boldsymbol{\theta})}$$

$$\mathrm{P}(\mathbf{y})\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y}) \qquad \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) = \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y})$$

$$\begin{aligned} \mathrm{P}(\mathbf{y})\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})} \end{aligned}$$

This famous result is known as *Bayes Rule*

It shows how to write a conditional probability $\mathrm{P}(a|b)$ in terms of its inverse, $\mathrm{P}(b|a)$

From models to inference

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})} \qquad (\textit{Bayes Rule})$$

So to infer the probability of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$, we need to know $\mathrm{P}(\boldsymbol{\theta})$ and $\mathrm{P}(\mathbf{y})$ *a priori*

From models to inference

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})} \qquad (\textit{Bayes Rule})$$

So to infer the probability of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$,
we need to know $\mathrm{P}(\boldsymbol{\theta})$ and $\mathrm{P}(\mathbf{y})$ *a priori*

We can rewrite Bayes rule to replace $\mathrm{P}(\mathbf{y})$ with other quantities (integrate it out):

$$\mathrm{P}(\mathbf{y}) \quad = \quad \int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y}) \mathrm{d}\boldsymbol{\theta}$$

From models to inference

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})} \qquad (\textit{Bayes Rule})$$

So to infer the probability of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$,
we need to know $\mathrm{P}(\boldsymbol{\theta})$ and $\mathrm{P}(\mathbf{y})$ *a priori*

We can rewrite Bayes rule to replace $\mathrm{P}(\mathbf{y})$ with other quantities (integrate it out):

$$\begin{aligned} \mathrm{P}(\mathbf{y}) &= \int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \end{aligned}$$

From models to inference

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})} \qquad (\textit{Bayes Rule})$$

So to infer the probability of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$,
we need to know $\mathrm{P}(\boldsymbol{\theta})$ and $\mathrm{P}(\mathbf{y})$ *a priori*

We can rewrite Bayes rule to replace $\mathrm{P}(\mathbf{y})$ with other quantities (integrate it out):

$$\begin{aligned} \mathrm{P}(\mathbf{y}) &= \int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta} \cap \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \end{aligned}$$

But $\mathrm{P}(\boldsymbol{\theta})$ is not known objectively

From models to inference

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \qquad (\textit{Bayes Rule})$$

$\mathrm{P}(\boldsymbol{\theta})$ is not known objectively, but we need it to compute $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

This creates a fork in the road of inference, with two major schools of thought:

Bayesian inference

1. Make a *subjective* guess of the *a priori* $\mathrm{P}(\boldsymbol{\theta})$
2. Then use $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$ to calculate $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

From models to inference

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{\theta})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \qquad (\textit{Bayes Rule})$$

$\mathrm{P}(\boldsymbol{\theta})$ is not known objectively, but we need it to compute $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

This creates a fork in the road of inference, with two major schools of thought:

Bayesian inference

1. Make a *subjective* guess of the *a priori* $\mathrm{P}(\boldsymbol{\theta})$
2. Then use $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$ to calculate $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

Likelihood inference:

1. Give up on calculaing $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ to avoid making subjective guesses of $\mathrm{P}(\boldsymbol{\theta})$
2. Instead focus on making inferences directly from $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$

Understanding Bayesian inference: An example

To understand likelihood inference,
it helps to start with Bayesian inference and strip pieces away

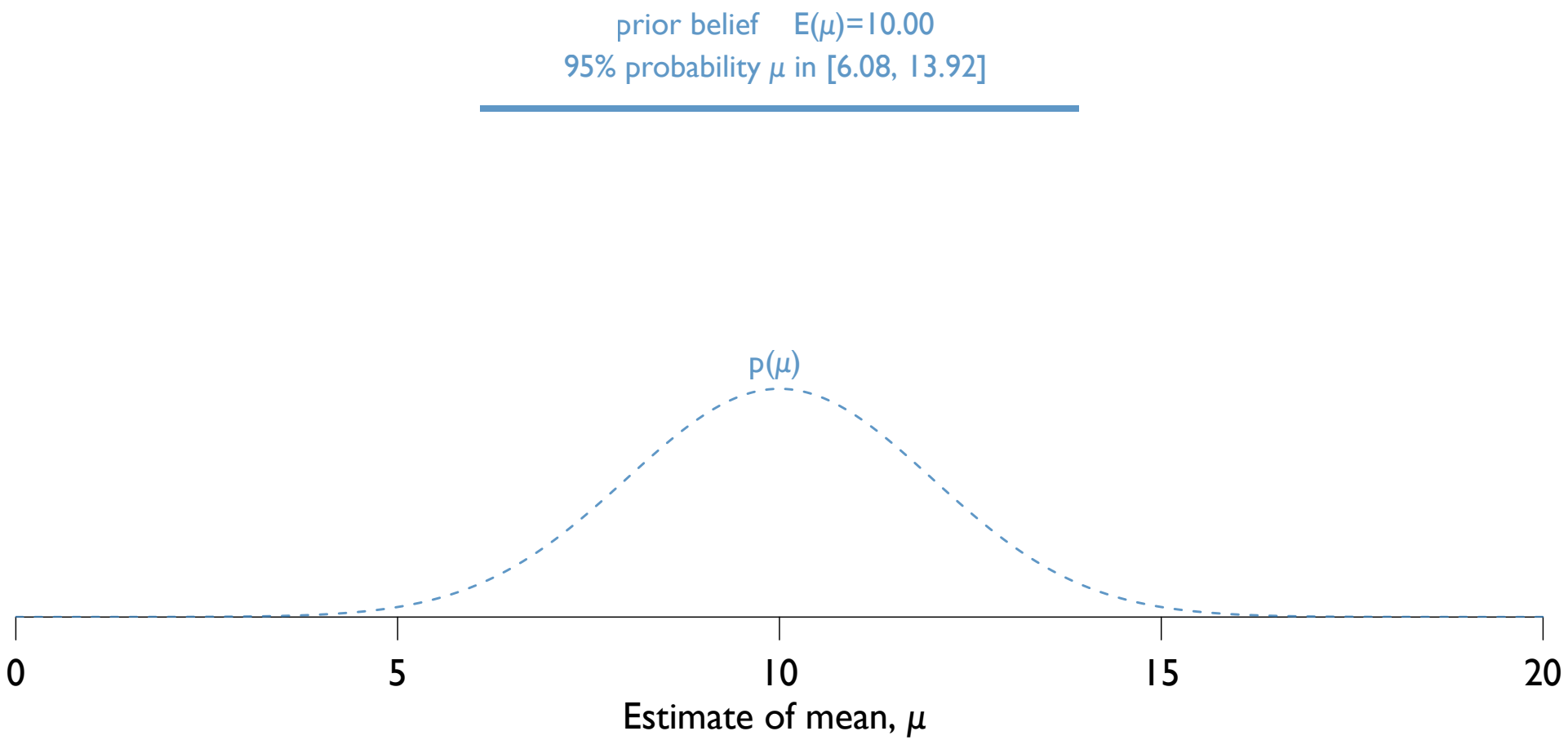
To understand the Bayesian logic of inference,
it helps to have an example

Suppose an statistics instructor wants to know
how many hours of work his homeworks take on average

Based on many years of teaching,
he believes (subjectively) that this average is most likely 10 hours per assignment

And he is 95% confident that the average student spends between 6 and 14 hours

This is his prior belief – can he improve on it by gathering a little data?

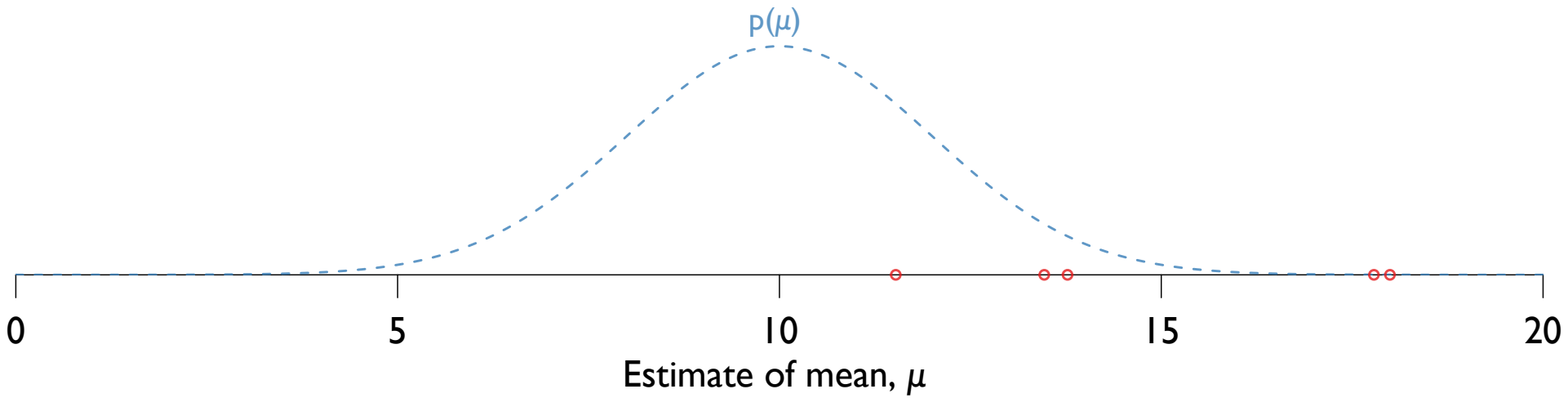


The instructor's uncertainty about the average workload corresponds to a Normal(10,4) *prior distribution*

To complement these subjective prior beliefs, he surveys 5 random students

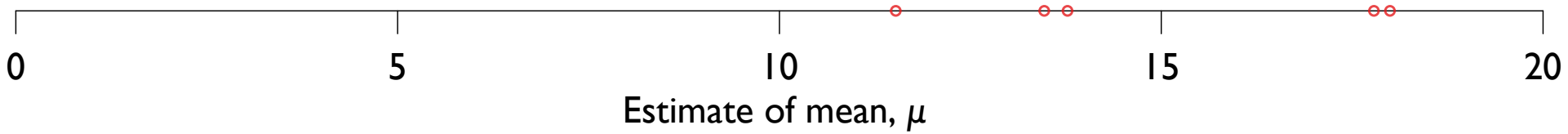
prior belief E(μ)=10.00

95% probability μ in [6.08, 13.92]

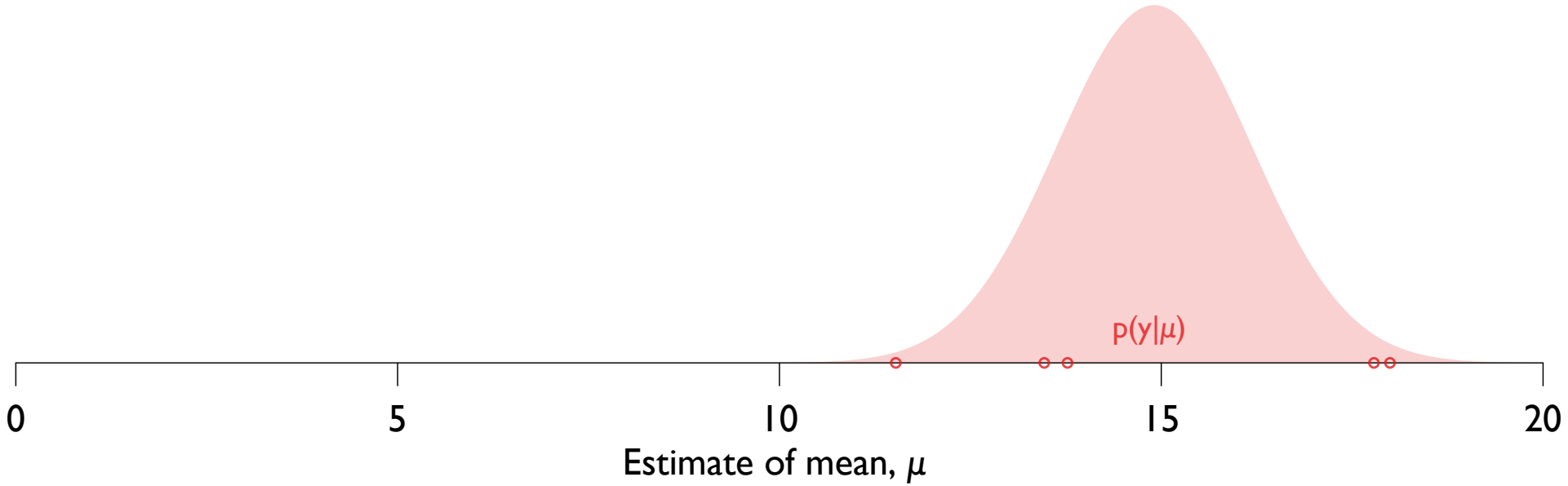


$\mathbf{y} = \{11.5, 13.5, 13.8, 17.8, 18.0\}$ mean=14.9, sd=2.9 *higher than expected!*

The instructor is reluctant to discard his prior knowledge on the basis of a tiny sample. Can he combine his insights with the data?

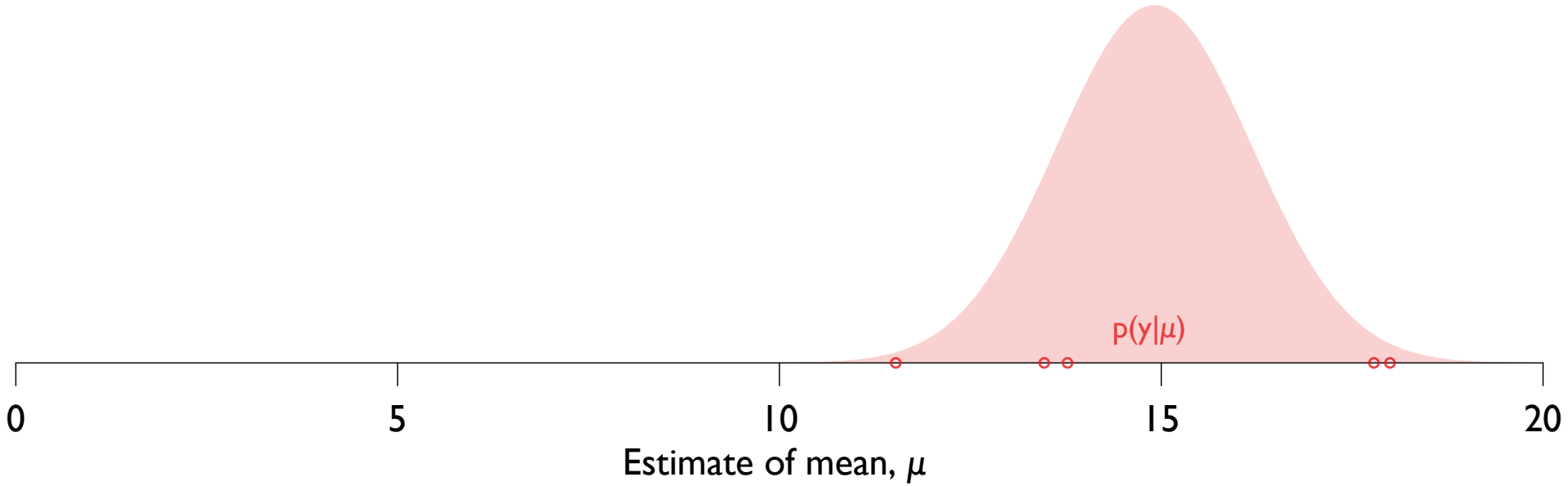


Assuming they came from a Normal distribution, what parameters were *most likely* to produce a sample with a mean of 14.9 and variance of 8.4?



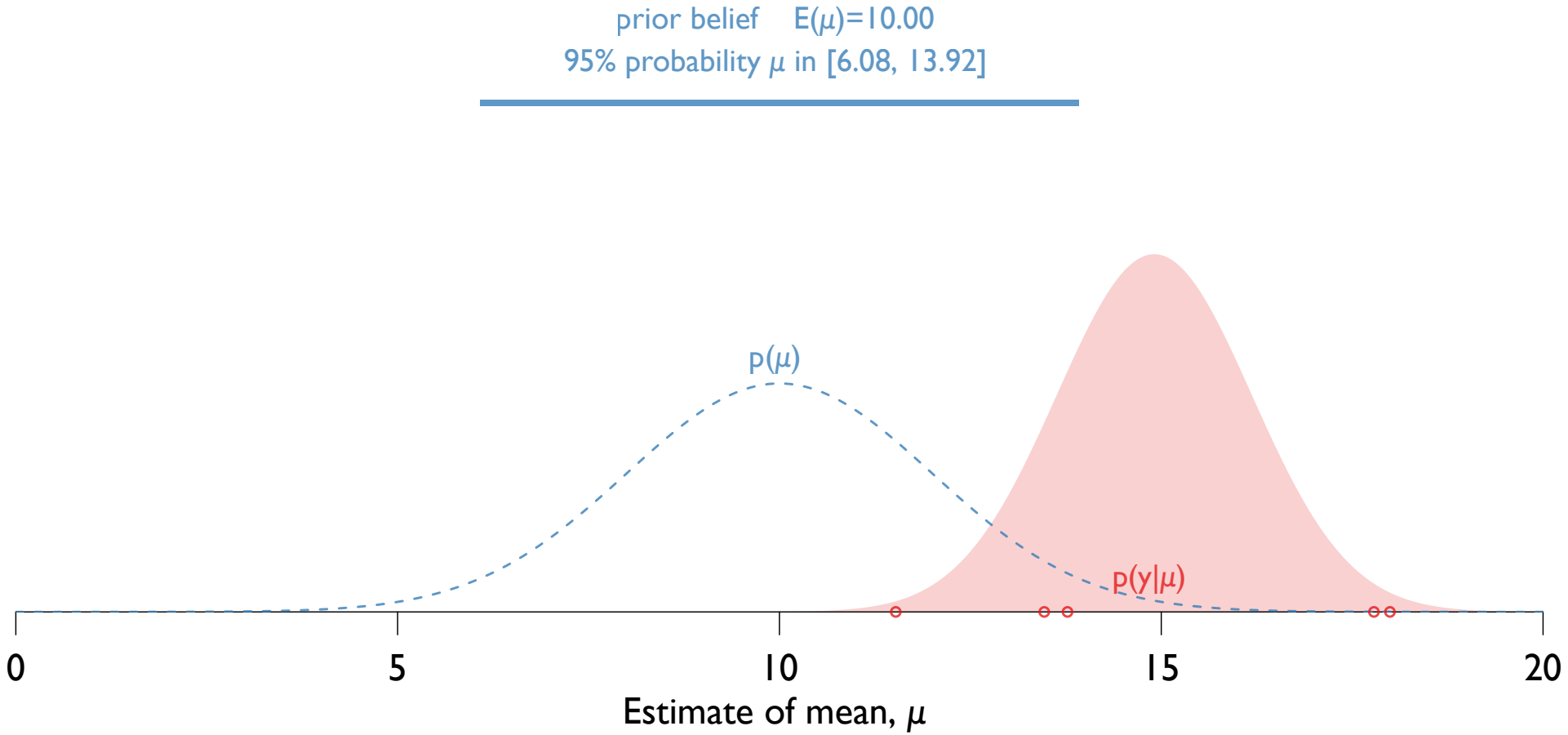
Assuming they came from a Normal distribution, what parameters were *most likely* to produce a sample with a mean of 14.9 and variance of 8.4?

The *most likely* distribution turns out to be Normal(μ =14.9, σ^2 =8.4/n)



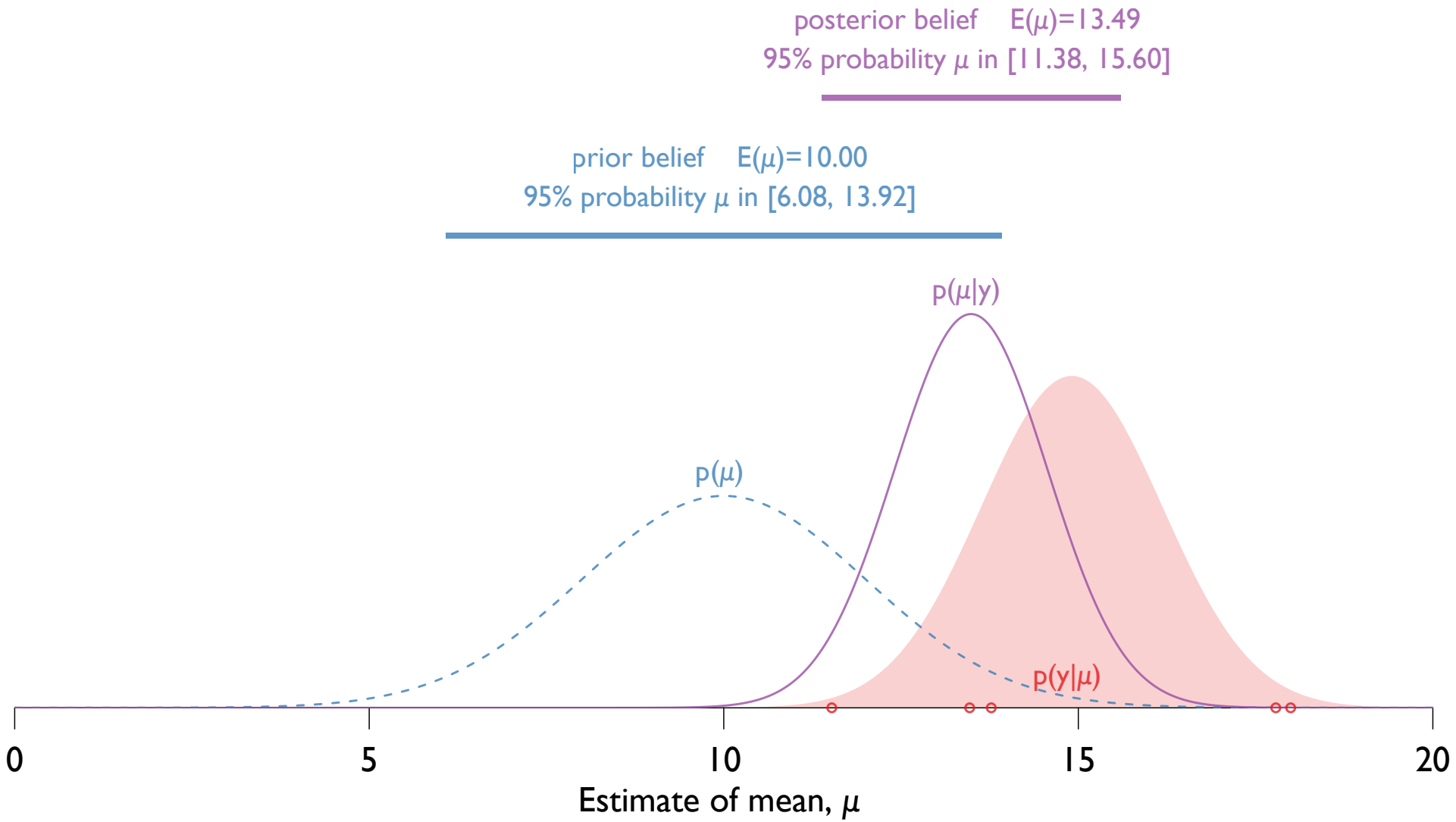
The distribution *most likely* to produce the sample is Normal(μ =14.9, σ^2 =8.4/n)

We call this distribution, $p(\mathbf{y}|\mu)$ – the probability of seeing the sample $\mathbf{y}$ given the value of the parameter μ – the *likelihood*



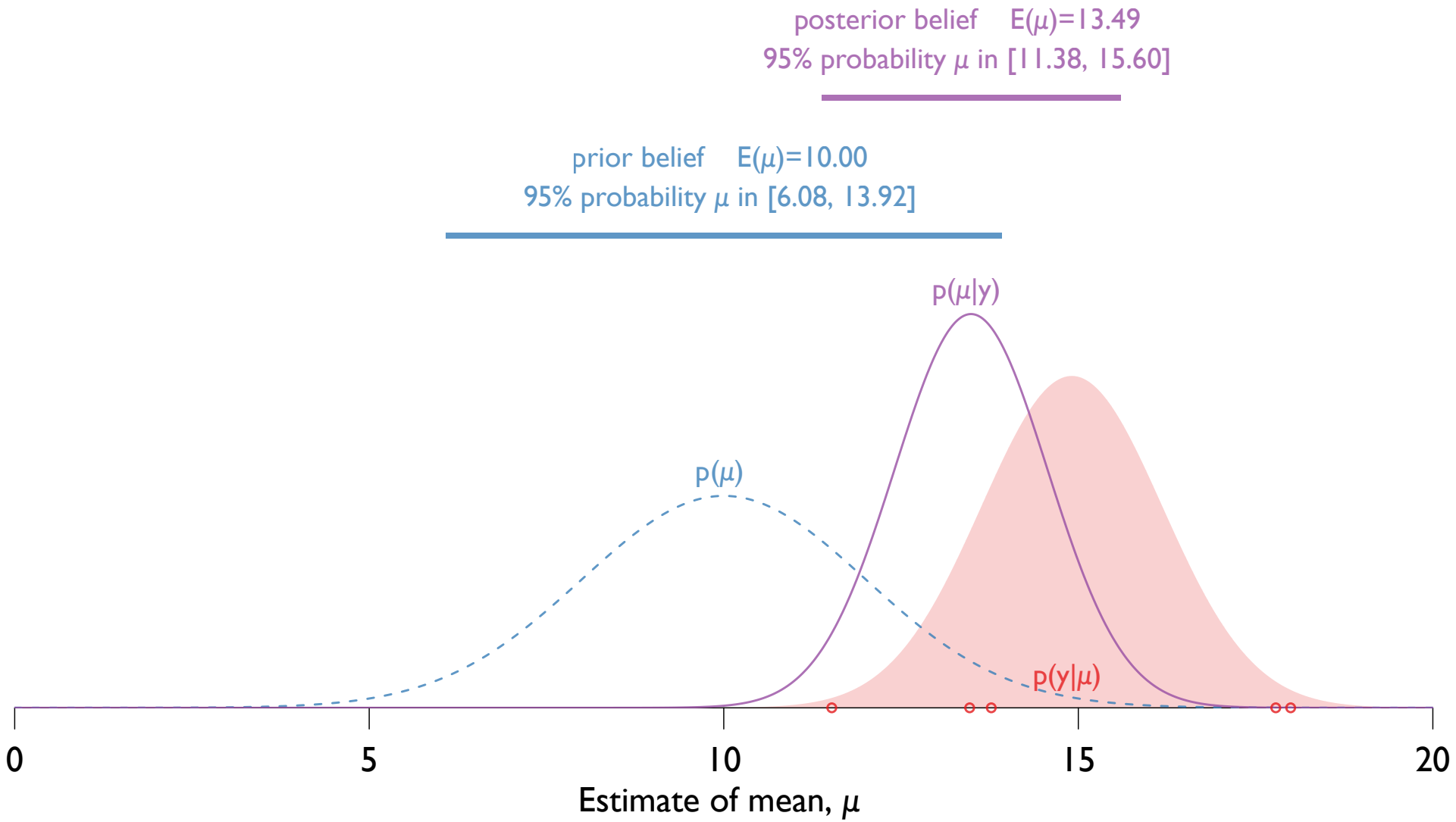
Bayesian inference:
using Bayes Theorem to *combine* the prior distribution and the likelihood

Multiplying these two distributions together & dividing by $p(\mathbf{y})$ yields. . .



The posterior distribution: our subjective beliefs about student workload *updated* to account for the objective new data we sampled

We now think there's a 95% probability students work between 11.38 & 15.60 hours

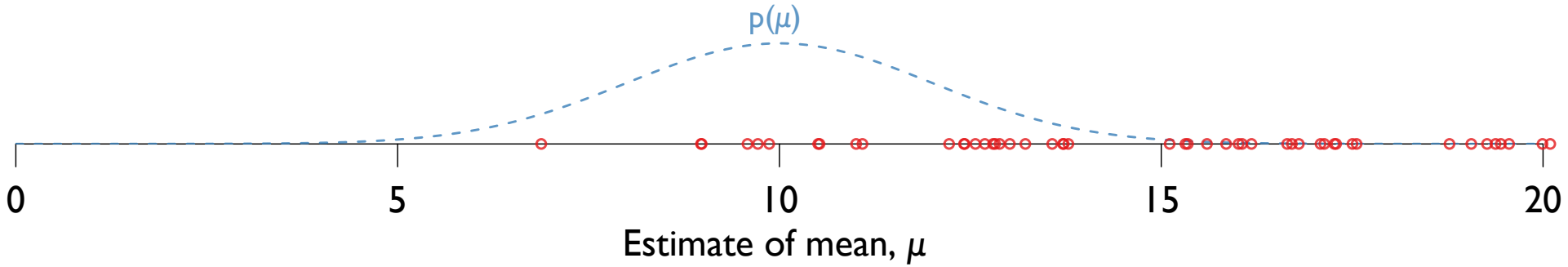


Note that our new beliefs compromise between our old beliefs and the data

Also, note that we draw clear but *subjective* conclusions about the probability distribution of the sample mean

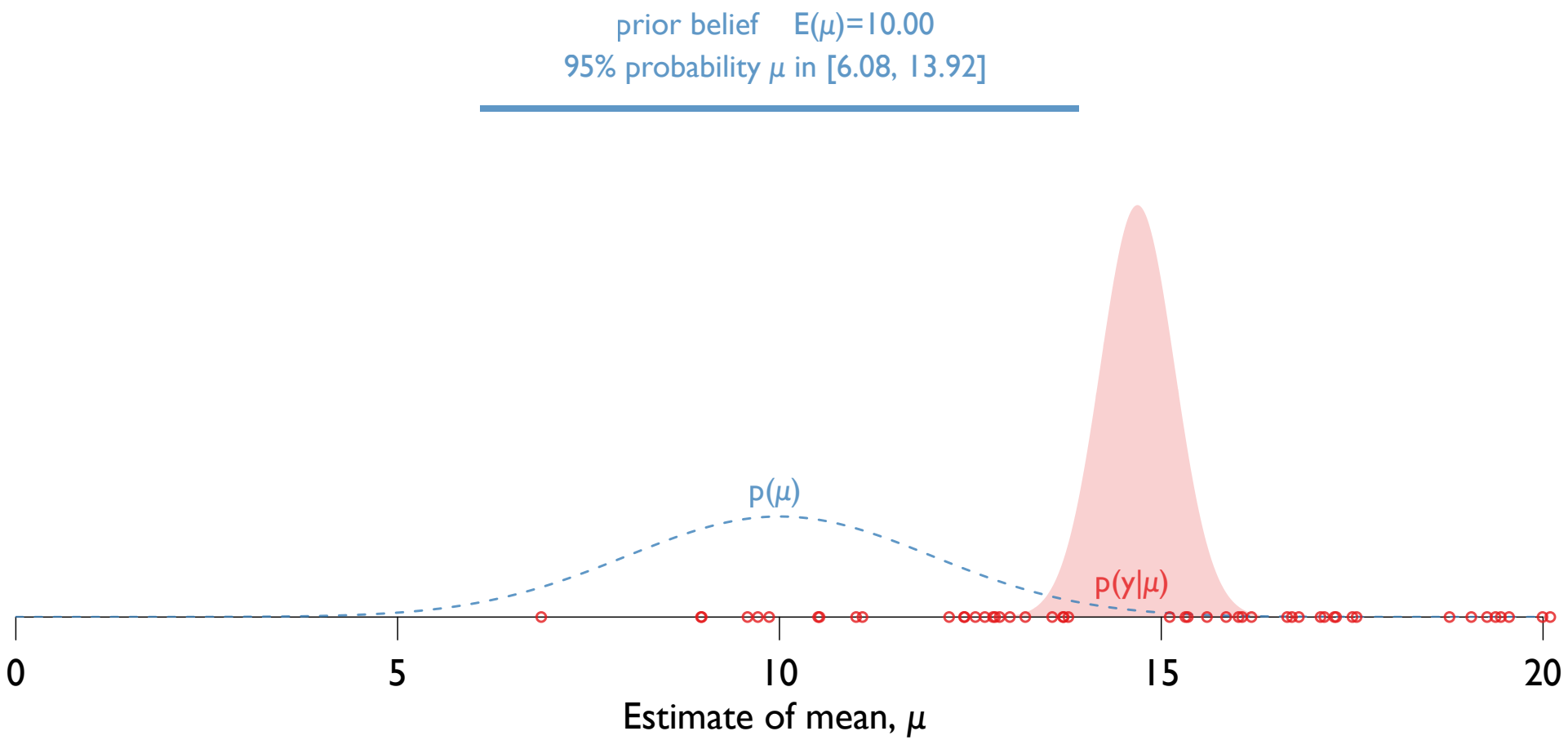
prior belief E(μ)=10.00

95% probability μ in [6.08, 13.92]

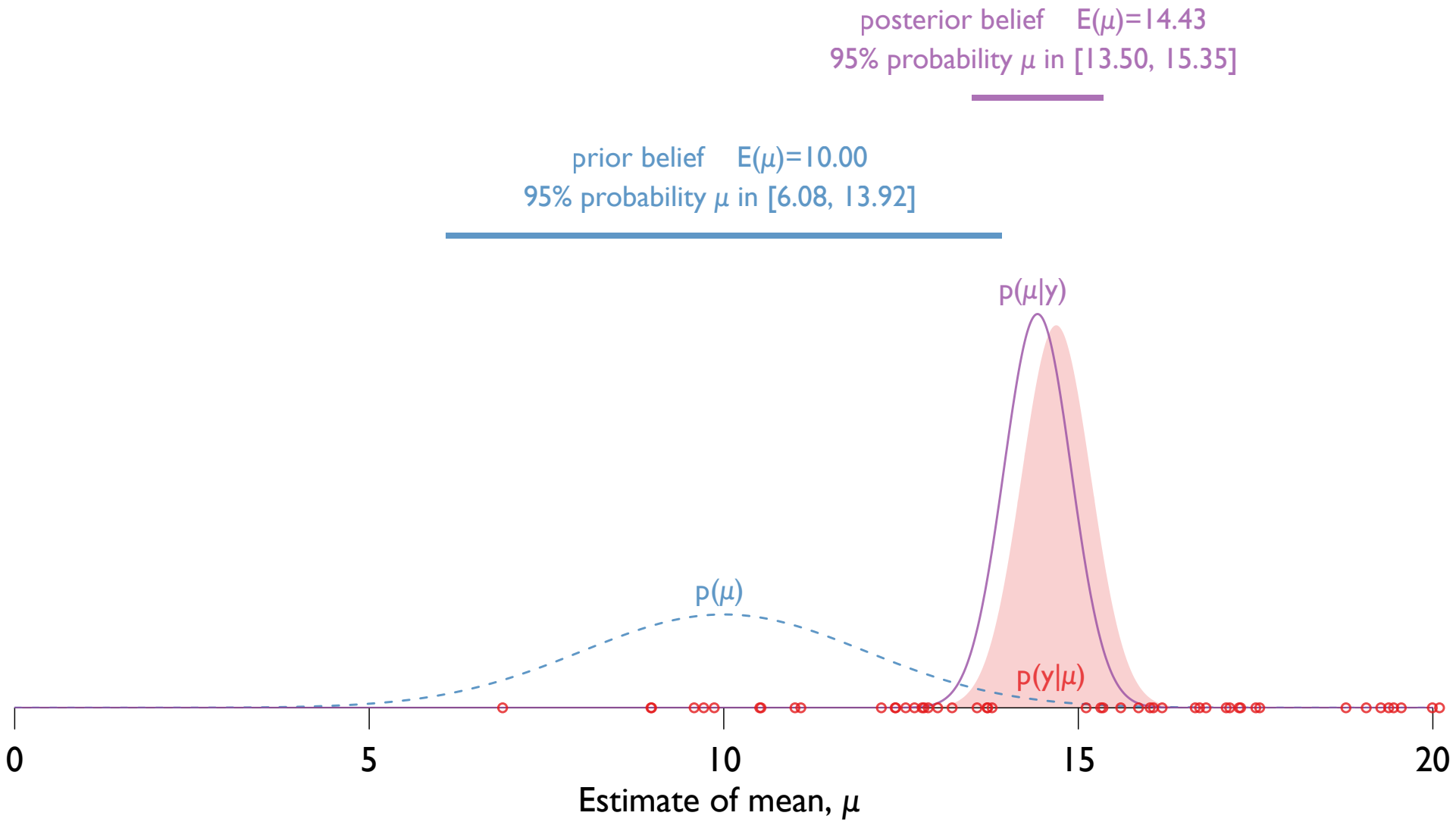


What if we had drawn a larger sample – say, 50 students

Now we obtain a sample mean of 14.7 and an sd of 3.5

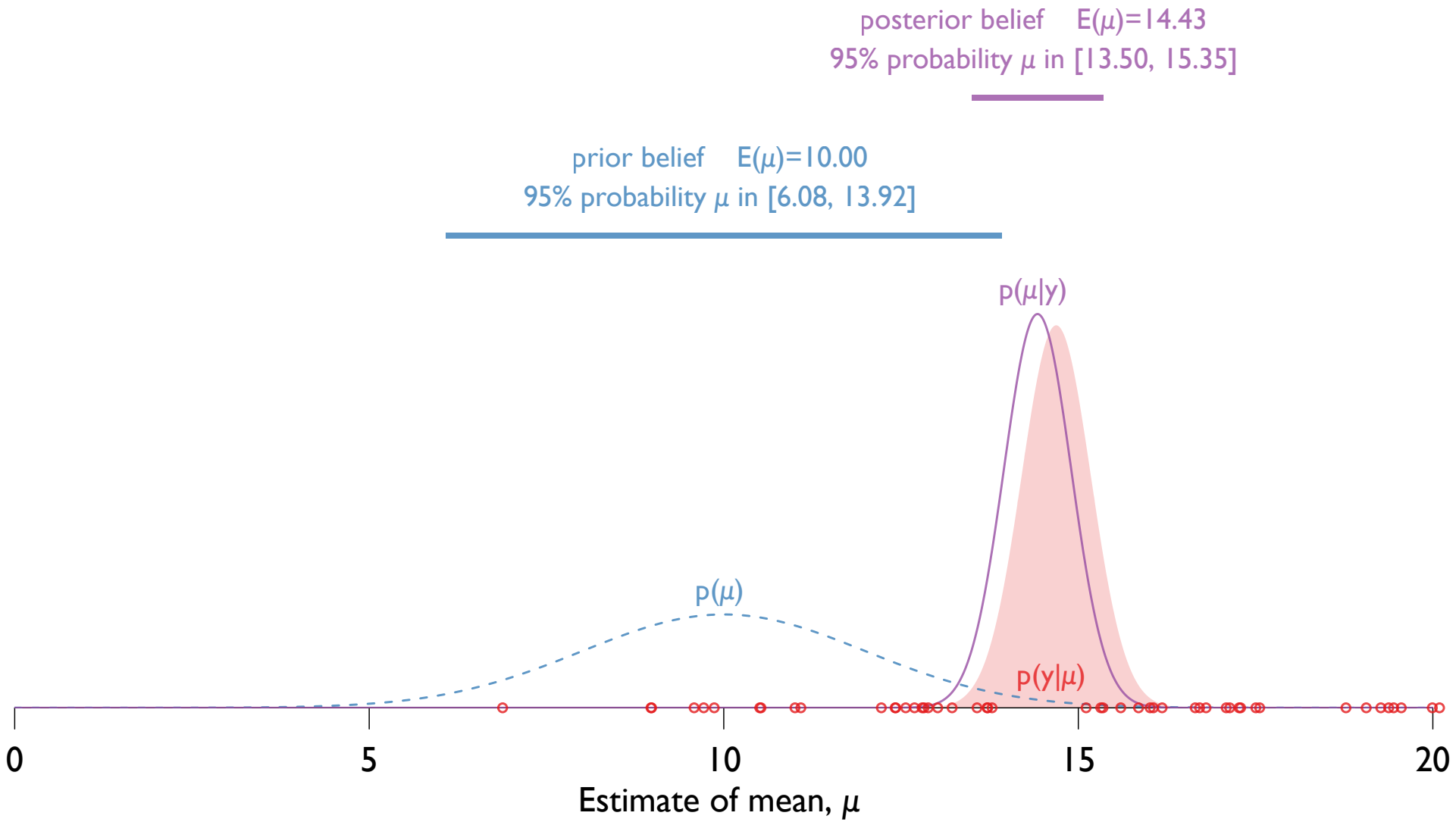


Our likelihood is now sharper, because we have a larger sample to work with



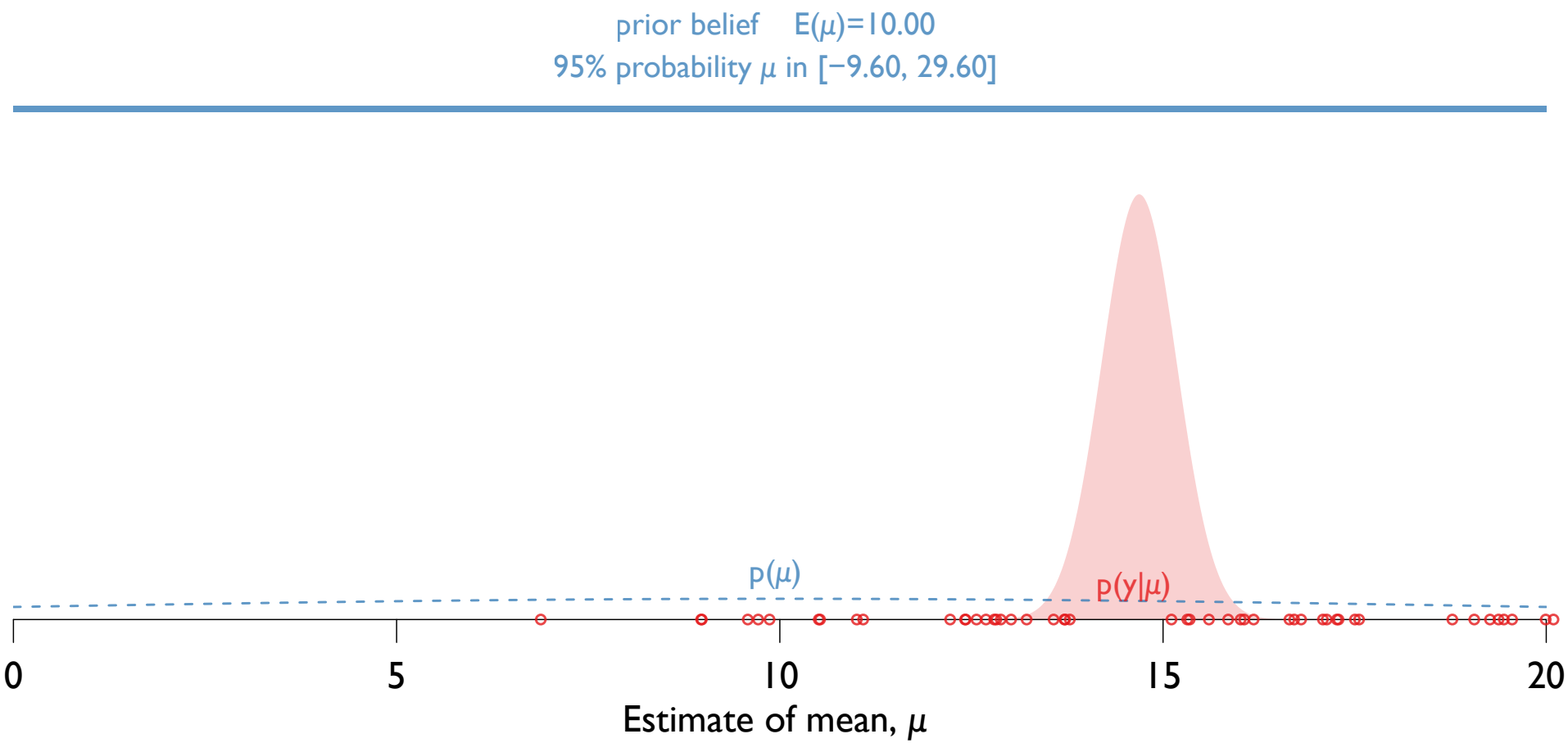
This more *informative* likelihood has a stronger influence on the posterior

Our posterior beliefs are closer to the sample mean and more certain than before



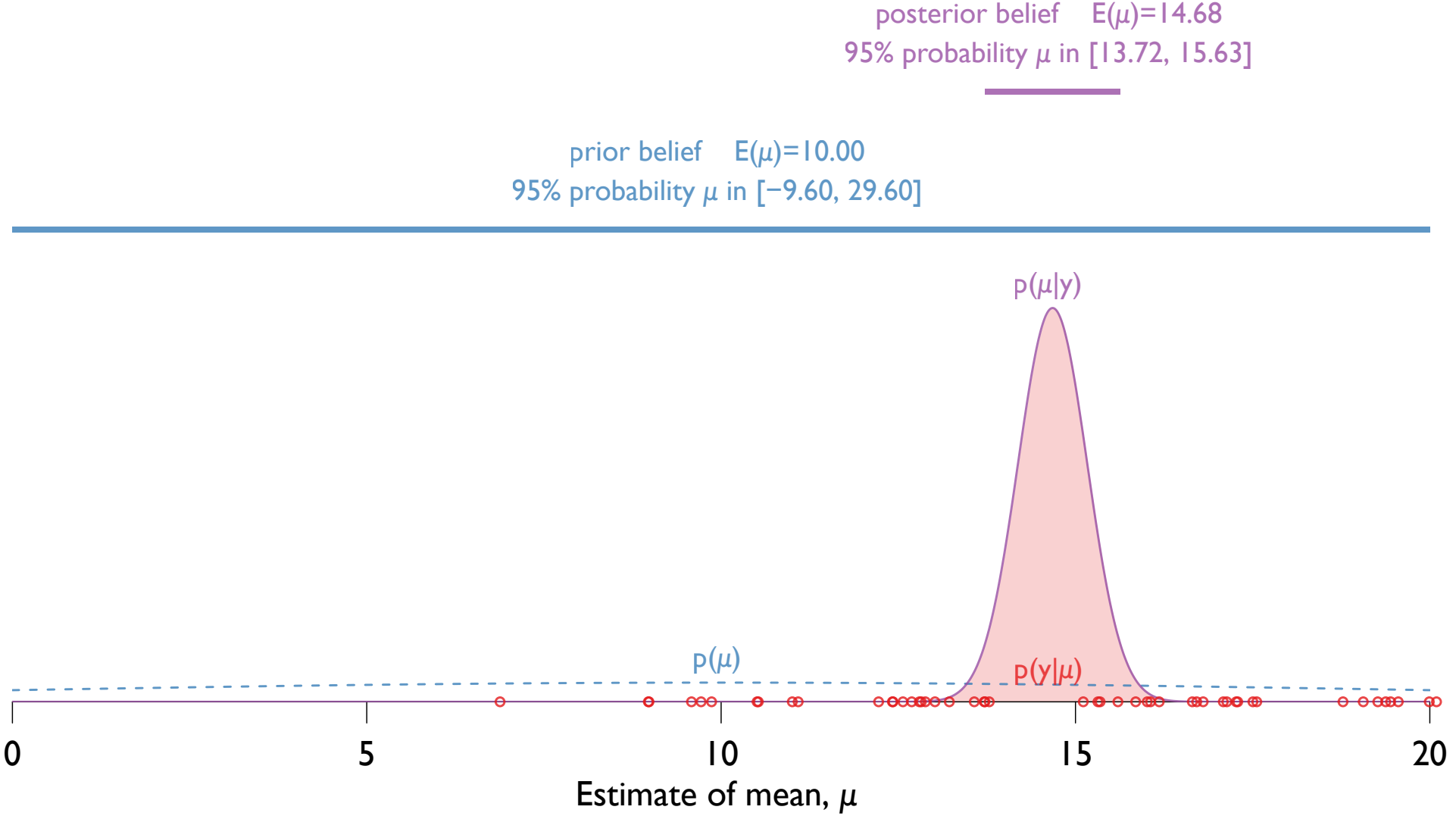
So far, we have specified *informative* priors to capture our subjective beliefs

But what if we wanted to be more agnostic?

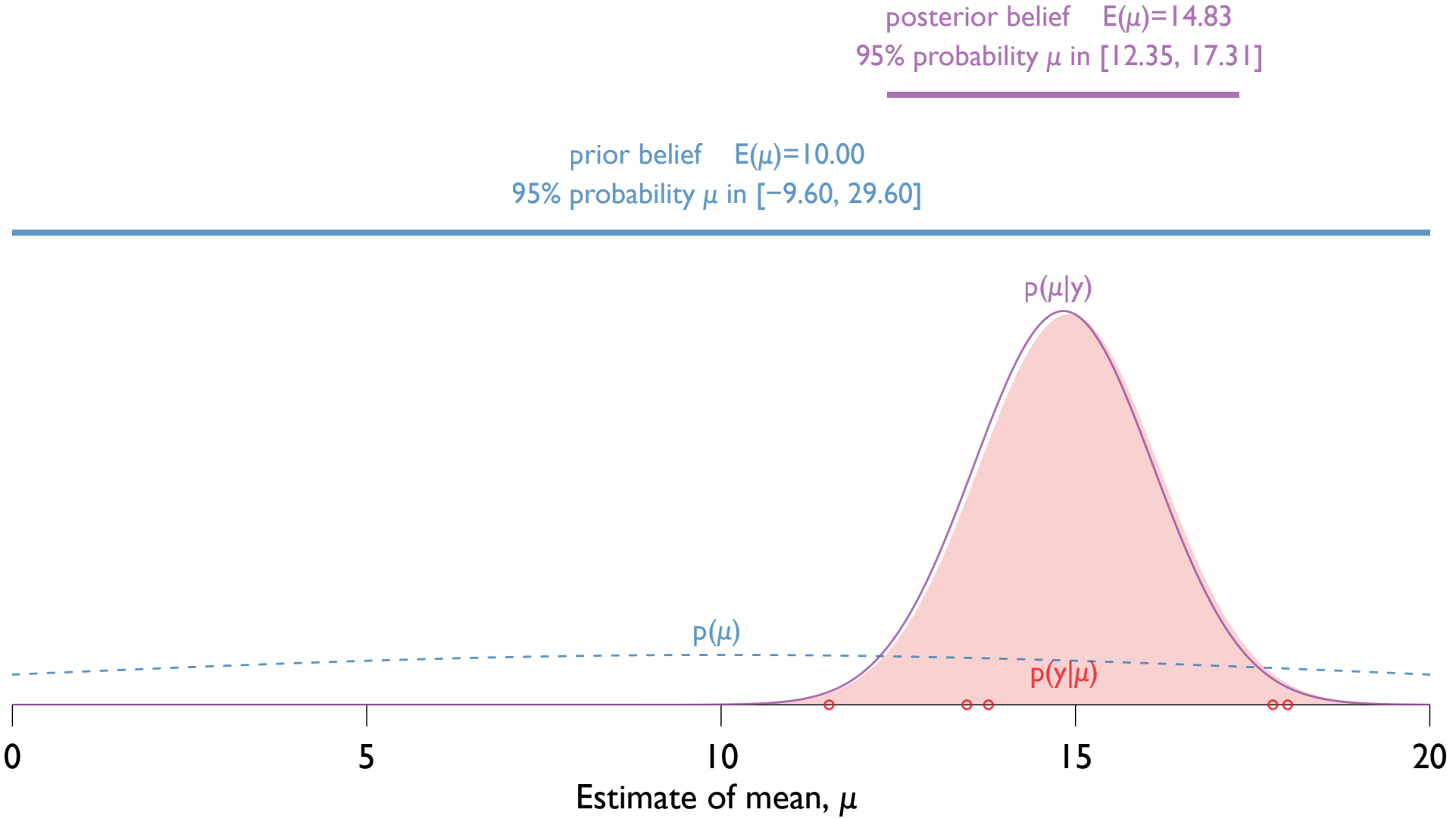


We could instead set a *diffuse* prior with a high variance

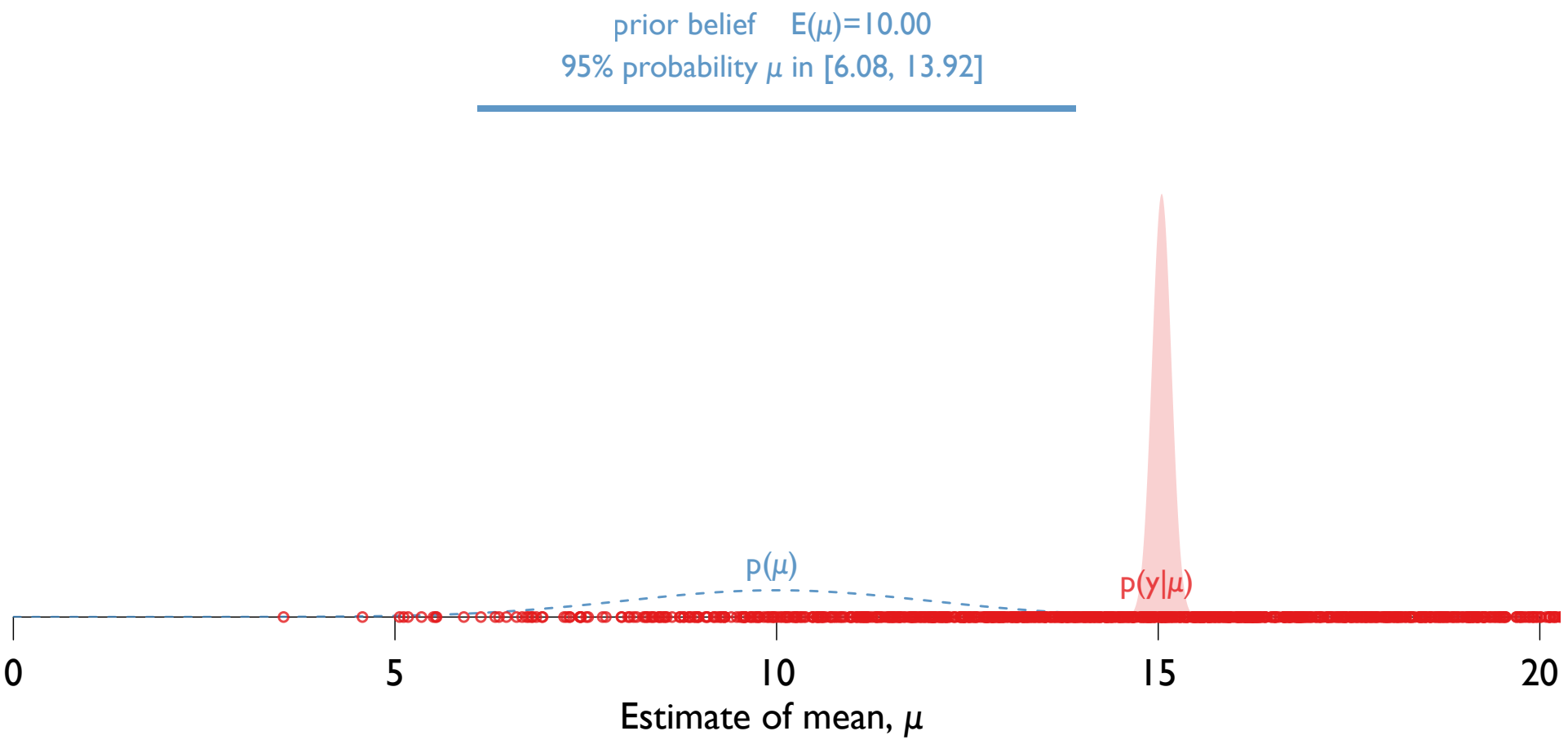
We still believe *a priori* that the most probable average workload is 10 hours, but now we consider a wide range of workloads almost as likely



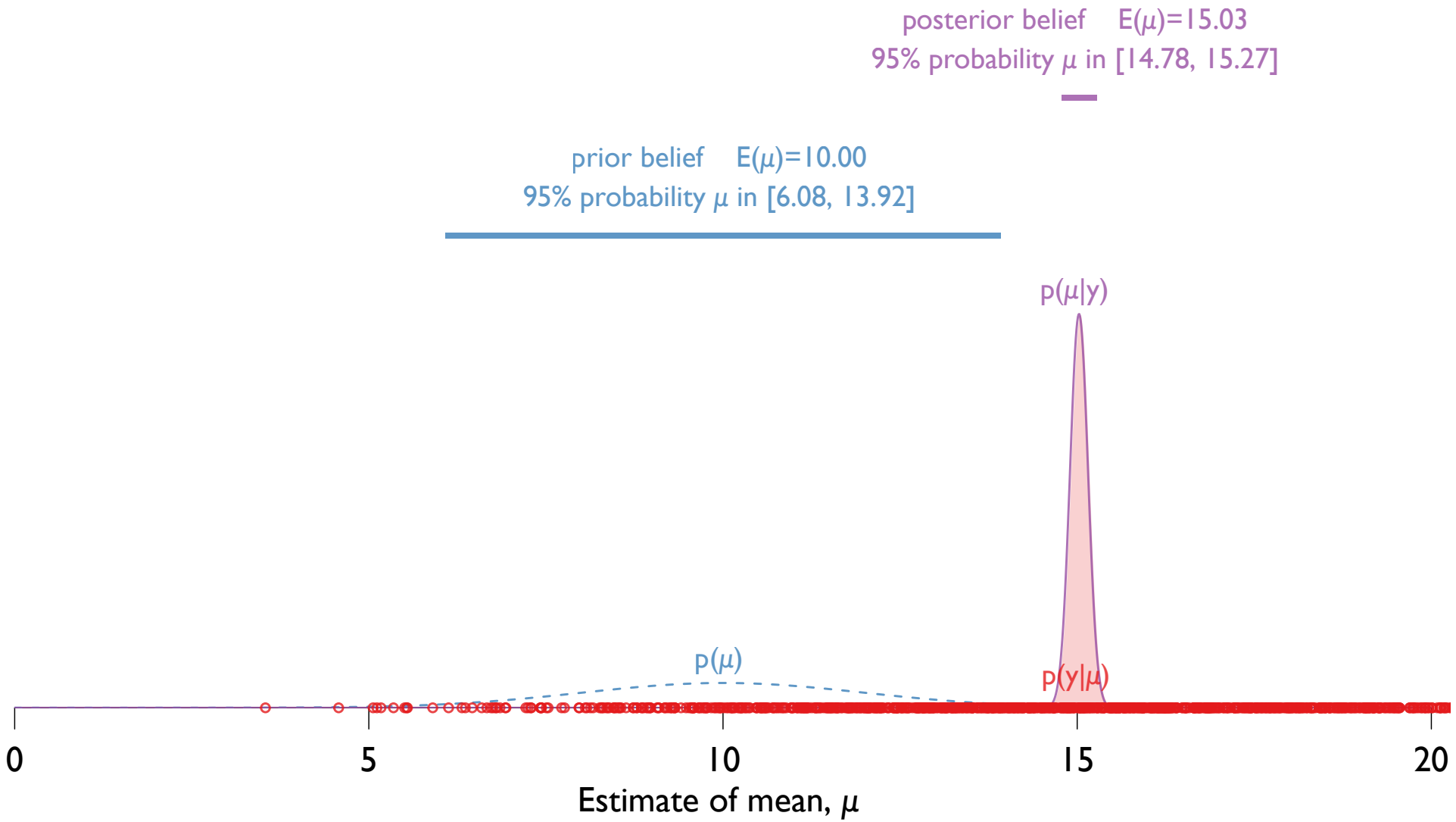
Because the prior now offers so little information, the posterior is dominated by the sampled data



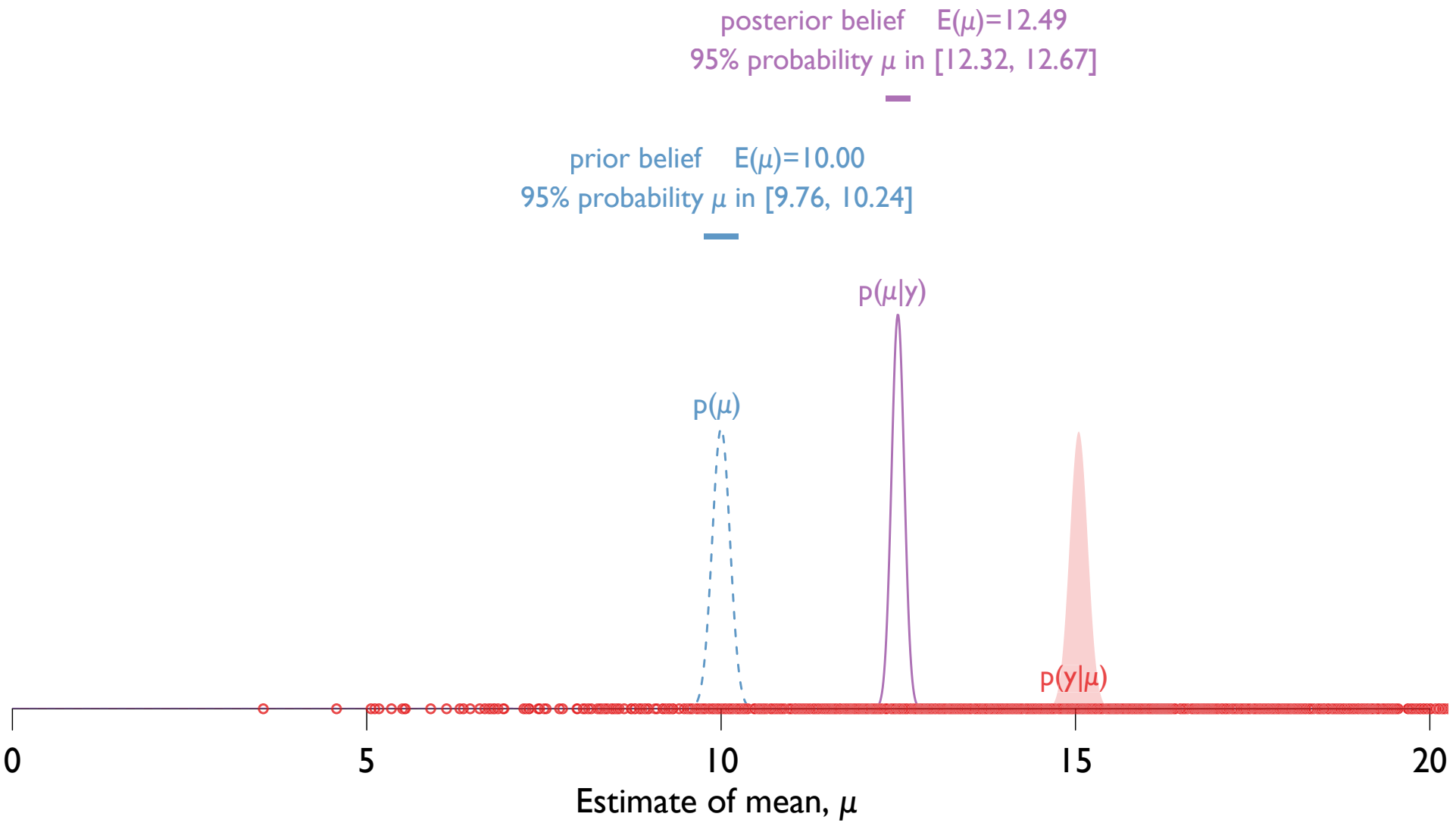
But if the sample is very small, even a diffuse prior can influence the posterior a little



What do you think will happen
if we combine our original, informative prior with our larger sample?

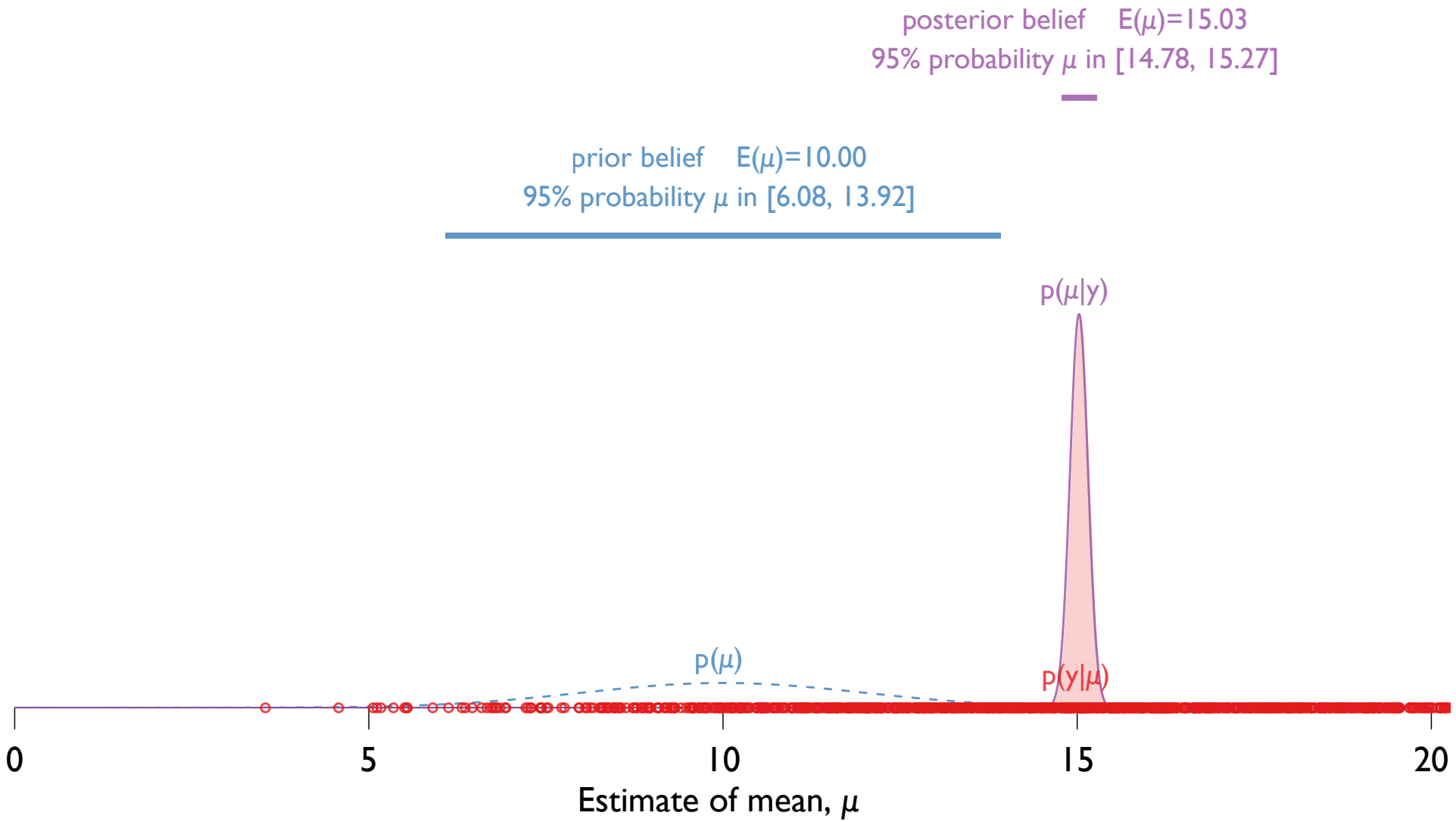


In this case, the data dominates almost completely



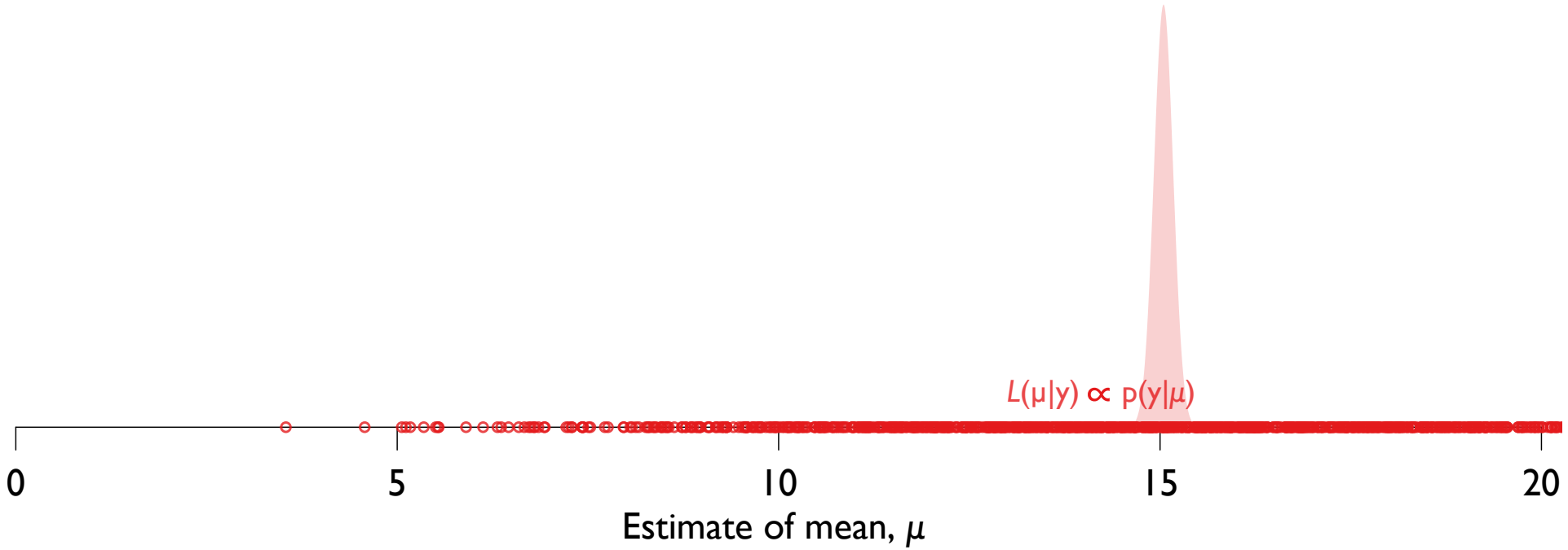
In this case, the data dominates almost completely

Of course, a sharp enough prior would force the posterior back into a compromise,

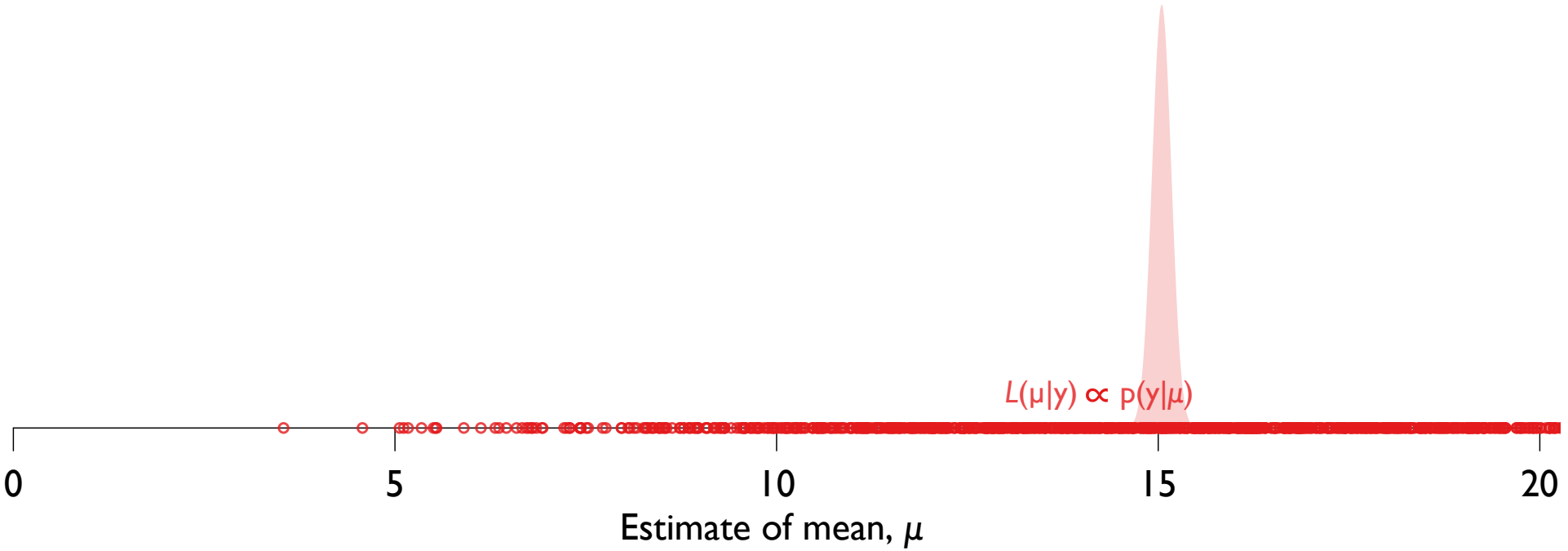


In this case, the data dominates almost completely

Of course, a sharp enough prior would force the posterior back into a compromise, but it's not likely we'd be so certain *a priori*



This example suggests that if we have a large enough sample,
the likelihood by itself is a good summary of where the most likely values of μ are



This example suggests that if we have a large enough sample,
the likelihood by itself is a good summary of where the most likely values of μ are

The downside of using the likelihood by itself is the loss of clear probability statements

Likelihood inference

What happens if you treat the prior as an unknown constant?

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) \quad = \quad \frac{\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}\mathrm{P}(\mathbf{y}|\boldsymbol{\theta})$$

Likelihood inference

What happens if you treat the prior as an unknown constant?

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= k(\mathbf{y})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

Likelihood inference

What happens if you treat the prior as an unknown constant?

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= k(\mathbf{y})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &\propto \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

The likelihood of the parameters given the data is proportional to the probability of the data given the parameters

Likelihood inference

What happens if you treat the prior as an unknown constant?

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= k(\mathbf{y})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &\propto \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

The likelihood of the parameters given the data is proportional to the probability of the data given the parameters

Though we can't objectively state the probability of a particular $\hat{\boldsymbol{\theta}}$ given $\mathbf{y}$, we can objectively state the (relative) likelihood of $\hat{\boldsymbol{\theta}}$ over some other $\hat{\boldsymbol{\theta}}'$:

Likelihood inference

What happens if you treat the prior as an unknown constant?

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= k(\mathbf{y})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &\propto \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

The likelihood of the parameters given the data is proportional to the probability of the data given the parameters

Though we can't objectively state the probability of a particular $\hat{\boldsymbol{\theta}}$ given $\mathbf{y}$, we can objectively state the (relative) likelihood of $\hat{\boldsymbol{\theta}}$ over some other $\hat{\boldsymbol{\theta}}'$:

$\mathcal{L}$ is a surface in $\boldsymbol{\theta}$ space showing which parameter values are more likely than others

Likelihood inference

What happens if you treat the prior as an unknown constant?

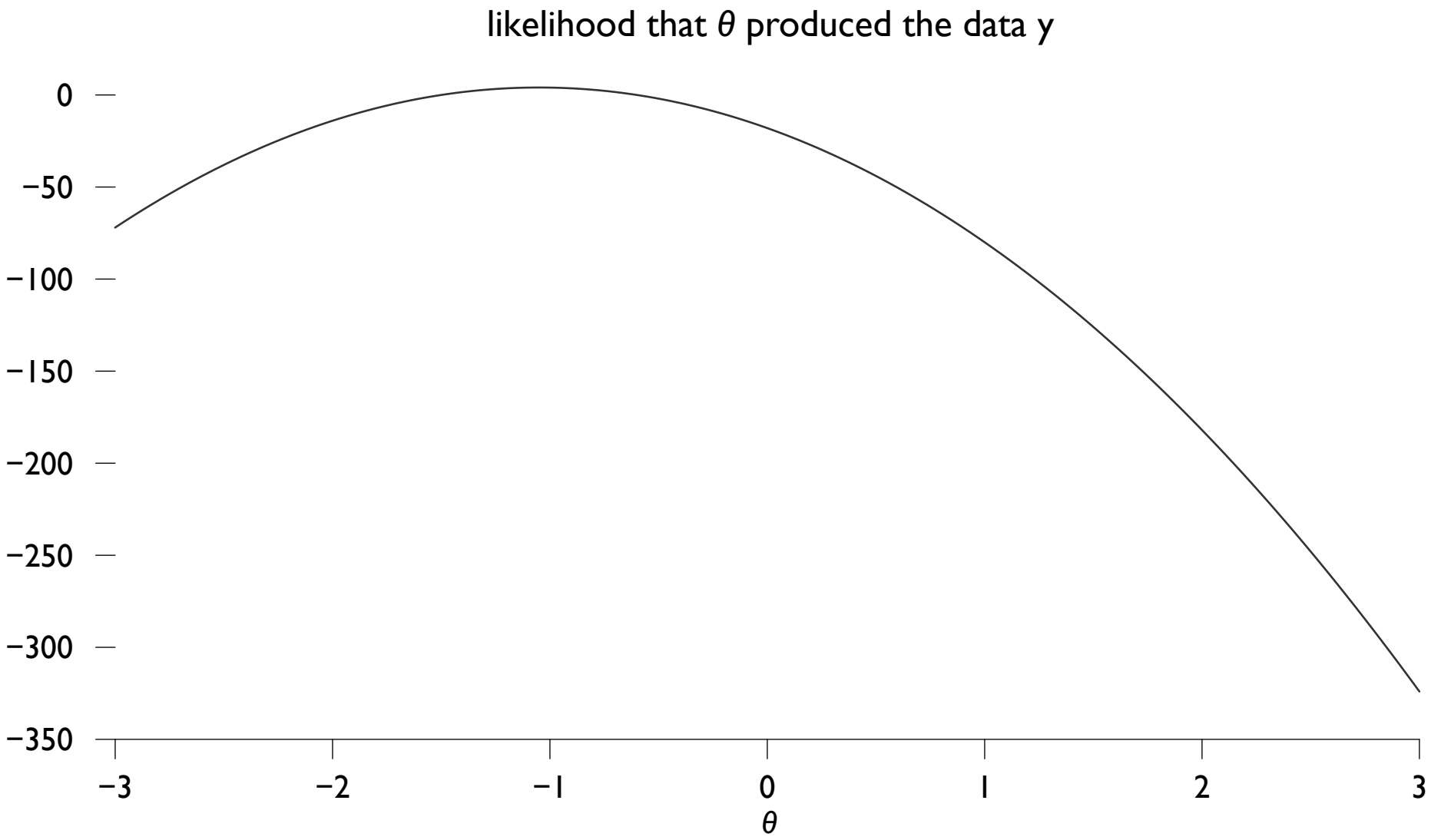
$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y})}\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= k(\mathbf{y})\mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &\propto \mathrm{P}(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

The likelihood of the parameters given the data is proportional to the probability of the data given the parameters

Though we can't objectively state the probability of a particular $\hat{\boldsymbol{\theta}}$ given $\mathbf{y}$, we can objectively state the (relative) likelihood of $\hat{\boldsymbol{\theta}}$ over some other $\hat{\boldsymbol{\theta}}'$:

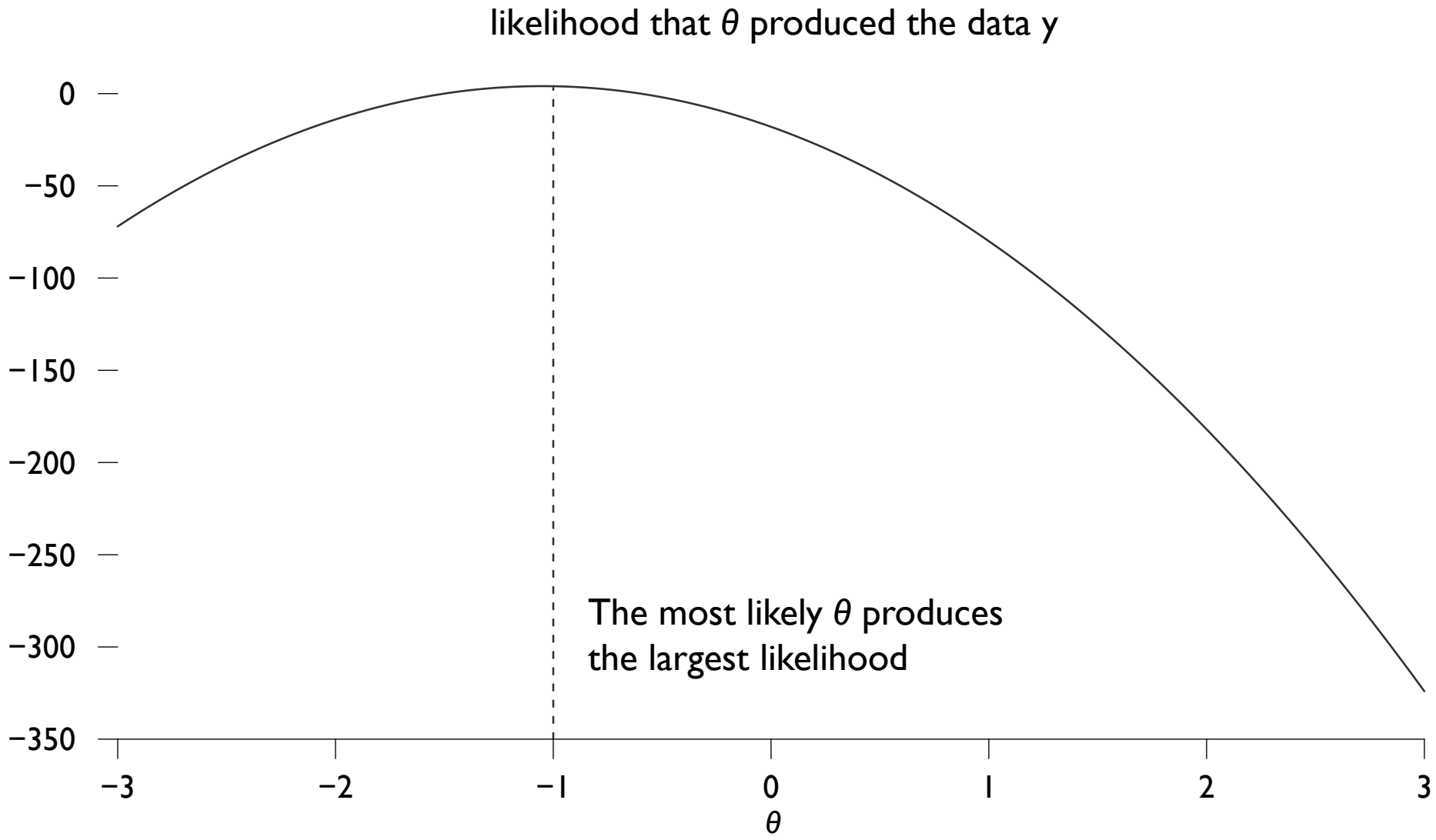
$\mathcal{L}$ is a surface in $\boldsymbol{\theta}$ space showing which parameter values are more likely than others

We can look at the profile of the likelihood function against each parameter in $\boldsymbol{\theta}$ to see which $\hat{\boldsymbol{\theta}}$'s are likely



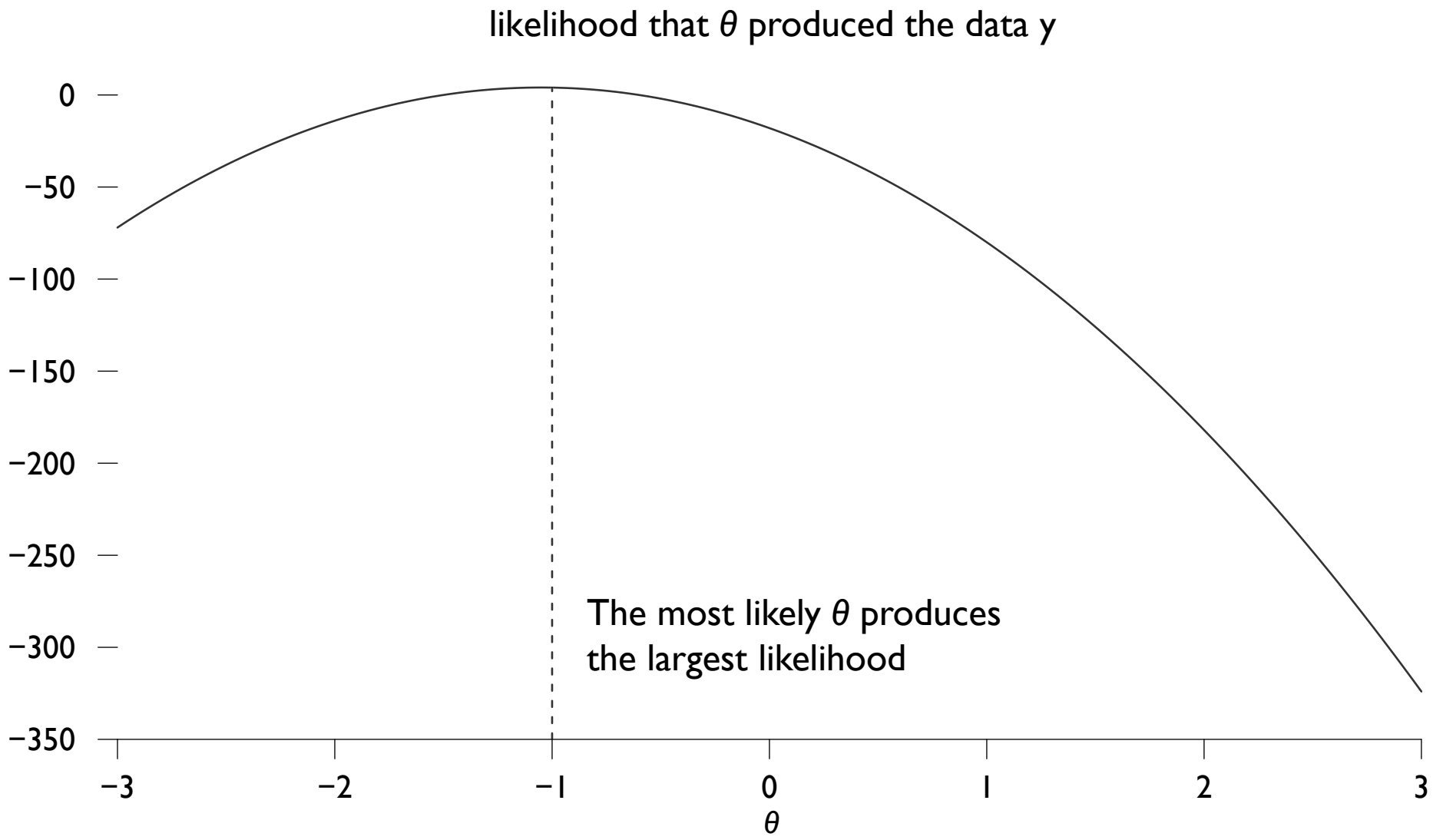
In many cases, the likelihood will approx quadratic, with a single maximum

In this case, the most likely θ appears to be -1.



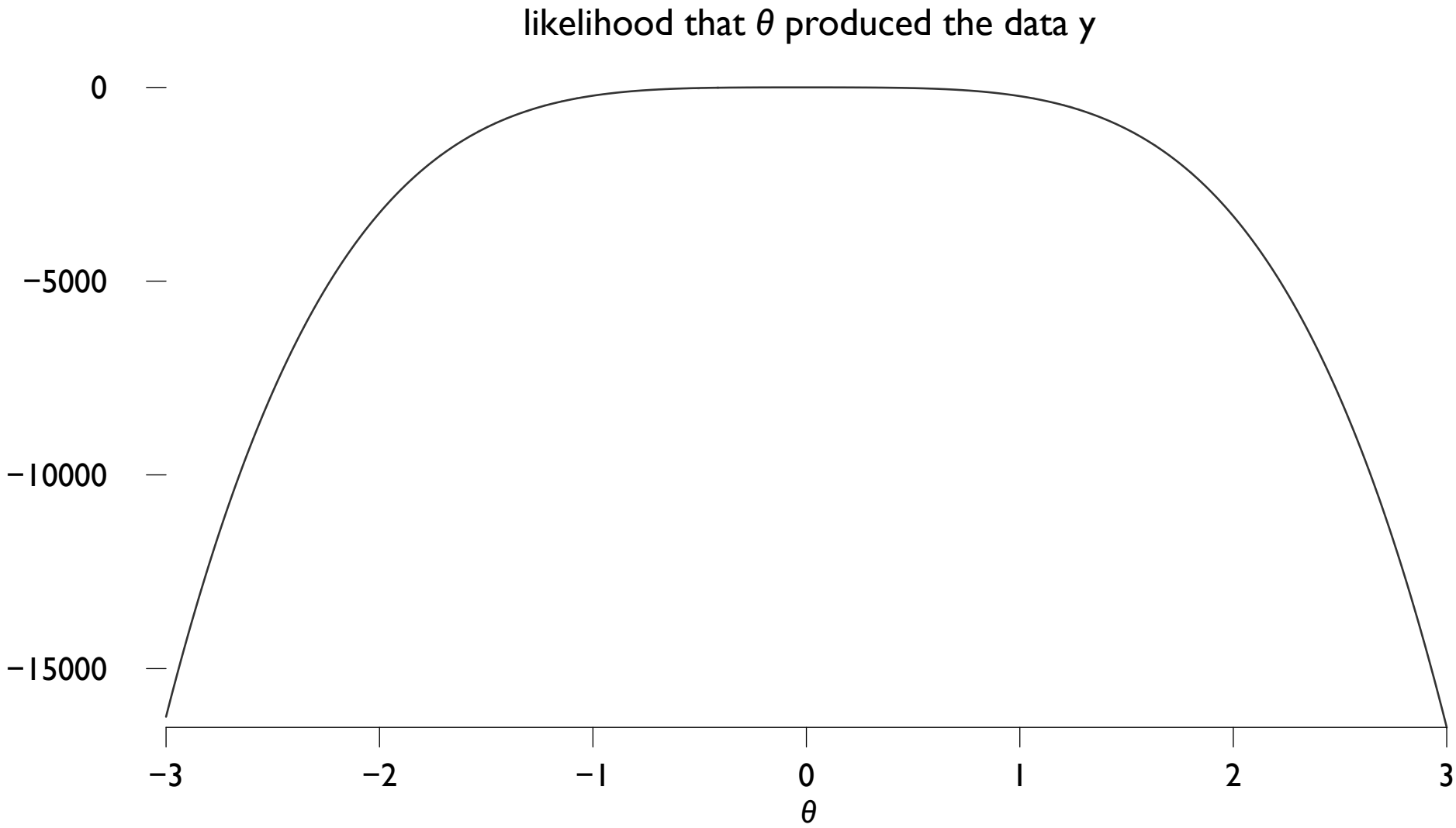
θ's get less likely as the get farther from -1,
but the likelihood profile reminds us values near -1 are almost as likely to be true

And even somewhat distant values could be the true θ



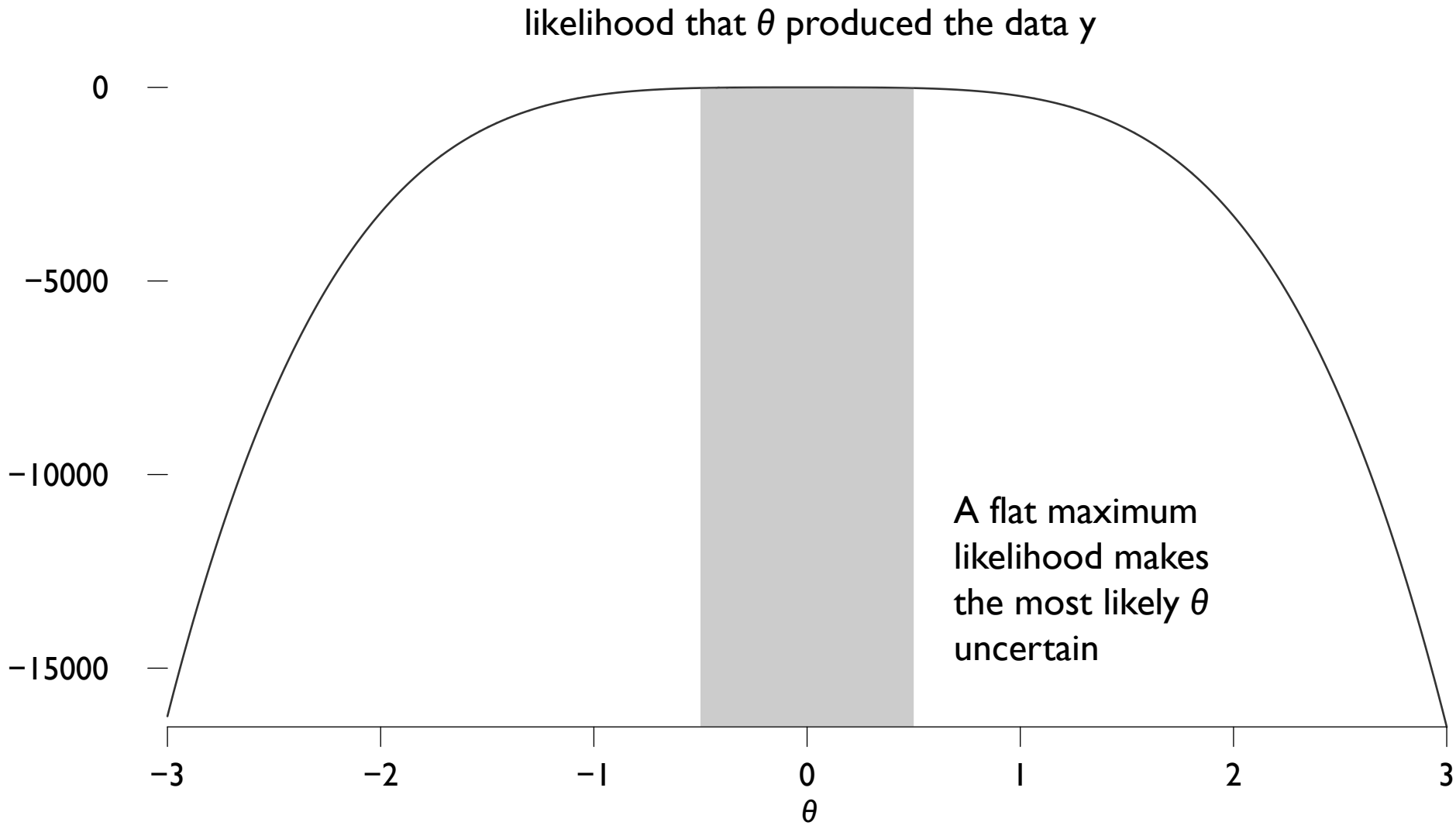
Note we cannot attach probabilities to each θ (the vertical axis is $\mathcal{L}$, not P)

This is the main difference between likelihood inference and Bayesian inference



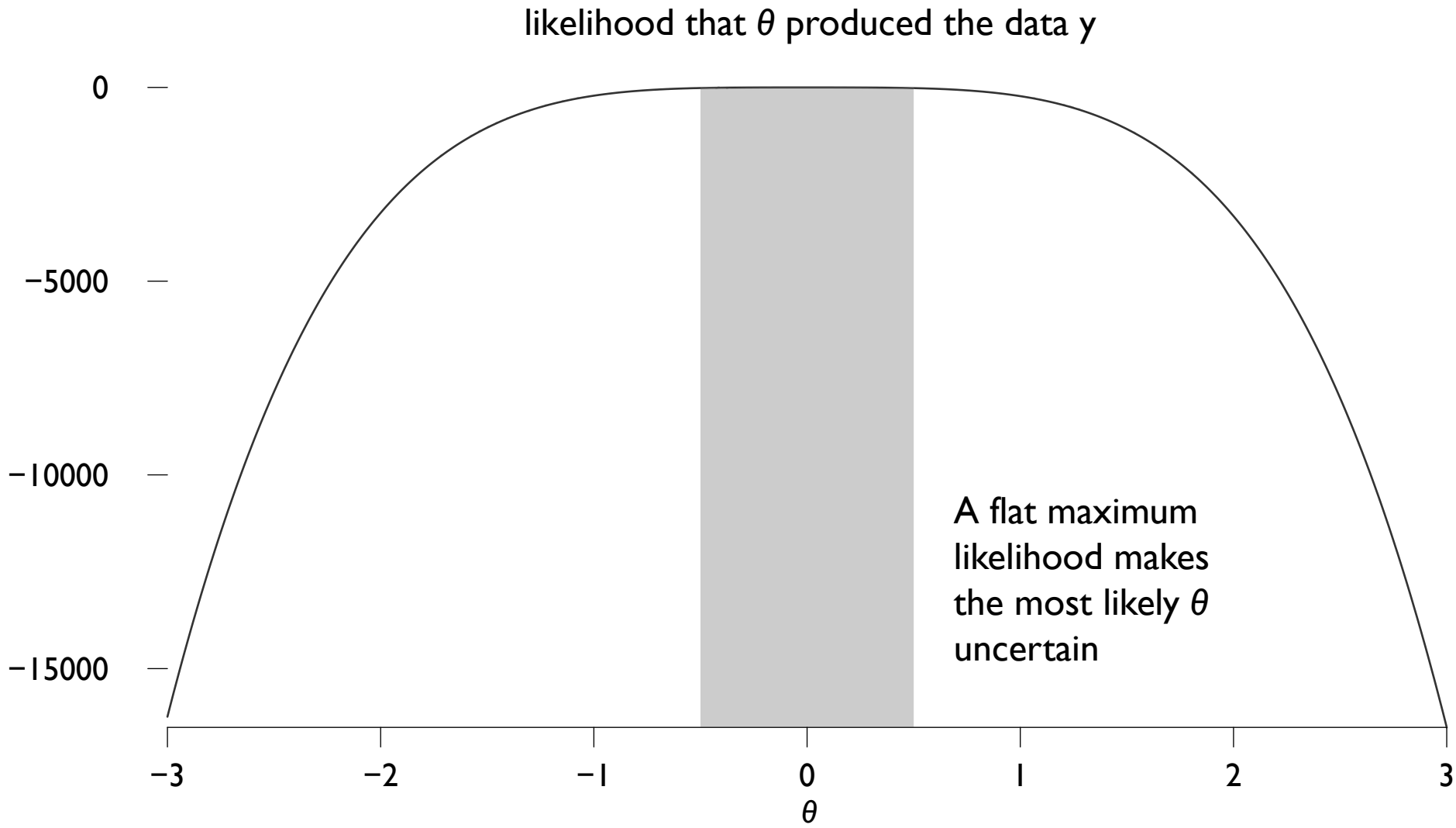
What if the likelihood is "flat" around the mean?

Here, the most likely θ appears to be 0. . .



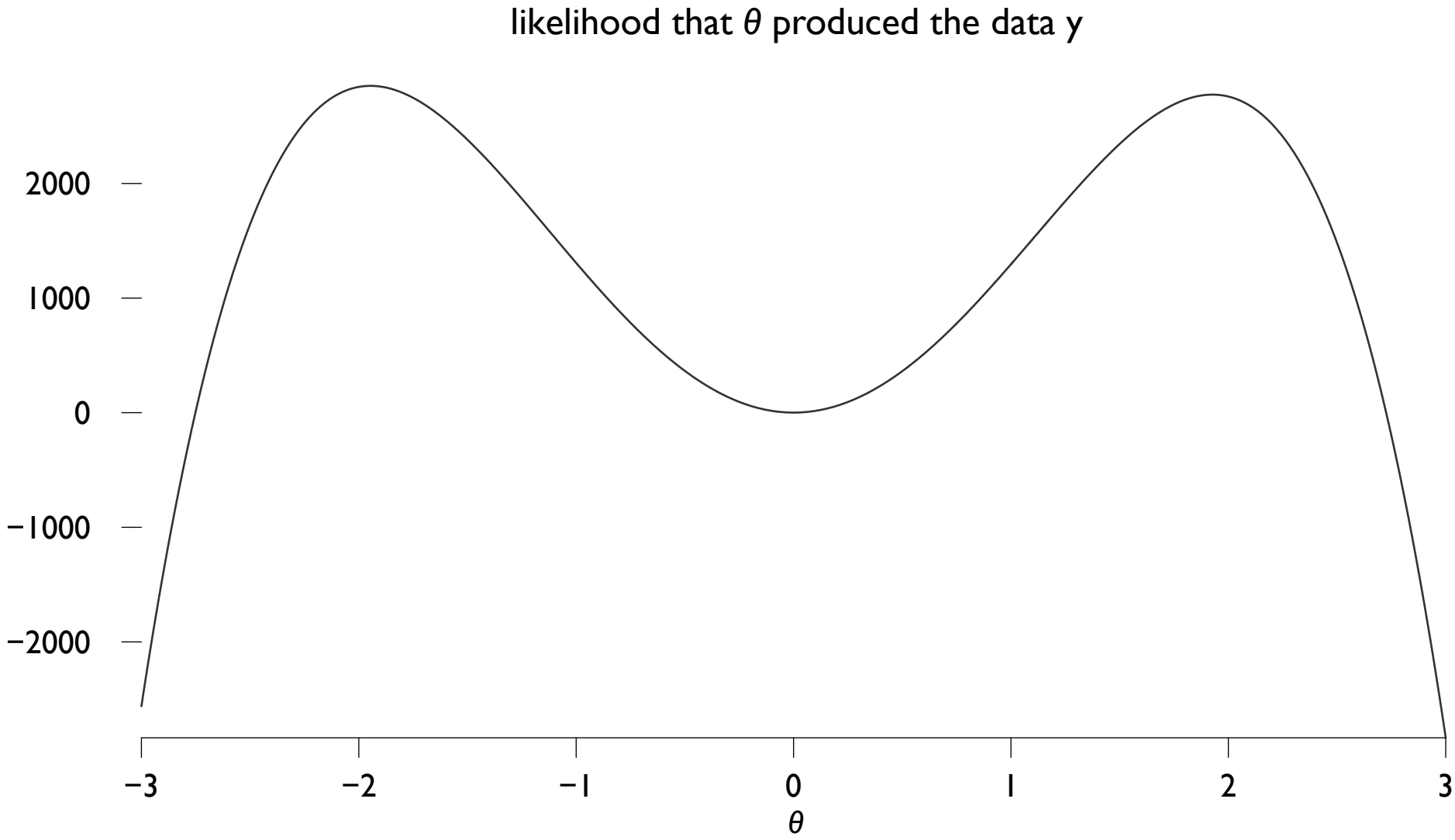
But θ's as far away as -0.5 and 0.5 seem equally likely

(What does this remind you of from your past methods classes?)

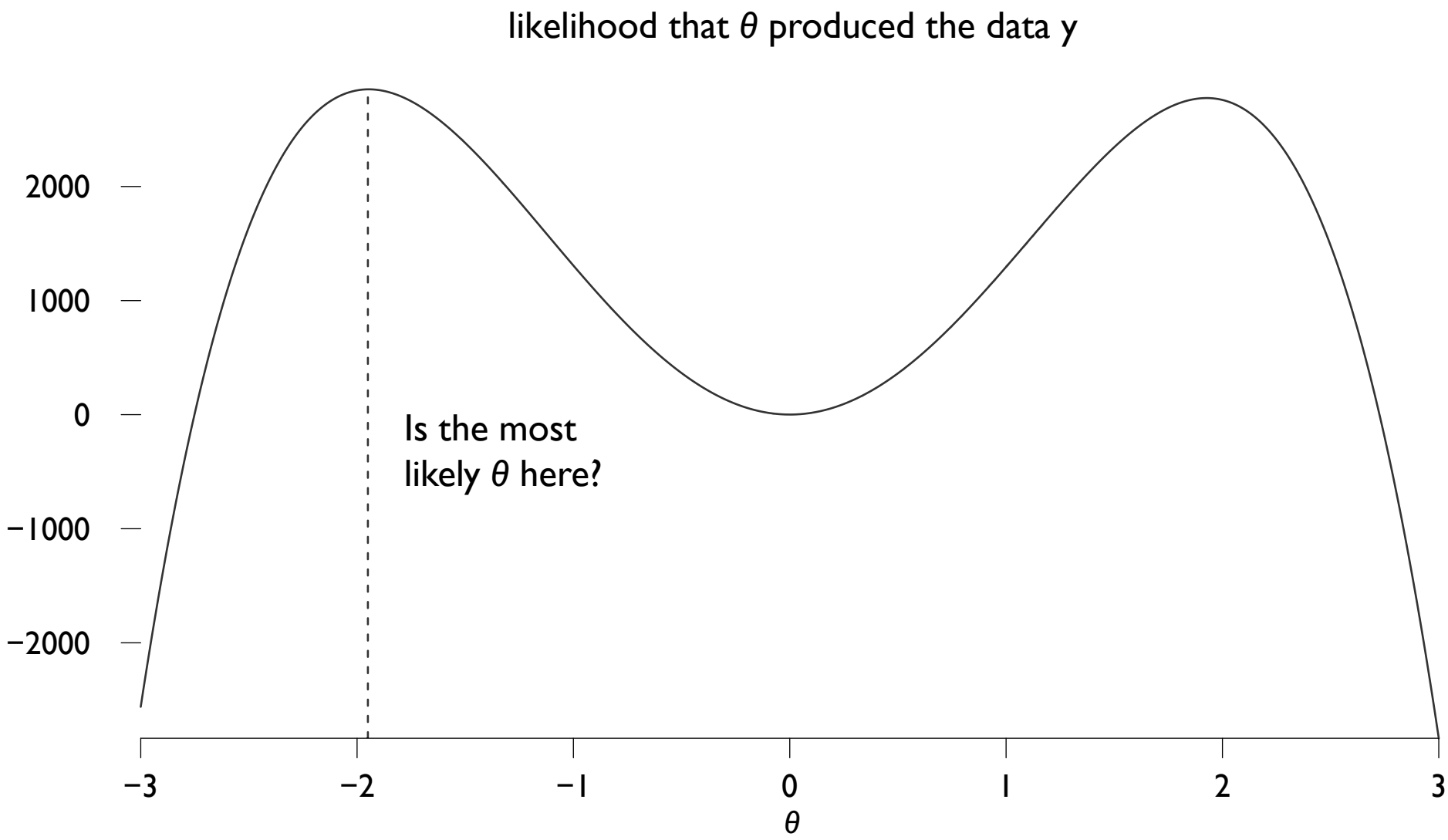


Flat $\mathcal{L} \Rightarrow$ insufficient information to discriminate among parameter values

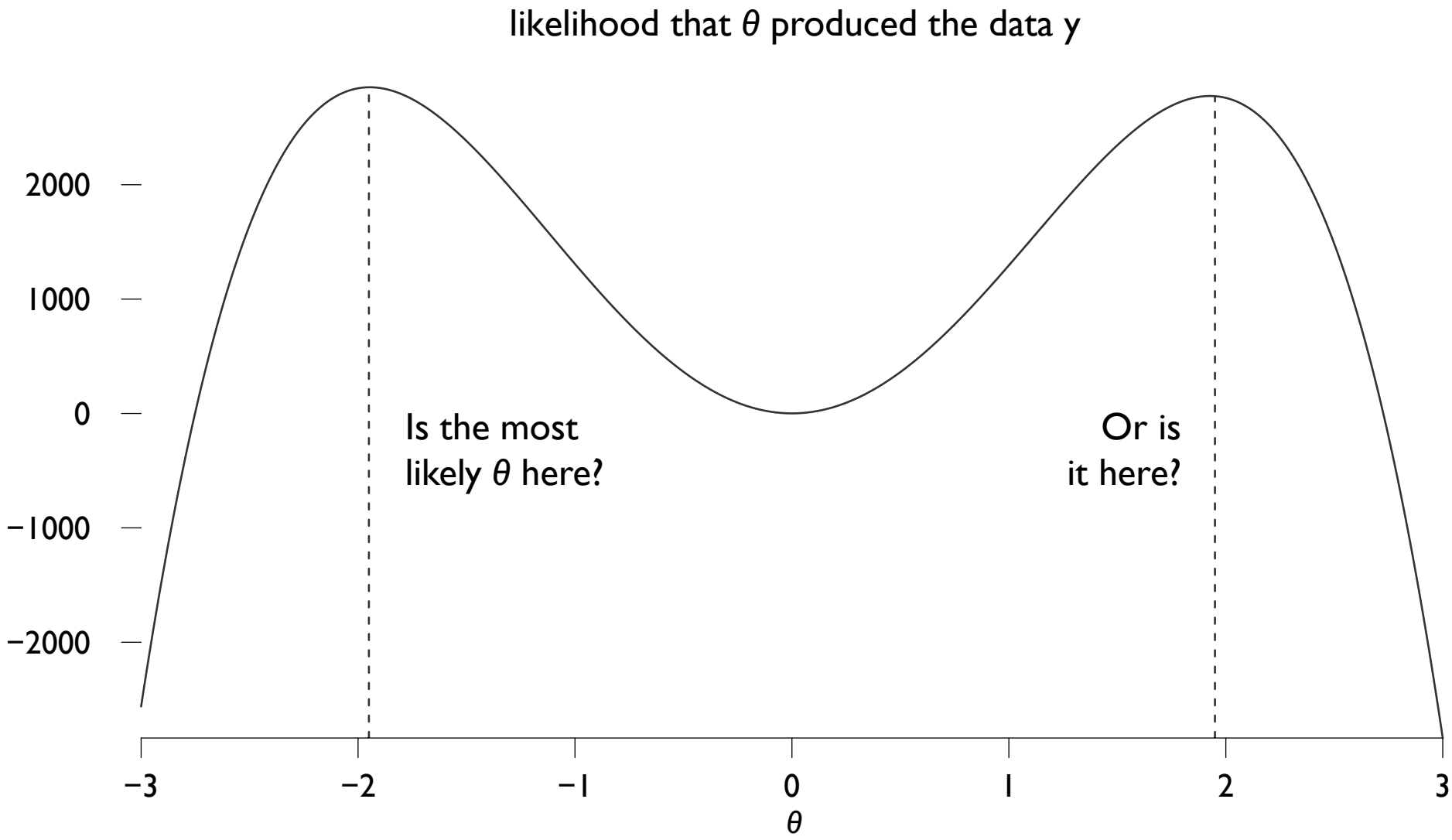
Mean by itself will be a misleading summary here (what would be better?)



How would you summarize this case?

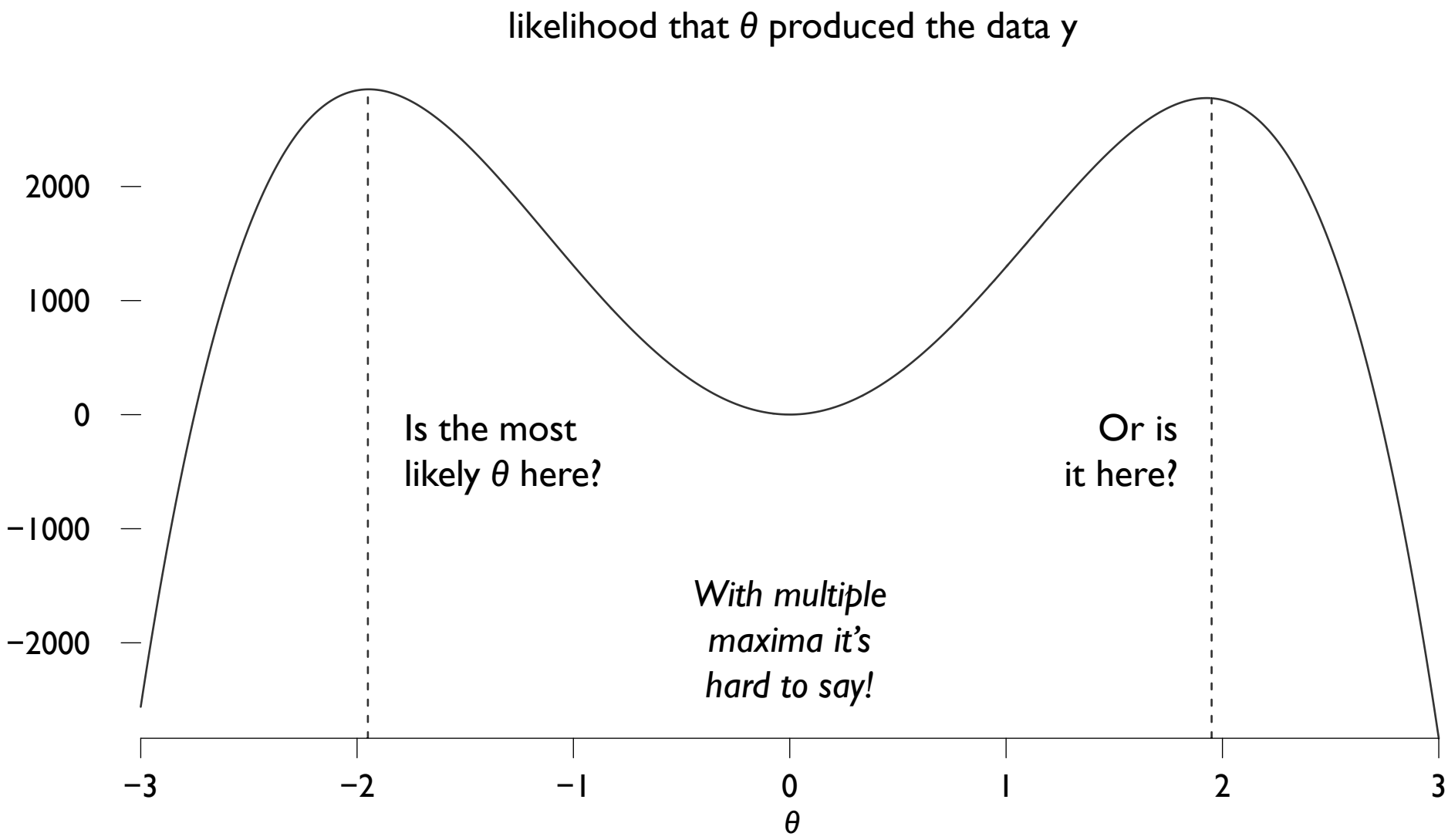


How would you summarize this case?



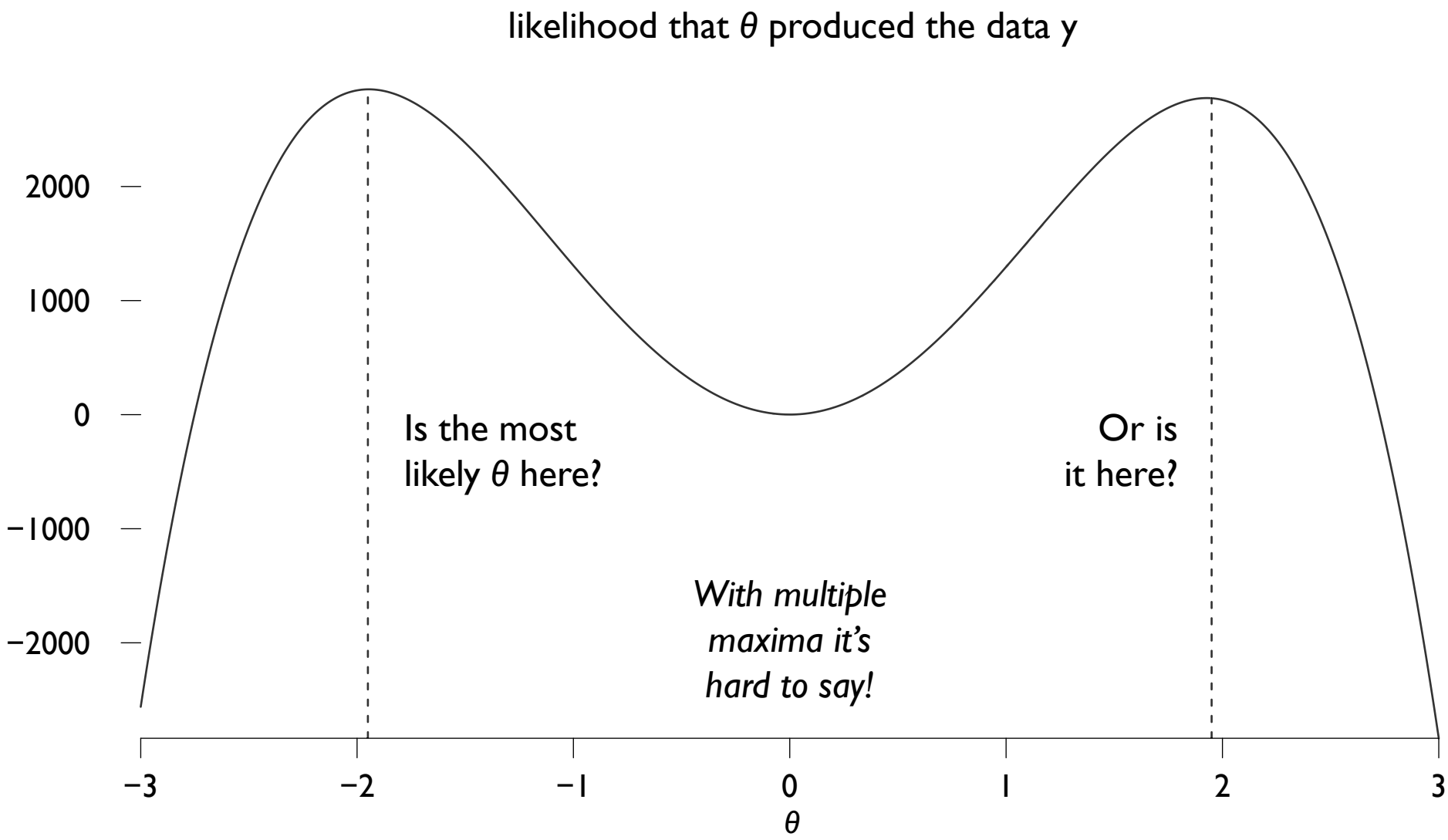
In unusual cases, the likelihood may have multiple modes or maxima

In this example, the most likely θ appears to be *either* around 2 or -2



But θ's in between are *less* likely

The mean by itself will be a very bad summary here – it's clearly not a particularly likely value



If we have many parameters, multimodal surfaces can be very hard to summarize

Fortunately, we won't encounter such likelihoods in *this* class,
but they are could occur in complex or unusual models

Profile Likelihoods: Example

Let's look at some real data

Profile Likelihoods: Example

Let's look at some real data

The turnout in 39 counties for the 2004 Washington State gubernatorial election

Profile Likelihoods: Example

Let's look at some real data

The turnout in 39 counties for the 2004 Washington State gubernatorial election

A good model of turnout would incorporate each county's unique features

We're going to estimate an oversimplified model for pedagogical purposes

(This is *not* a model we would want to use for anything important)

Profile Likelihoods: Example

Let's look at some real data

The turnout in 39 counties for the 2004 Washington State gubernatorial election

A good model of turnout would incorporate each county's unique features

We're going to estimate an oversimplified model for pedagogical purposes

(This is *not* a model we would want to use for anything important)

We'll assume voters in each county have the same probability of turning out

Profile Likelihoods: Example

Let's look at some real data

The turnout in 39 counties for the 2004 Washington State gubernatorial election

A good model of turnout would incorporate each county's unique features

We're going to estimate an oversimplified model for pedagogical purposes

(This is *not* a model we would want to use for anything important)

We'll assume voters in each county have the same probability of turning out

Under this assumption, each county's turnout can be treated as a binomial RV

Profile Likelihoods: Example

Let's look at some real data

The turnout in 39 counties for the 2004 Washington State gubernatorial election

A good model of turnout would incorporate each county's unique features

We're going to estimate an oversimplified model for pedagogical purposes

(This is *not* a model we would want to use for anything important)

We'll assume voters in each county have the same probability of turning out

Under this assumption, each county's turnout can be treated as a binomial RV

To find the ML estimate of the common turnout rate (i.e., π), we calculate $\mathcal{L}$

Profile Likelihoods: Example

For now, I'll give you the likelihood for the binomial (you'll derive as HW)

$$\mathcal{L}\left(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}\right) \propto \prod_{i=1}^{n} \frac{M_i!}{y_i!(M_i-y_i)!}\pi^{y_i}(1-\pi)^{M_i-y_i}$$

Profile Likelihoods: Example

For now, I'll give you the likelihood for the binomial (you'll derive as HW)

$$\mathcal{L}\left(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}\right) \propto \prod_{i=1}^{n} \frac{M_i!}{y_i!(M_i-y_i)!}\pi^{y_i}(1-\pi)^{M_i-y_i}$$

A practical problem: raising π's to large numbers will give R a headache

Profile Likelihoods: Example

For now, I'll give you the likelihood for the binomial (you'll derive as HW)

$$\mathcal{L}\left(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}\right) \propto \prod_{i=1}^{n} \frac{M_i!}{y_i!(M_i-y_i)!}\pi^{y_i}(1-\pi)^{M_i-y_i}$$

A practical problem: raising π's to large numbers will give R a headache

Transform $\mathcal{L}$ to maintain the same maximum, but less extreme values?

Profile Likelihoods: Example

For now, I'll give you the likelihood for the binomial (you'll derive as HW)

$$\mathcal{L}\left(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}\right) \propto \prod_{i=1}^{n} \frac{M_i!}{y_i!(M_i-y_i)!}\pi^{y_i}(1-\pi)^{M_i-y_i}$$

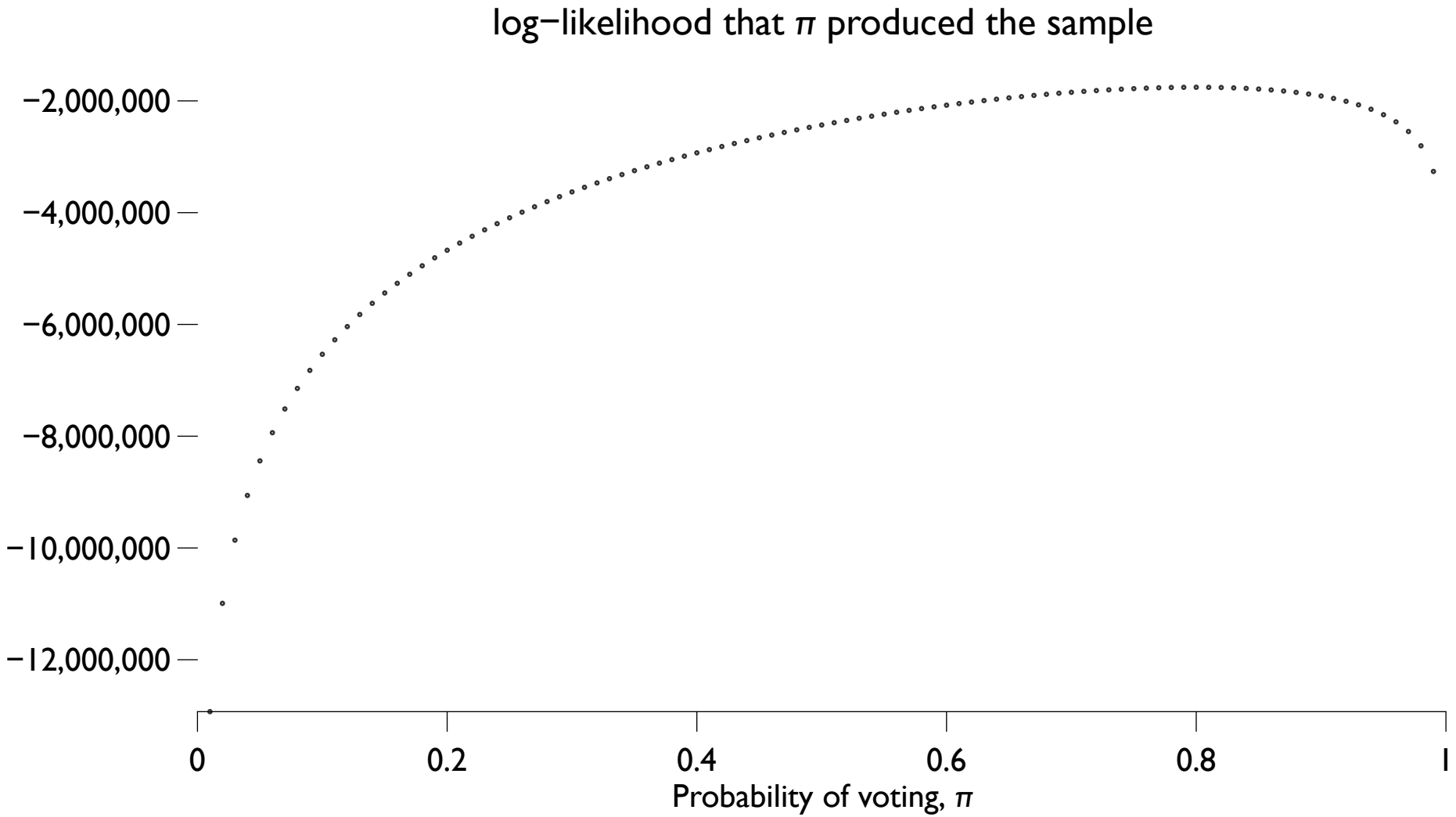
A practical problem: raising π's to large numbers will give `R` a headache

Transform $\mathcal{L}$ to maintain the same maximum, but less extreme values?

Let's try $\log \mathcal{L}$. In this case

$$\log \mathcal{L}\left(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}\right) \propto \sum_{i=1}^{n} y_i \log \pi + \sum_{i=1}^{n} (M_i - y_i)\log(1-\pi)$$

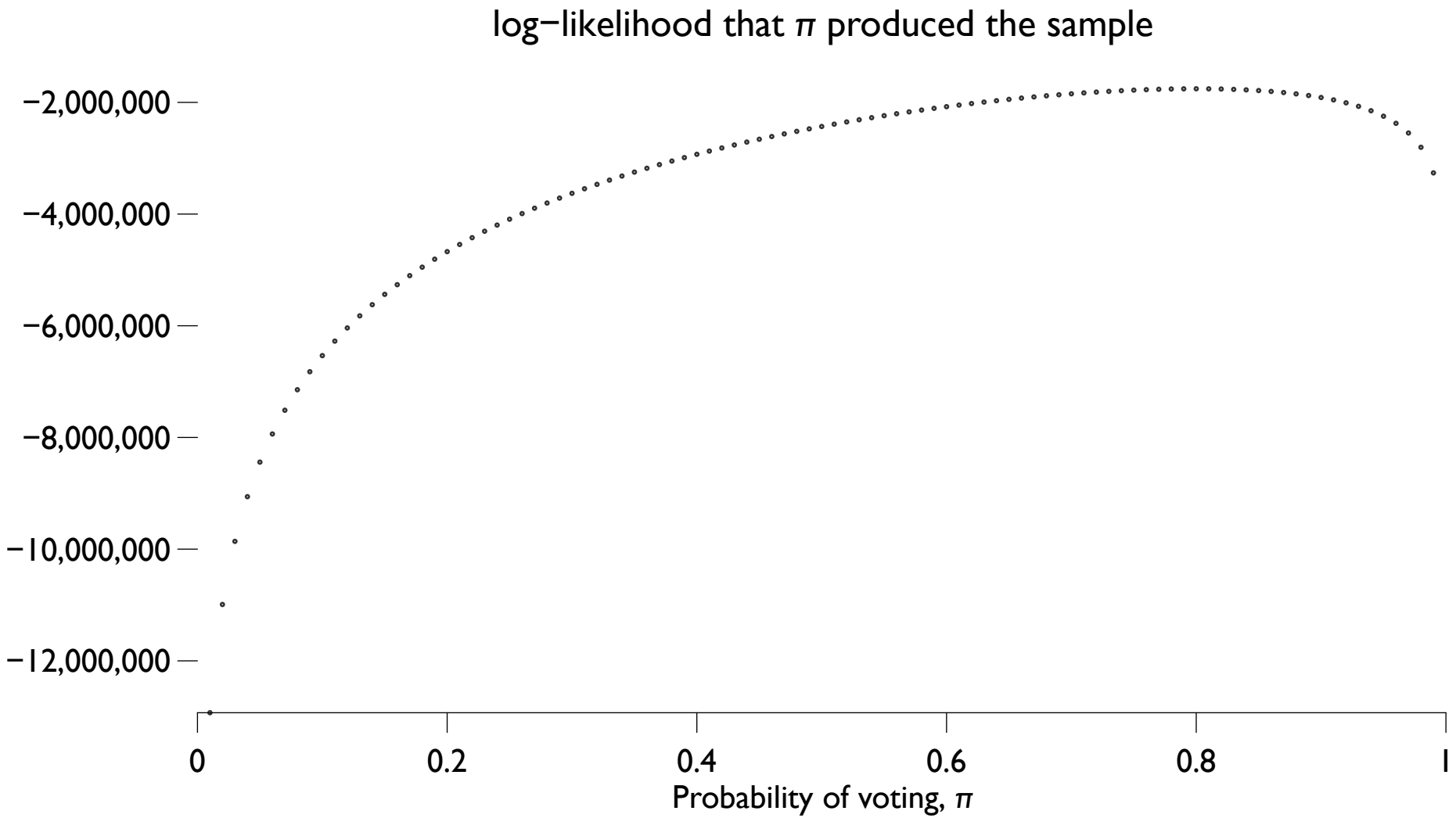
Because likelihoods are a relative measure only,
we're allowed to drop any terms that do not depend on estimated parameters;
what remains are *sufficient statistics* of $\log \mathcal{L}$



How do we use the likelihood to learn about the unknown parameter π?

Consider different π's in the possible range $[0, 1]$ and calculate $\log L$ for each

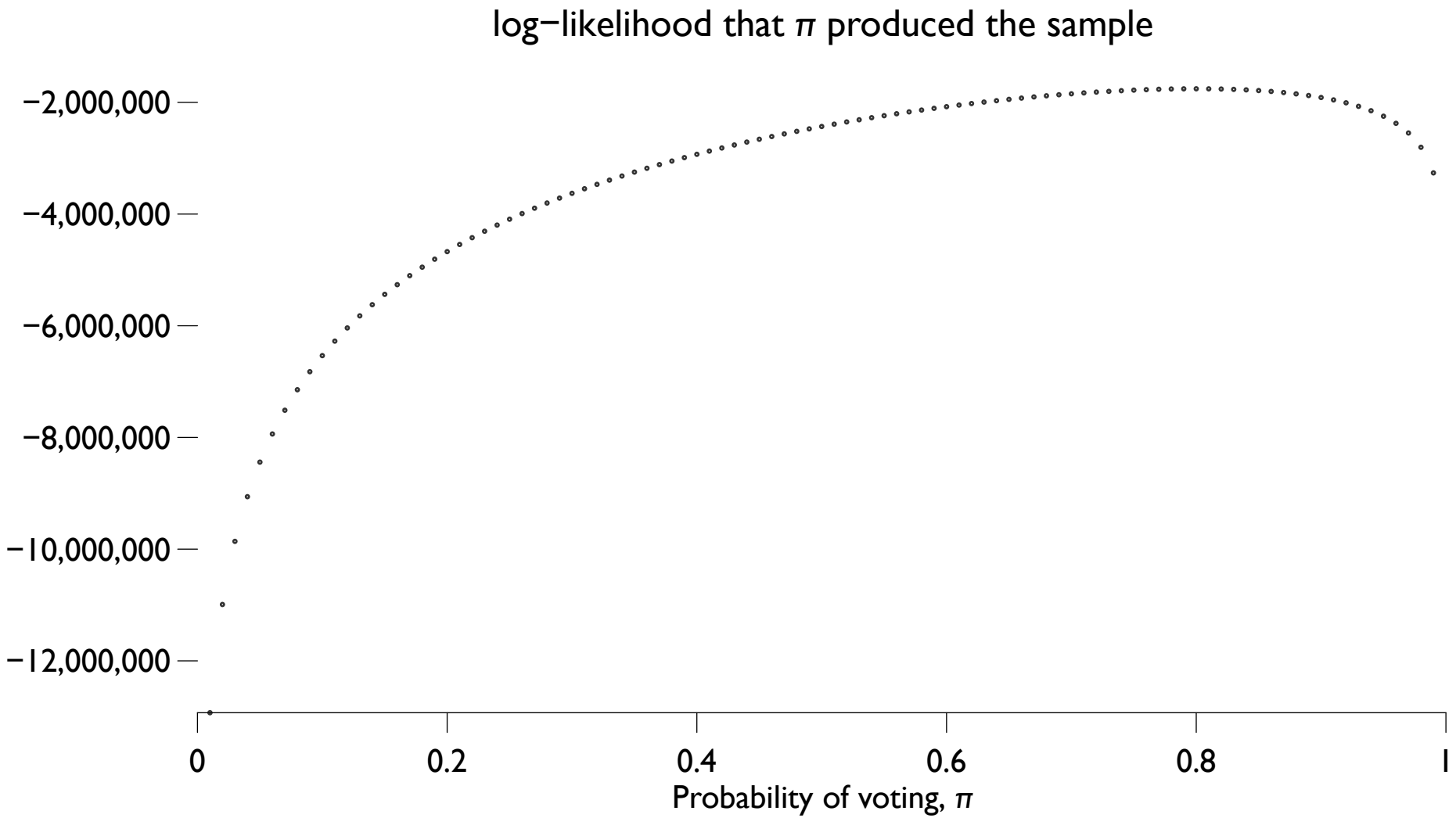
Then plot the $\log \mathcal{L}$'s against the π's to produce a *profile likelihood*



Remember that likelihoods (and log-likelihoods) are relative measures only

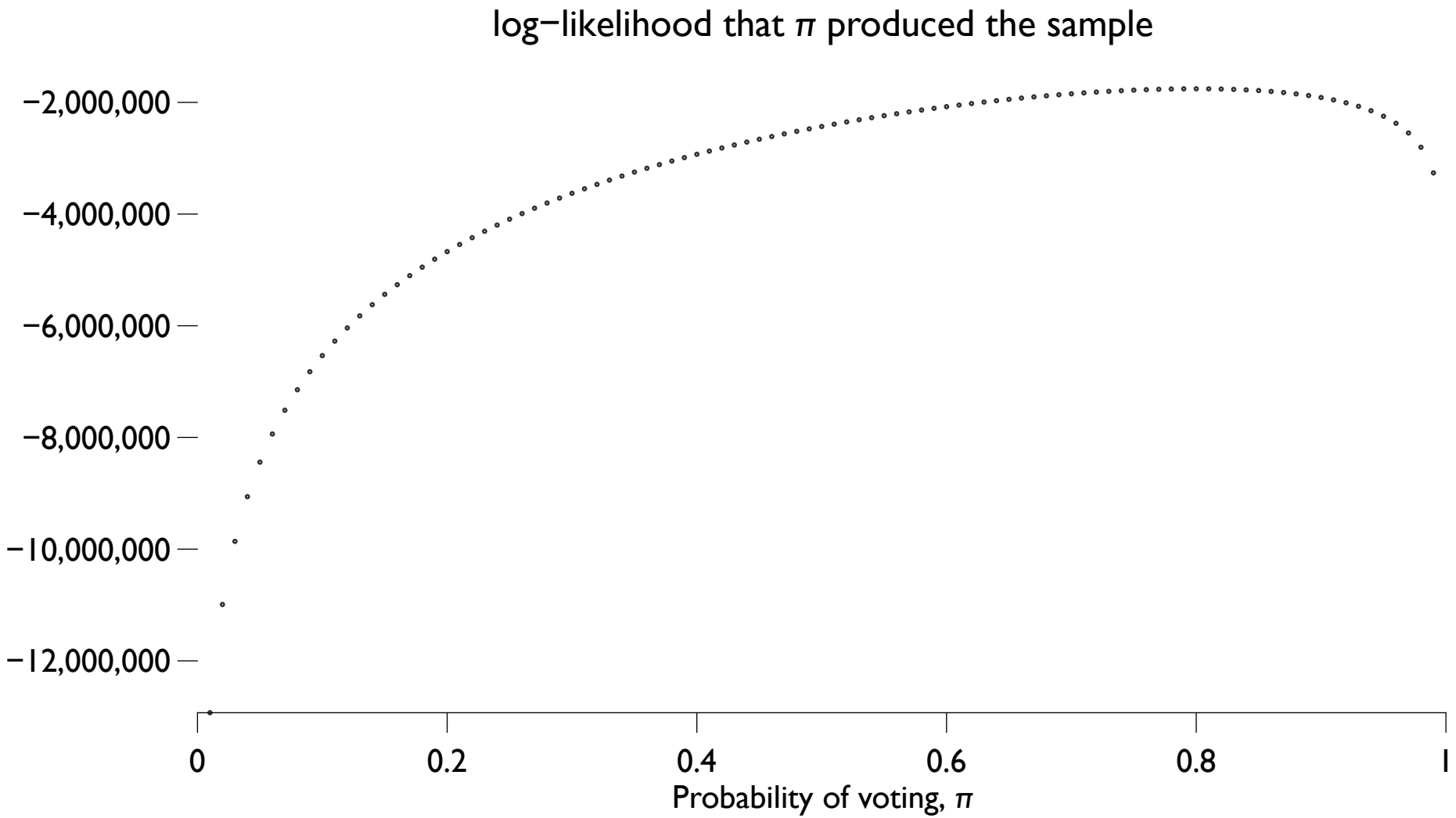
Higher likelihoods indicate more likely parameter values

But we don't know the probability an estimate of the parameter is correct



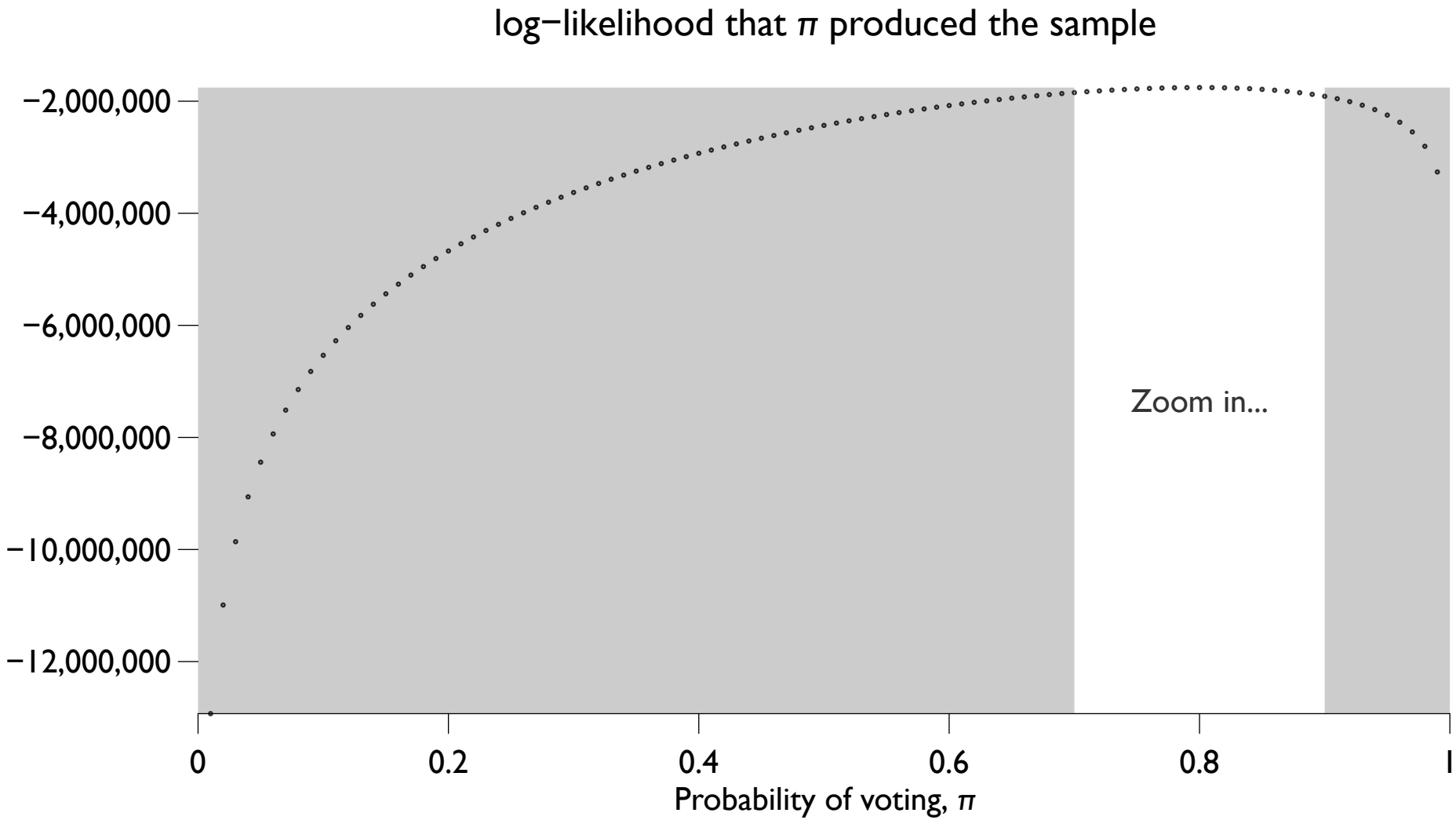
Why do I show log-likelihoods here, instead of the likelihood itself?

Does it make a difference for assessing the most likely parameter values?



Using the likelihood or its log makes no difference statistically

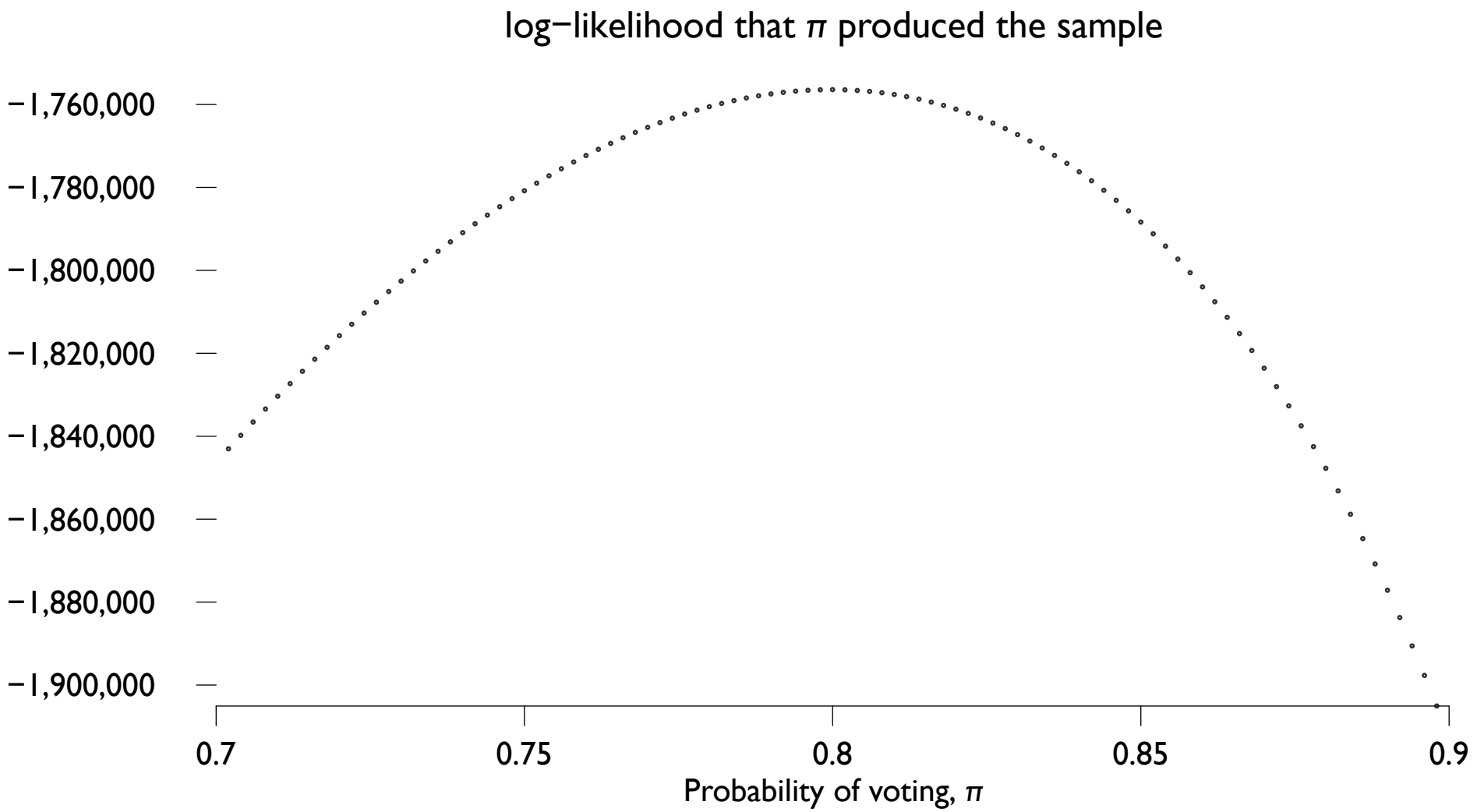
The likelihood and log-likelihood have the same maximum
and the same ordering of likely parameter values,
so we can use whichever is more convenient for our computers



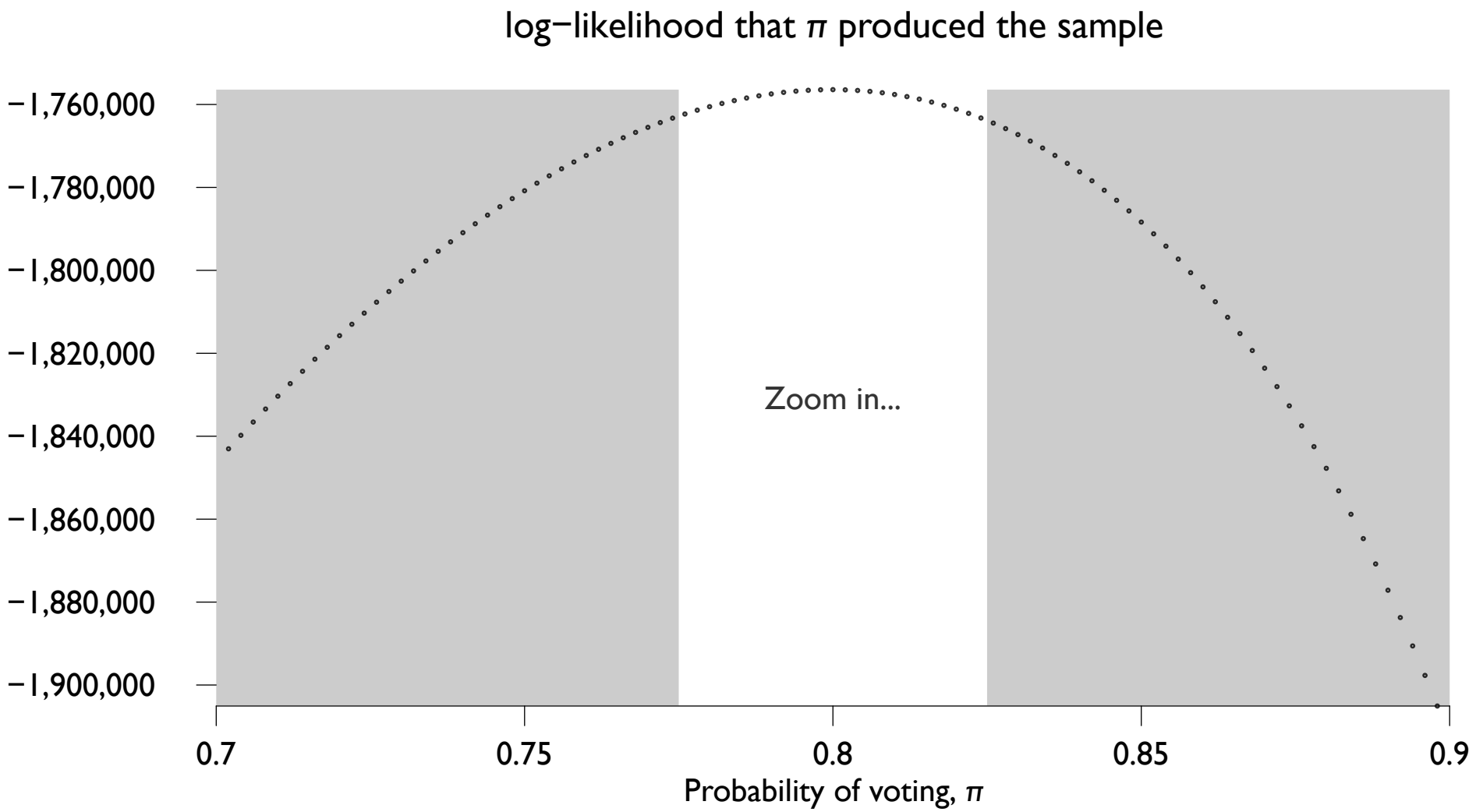
Is there a clear maximum of this likelihood?

What parameter value does the maximum indicate as most likely?

Let's zoom in and see. . .

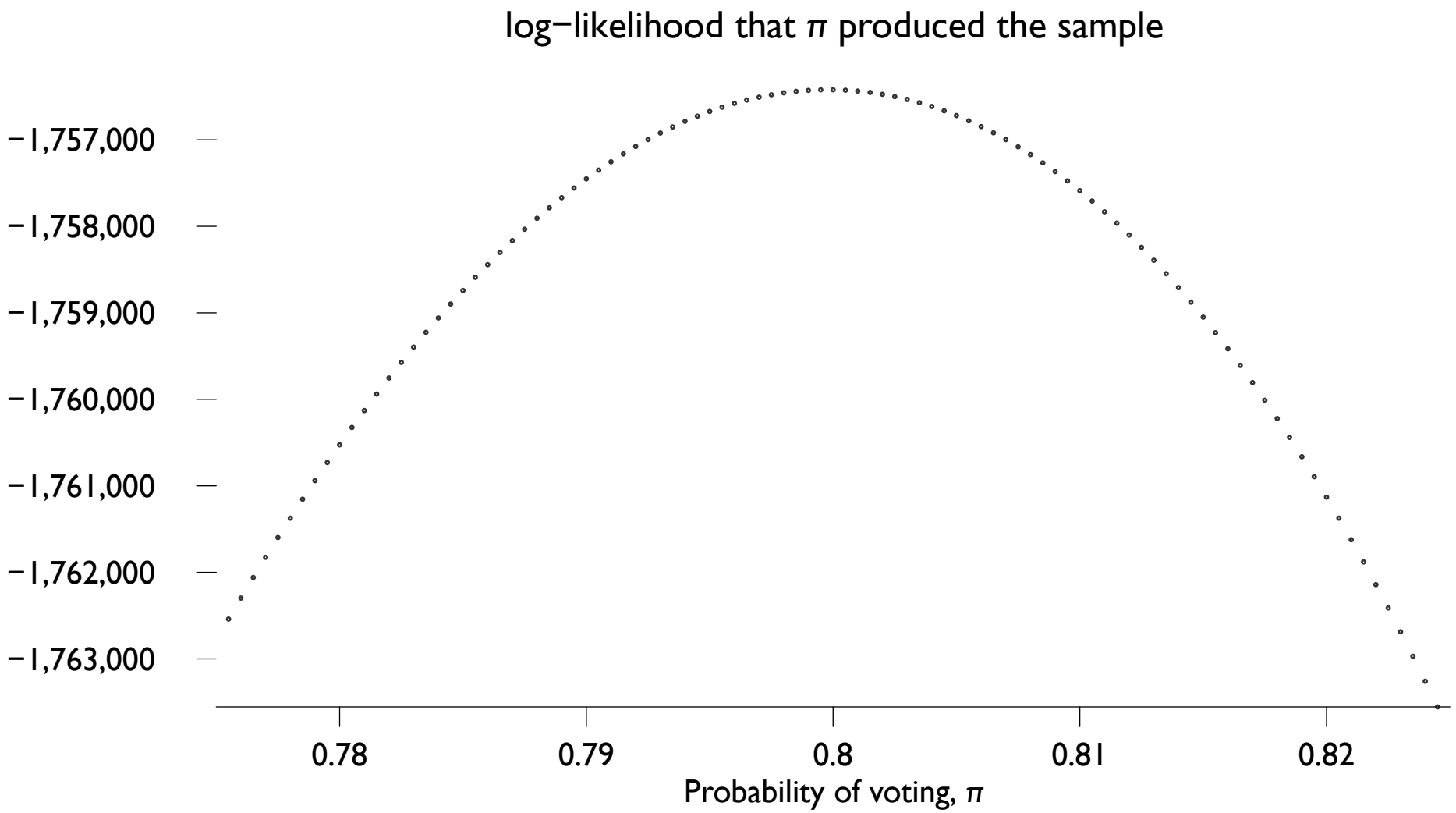


The maximum likelihood occurs near $\pi = 0.8$



The maximum likelihood occurs near $\pi = 0.8$

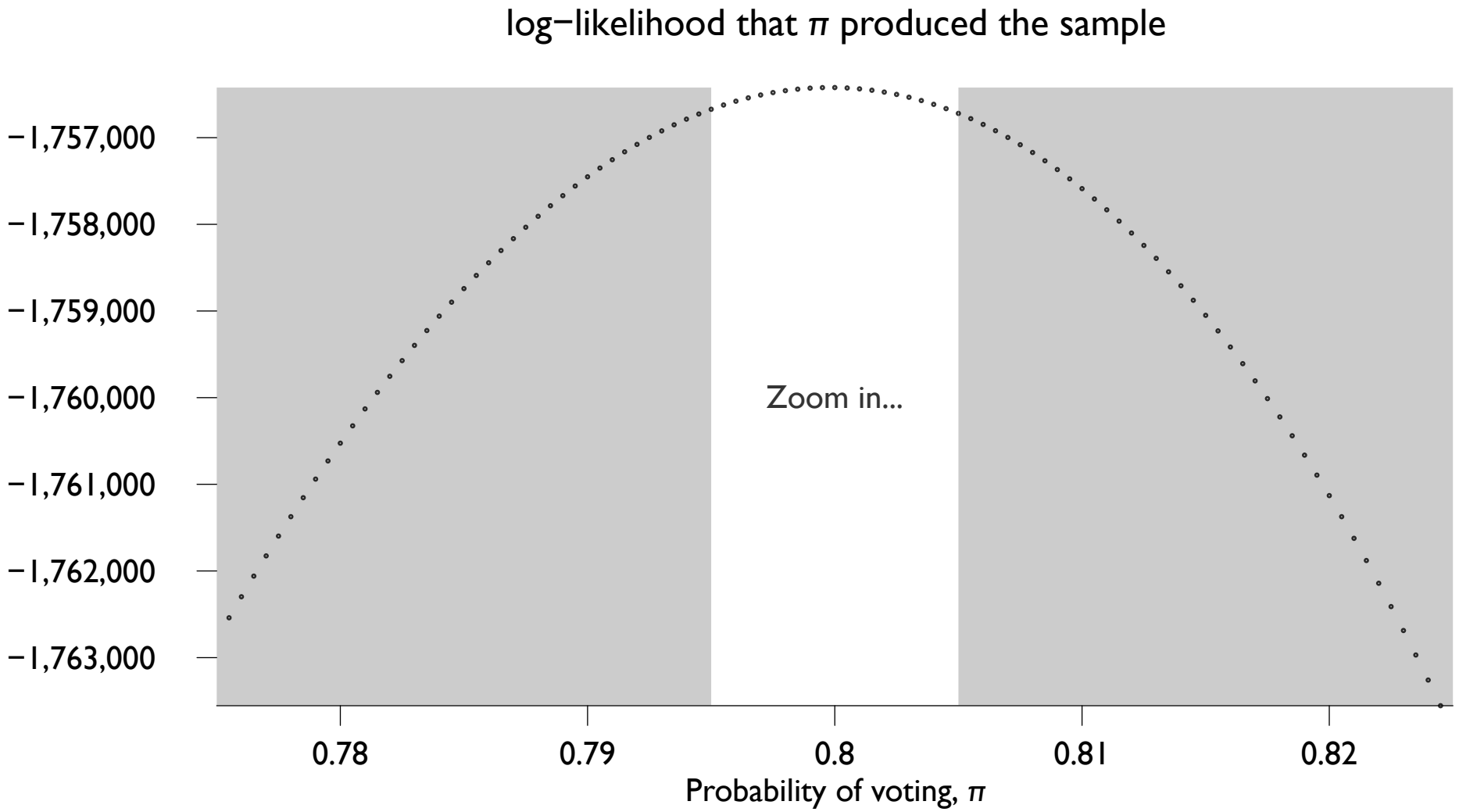
Let's sharpen the estimate by zooming in closer



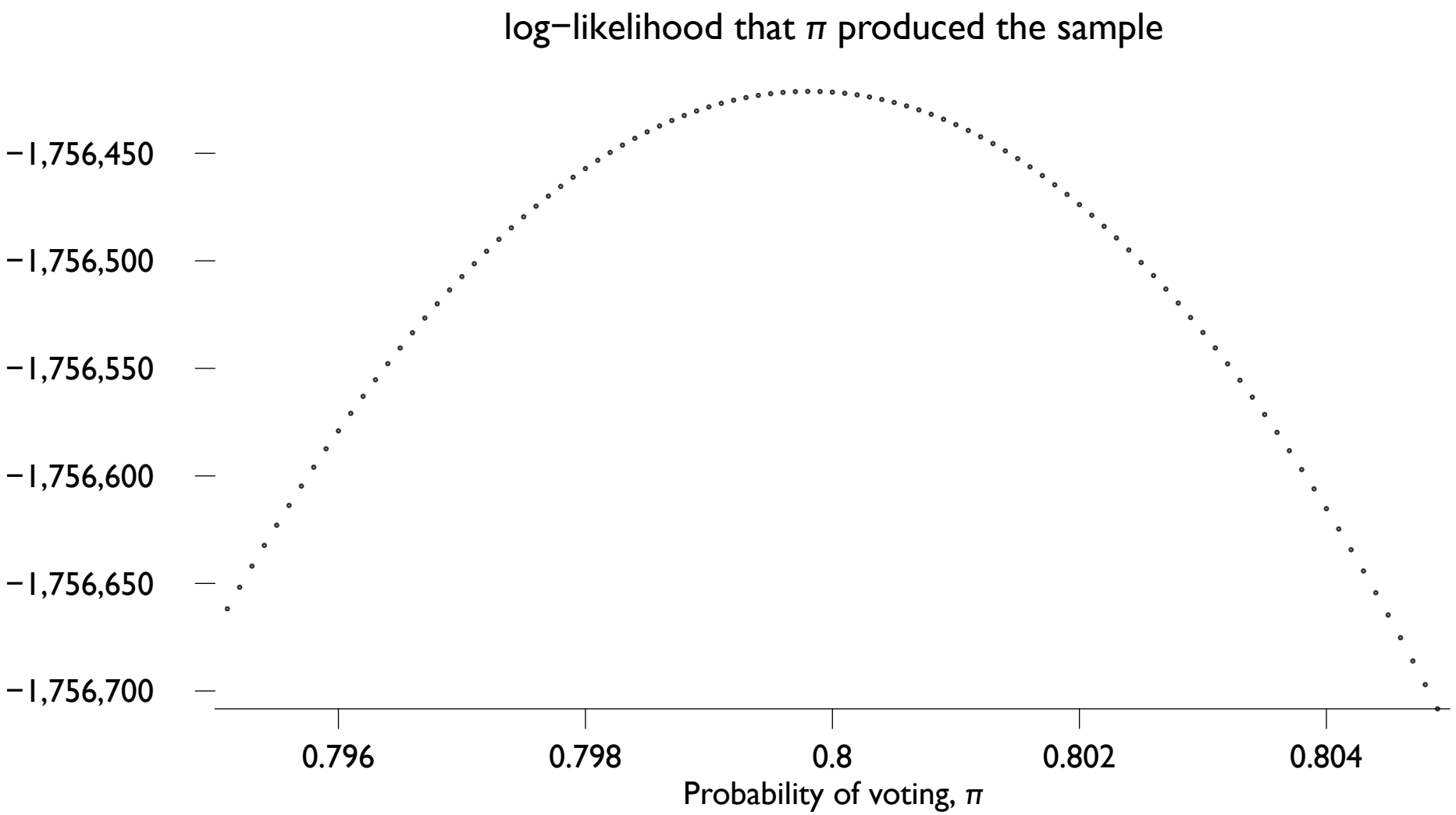
What happens when I zoom in?

I calculate the likelihood again for a finer set of π's near the maximum likelihood

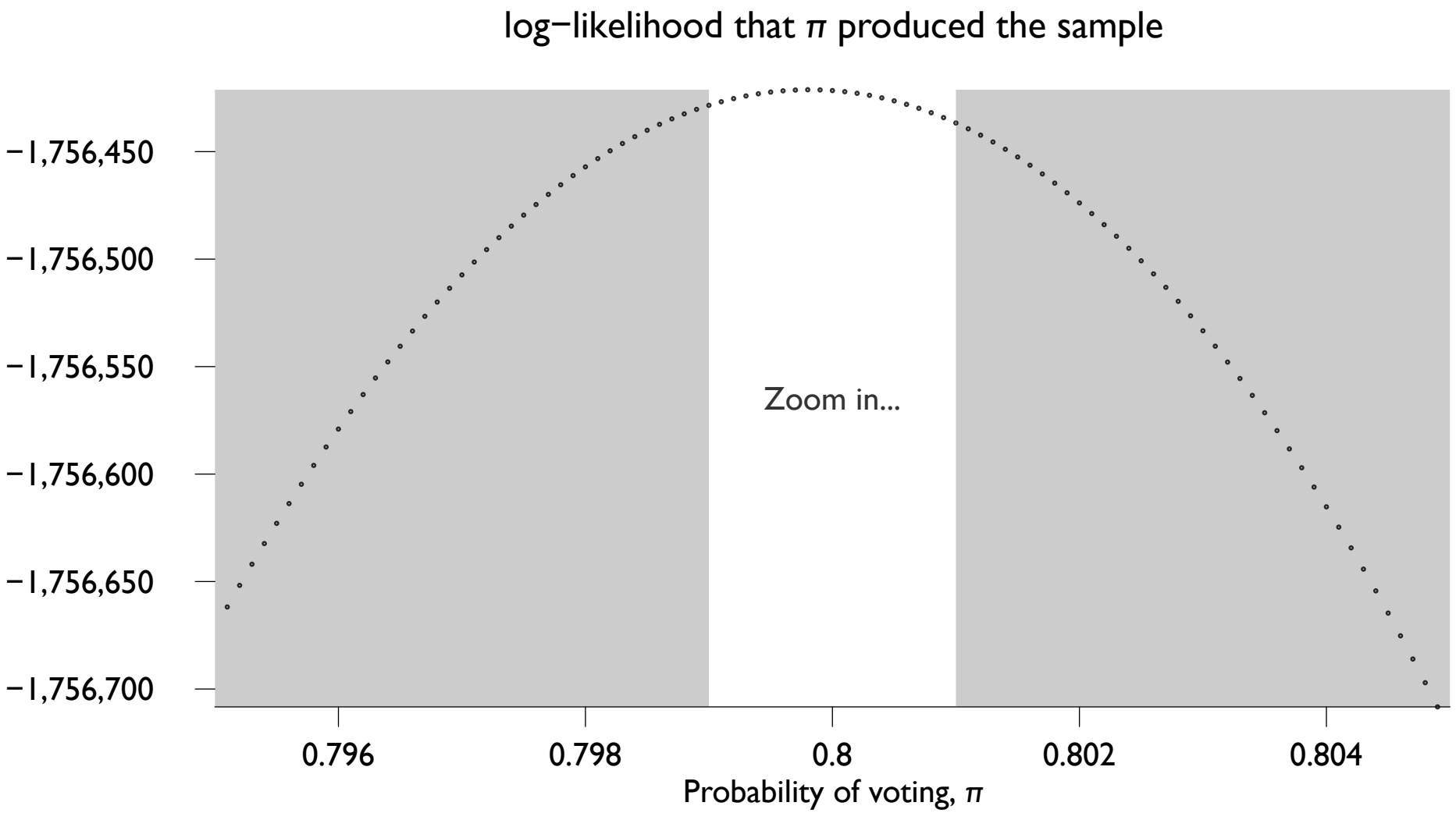
This is known as an iterative search using a grid method



We can iterate the search yet again,
computing the likelihood for a still finer grid of candidate parameter values

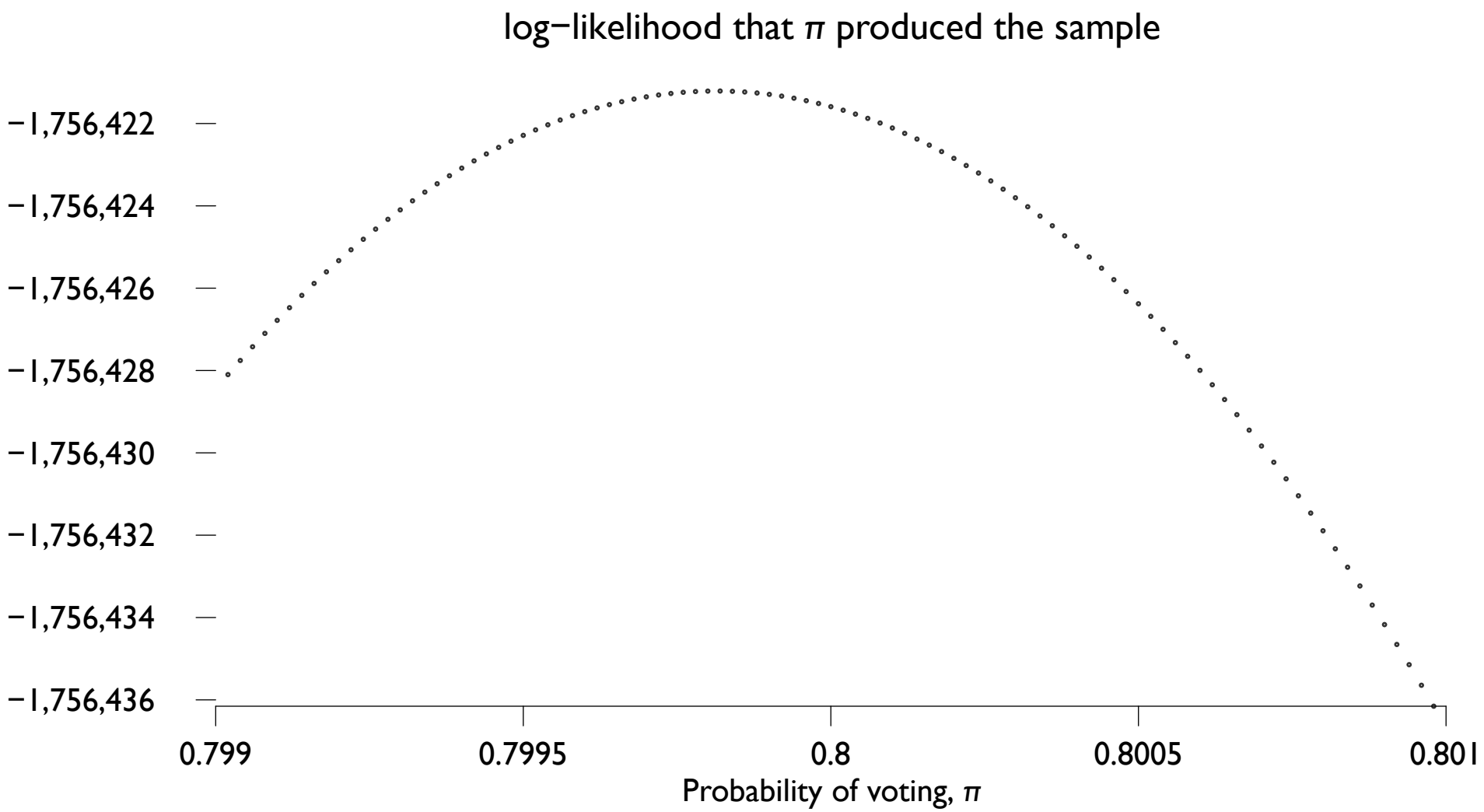


Repeating the grid search helps us find the precise value of π than maximizes $\mathcal{L}$



Repeating the grid search helps us find the precise value of π than maximizes $\mathcal{L}$

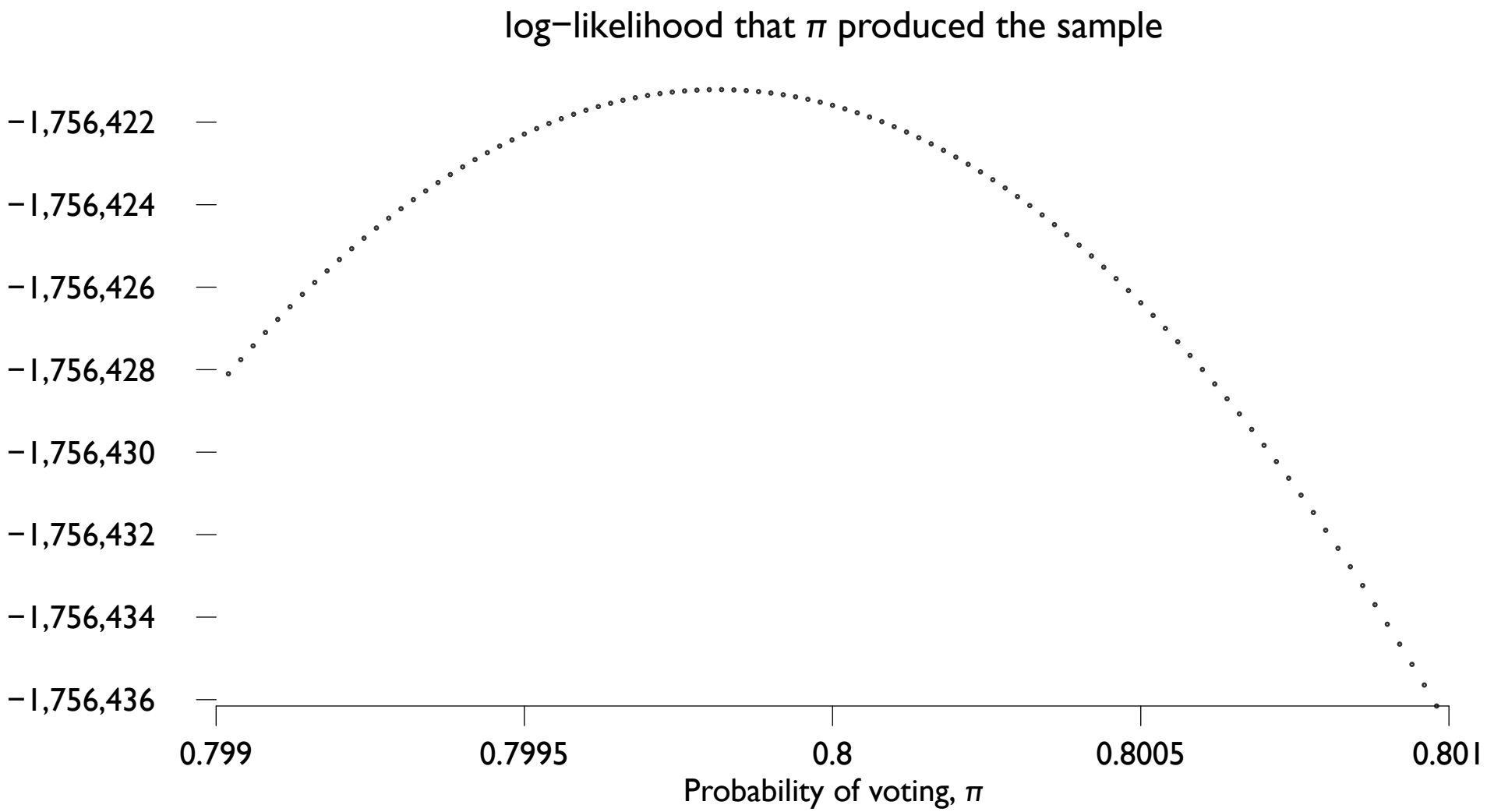
We'll iterate once more, but notice that the differences in $\mathcal{L}$ are getting very small



Our final iteration suggests that a π around 0.7998 maximizes the likelihood

But we shouldn't trust all those digits: the likelihood is very flat for $\pi \approx 0.80$

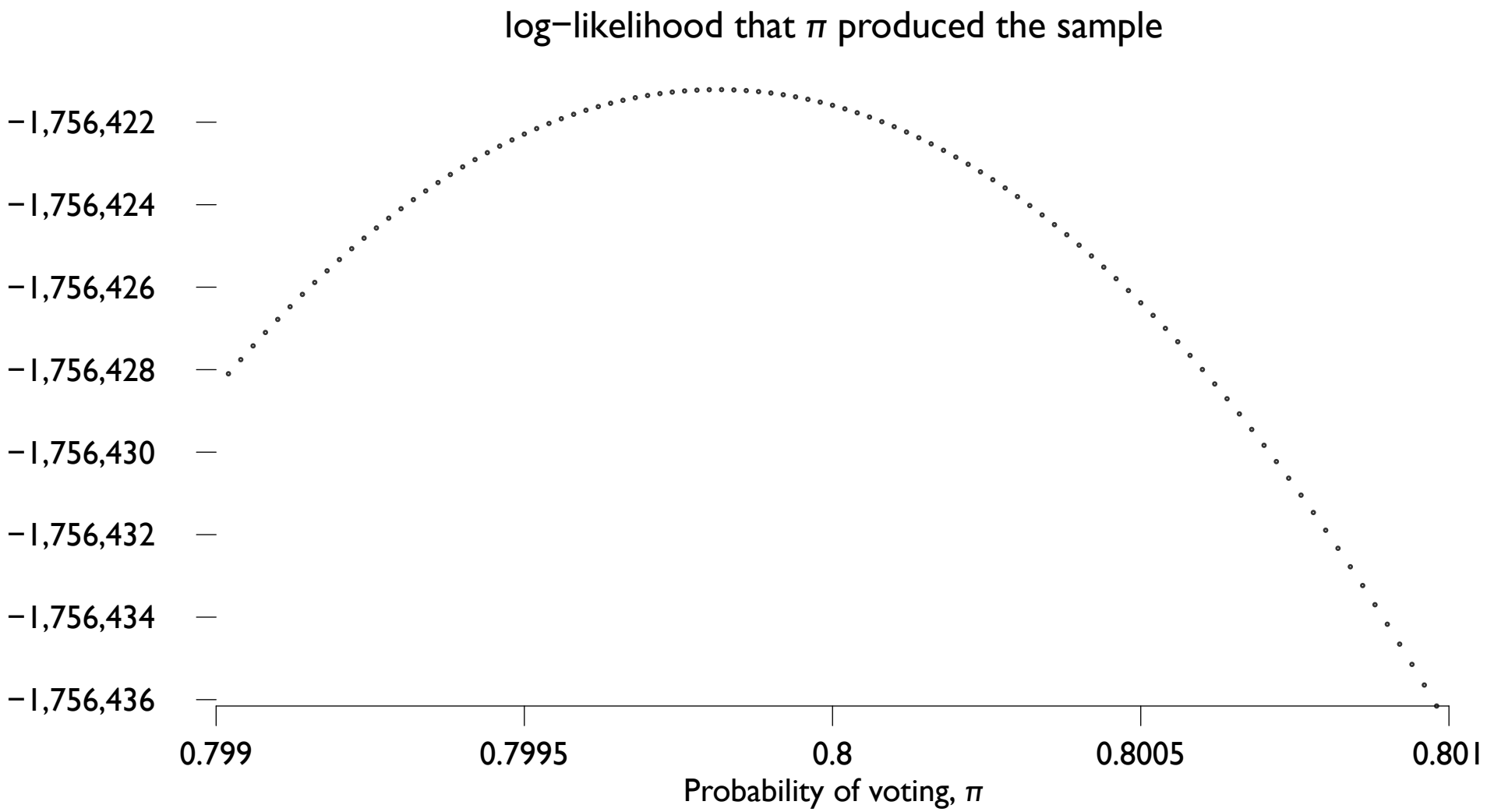
We can, however, be confident π is not *too far* from 0.80



In fact, the mean turnout rate for the state was 0.7998

So have we just found a fancy way to calculate the mean?

We've learned something else: Relative likelihood of different values of π



We have also tested the idea of using likelihood to estimate an unknown

If it works for estimating means,
it may also work to estimate unknown regression coefficients

Maximum likelihood

If we are reasonably sure that our likelihood is unimodal

Or we find the global maximum to be much higher than other modes

Maximum likelihood

If we are reasonably sure that our likelihood is unimodal

Or we find the global maximum to be much higher than other modes

And we find the surface to be narrowly peaked around the max, then

Maximum likelihood

If we are reasonably sure that our likelihood is unimodal

Or we find the global maximum to be much higher than other modes

And we find the surface to be narrowly peaked around the max, then

An attractive summary of $\mathcal{L}$ is its maximum,
and in particular,
the values of θ at the maximum $\mathcal{L}$

Maximum likelihood

If we are reasonably sure that our likelihood is unimodal

Or we find the global maximum to be much higher than other modes

And we find the surface to be narrowly peaked around the max, then

An attractive summary of $\mathcal{L}$ is its maximum,
and in particular,
the values of θ at the maximum $\mathcal{L}$

Maximum likelihood estimation entails finding those $\hat{\theta}_{\mathrm{ML}}$'s

Maximum likelihood

If we are reasonably sure that our likelihood is unimodal

Or we find the global maximum to be much higher than other modes

And we find the surface to be narrowly peaked around the max, then

An attractive summary of $\mathcal{L}$ is its maximum,
and in particular,
the values of $\boldsymbol{\theta}$ at the maximum $\mathcal{L}$

Maximum likelihood estimation entails finding those $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$'s

In practice, it will prove easier (& equivalent) to find the max of $\log \mathcal{L}$

How to derive maximum likelihood estimators

. . . in four easy steps:

How to derive maximum likelihood estimators

. . . in four easy steps:

1. Express the joint probability of the data,
   using the chosen probability distribution

How to derive maximum likelihood estimators

. . . in four easy steps:

1. Express the joint probability of the data,
   using the chosen probability distribution

2. Convert the joint probability to the likelihood
   (trivial, as they are proportional)

How to derive maximum likelihood estimators

. . . in four easy steps:

1. Express the joint probability of the data,
   using the chosen probability distribution

2. Convert the joint probability to the likelihood
   (trivial, as they are proportional)

3. Simplify the likelihood for easy maximization
   (take logs and reduce to "sufficient statistics")

How to derive maximum likelihood estimators

. . . in four easy steps:

1. Express the joint probability of the data,
   using the chosen probability distribution

2. Convert the joint probability to the likelihood
   (trivial, as they are proportional)

3. Simplify the likelihood for easy maximization
   (take logs and reduce to "sufficient statistics")

4. Substitute in the systematic component

How to derive maximum likelihood estimators

. . . in four easy steps:

1. Express the joint probability of the data,
   using the chosen probability distribution

2. Convert the joint probability to the likelihood
   (trivial, as they are proportional)

3. Simplify the likelihood for easy maximization
   (take logs and reduce to "sufficient statistics")

4. Substitute in the systematic component

Now we have something easy to maximize,
and will be able to estimate the parameters given the data

MLE for a Normally distributed response

Step 1: Express the joint probability of the data using the Normal distribution

MLE for a Normally distributed response

Step 1: Express the joint probability of the data using the Normal distribution

$$\mathrm{P}\left(y_1|\mu_1,\sigma^2\right) \quad = \quad f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right)$$

MLE for a Normally distributed response

Step 1: Express the joint probability of the data using the Normal distribution

$$\begin{aligned} \mathrm{P}\left(y_1|\mu_1,\sigma^2\right) &= f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right) \\ \mathrm{P}\left(y_1,y_2|\mu_1,\mu_2,\sigma^2\right) &= f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right) \times f_{\mathcal{N}}\left(y_2|\mu_2,\sigma^2\right) \end{aligned}$$

MLE for a Normally distributed response

Step 1: Express the joint probability of the data using the Normal distribution

$$
\begin{aligned}
\mathrm{P}\left(y_1|\mu_1,\sigma^2\right) &= f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right) \\
\mathrm{P}\left(y_1,y_2|\mu_1,\mu_2,\sigma^2\right) &= f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right) \times f_{\mathcal{N}}\left(y_2|\mu_2,\sigma^2\right) \\
\mathrm{P}\left(y_1,y_2,\ldots,y_i,\ldots y_n|\mu_i,\sigma^2\right) &= \prod_{i=1}^{n} f_{\mathcal{N}}\left(y_i|\mu_i,\sigma^2\right)
\end{aligned}
$$

MLE for a Normally distributed response

Step 1: Express the joint probability of the data using the Normal distribution

$$
\begin{aligned}
\mathrm{P}\left(y_1|\mu_1,\sigma^2\right) &= f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right) \\
\mathrm{P}\left(y_1,y_2|\mu_1,\mu_2,\sigma^2\right) &= f_{\mathcal{N}}\left(y_1|\mu_1,\sigma^2\right) \times f_{\mathcal{N}}\left(y_2|\mu_2,\sigma^2\right) \\
\mathrm{P}\left(y_1,y_2,\ldots,y_i,\ldots y_n|\mu_i,\sigma^2\right) &= \prod_{i=1}^{n} f_{\mathcal{N}}\left(y_i|\mu_i,\sigma^2\right) \\
\mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\sigma^2\right) &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i-\mu_i)^2}{2\sigma^2}\right)
\end{aligned}
$$

Note that we assume $y_i,\ldots,y_n$ *are* **iid***:*
Our biggest assumption to date

MLE for a Normally distributed response

Step 2: Convert the joint probability to the likelihood

MLE for a Normally distributed response

Step 2: Convert the joint probability to the likelihood

$$\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) \quad \propto \quad \mathrm{P}\left(\mathbf{y} | \boldsymbol{\mu}, \sigma^2\right)$$

MLE for a Normally distributed response

Step 2: Convert the joint probability to the likelihood

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &\propto \mathrm{P}\left(\mathbf{y} | \boldsymbol{\mu}, \sigma^2\right) \\ \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \mathrm{P}\left(\mathbf{y} | \boldsymbol{\mu}, \sigma^2\right) \end{aligned}$$

MLE for a Normally distributed response

Step 2: Convert the joint probability to the likelihood

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &\propto \mathrm{P}\left(\mathbf{y} | \boldsymbol{\mu}, \sigma^2\right) \\
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \mathrm{P}\left(\mathbf{y} | \boldsymbol{\mu}, \sigma^2\right) \\
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)
\end{aligned}
$$

Step 3: Simplify the likelihood for easy maximization

Step 3: Simplify the likelihood for easy maximization

$$\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) = k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)$$

Step 3: Simplify the likelihood for easy maximization

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \log \prod_{i=1}^{n} \left(k(y_i) \times \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)\right)
\end{aligned}$$

Step 3: Simplify the likelihood for easy maximization

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \log \prod_{i=1}^{n} \left(k(y_i) \times \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \left(\log k(y_i) - \frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{(y_i - \mu_i)^2}{2\sigma^2}\right)
\end{aligned}$$

Step 3: Simplify the likelihood for easy maximization

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \log \prod_{i=1}^{n} \left(k(y_i) \times \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \left(\log k(y_i) - \frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \log k(y_i) - \frac{1}{2} \sum_{i=1}^{n} \log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2}
\end{aligned}
$$

Step 3: Simplify the likelihood for easy maximization

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \log \prod_{i=1}^{n} \left(k(y_i) \times \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \left(\log k(y_i) - \frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \log k(y_i) - \frac{1}{2} \sum_{i=1}^{n} \log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2} \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \log k(y_i) - \frac{1}{2} \sum_{i=1}^{n} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2}
\end{aligned}
$$

Step 3: Simplify the likelihood for easy maximization

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= k(\mathbf{y}) \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \log \prod_{i=1}^{n} \left(k(y_i) \times \left(2\pi\sigma^2\right)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \left(\log k(y_i) - \frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{(y_i - \mu_i)^2}{2\sigma^2}\right) \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \log k(y_i) - \frac{1}{2} \sum_{i=1}^{n} \log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2} \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &= \sum_{i=1}^{n} \log k(y_i) - \frac{1}{2} \sum_{i=1}^{n} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2} \\
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &\propto -\frac{1}{2} \sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2}
\end{aligned}
$$

Note the last step reduces to sufficient statistics for $\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right)$

MLE for a Normally distributed response

Step 4: Substitute in the systematic component

MLE for a Normally distributed response

Step 4: Substitute in the systematic component

$$\begin{aligned} \log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &\propto -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2} \\ \mu_i &= \mathbf{x}_i \boldsymbol{\beta} \end{aligned}$$

MLE for a Normally distributed response

Step 4: Substitute in the systematic component

$$
\begin{aligned}
\log \mathcal{L}\left(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}\right) &\propto -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2} \\
\mu_i &= \mathbf{x}_i \boldsymbol{\beta} \\
\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) &\propto -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\sigma^2}
\end{aligned}
$$

MLE for a Normally distributed response

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) = -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\sigma^2}$$

Note some interesting feature of this MLE:

- $\log \mathcal{L} \uparrow$ as the sum of squared errors $\downarrow$

MLE for a Normally distributed response

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) = -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\sigma^2}$$

Note some interesting feature of this MLE:

- $\log \mathcal{L} \uparrow$ as the sum of squared errors $\downarrow$
- MLE for normal data is the estimator that minimizes the squared errors

MLE for a Normally distributed response

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) = -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\sigma^2}$$

Note some interesting feature of this MLE:

- $\log \mathcal{L} \uparrow$ as the sum of squared errors $\downarrow$
- MLE for normal data is the estimator that minimizes the squared errors
- In the Normal case, least squares (LS) is the MLE

MLE for a Normally distributed response

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) = -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\sigma^2}$$

Note some interesting feature of this MLE:

- $\log \mathcal{L} \uparrow$ as the sum of squared errors $\downarrow$
- MLE for normal data is the estimator that minimizes the squared errors
- In the Normal case, least squares (LS) is the MLE
- Note that we now have a justification for LS over, say, minimizing absolute error
- In other words, we have derived LS from first principles

So what?

We already know least squares, so has all this theory gotten us anywhere?

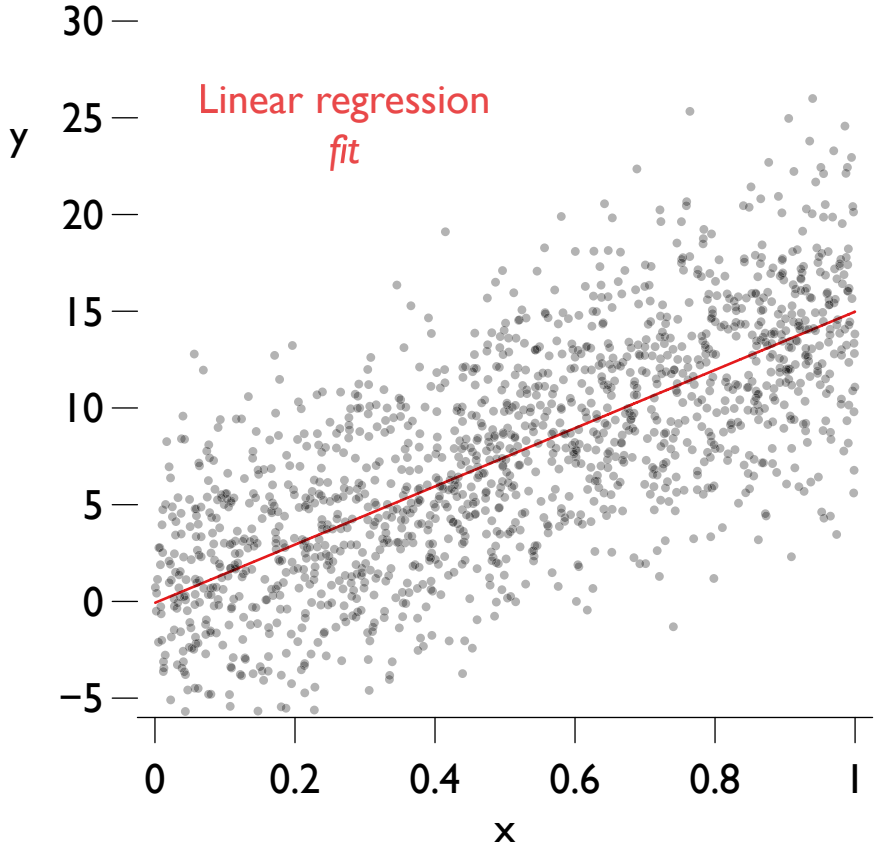
Yes

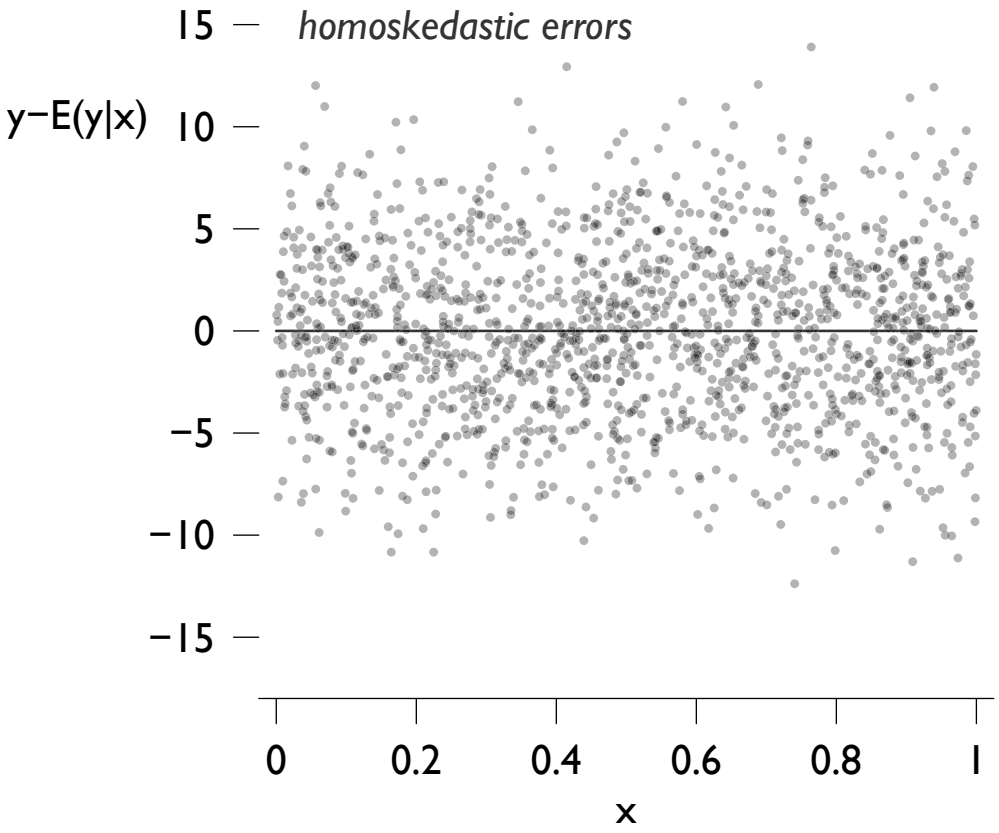
Use the same steps to derive an MLE for **any** probability distribution

Can produce & use models closer to how we, as scientists, think our data behaves

Only limit now is our creativity

So let's derive something interesting, but not too different from LS

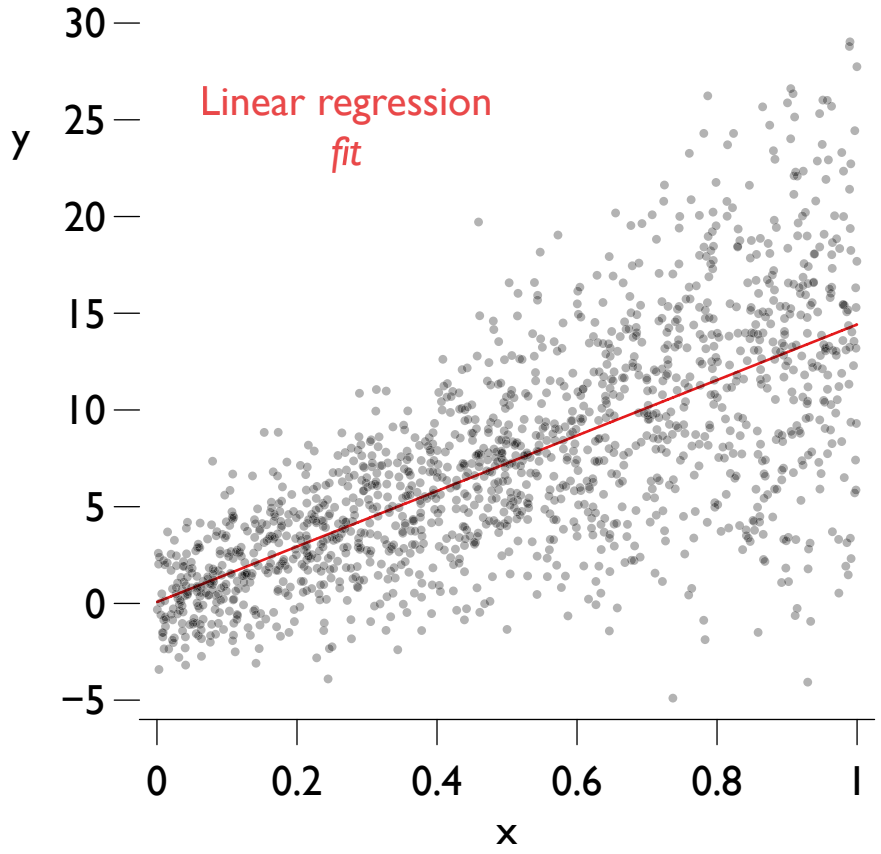


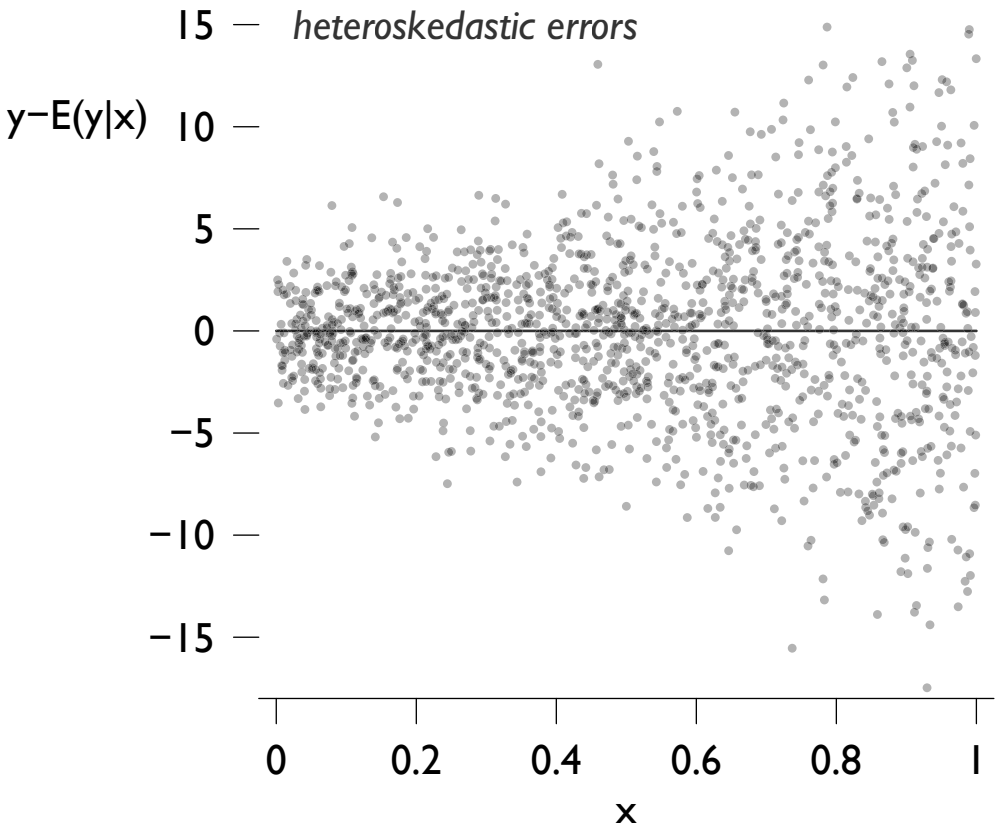


Linear regression models assume errors are homoskedastic

Homoskedastic = constant error variance

The model assumes same σ^2 for all cases: $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$



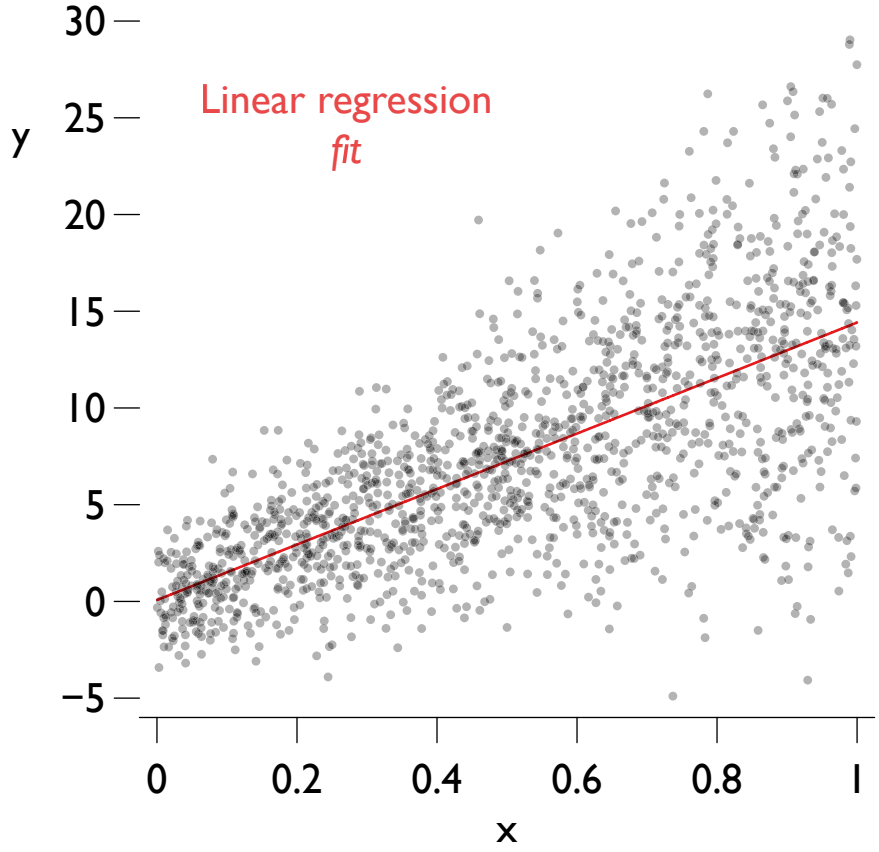


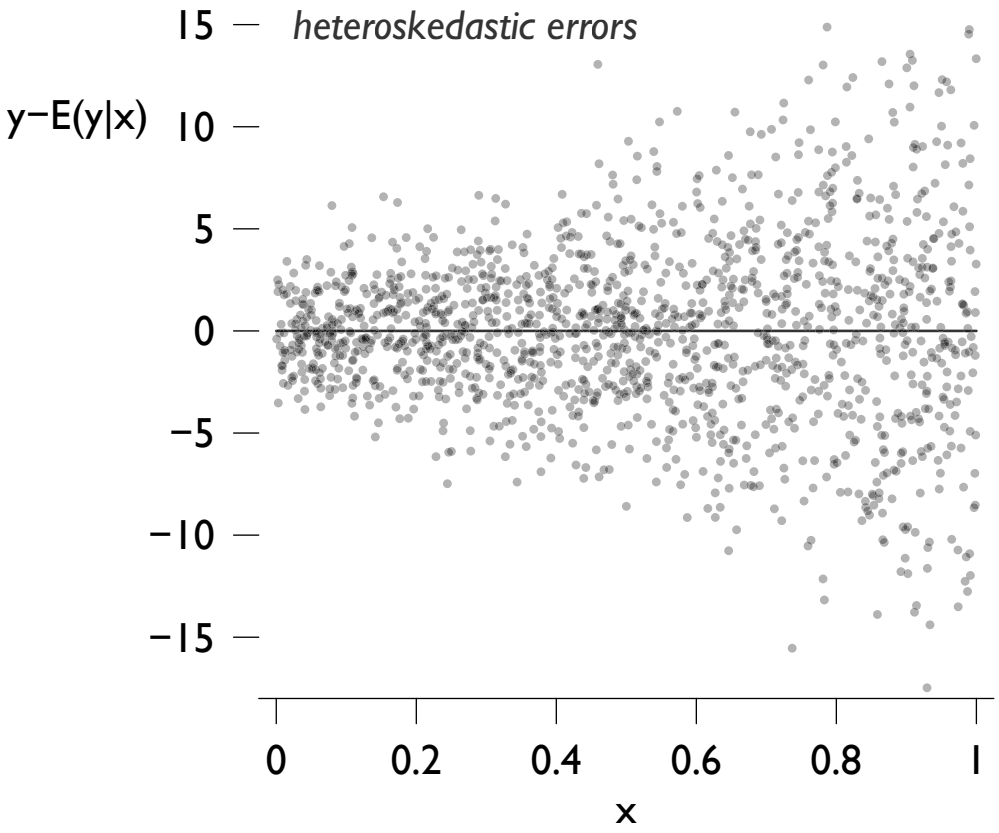
What if errors are heteroskedastic, $y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$? Two problems arise:

1. $\mathrm{se}(\hat{\beta})$ may be biased
2. Estimates of $\hat{\beta}$ will be inefficient

robust standard errors attempt to fix

robust standard errors do not *fix*

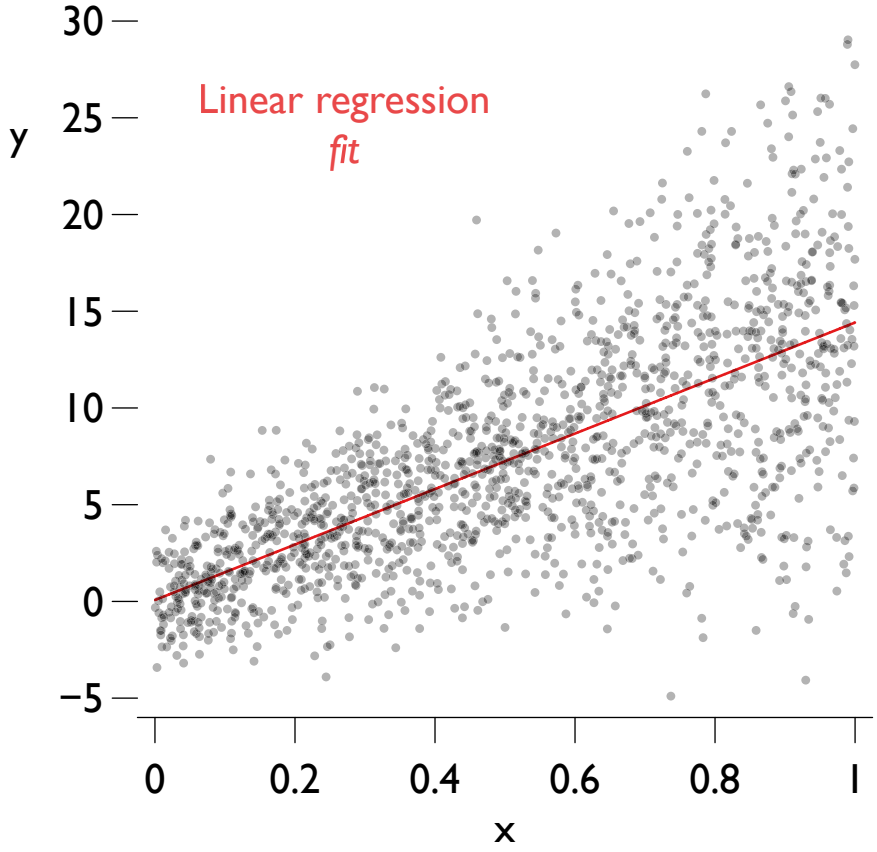


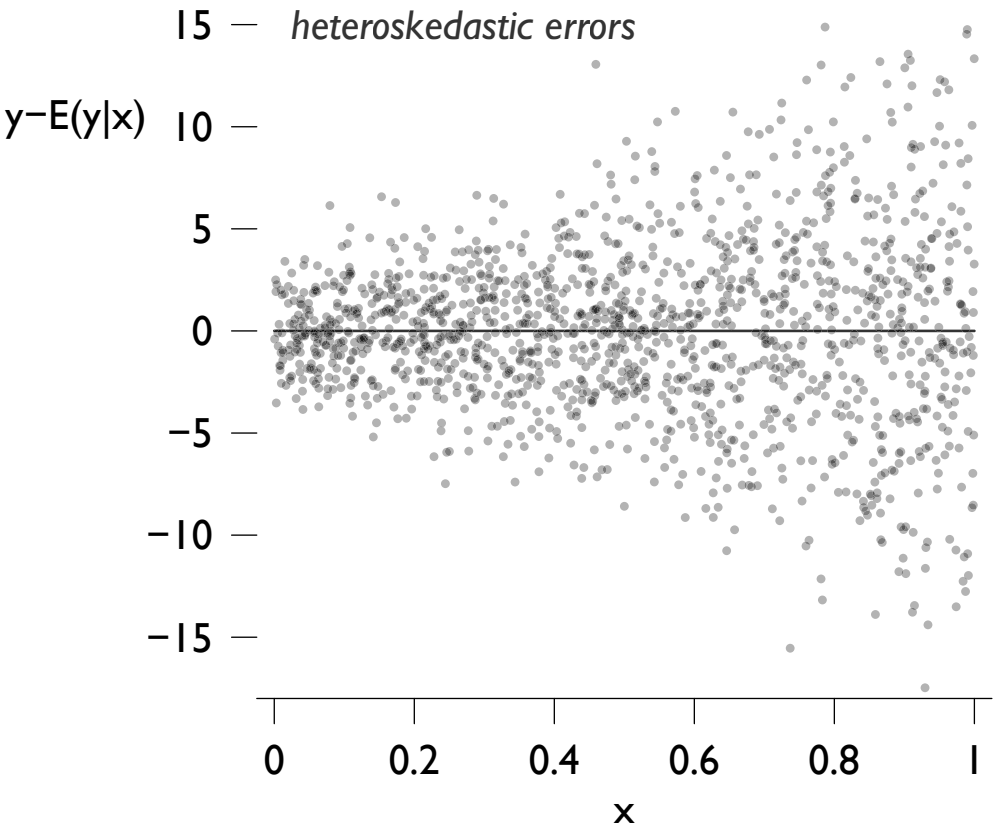


Why does heteroskedasticity make linear regression inefficient?

Observations with higher variance in errors contain less information

Observations with lower variance tend to be very close to the regression line





But heteroskedasticity is a "problem" only because we assumed it didn't exist

We don't ever talk about the problem of "non-constant means" because we have μ_i

What if included σ_i^2 as part of the model?

Bigger question

What if the heteroskedasticity *is* the interesting part of the data?

Suppose. . .

1. Roughly balanced powers are a necessary (but not sufficient) condition for war, making war/peace more variable?

2. Privatizing social services doesn't lower average welfare (much), but increases the variability of (say) health outcomes by increasing risk of non-coverage?

Linear regression won't answer these questions well

Can we model variance directly using maximum likelihood?

MLE for a heteroskedastic Normal response

Heteroskedasticity isn't a flaw but a real feature of the data

MLE for a heteroskedastic Normal response

Heteroskedasticity isn't a flaw but a real feature of the data

With maximum likelihood, we can now model it explicitly:

- I.e., derive a model that explicitly allows for heteroskedasticity

MLE for a heteroskedastic Normal response

Heteroskedasticity isn't a flaw but a real feature of the data

With maximum likelihood, we can now model it explicitly:

- I.e., derive a model that explicitly allows for heteroskedasticity
- and parameterize it (model heteroskedasticity as a function of covariates)

MLE for a heteroskedastic Normal response

Heteroskedasticity isn't a flaw but a real feature of the data

With maximum likelihood, we can now model it explicitly:

- I.e., derive a model that explicitly allows for heteroskedasticity
- and parameterize it (model heteroskedasticity as a function of covariates)
- example: we could show that x_1 not only $\uparrow$'s the mean, it also $\uparrow$'s the variance

MLE for a heteroskedastic Normal response

To derive the MLE for a heteroskedastic Normal model, we need to specify the

stochastic component

$$y_i \sim f_{\mathcal{N}}(\mu_i, {\sigma_i}^2)$$

MLE for a heteroskedastic Normal response

To derive the MLE for a heteroskedastic Normal model, we need to specify the

stochastic component

$$y_i \sim f_{\mathcal{N}}(\mu_i, {\sigma_i}^2)$$

systematic components

$$\mu_i \quad = \quad \mathbf{x}_i \boldsymbol{\beta}$$

MLE for a heteroskedastic Normal response

To derive the MLE for a heteroskedastic Normal model, we need to specify the

stochastic component

$$y_i \sim f_{\mathcal{N}}(\mu_i, {\sigma_i}^2)$$

systematic components

$$\begin{aligned} \mu_i &= \mathbf{x}_i\boldsymbol{\beta} \\ \sigma_i^2 &= \exp(\mathbf{z}_i\boldsymbol{\gamma}) \end{aligned}$$

Notice the difference from linear regression:
σ_i^2 has an extra systematic component

Why do we model σ_i^2 as exponential?

MLE for a heteroskedastic Normal response

The derivation of the heteroskedastic MLE largely reproduces the homoskedastic case

Just add subscripts to the σ^2's!

MLE for a heteroskedastic Normal response

The derivation of the heteroskedastic MLE largely reproduces the homoskedastic case

Just add subscripts to the σ^2's!

$$\mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) = \prod_{i=1}^{n} f_{\mathcal{N}}\left(y_i|\mu_i,\sigma_i^2\right)$$

MLE for a heteroskedastic Normal response

The derivation of the heteroskedastic MLE largely reproduces the homoskedastic case

Just add subscripts to the σ^2's!

$$\begin{aligned} \mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} f_{\mathcal{N}}\left(y_i|\mu_i,\sigma_i^2\right) \\ \mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left[\frac{-(y_i-\mu_i)^2}{2\sigma_i^2}\right] \\ &\dots \end{aligned}$$

MLE for a heteroskedastic Normal response

The derivation of the heteroskedastic MLE largely reproduces the homoskedastic case

Just add subscripts to the σ^2's!

$$
\begin{aligned}
\mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} f_{\mathcal{N}}\left(y_i|\mu_i,\sigma_i^2\right) \\
\mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left[\frac{-(y_i-\mu_i)^2}{2\sigma_i^2}\right] \\
&\dots \\
\log \mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\sigma}^2|\mathbf{y}\right) &\propto -\frac{1}{2}\sum_{i=1}^{n}\log\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma_i^2}
\end{aligned}
$$

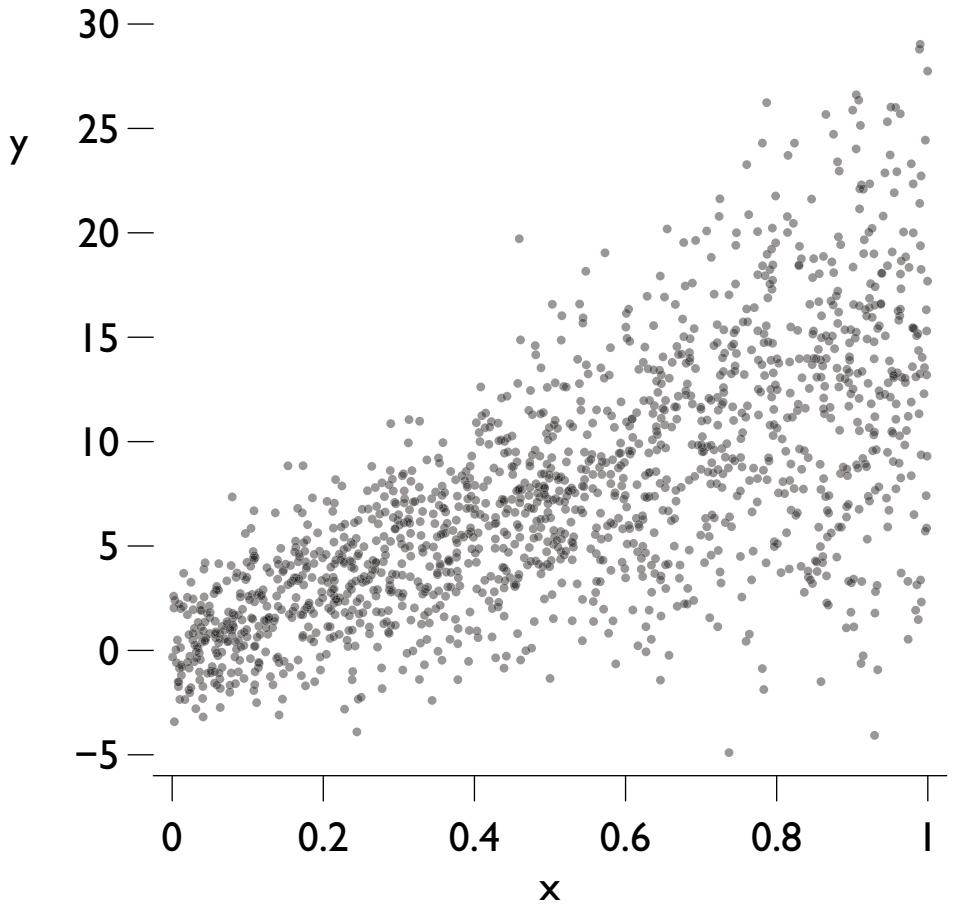
MLE for a heteroskedastic Normal response

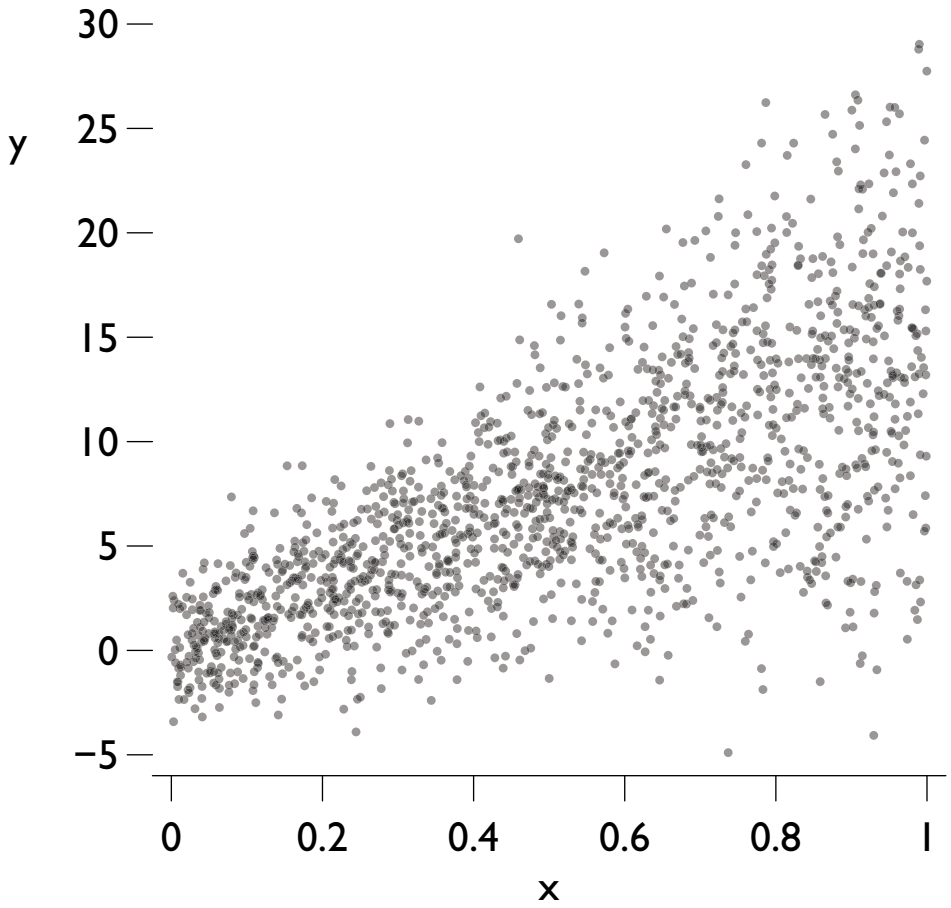
The derivation of the heteroskedastic MLE largely reproduces the homoskedastic case

Just add subscripts to the σ^2's!

$$
\begin{aligned}
\mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} f_{\mathcal{N}}\left(y_i|\mu_i,\sigma_i^2\right) \\
\mathrm{P}\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left[\frac{-(y_i-\mu_i)^2}{2\sigma_i^2}\right] \\
&\dots \\
\log \mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\sigma}^2|\mathbf{y}\right) &\propto -\frac{1}{2}\sum_{i=1}^{n}\log\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma_i^2} \\
\log \mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\gamma}|\mathbf{y}\right) &\propto -\frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i\boldsymbol{\gamma} - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mathbf{x}_i\boldsymbol{\beta})^2}{\exp(\mathbf{z}_i\boldsymbol{\gamma})}
\end{aligned}
$$

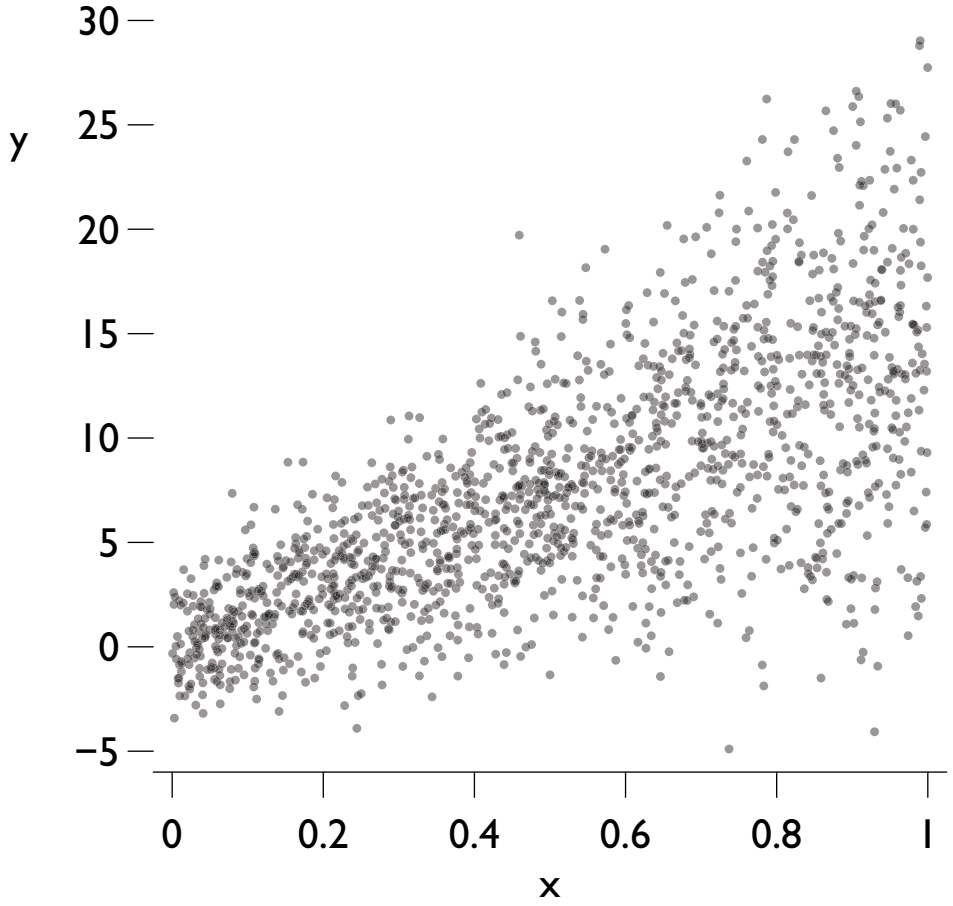
Now we just find the parameters ($\boldsymbol{\beta}$'s and $\boldsymbol{\gamma}$'s) that maximize this likelihood

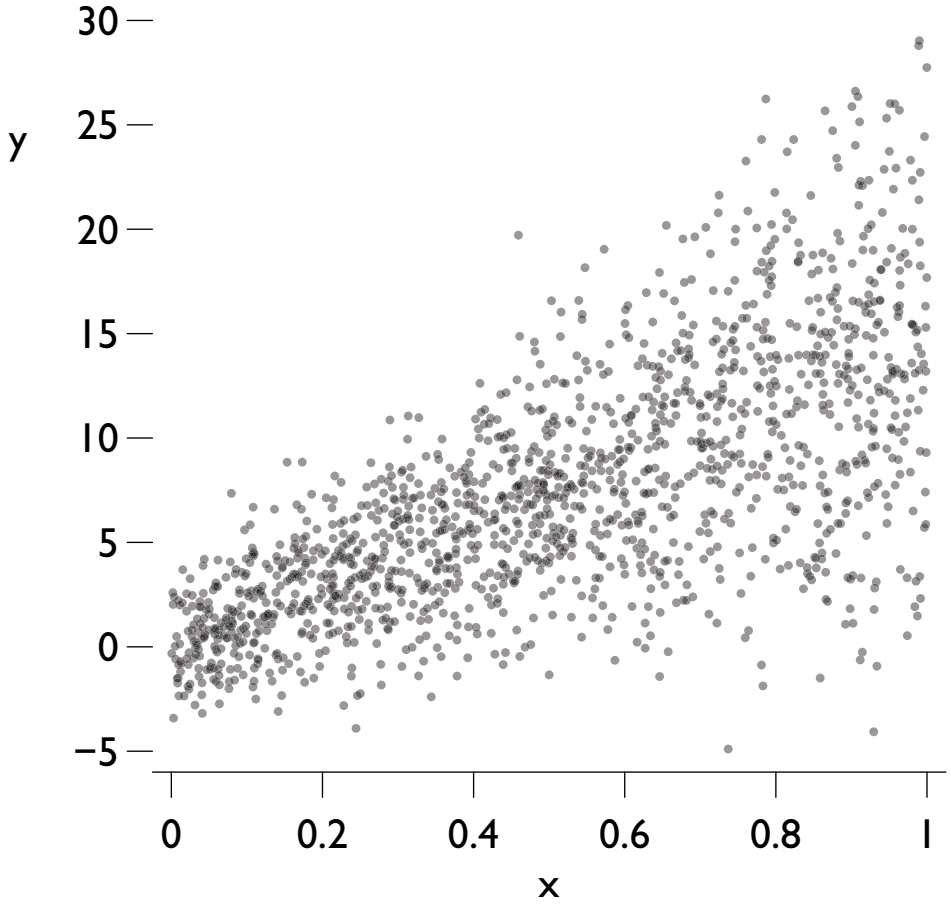




A good way to test a new model: use it on Monte Carlo data

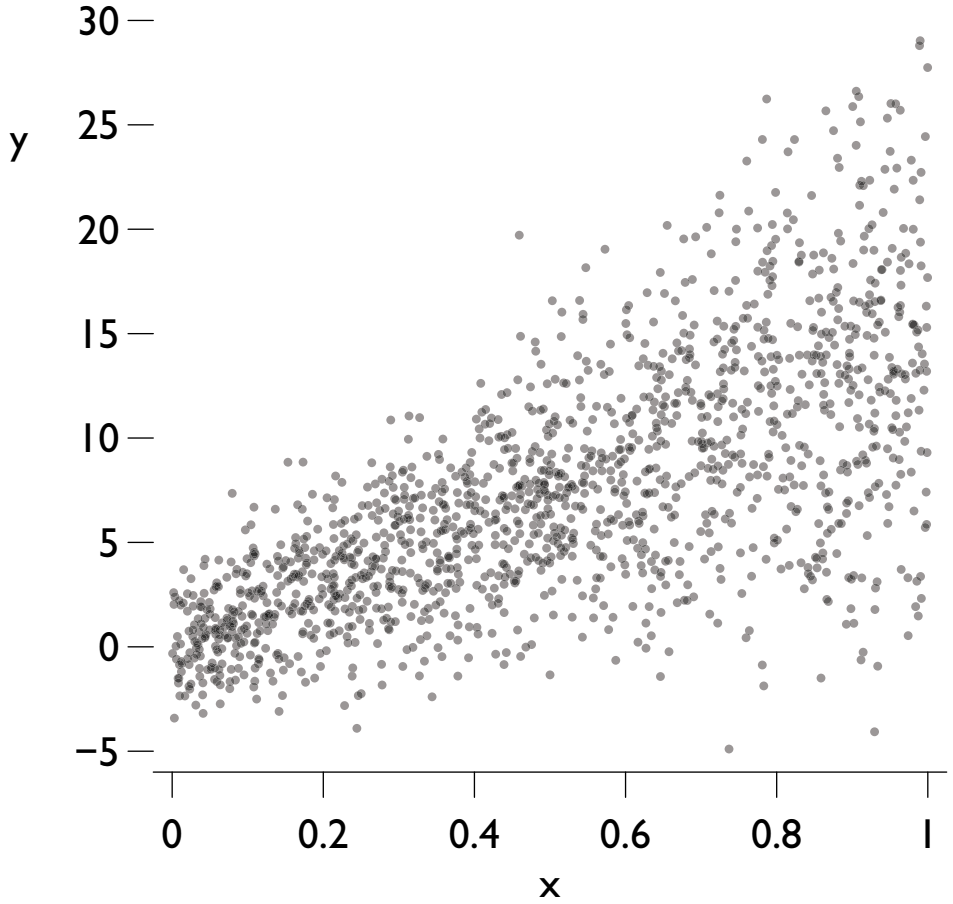
1. Simulate data with known parameters and an appropriate distribution
2. Attempt to recover the true parameters with the model

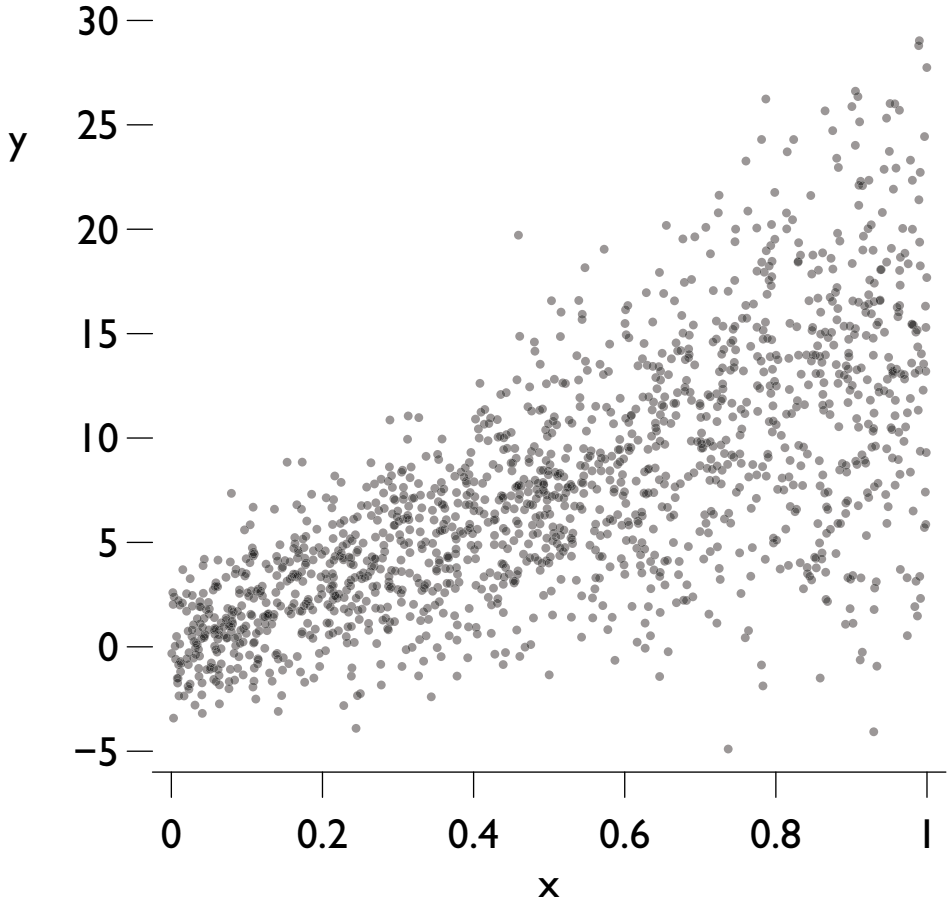




Above data ($N = 1500$) are drawn from this heteroskedastic distribution

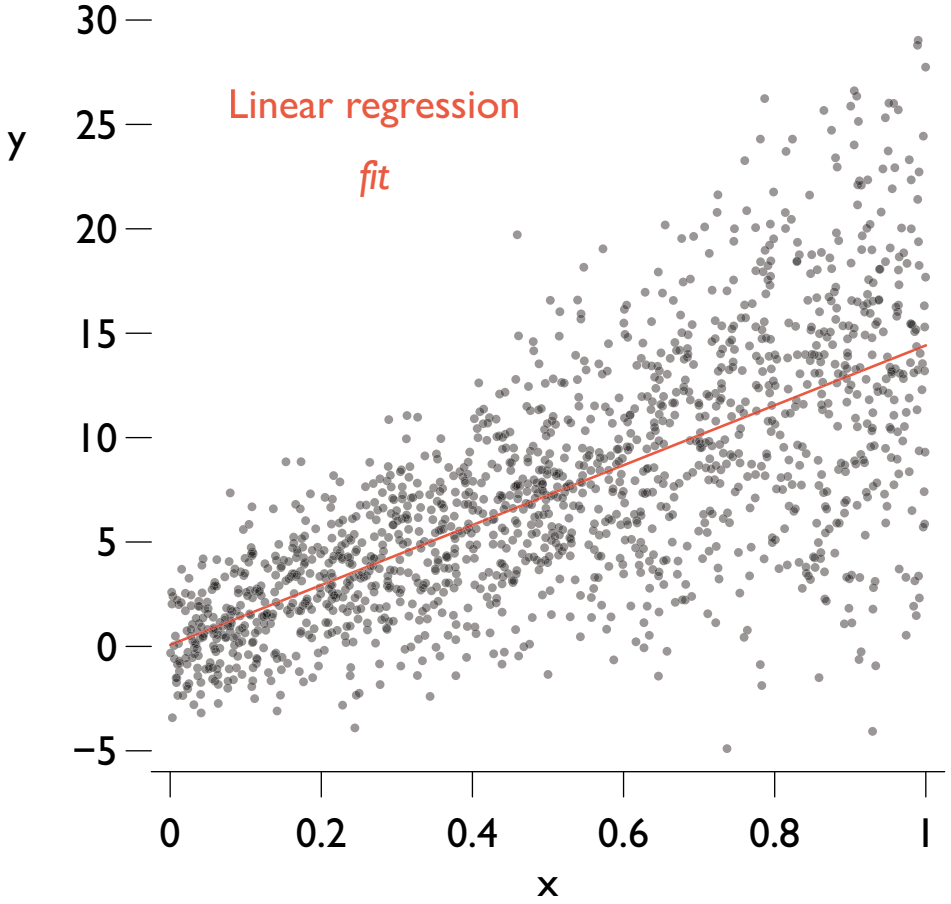
$$
\begin{aligned}
y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\
\mu_i &= \beta_0 + \beta_1 x_i \\
\sigma_i^2 &= \exp(\gamma_0 + \gamma_1 x_i)
\end{aligned}
$$

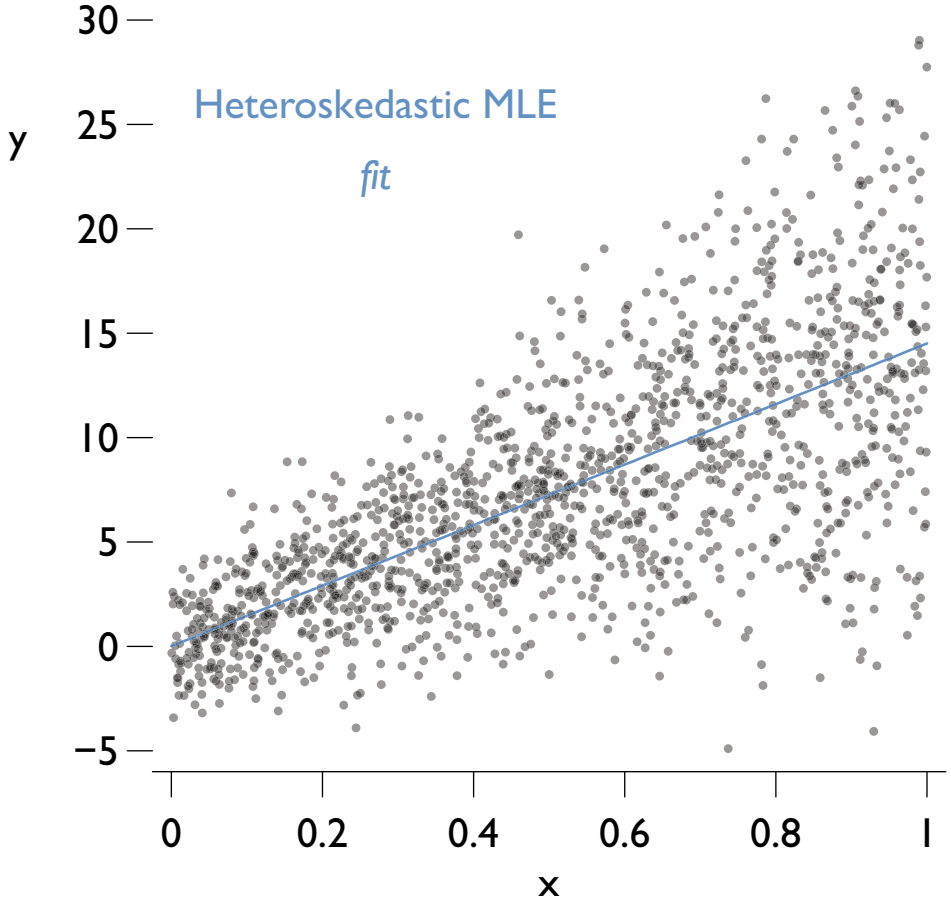




Above data ($N = 1500$) are drawn from this heteroskedastic distribution

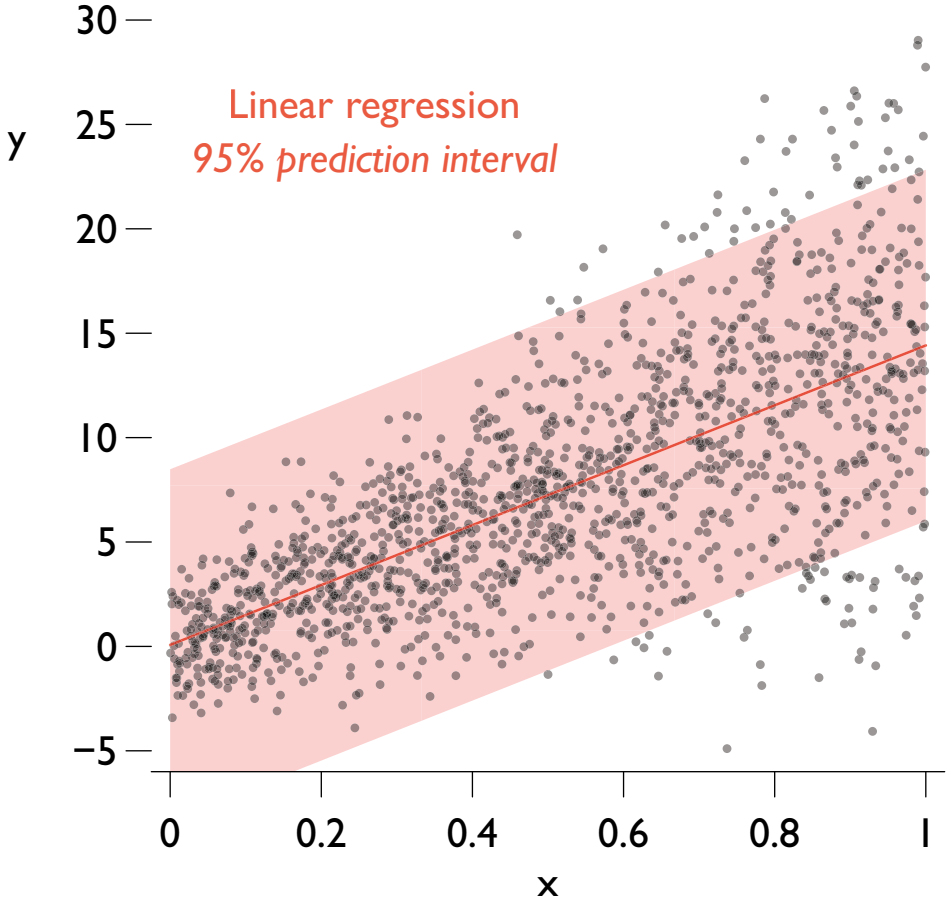
$$
\begin{aligned}
y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\
\mu_i &= 0 + 15x_i \\
\sigma_i^2 &= \exp(1 + 3x_i)
\end{aligned}
$$

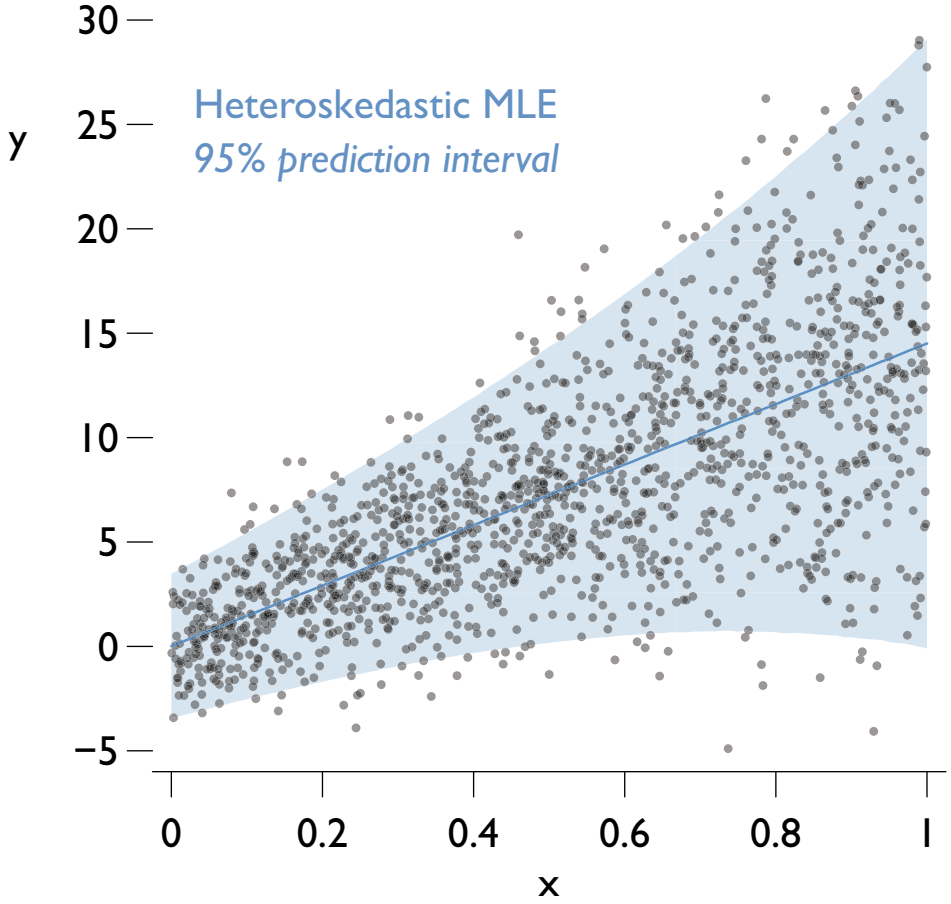




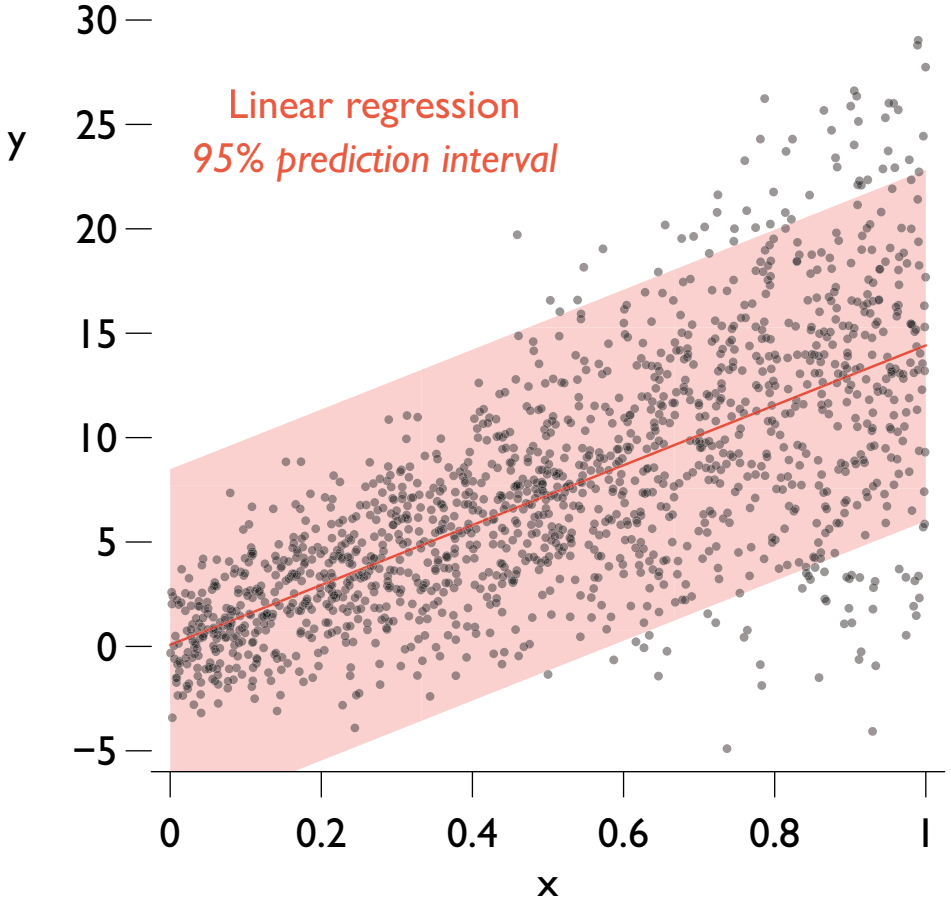
Expected values from the heteroskedastic MLE closely match those from linear regression

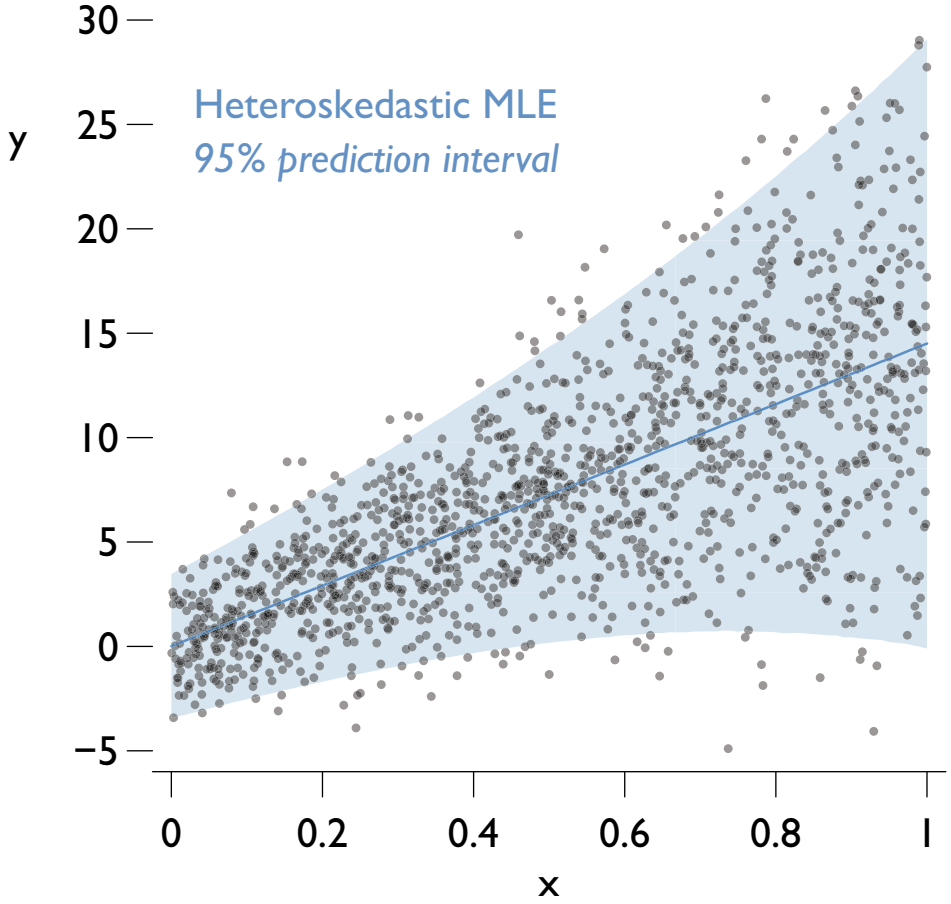
Not a surprise: these models model the mean of $y_i|x_i$ in the same way





Key test is to compare *prediction intervals* (*not* confidence intervals – why?)

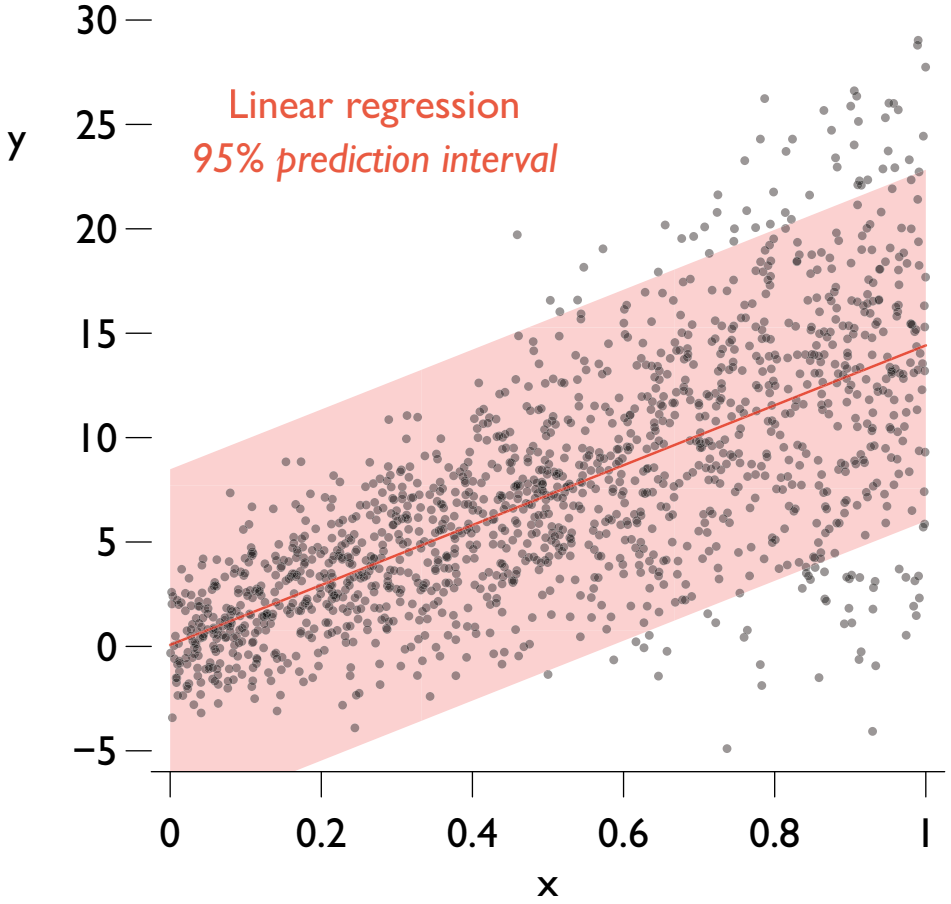


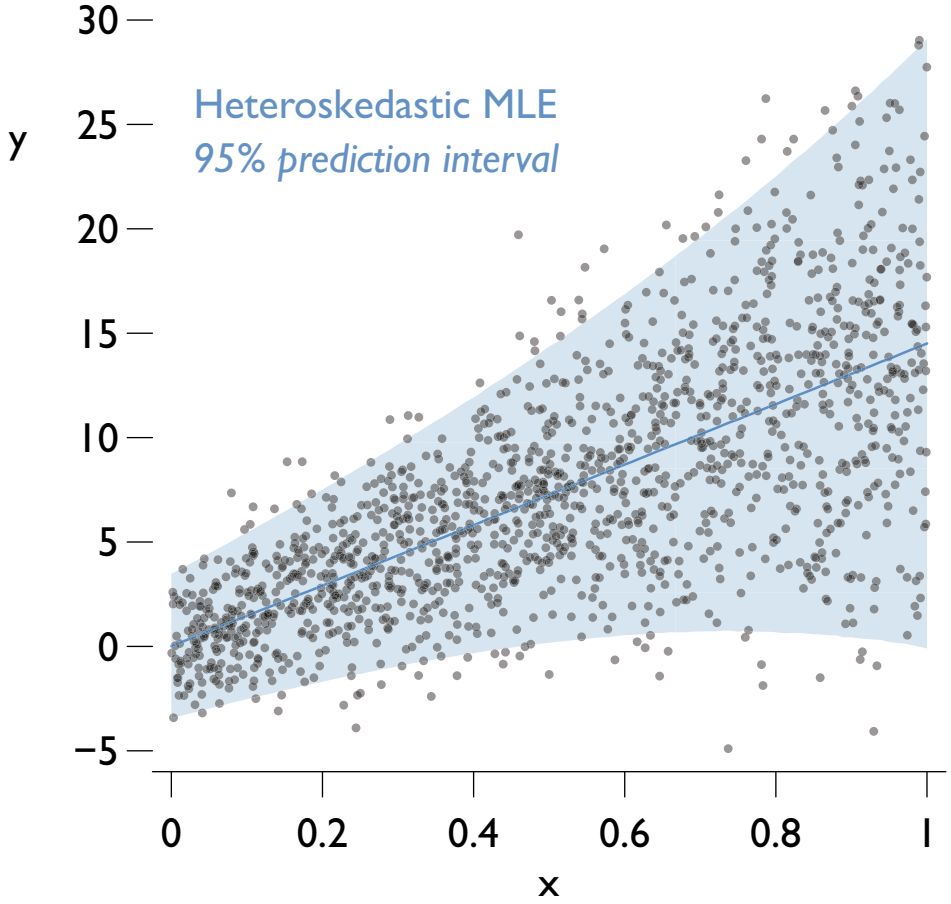


Key test is to compare *prediction intervals* (*not* confidence intervals – why?)

Predicted values are draws from the stochastic component of the model

95% of the data should lie in the 95% prediction interval





Unlike linear regression, the heteroskedatic MLE accurately captures relationships between the variance of y_i and the levels of covariates x_i

We've built a "new" model to better fit the substantive behavior of our data, and estimated it using maximum likelihood

Finding maximum likelihood estimates: analytical solutions

We've turned a hard problem — *finding most likely parameter values*

. . . into an easy one — *maximizing a single function*

Ideally, we'd just use calculus to find the maximum

Finding maximum likelihood estimates: analytical solutions

We've turned a hard problem — *finding most likely parameter values*

. . . into an easy one — *maximizing a single function*

Ideally, we'd just use calculus to find the maximum

For the turnout example,
we just need the derivative of a binomial distribution with a fixed π

$$\frac{\mathrm{d}\log\mathcal{L}(\pi|\mathbf{y})}{\mathrm{d}\,\pi} = \frac{1}{\pi}\sum_{i}^{n} y_i + \frac{1}{\pi - 1}\sum_{i}^{n} M_i - y_i$$

Plugging in $\mathbf{y}$ and $\mathbf{M}$, setting equal to 0, and solving for π reveals
$\hat{\pi}_{\mathrm{ML}} = 0.7998137511\ldots$

Finding maximum likelihood estimates: analytical solutions

We've turned a hard problem — *finding most likely parameter values*
. . . into an easy one — *maximizing a single function*

Ideally, we'd just use calculus to find the maximum

For the turnout example,
we just need the derivative of a binomial distribution with a fixed π

$$\frac{\mathrm{d}\log\mathcal{L}(\pi|\mathbf{y})}{\mathrm{d}\,\pi} = \frac{1}{\pi}\sum_i^n y_i + \frac{1}{\pi - 1}\sum_i^n M_i - y_i$$

Plugging in $\mathbf{y}$ and $\mathbf{M}$, setting equal to 0, and solving for π reveals
$\hat{\pi}_{\mathrm{ML}} = 0.7998137511\ldots$

To confirm this is a maximum, check the second derivative

$$\frac{\mathrm{d}^2\log\mathcal{L}(\pi|\mathbf{y})}{\mathrm{d}\,\pi^2} = -\frac{1}{\pi^2}\sum_i^n y_i + \frac{1}{(\pi - 1)^2}\sum_i^n M_i - y_i$$

Plugging in $\mathbf{y}$, $\mathbf{M}$, and $\hat{\pi}_{\mathrm{ML}}$ yields -21911001, confirming a maximum

Finding maximum likelihood estimates: analytical solutions

With a few exceptions (such as linear regression),
we lack analytic solutions for MLEs

Instead, we use *numerical* methods:
Have the computer search and test many possible solutions iteratively

Iterative search

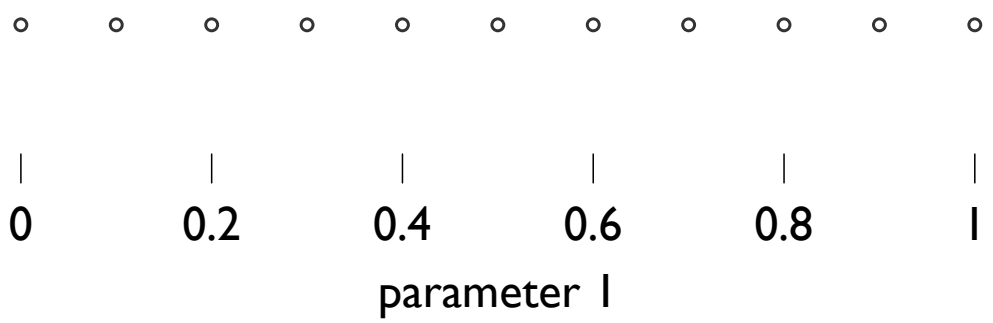
1. Start with an initial guess
2. Use your current guess to seek a new best guess
3. Repeat step 2 until "convergence": e.g., the local derivative of $\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \approx 0$

Many search algorithms are available,
ranging from brute force to inspired and elegant approaches

Numerical Methods of Optimization

Grid search brute force: casting ever-finer nets

1D grid search: compute L(θ|y) for each



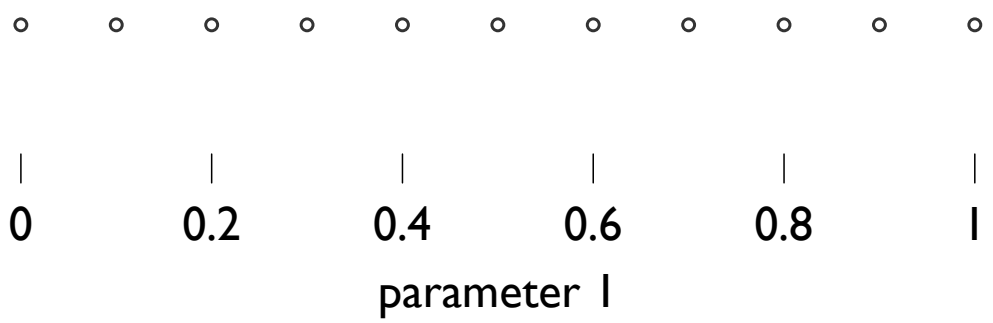
We've already seen the grid search applied to the turnout example

Grid search works well for maximizing a single unknown parameter,
especially when the likelihood is globally concave

If you have doubts about concavity,
could use a very fine mesh and check to see if there is more than a single peak

This adds computation time: number of points x number of iterations

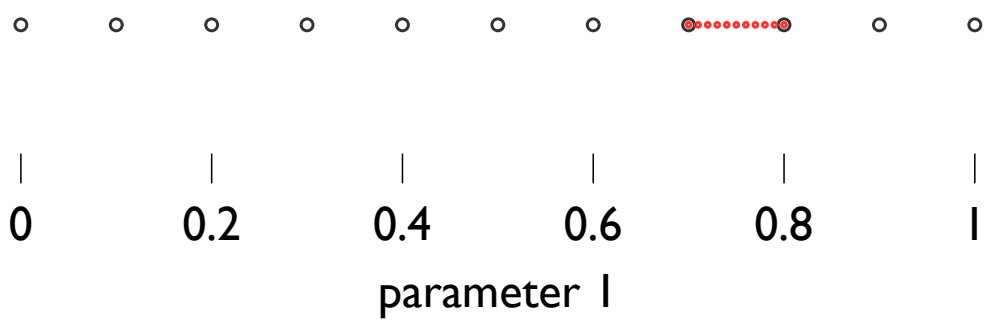
1D grid search: compute L(θ|y) for each



We compute the above 11 values of the likelihood given hypothetical values of π

. . . select the highest pair

1D grid search: compute L(θ|y) for each



We compute the above 11 values of the likelihood given hypothetical values of π

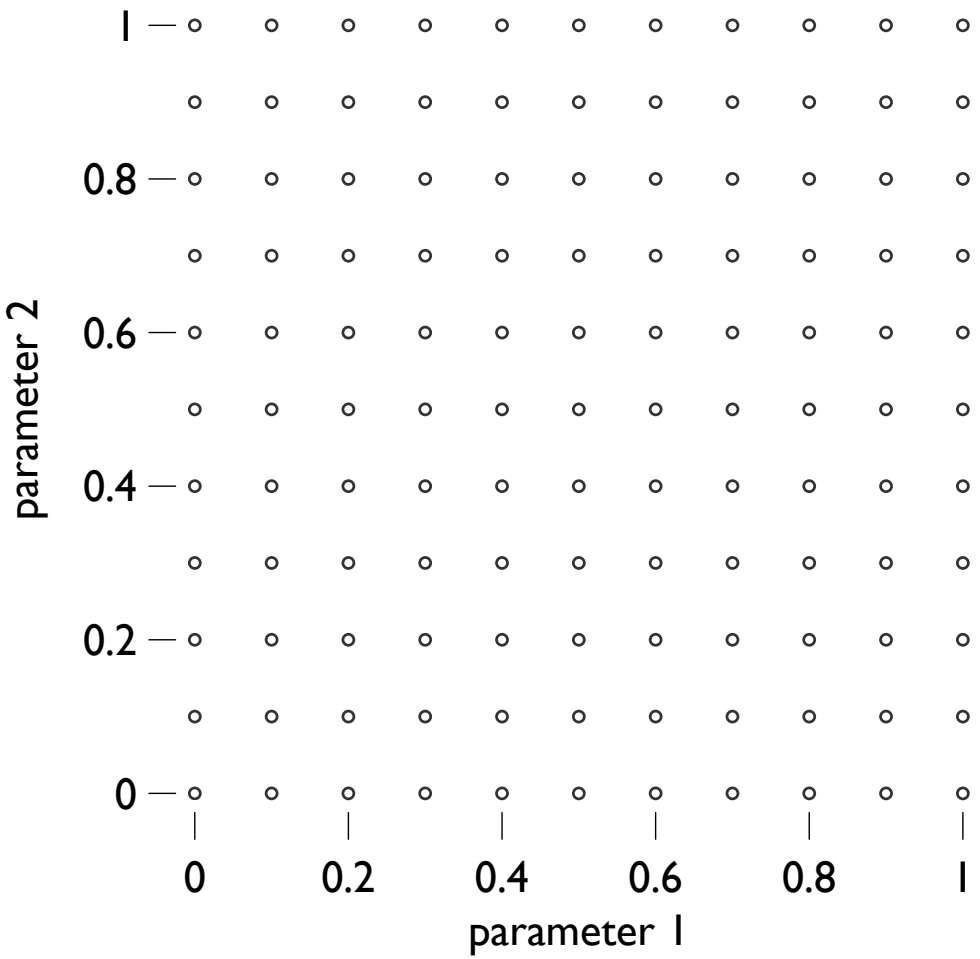
. . . select the highest pair

. . . then repeat the exercise between those two π's

. . . and iterate until the desired precision is reached (convergence)

Earlier, 5 iterations provided us with convergence to 3 digits (0.799. . .)

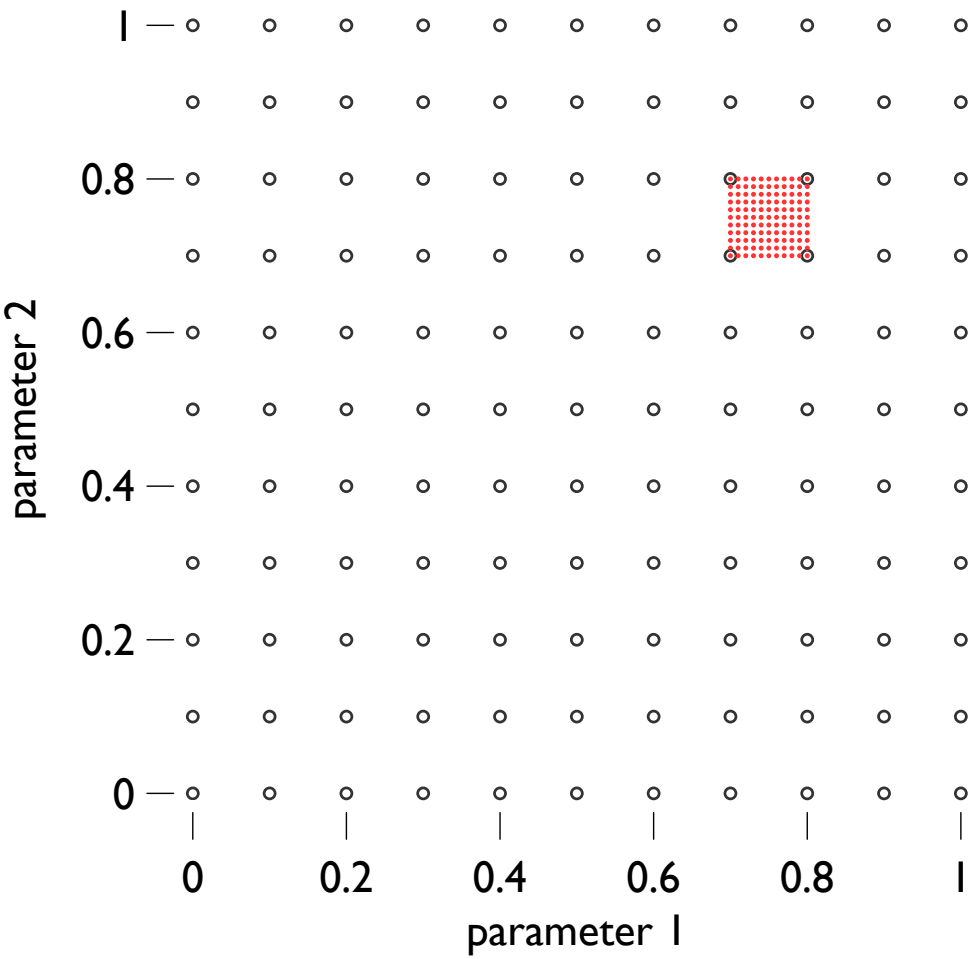
2D grid search: compute L(θ|y) for each



But what if you have 2 unknown parameters?

Now you need to compute every pair of possible values:
$11^2 = 121$ calculations per iteration

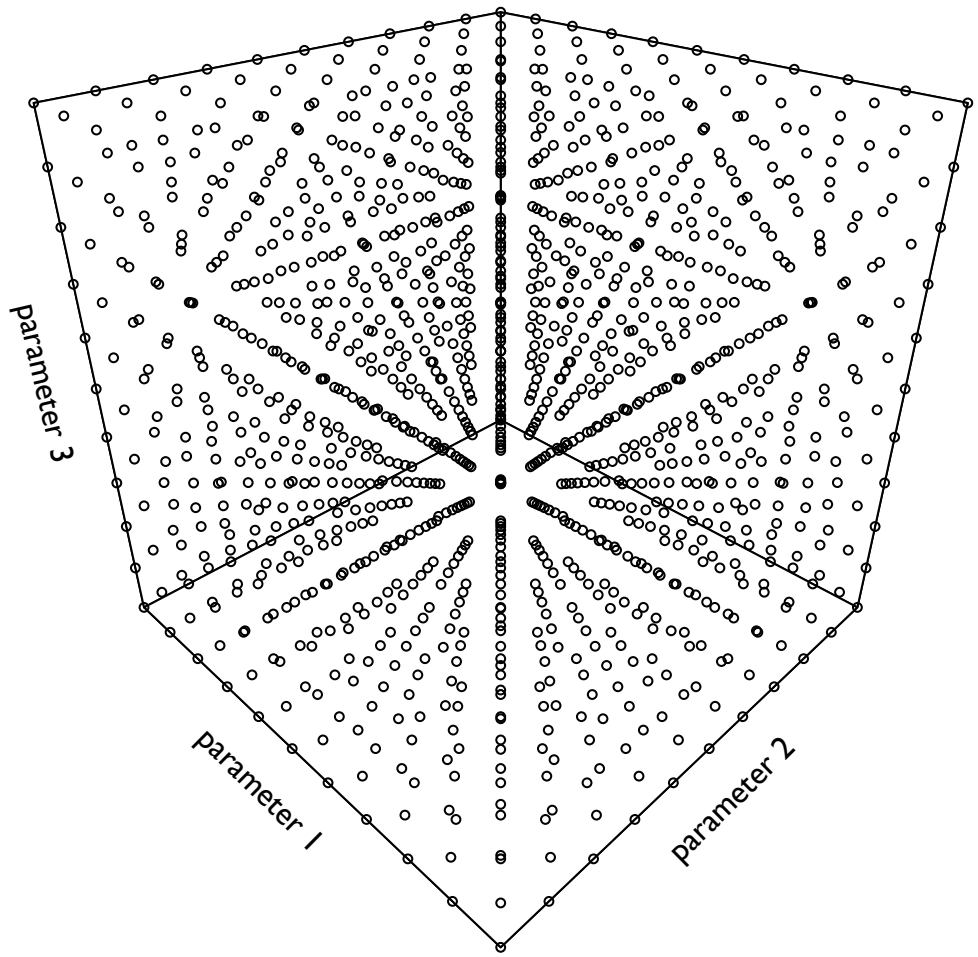
2D grid search: compute L(θ|y) for each



Now we choose the square of the grid with the highest values, and repeat the grid inside the square

If we needed 5 iterations here, we'd do a total of 605 computations

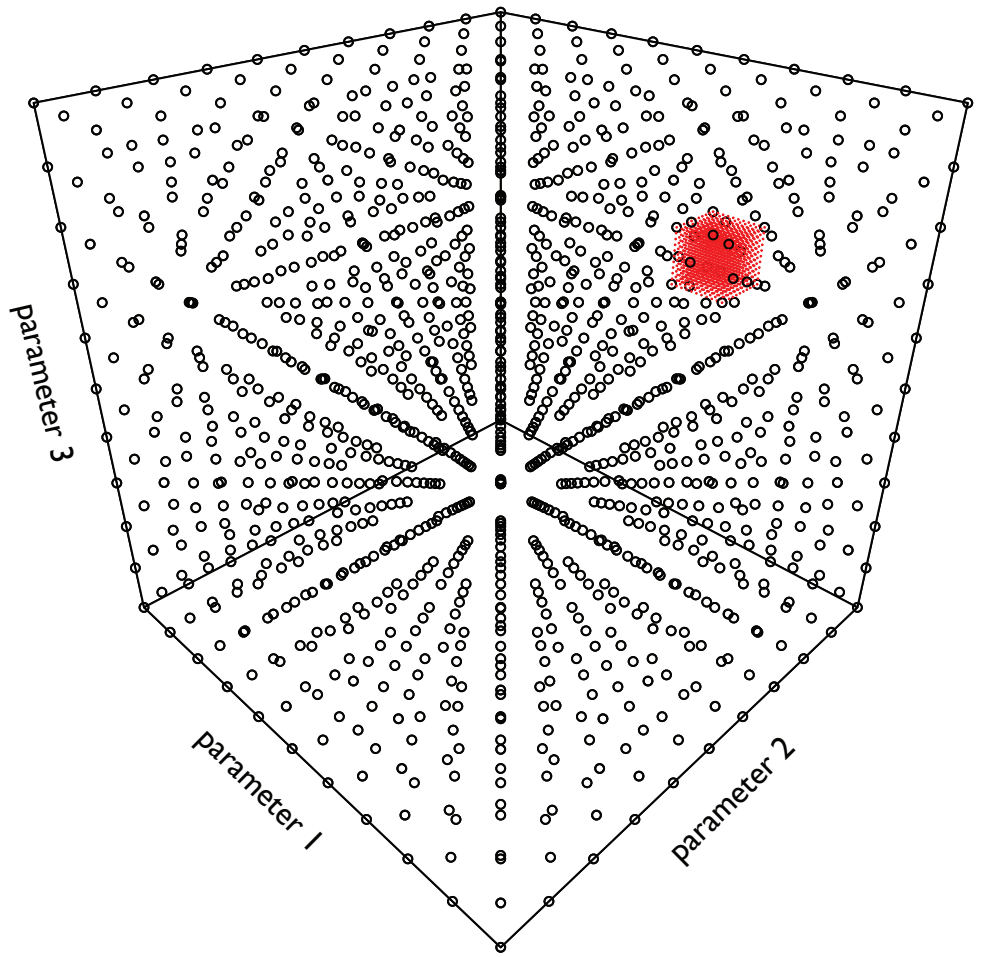
3D grid search: compute L(θ|y) for each



What if there are three unknown parameters?

Now there are $11^3 = 1331$ computations per iteration

3D grid search: compute L(θ|y) for each



As the number of parameters rises, grid search becomes computationally infeasible

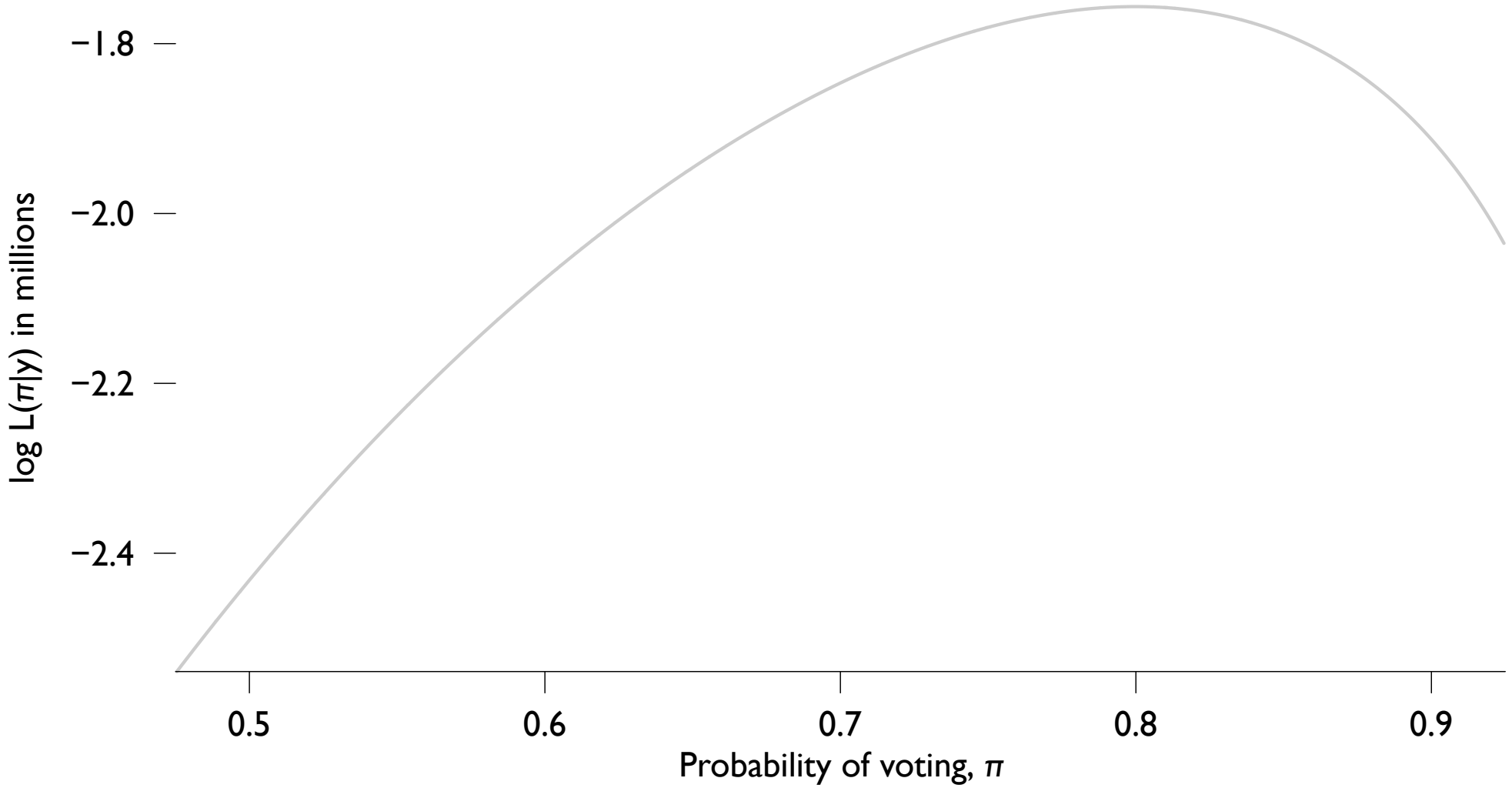
A regression model with 10 unknown parameters is hardly unusual,
but would take $11^{10} = 25,937,424,601$ computations per iteration to estimate

Numerical Methods of Optimization

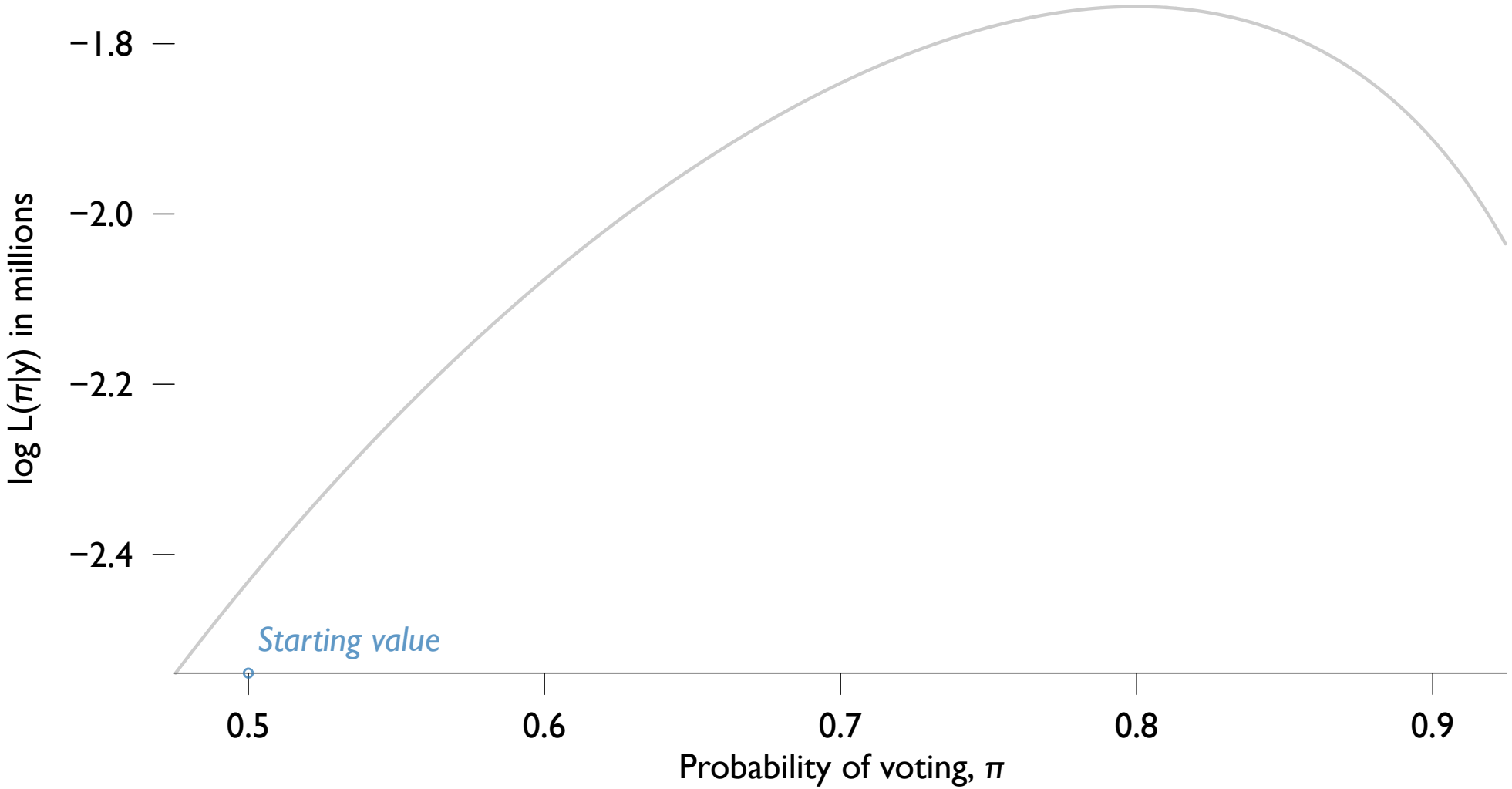
Grid search — brute force: casting ever-finer nets

Gradient descent/ascent — step-by-step hill-climbing

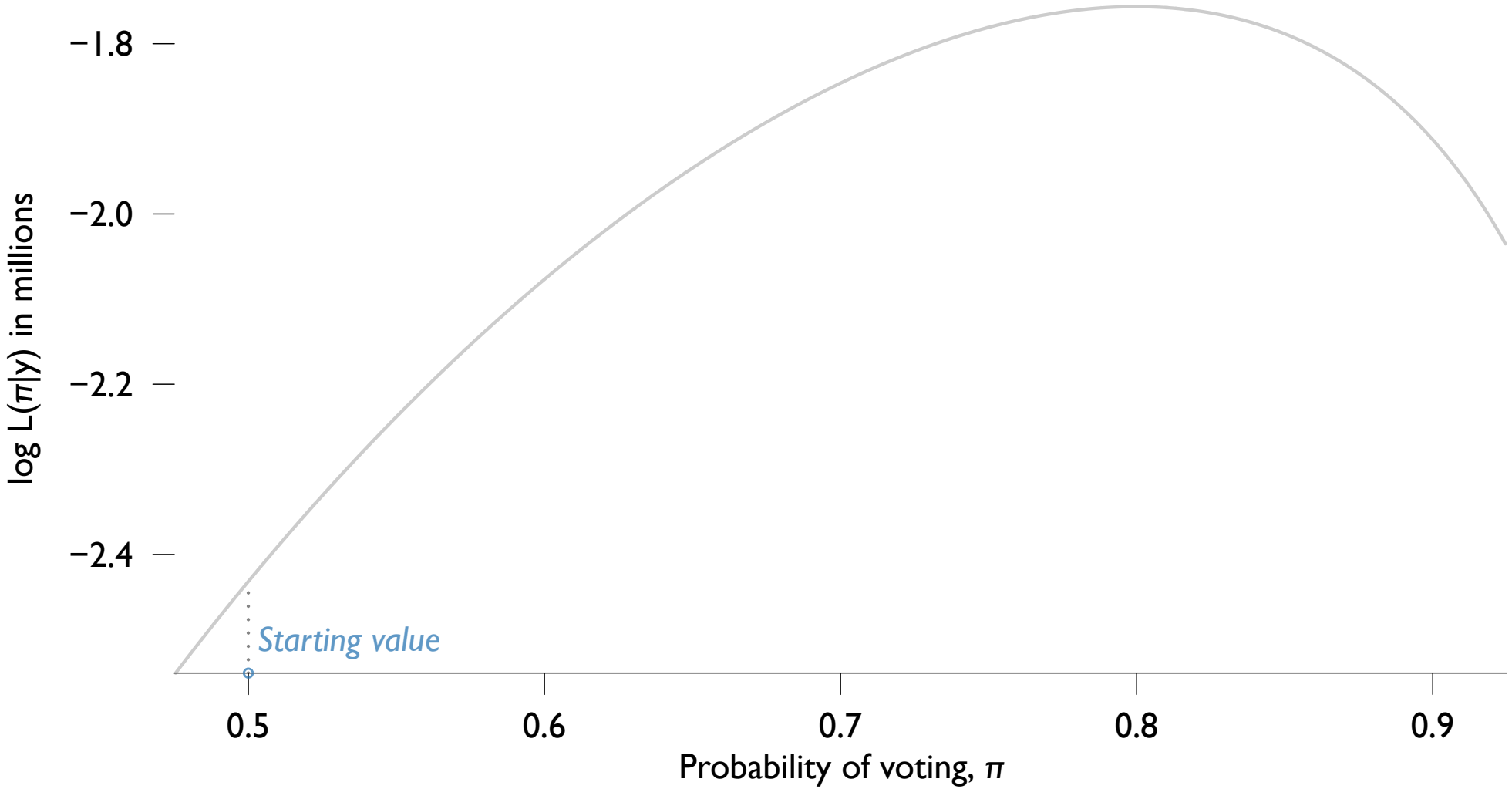
Newton-Raphson, Nelder-Mead, BFGS



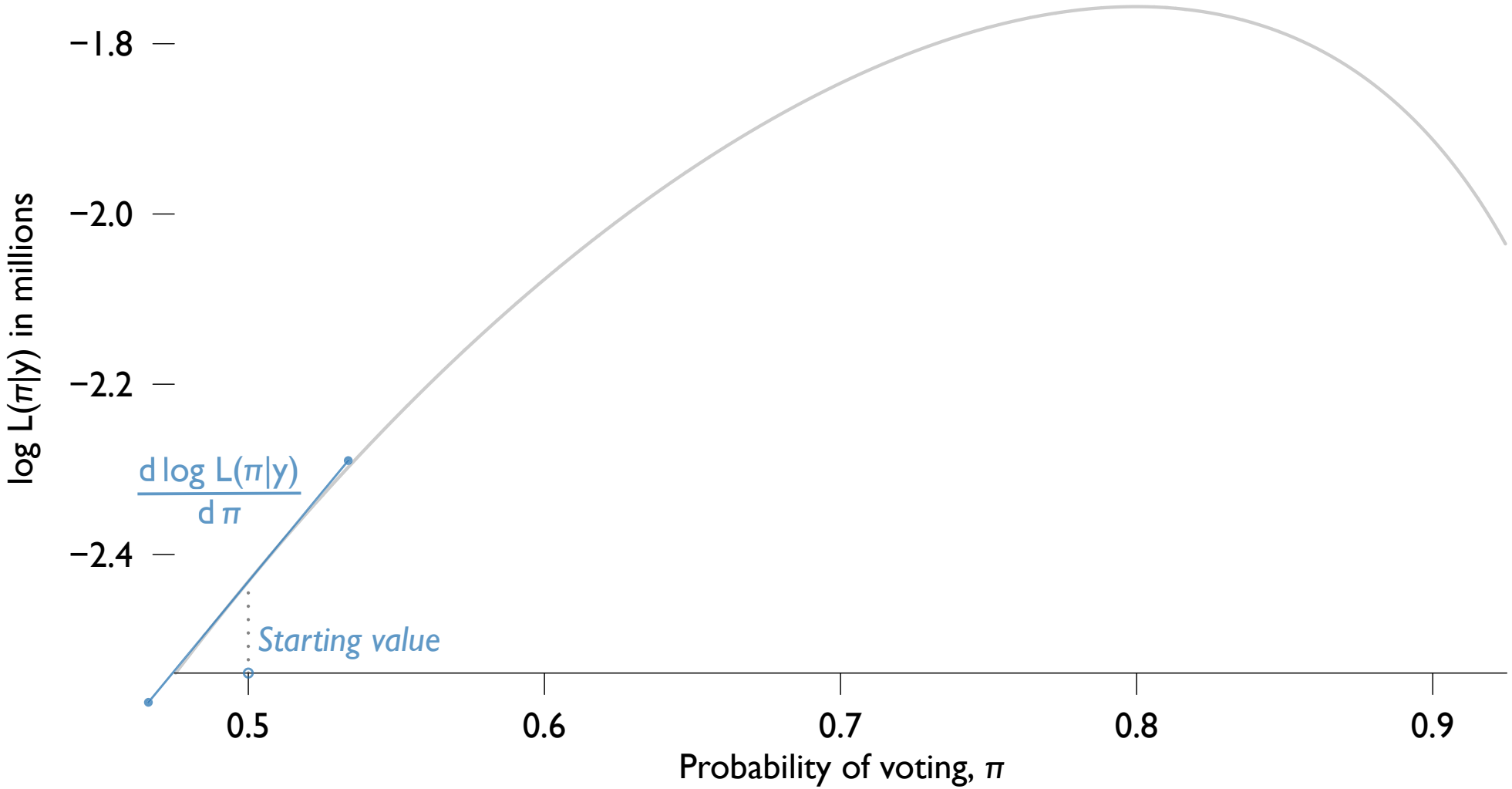
Suppose we want to find the maximum of this binomial likelihood without using analytic derivatives or brute force



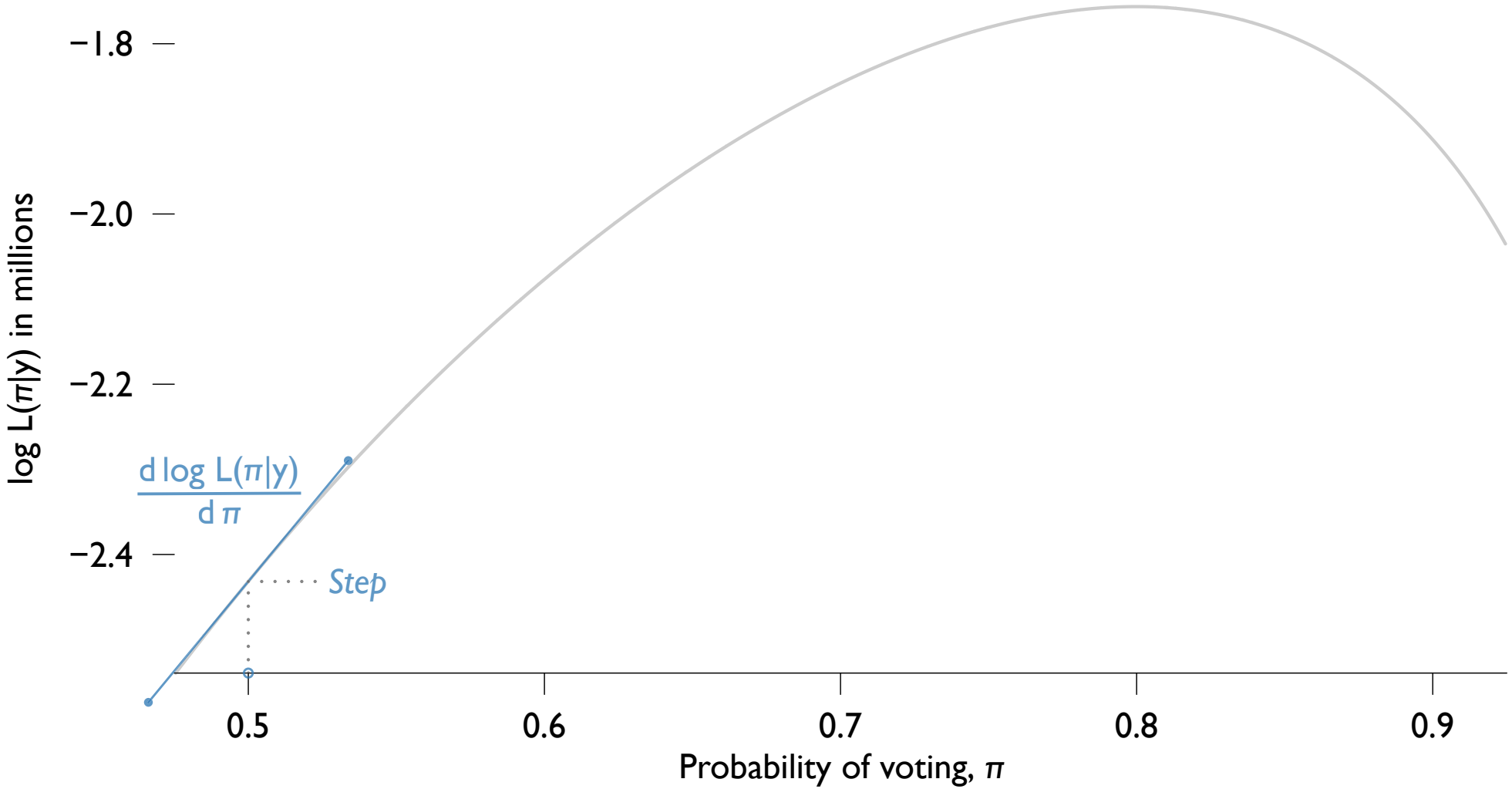
Instead, let's *assume* it's twice differentiable and globally concave,
and propose a randomly selected point π as a candidate maximum



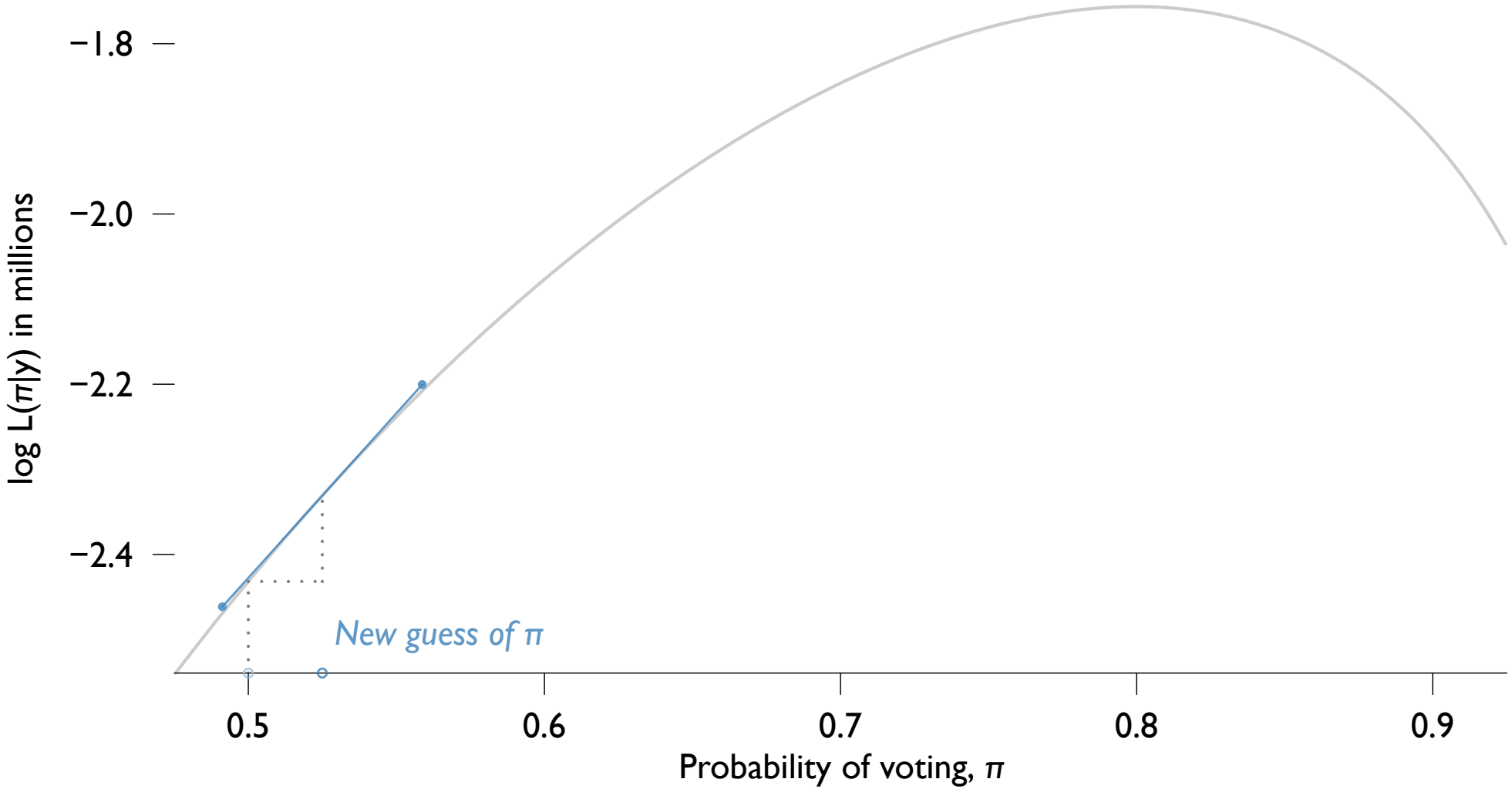
Using the data, we calculate the log likelihood at the candidate π. . .



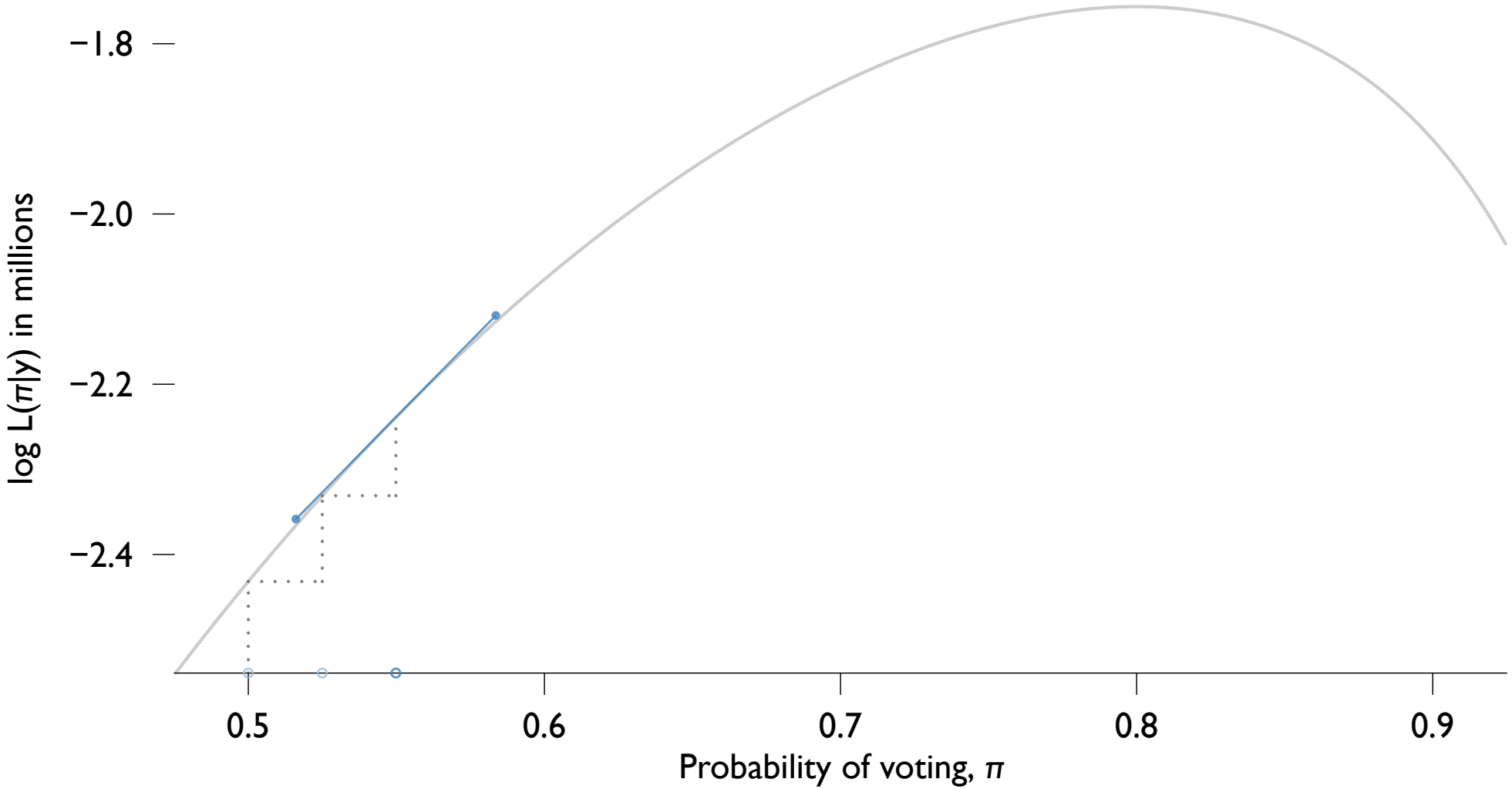
And compute a local approximation of the derivative,
which turns out to be positive



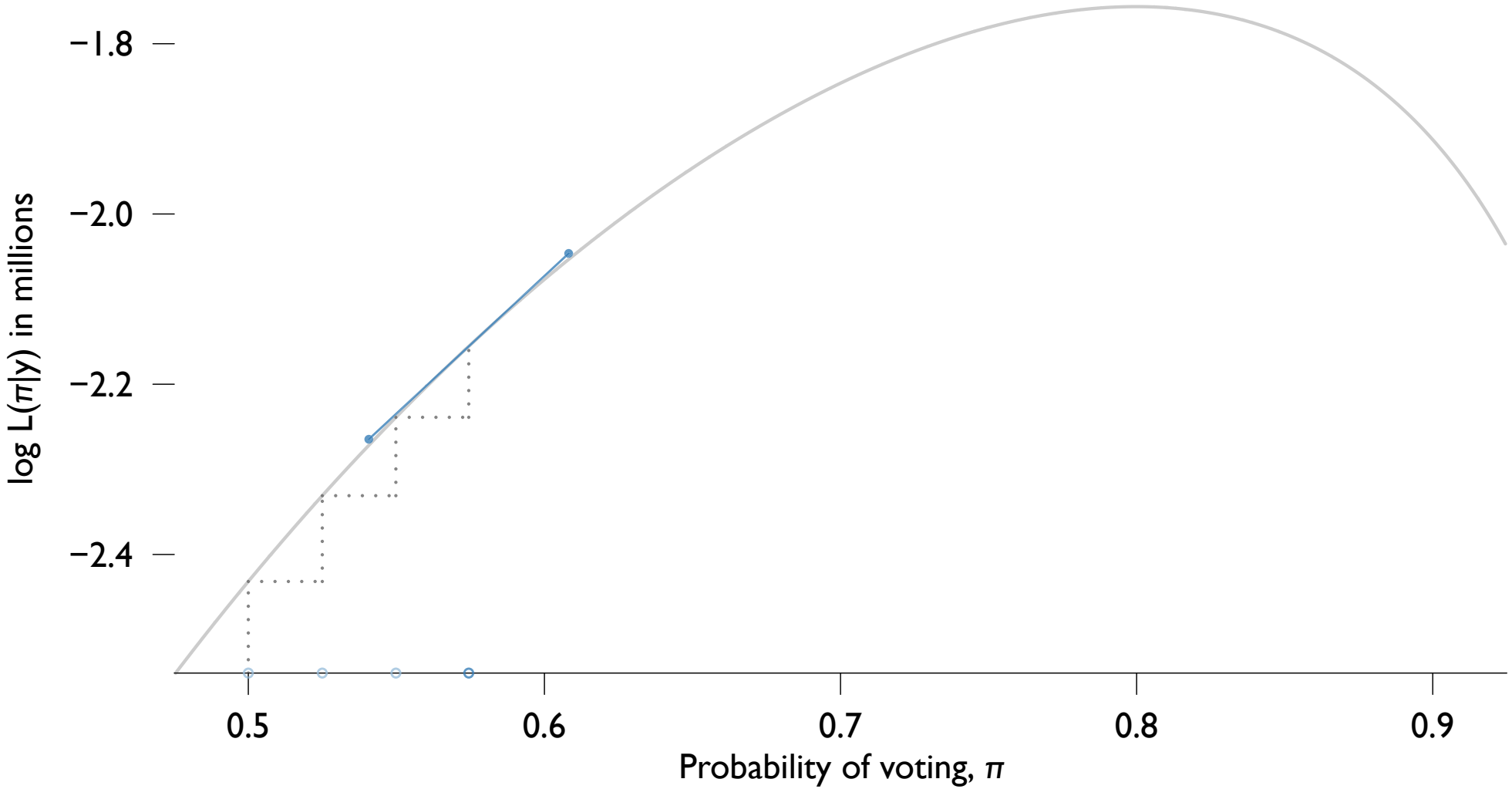
This suggests we should step to the right to find a new candidate π



. . . and repeat the process

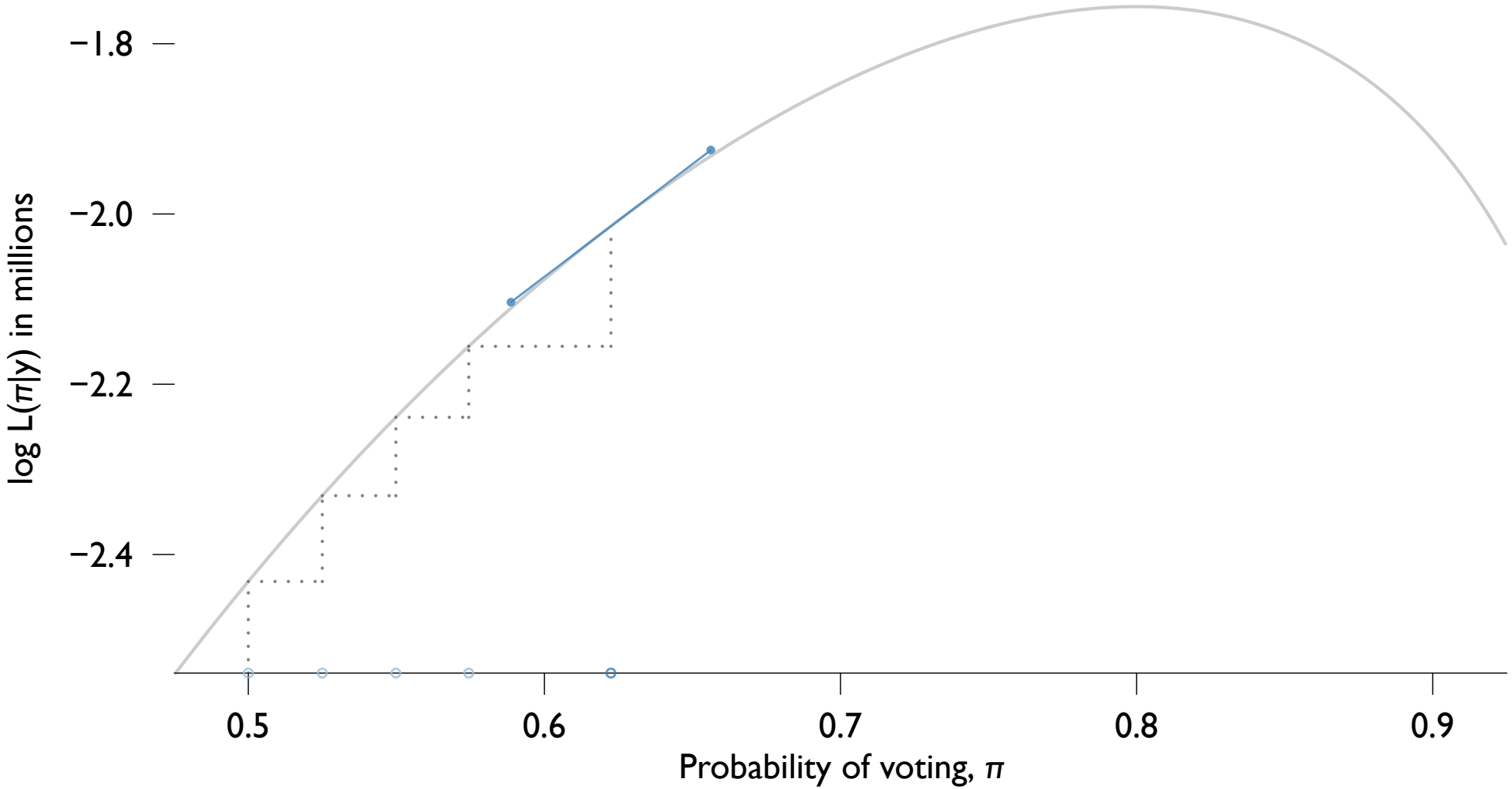


We iterate, taking bigger or smaller steps,
as suggested by our gradient search algorithm

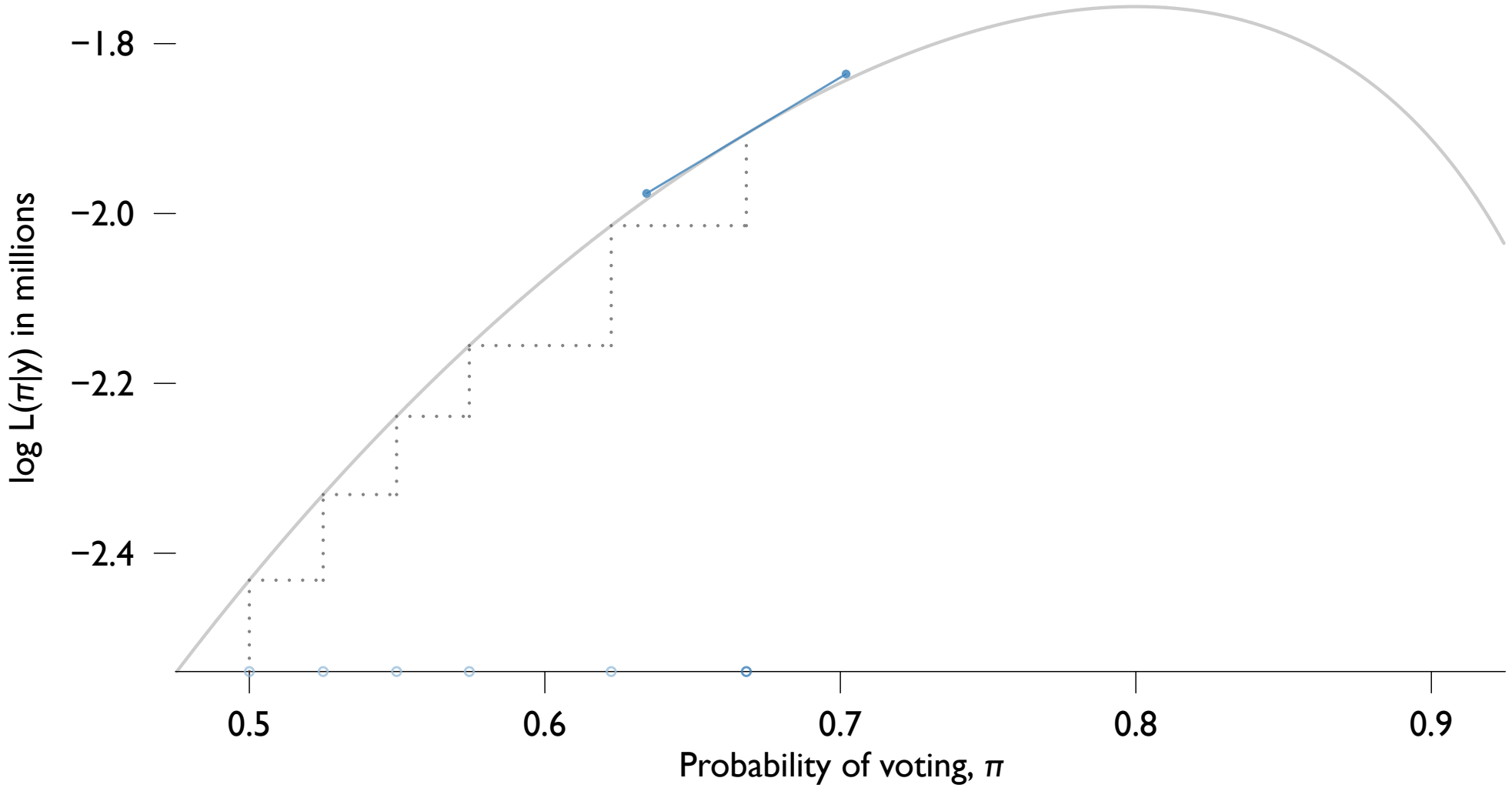


Gradient = a derivative computed across one *or more* dimensions

This algorithm can climb hills or 1, 2, or many dimensions

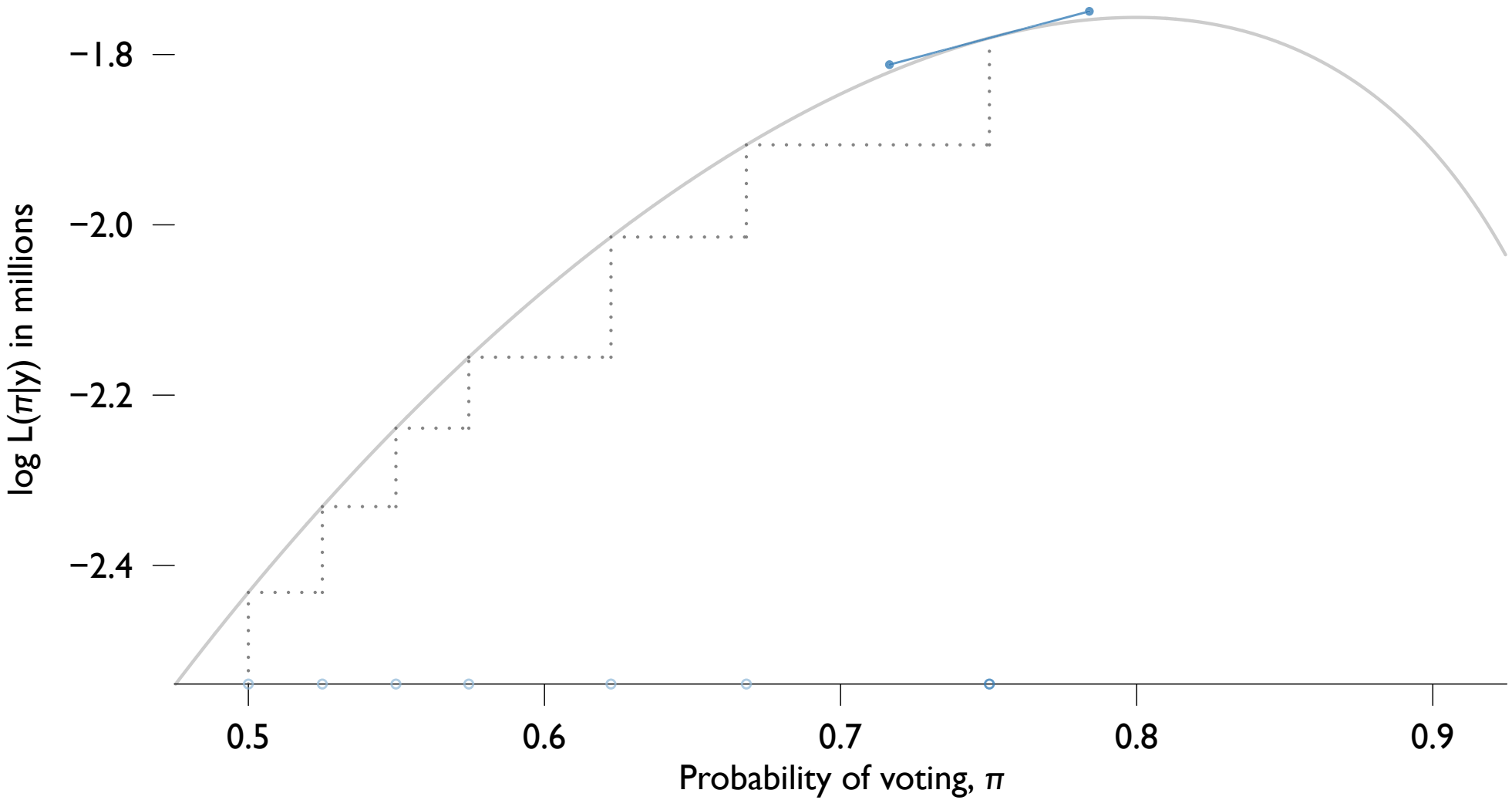


At each step, we shift our candidate along each dimension (parameter), tending to take bigger steps in directions that are "steep"

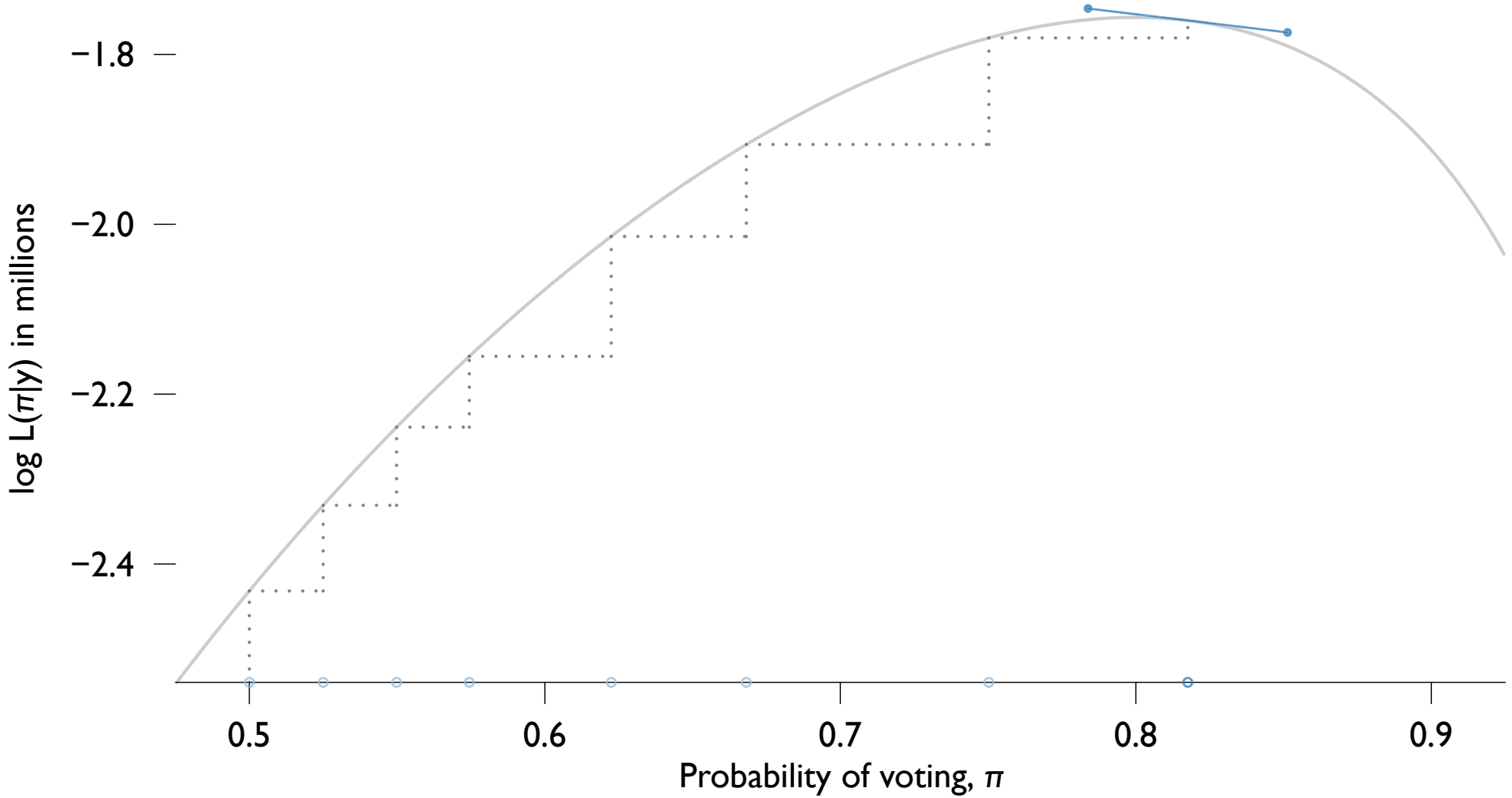


Like climbing a hill blindfolded:

To reach the top fast, step in whichever direction rises fastest, turning as needed

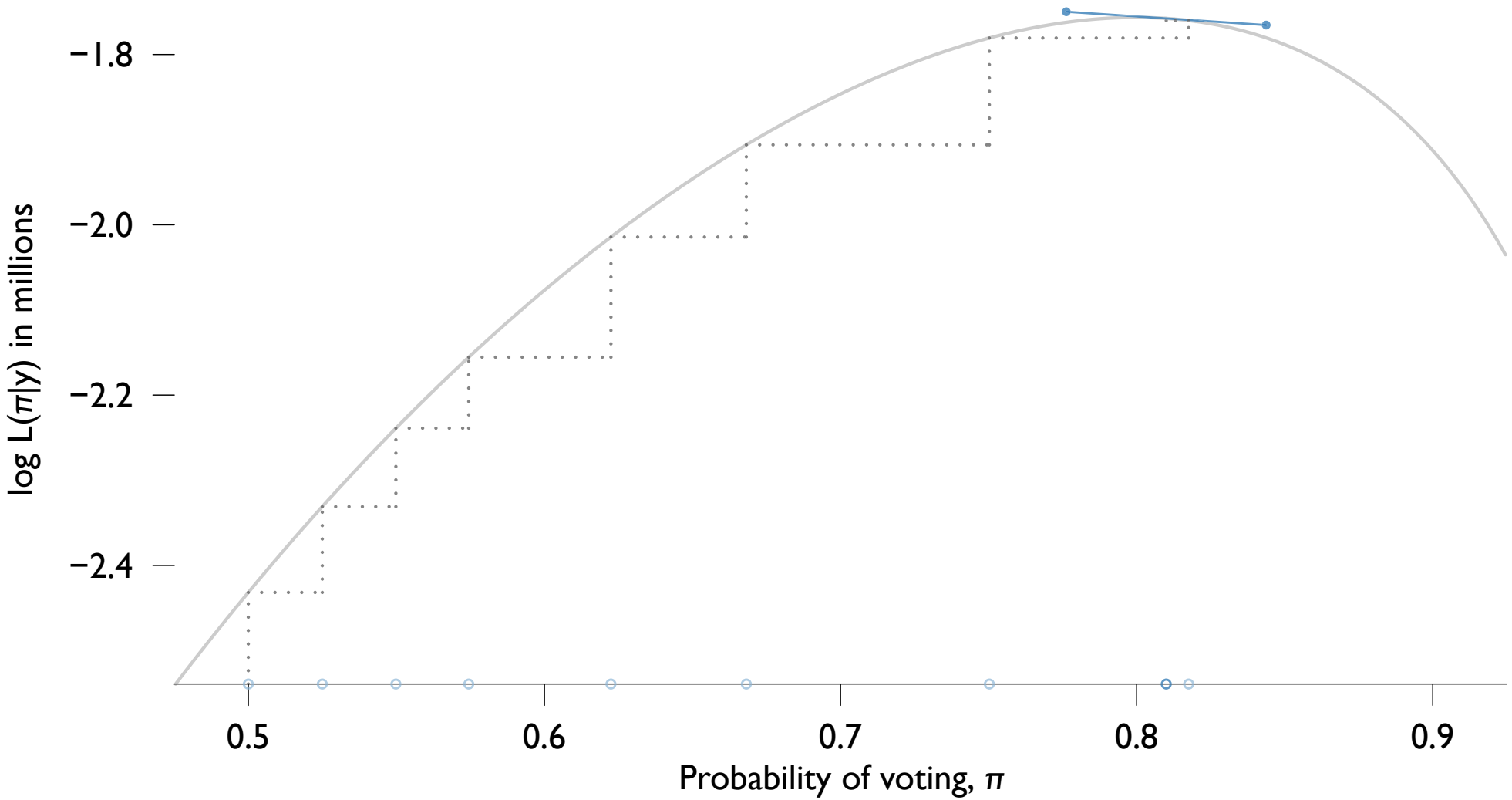


So far, each step has brought us closer to the top of the hill

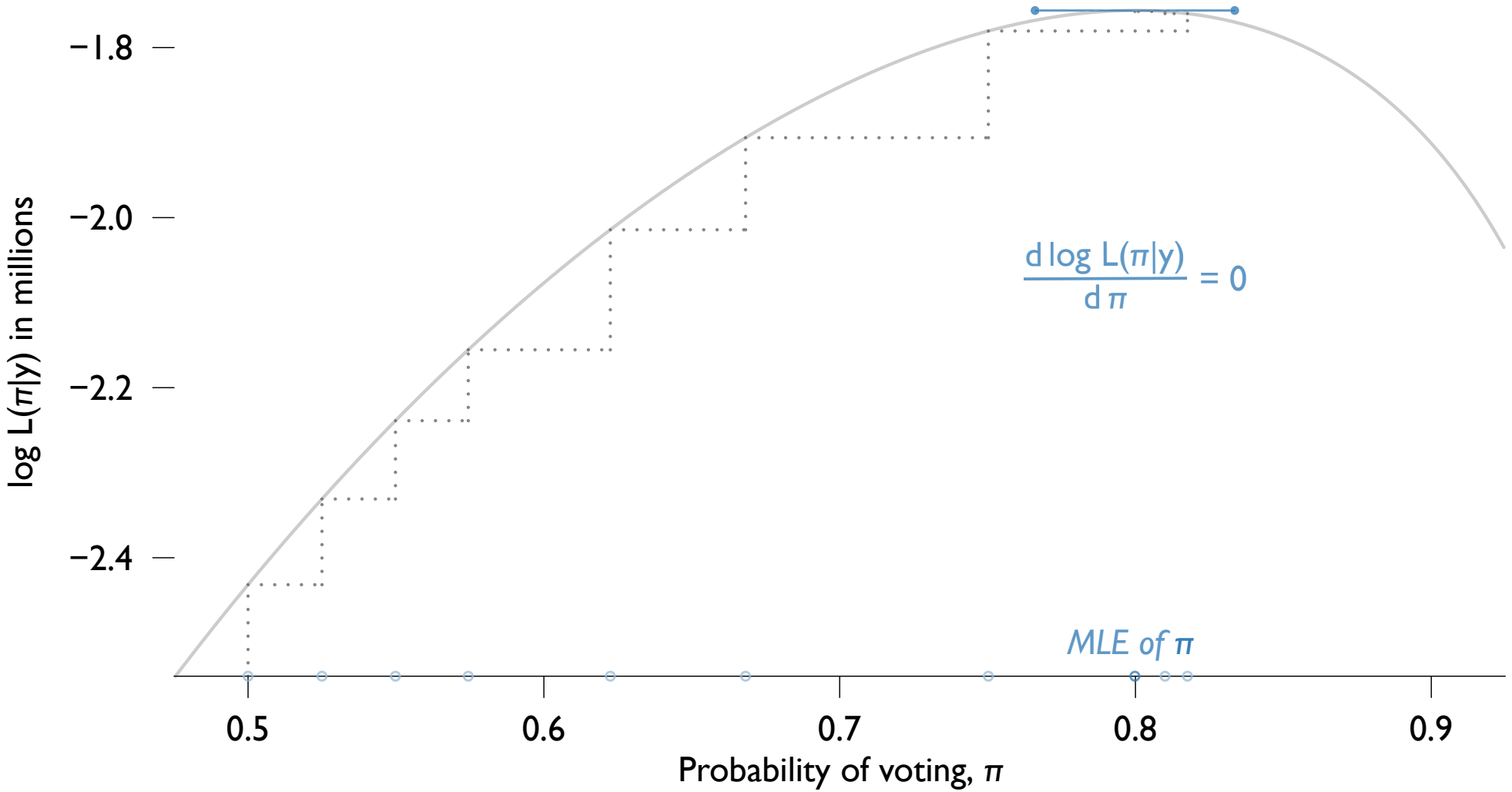


But overstepping is likely to happen eventually. . .

The derivative now suggests we should move back to the left, just a little

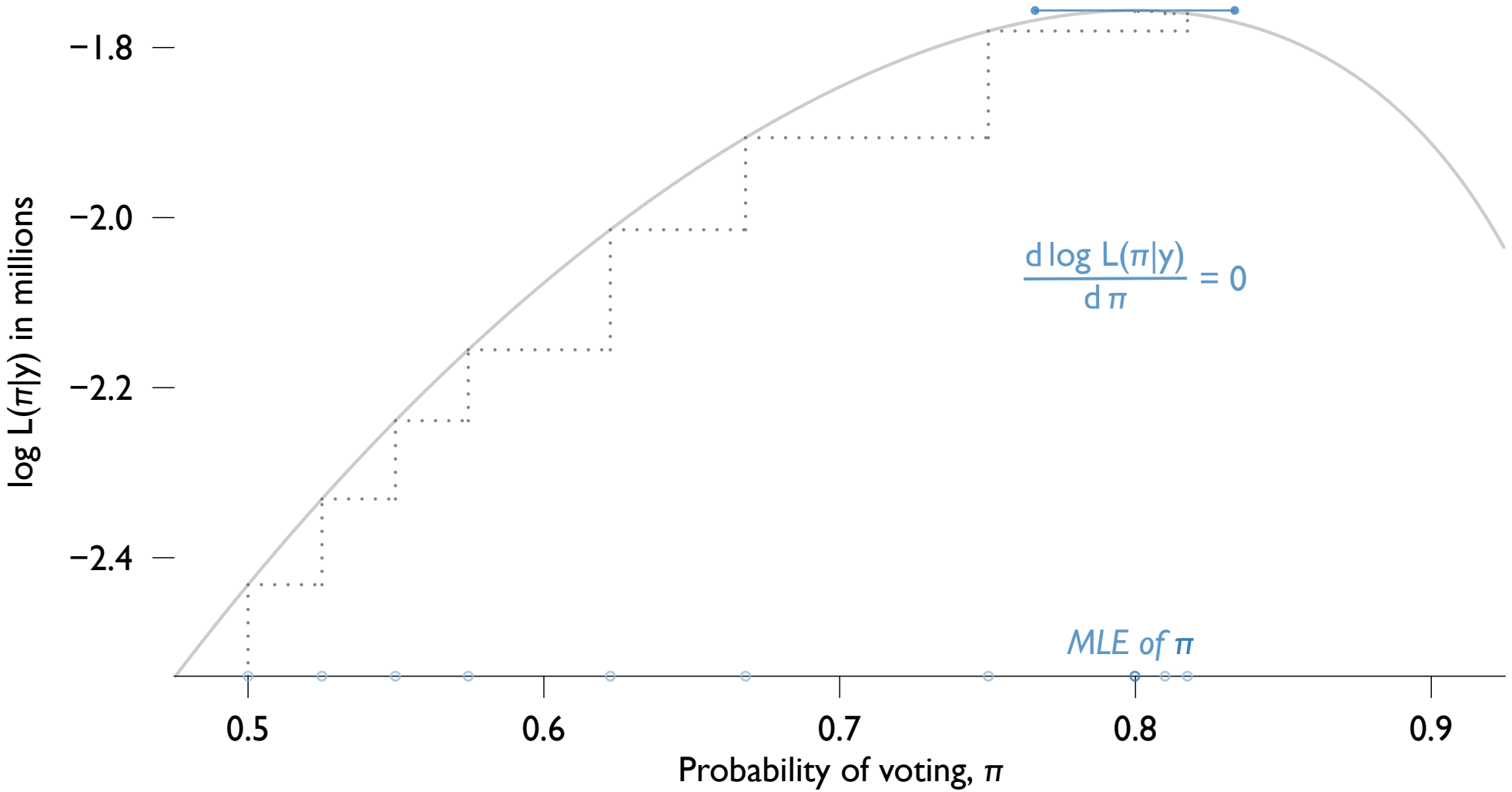


From here, I've omitted dozens of small steps zeroing in on the maximum

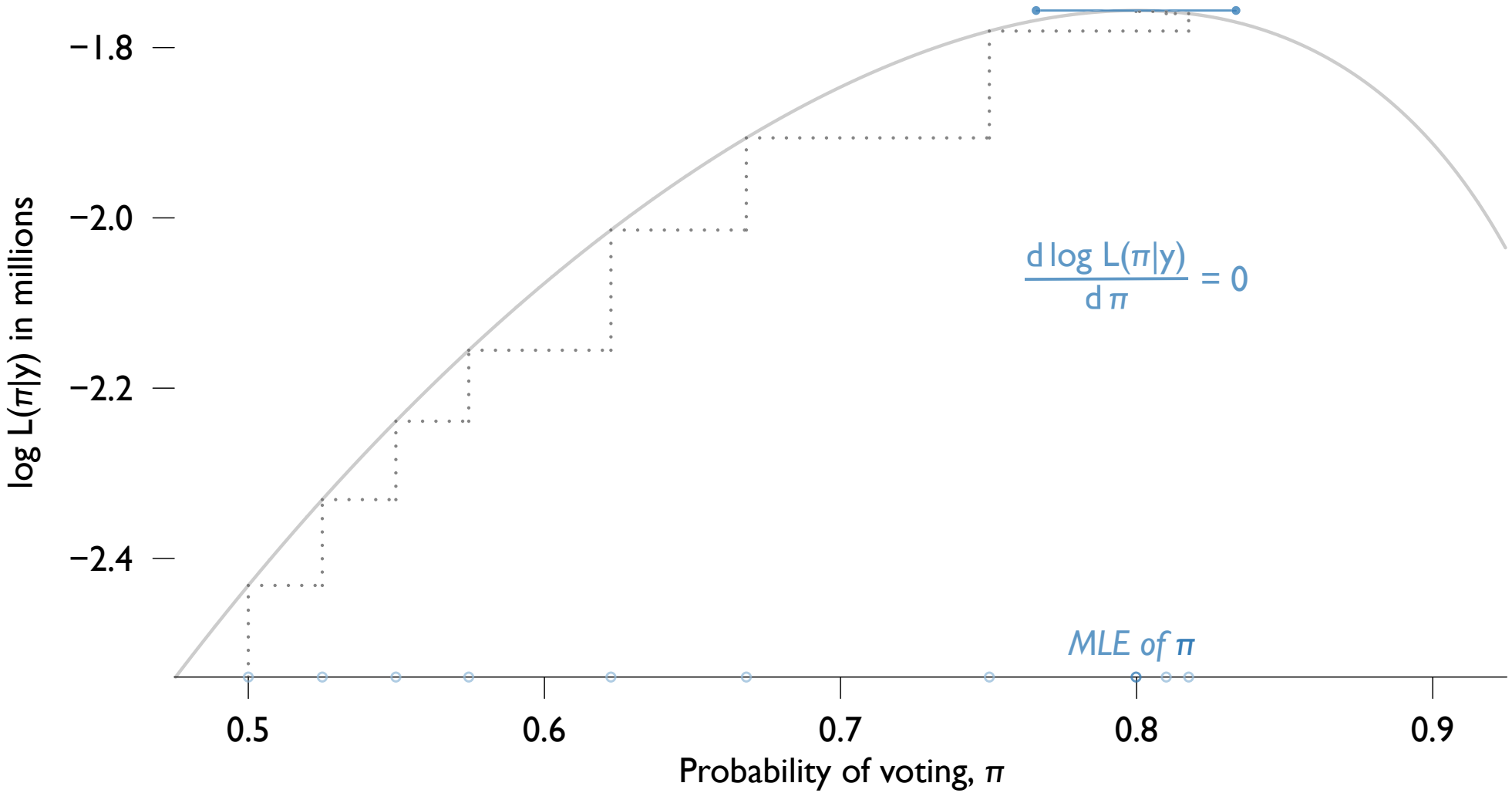


Within 50 or so steps, we find a parameter that produces a zero gradient

We treat the final candidate π as our MLE

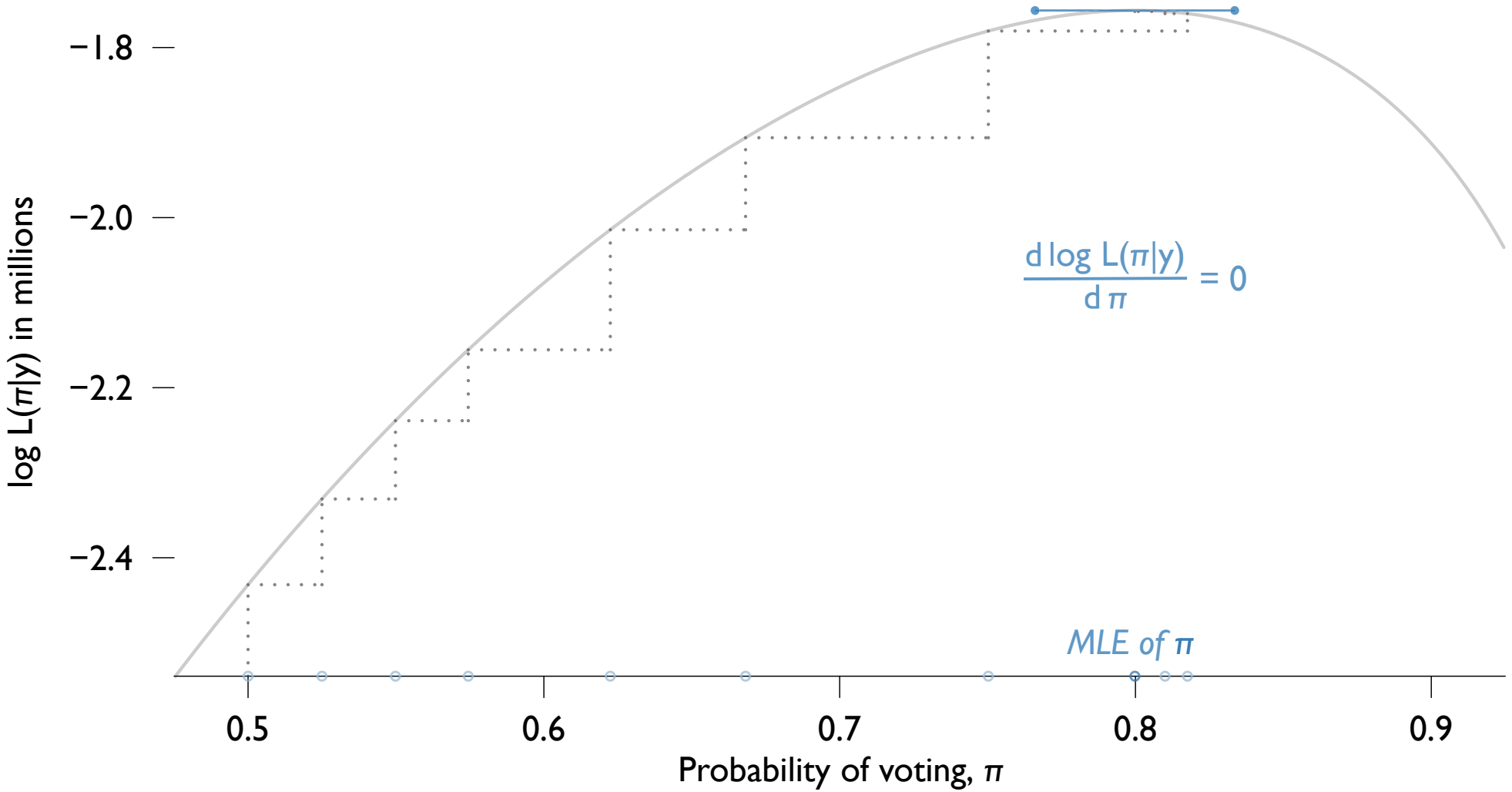


The same logic applies even if there are many k parameters $\boldsymbol{\theta}$ to estimate

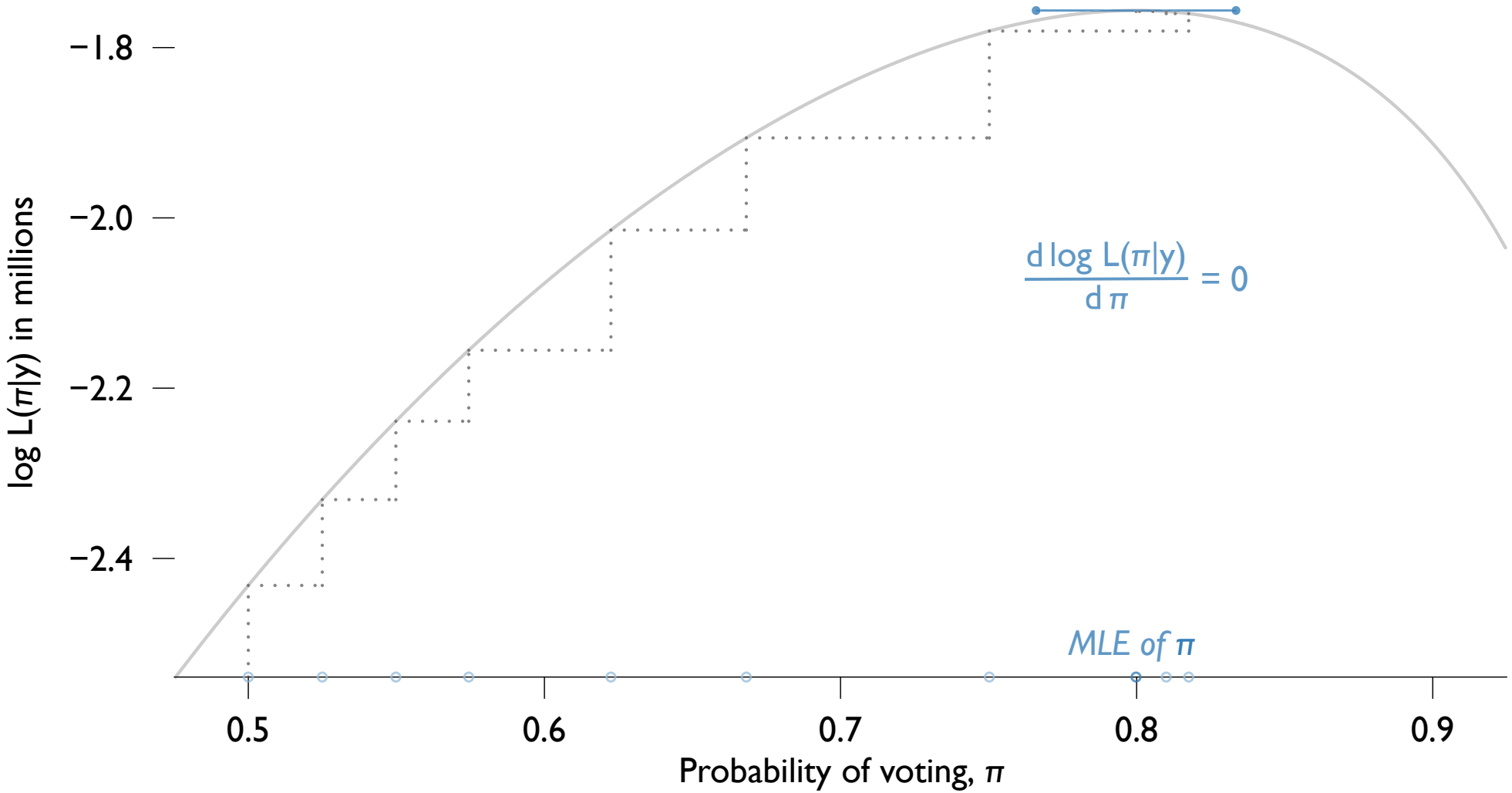


In that case, the "hill" of the likelihood exists in $k+1$ space

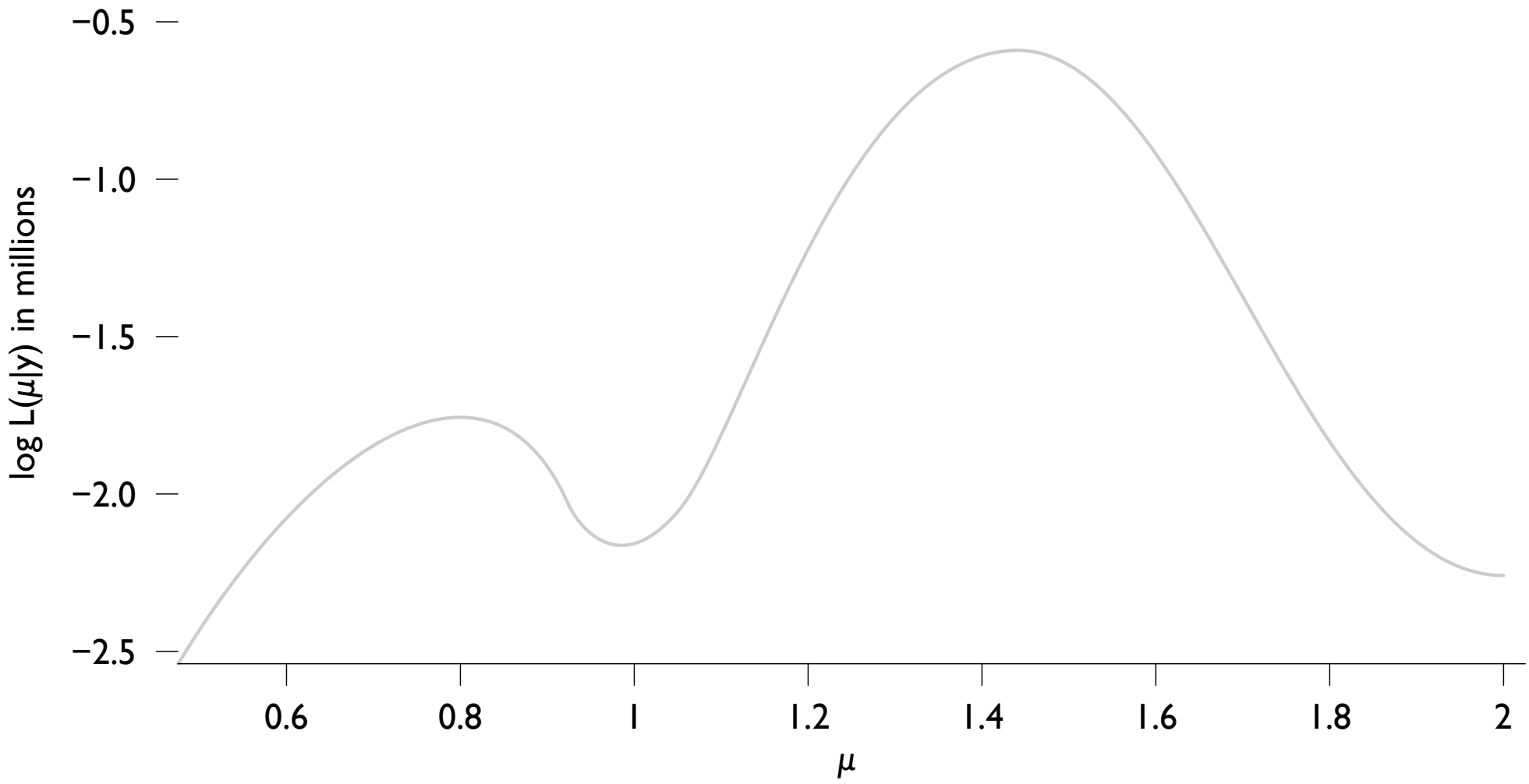
The gradient is a k-vector of derivatives, $\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})/\partial\boldsymbol{\theta}$



In general, another name for the gradient of the likelihood wrt some particular θ is the *score* of the likelihood with respect to θ

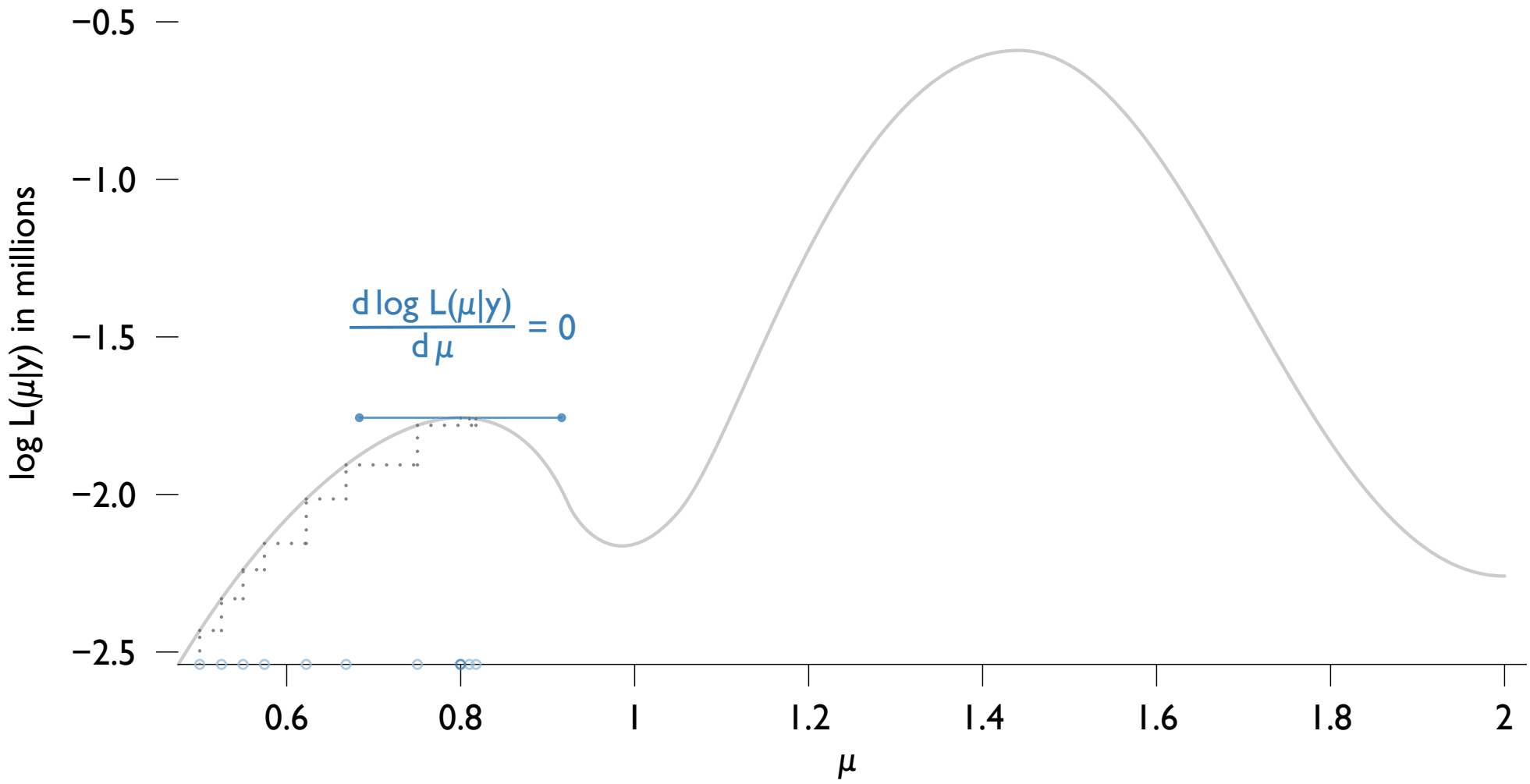


At the MLE, the score is 0;
large scores away from the MLE suggest the likelihood is *sensitive* to the parameter θ

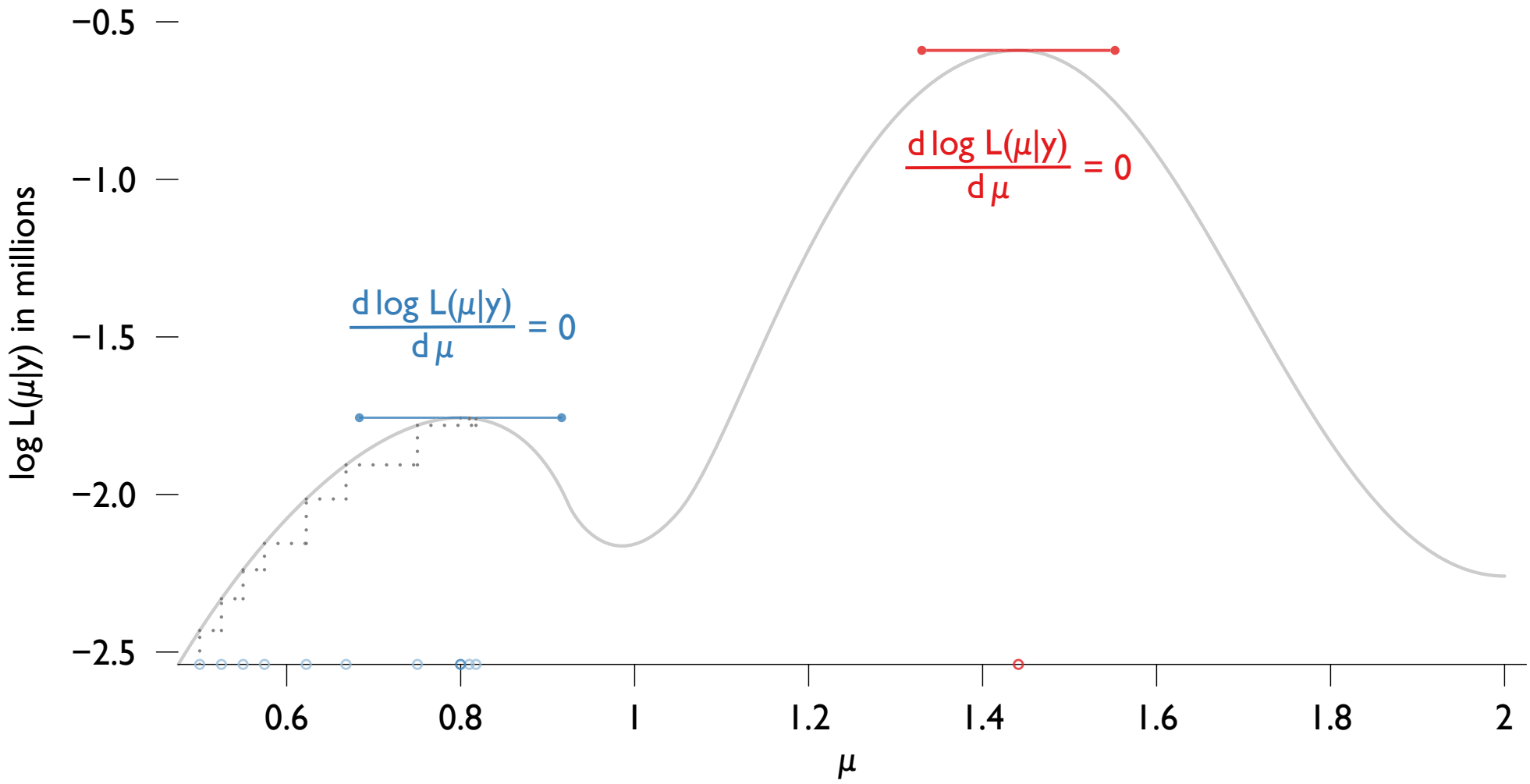


We assumed our likelihood surface was globally concave – what if it isn't?

(Note this is an invented curve, not a likelihood from a particular distribution)

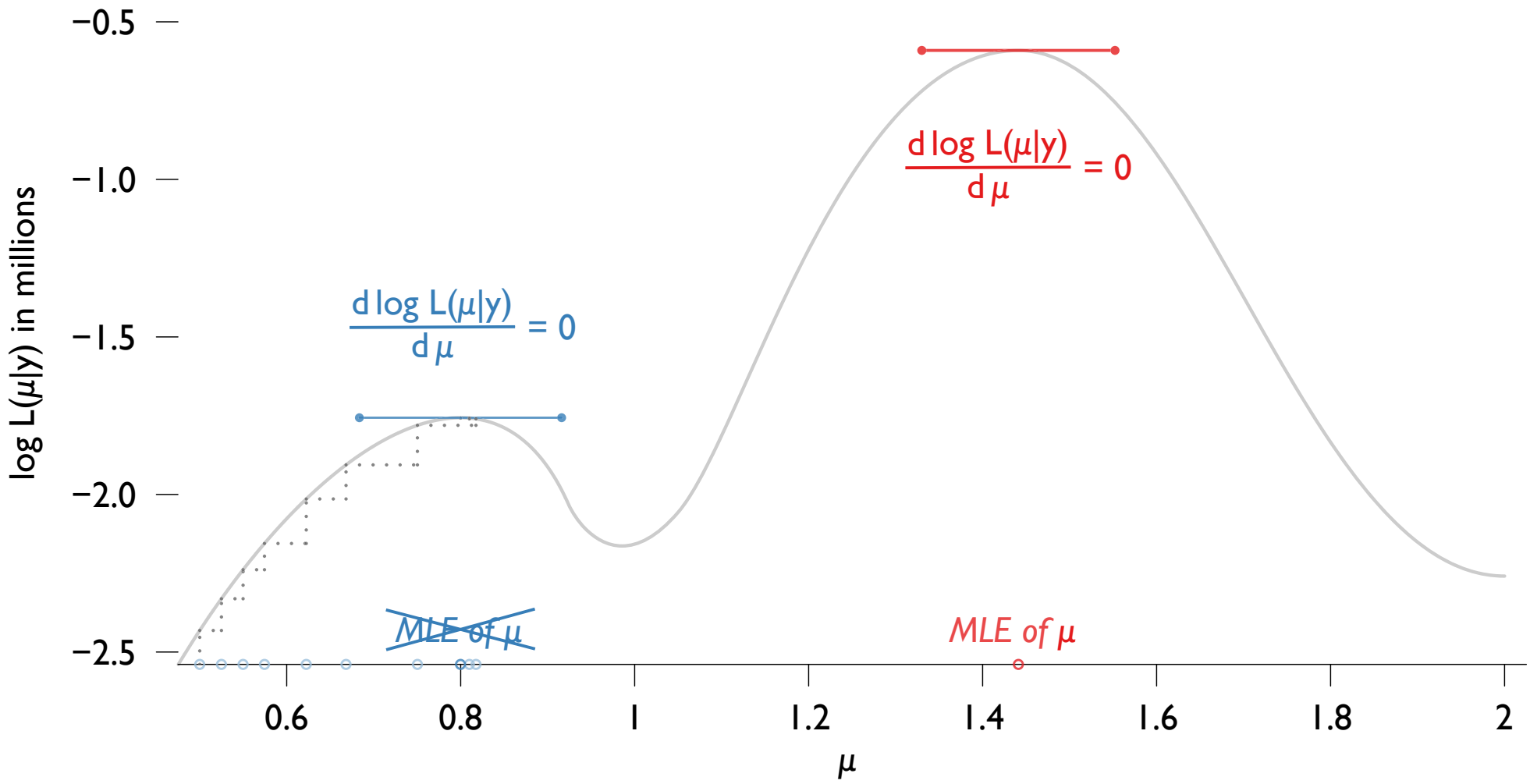


If we had started at $\mu = 0.5$, or any $\mu < 1$,
we'd have found this local maximum as our "MLE"



But the global maximum is far to the right, near $\mu = 1.4$

Starting values above $\mu = 1$ would find this maximum



Only the global maximum is the MLE of μ

If you suspect local maxima, try multiple starting values

Numerical Methods of Optimization

Grid search — brute force: casting ever-finer nets

Gradient descent/ascent — step-by-step hill-climbing

Newton-Raphson, Nelder-Mead, BFGS

Simulated annealing — metallurgy metaphor

melt-freeze-repeat

Genetic algorithms — natural selection on population of sol'ns

mutate-select-repeat

Particle swarm optimization — semi-autonomous agents

search and share solutions

Markov-chain Monte Carlo (MCMC) — draw correlated series

Gibbs sampler, Metropolis-Hastings, of random numbers

Hamiltonian Monte Carlo converging in probability

to a target distribution

Numerical Methods of Optimization

Method	Virtues	Limitations
Grid search		very slow can find local maxima
Gradient descent	very fast scales well	can find local maxima needs smooth surfaces
Simulated annealing	avoids local maxima works with discontinuities	imprecise scales poorly
Genetic algorithms	minimal assumptions	can find local maxima scales poorly
Particle swarm optimization	minimal assumptions	convergence uncertain scales poorly
Markov chain Monte Carlo	convergence guaranteed modest assumptions	slow hard to assess

Statistical properties of MLEs

1. Minimum variance unbiasedness

Statistical properties of MLEs

1. Minimum variance unbiasedness
   - MVU is the unbiased estimator with least variance (highest efficiency)

Statistical properties of MLEs

1. Minimum variance unbiasedness
   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE

Statistical properties of MLEs

1. Minimum variance unbiasedness
   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

Statistical properties of MLEs

1. Minimum variance unbiasedness
   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization

Statistical properties of MLEs

1. Minimum variance unbiasedness
   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization
   - If est of α is $\hat{\alpha}_{\mathrm{ML}}$, then est of $\beta = f(\alpha)$ is $\hat{\beta}_{\mathrm{ML}} = f(\hat{\alpha}_{\mathrm{ML}})$

Statistical properties of MLEs

1. Minimum variance unbiasedness

   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization

   - If est of α is $\hat{\alpha}_{\mathrm{ML}}$, then est of $\beta = f(\alpha)$ is $\hat{\beta}_{\mathrm{ML}} = f(\hat{\alpha}_{\mathrm{ML}})$
   - But $\mathbb{E}(f(\alpha)) \neq f(\mathbb{E}(\alpha))$ for most $f(\cdot)$,
     so choice of parameterization can cause bias in small samples (e.g., for σ^2)

Statistical properties of MLEs

1. Minimum variance unbiasedness

   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization

   - If est of α is $\hat{\alpha}_{\mathrm{ML}}$, then est of $\beta = f(\alpha)$ is $\hat{\beta}_{\mathrm{ML}} = f(\hat{\alpha}_{\mathrm{ML}})$
   - But $\mathbb{E}(f(\alpha)) \neq f(\mathbb{E}(\alpha))$ for most $f(\cdot)$,
     so choice of parameterization can cause bias in small samples (e.g., for σ^2)
   - Not a problem in the limit (see below)

3. Invariance to sampling plans

Statistical properties of MLEs

1. Minimum variance unbiasedness

   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization

   - If est of α is $\hat{\alpha}_{\mathrm{ML}}$, then est of $\beta = f(\alpha)$ is $\hat{\beta}_{\mathrm{ML}} = f(\hat{\alpha}_{\mathrm{ML}})$
   - But $\mathbb{E}(f(\alpha)) \neq f(\mathbb{E}(\alpha))$ for most $f(\cdot)$,
     so choice of parameterization can cause bias in small samples (e.g., for σ^2)
   - Not a problem in the limit (see below)

3. Invariance to sampling plans

   - Estimate depends on data only through likelihood

Statistical properties of MLEs

1. Minimum variance unbiasedness

   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization

   - If est of α is $\hat{\alpha}_{\mathrm{ML}}$, then est of $\beta = f(\alpha)$ is $\hat{\beta}_{\mathrm{ML}} = f(\hat{\alpha}_{\mathrm{ML}})$
   - But $\mathbb{E}(f(\alpha)) \neq f(\mathbb{E}(\alpha))$ for most $f(\cdot)$,
     so choice of parameterization can cause bias in small samples (e.g., for σ^2)
   - Not a problem in the limit (see below)

3. Invariance to sampling plans

   - Estimate depends on data only through likelihood
   - Estimator same regardless of sample size, n

Statistical properties of MLEs

1. Minimum variance unbiasedness

   - MVU is the unbiased estimator with least variance (highest efficiency)
   - If an unbiased minimum variance estimator exists, it's the MLE
   - Even if no unbiased estimator, ML picks an efficient one

2. Invariance to reparameterization

   - If est of α is $\hat{\alpha}_{\mathrm{ML}}$, then est of $\beta = f(\alpha)$ is $\hat{\beta}_{\mathrm{ML}} = f(\hat{\alpha}_{\mathrm{ML}})$
   - But $\mathbb{E}(f(\alpha)) \neq f(\mathbb{E}(\alpha))$ for most $f(\cdot)$,
     so choice of parameterization can cause bias in small samples (e.g., for σ^2)
   - Not a problem in the limit (see below)

3. Invariance to sampling plans

   - Estimate depends on data only through likelihood
   - Estimator same regardless of sample size, n
   - You can stop sampling anytime you are pleased with precision

Statistical properties of MLEs

Asymptotic properties:

Statistical properties of MLEs

Asymptotic properties:

1. Consistency

Statistical properties of MLEs

Asymptotic properties:

1. Consistency
   - MLE collapses to a spike over true parameter values as $n \to \infty$

Statistical properties of MLEs

Asymptotic properties:

1. Consistency
   - MLE collapses to a spike over true parameter values as $n \to \infty$
2. Asymtoptic Normality

Statistical properties of MLEs

Asymptotic properties:

1. Consistency
   - MLE collapses to a spike over true parameter values as $n \to \infty$
2. Asymtoptic Normality
   - For large n, sampling distribution of $\hat{\theta}$ becomes Normally distributed

Statistical properties of MLEs

Asymptotic properties:

1. Consistency
   - MLE collapses to a spike over true parameter values as $n \to \infty$
2. Asymtoptic Normality
   - For large n, sampling distribution of $\hat{\theta}$ becomes Normally distributed
   - Allows for easy caculation of standard errors, confidence intervals, etc

Statistical properties of MLEs

Asymptotic properties:

1. Consistency
   - MLE collapses to a spike over true parameter values as $n \to \infty$
2. Asymtoptic Normality
   - For large n, sampling distribution of $\hat{\theta}$ becomes Normally distributed
   - Allows for easy caculation of standard errors, confidence intervals, etc
3. Asymptotic Efficiency
   - As $n \to \infty$, the MLE tends to be the estimator with lowest error

Two kinds of precision

Using gradient search, we can find the maximum of the likelihood function to whatever level of precision (*number of computed digits*) we desire

But can we trust the maximum of the likelihood to be a good summary of the true parameter value?

This is a different kind of precision

How precise – in the sense of being *certain* to be correct – is our estimate of the population parameter?

Two kinds of precision

Using gradient search, we can find the maximum of the likelihood function to whatever level of precision (*number of computed digits*) we desire

But can we trust the maximum of the likelihood to be a good summary of the true parameter value?

This is a different kind of precision

How precise – in the sense of being *certain* to be correct – is our estimate of the population parameter?

Bayesian inference would answer this with a probability interval: 95% subjective probability the true parameter lies in $[\text{lower}, \text{upper}]$

But we'd need the posterior $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$ to compute this

In likelihood inference we don't attempt to estimate the posterior, just the likelihood

Precision of maximum likelihood estimators

We've given up (for now) on calculating $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

Precision of maximum likelihood estimators

We've given up (for now) on calculating $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

But we'd still like some idea of how certain are estimates are to be (approx) right

Precision of maximum likelihood estimators

We've given up (for now) on calculating $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

But we'd still like some idea of how certain are estimates are to be (approx) right

For example, measures of uncertainty we'd like to have include

- standard errors of $\hat{\boldsymbol{\theta}}$

Precision of maximum likelihood estimators

We've given up (for now) on calculating $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

But we'd still like some idea of how certain are estimates are to be (approx) right

For example, measures of uncertainty we'd like to have include

- standard errors of $\hat{\boldsymbol{\theta}}$
- confidence intervals around $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$, etc.

Precision of maximum likelihood estimators

We've given up (for now) on calculating $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y})$

But we'd still like some idea of how certain are estimates are to be (approx) right

For example, measures of uncertainty we'd like to have include

- standard errors of $\hat{\boldsymbol{\theta}}$
- confidence intervals around $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$, etc.

In general, these result from a description of the likelihood surface

Intuitively, the more $\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})$ looks like a tall peak around $\hat{\boldsymbol{\theta}}$,
the more certain we are about $\hat{\boldsymbol{\theta}}$ being right

The more $\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})$ is spread out, the less certain we are about $\hat{\boldsymbol{\theta}}$

MLE standard errors: Normal approximation

In the linear regression case & other cases asymtoptically,
we can summarize the curvature in $\mathcal{L}(\mu, \sigma^2|\mathbf{y})$ around the MLE as:

$$\log \mathcal{L}(\mu, \sigma^2|\mathbf{y}) \quad = \quad -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

MLE standard errors: Normal approximation

In the linear regression case & other cases asymtoptically,
we can summarize the curvature in $\mathcal{L}(\mu, \sigma^2|\mathbf{y})$ around the MLE as:

$$\begin{aligned}
\log \mathcal{L}(\mu, \sigma^2|\mathbf{y}) &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i^2 - 2y_i\mu + \mu^2 \right)
\end{aligned}$$

MLE standard errors: Normal approximation

In the linear regression case & other cases asymtoptically,
we can summarize the curvature in $\mathcal{L}(\mu, \sigma^2|\mathbf{y})$ around the MLE as:

$$\begin{aligned}
\log \mathcal{L}(\mu, \sigma^2|\mathbf{y}) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\mu + \mu^2\right) \\
&= -\frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}\mu - \frac{n}{2\sigma^2}\mu^2
\end{aligned}$$

MLE standard errors: Normal approximation

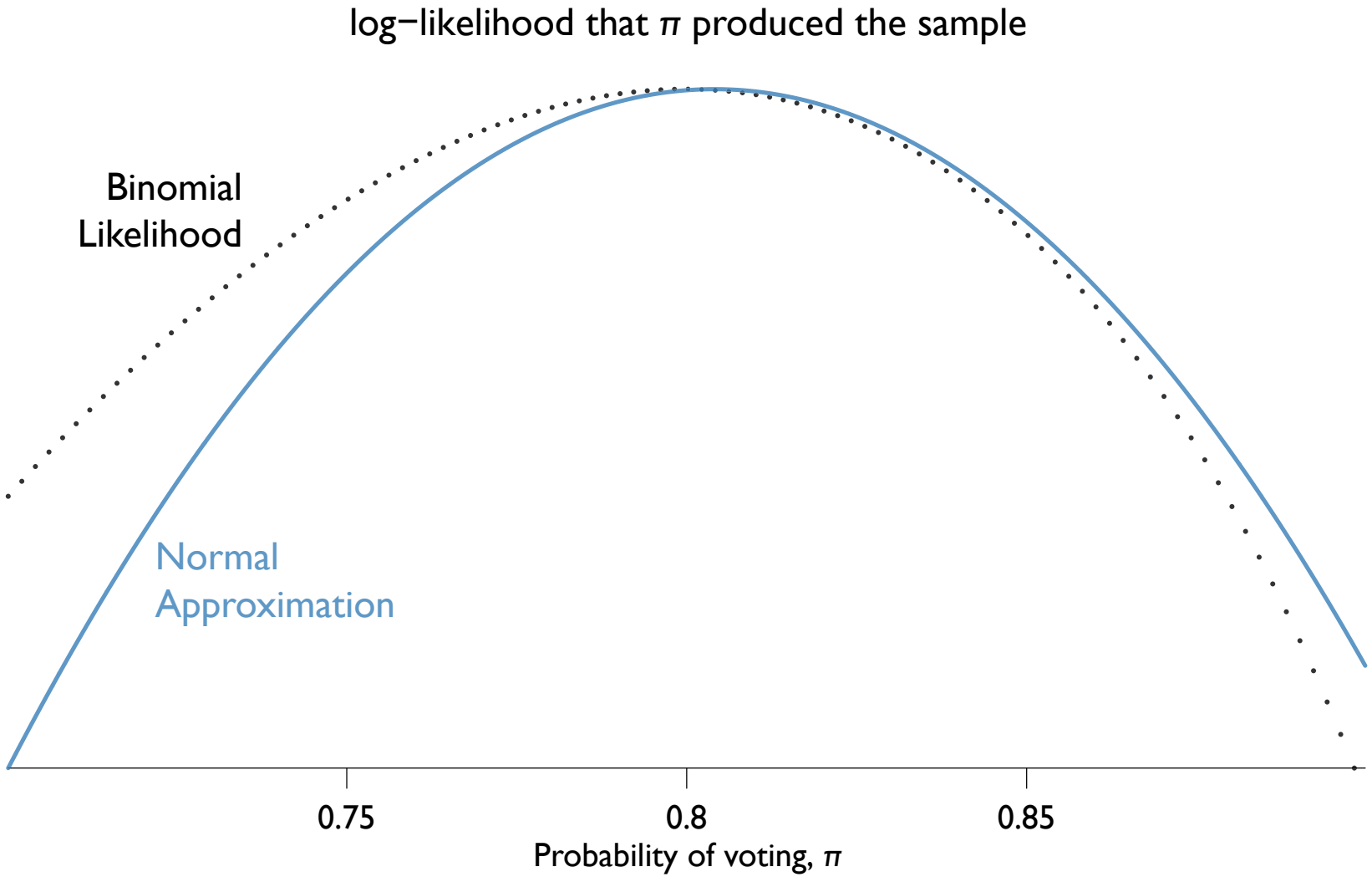
In the linear regression case & other cases asymtoptically,
we can summarize the curvature in $\mathcal{L}(\mu, \sigma^2|\mathbf{y})$ around the MLE as:

$$\begin{aligned}
\log \mathcal{L}(\mu, \sigma^2|\mathbf{y}) &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i^2 - 2y_i\mu + \mu^2 \right) \\
&= -\frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}\mu - \frac{n}{2\sigma^2}\mu^2
\end{aligned}$$

The key to curvature is the coefficient of μ^2, which is $-\frac{n}{2\sigma^2}$

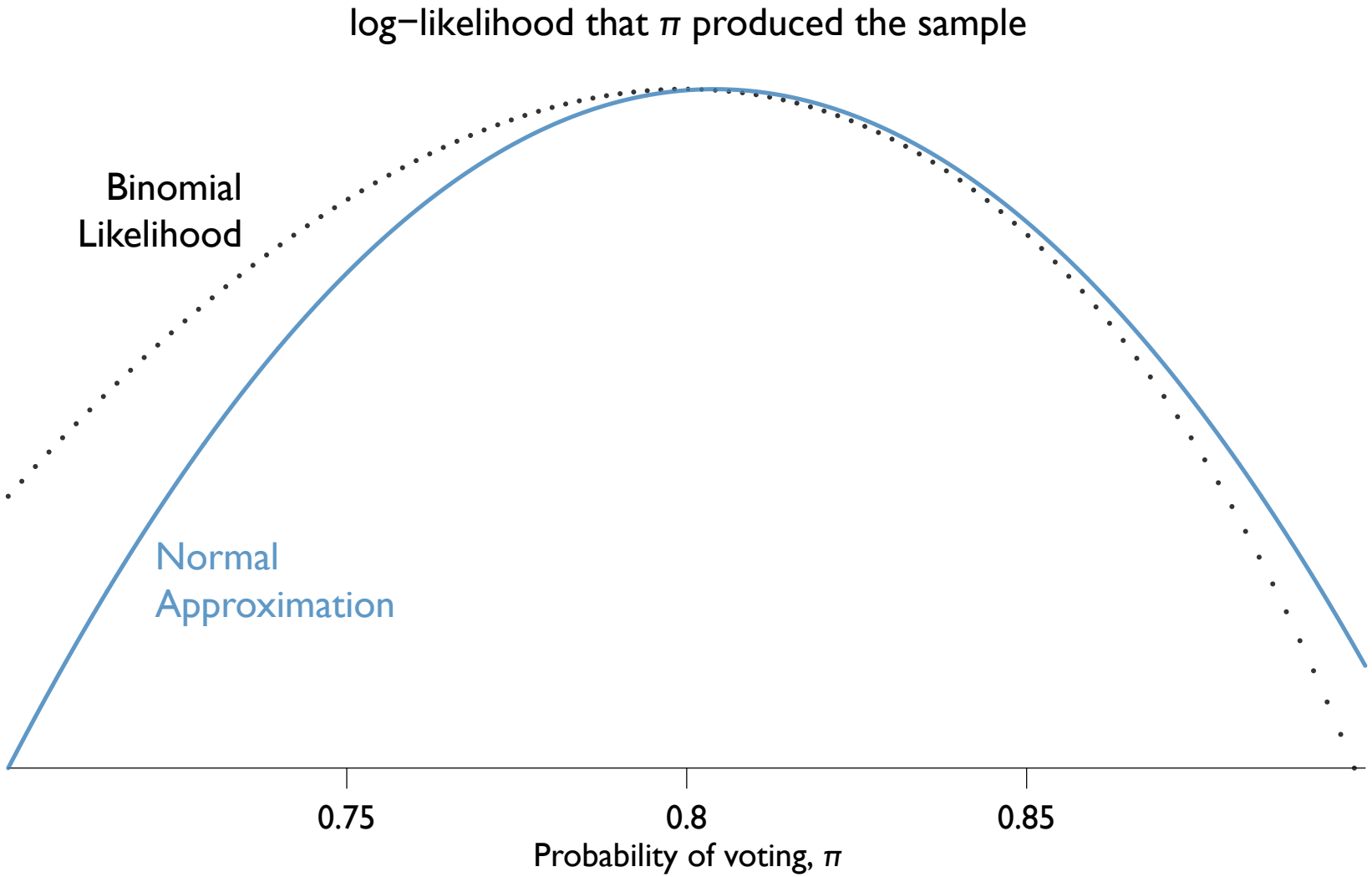
This implies a concave parabola descending faster as

n gets larger *and* σ^2 gets smaller



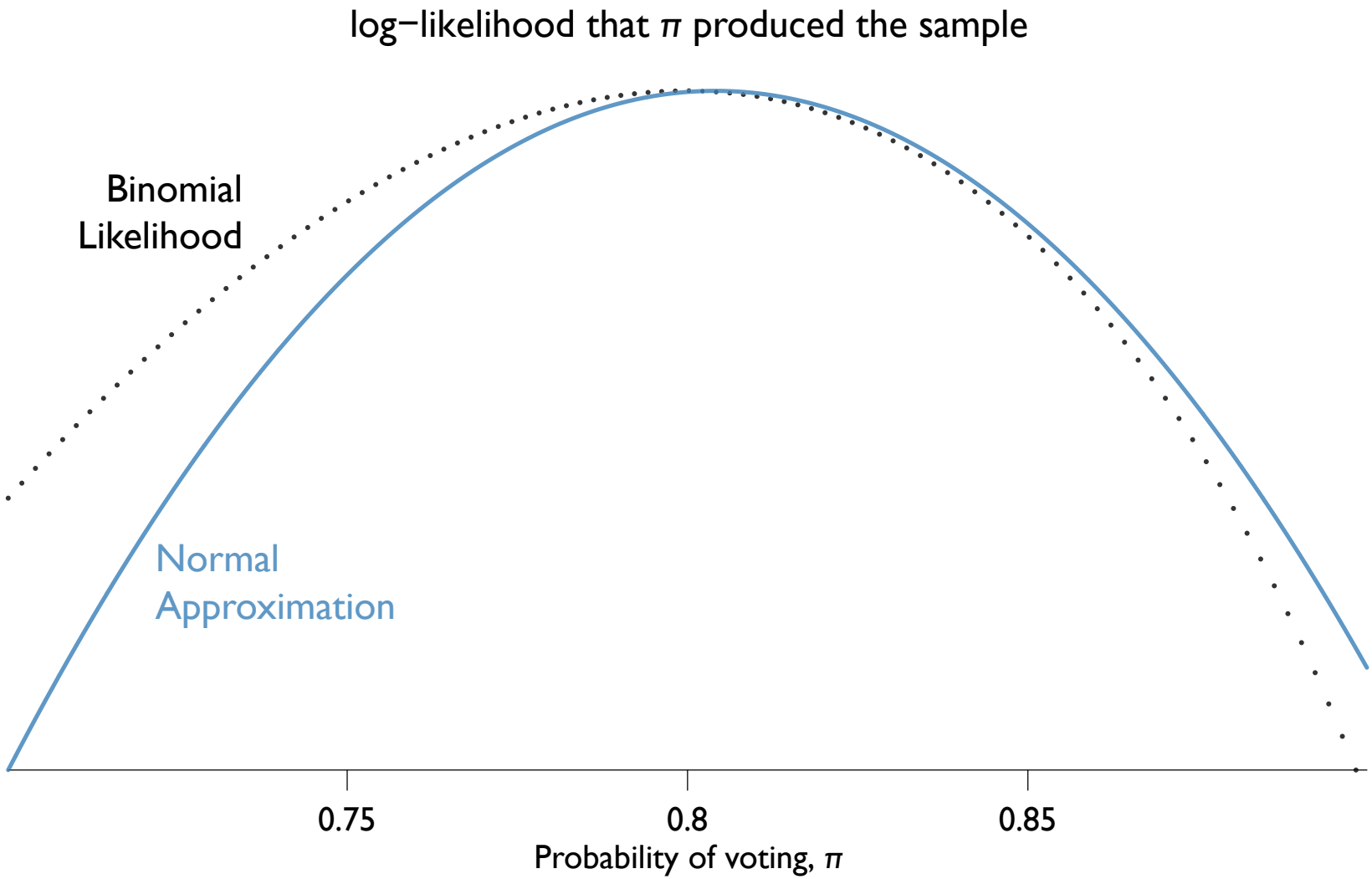
To see how a non-Normal likelihoods approximate the Normal, we return to the Binomial turnout example

The approximation of the Normal to the Binomial is not perfect



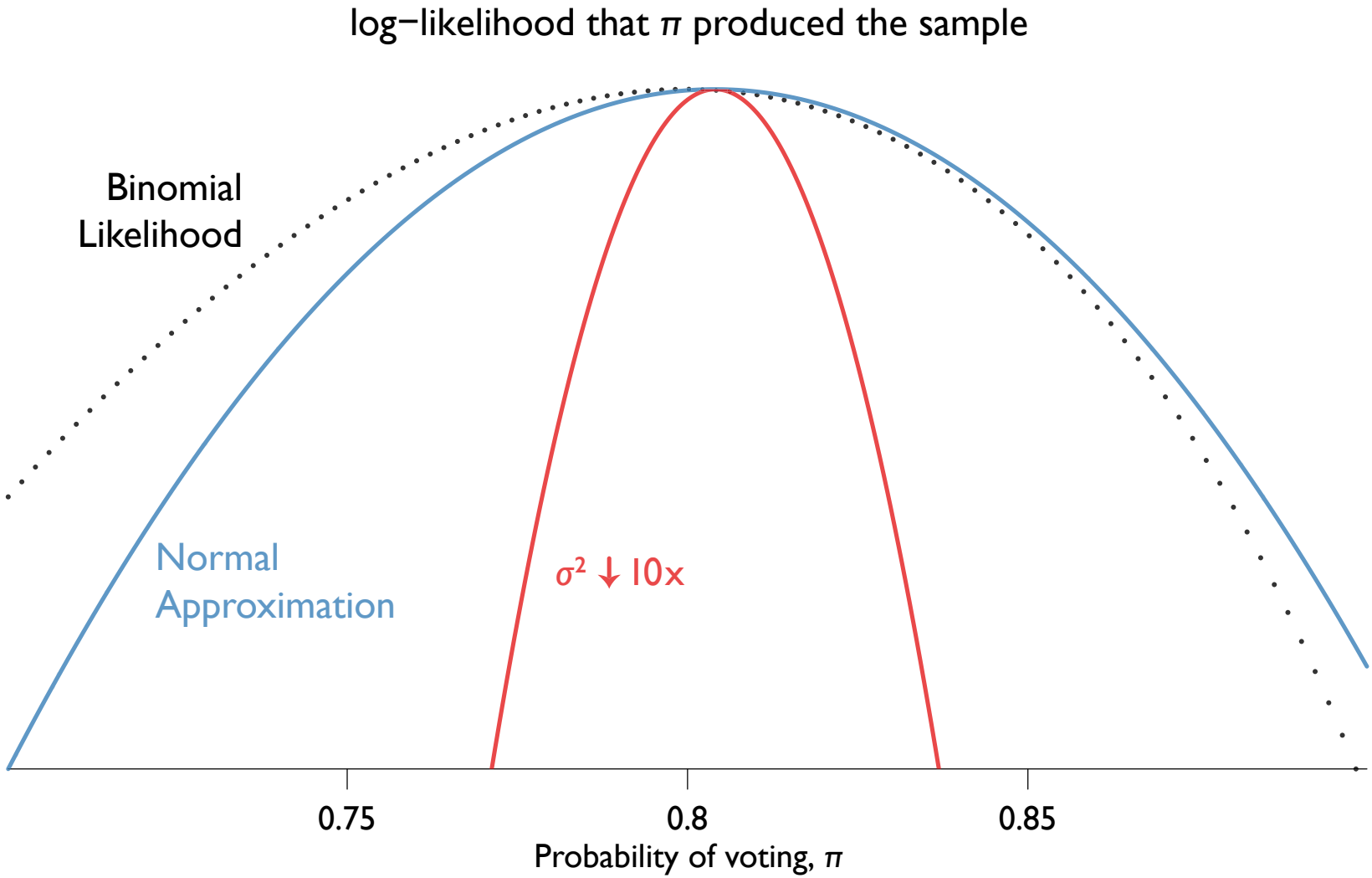
Two problems with the Normal approximation:

1. $\hat{\pi}_{\text{Normal}}$ is slightly high (peak of likelihood is too far right)
2. $\text{se}(\hat{\pi}_{\text{Normal}})$ is overconfident (curvature of likelihood is too steep)



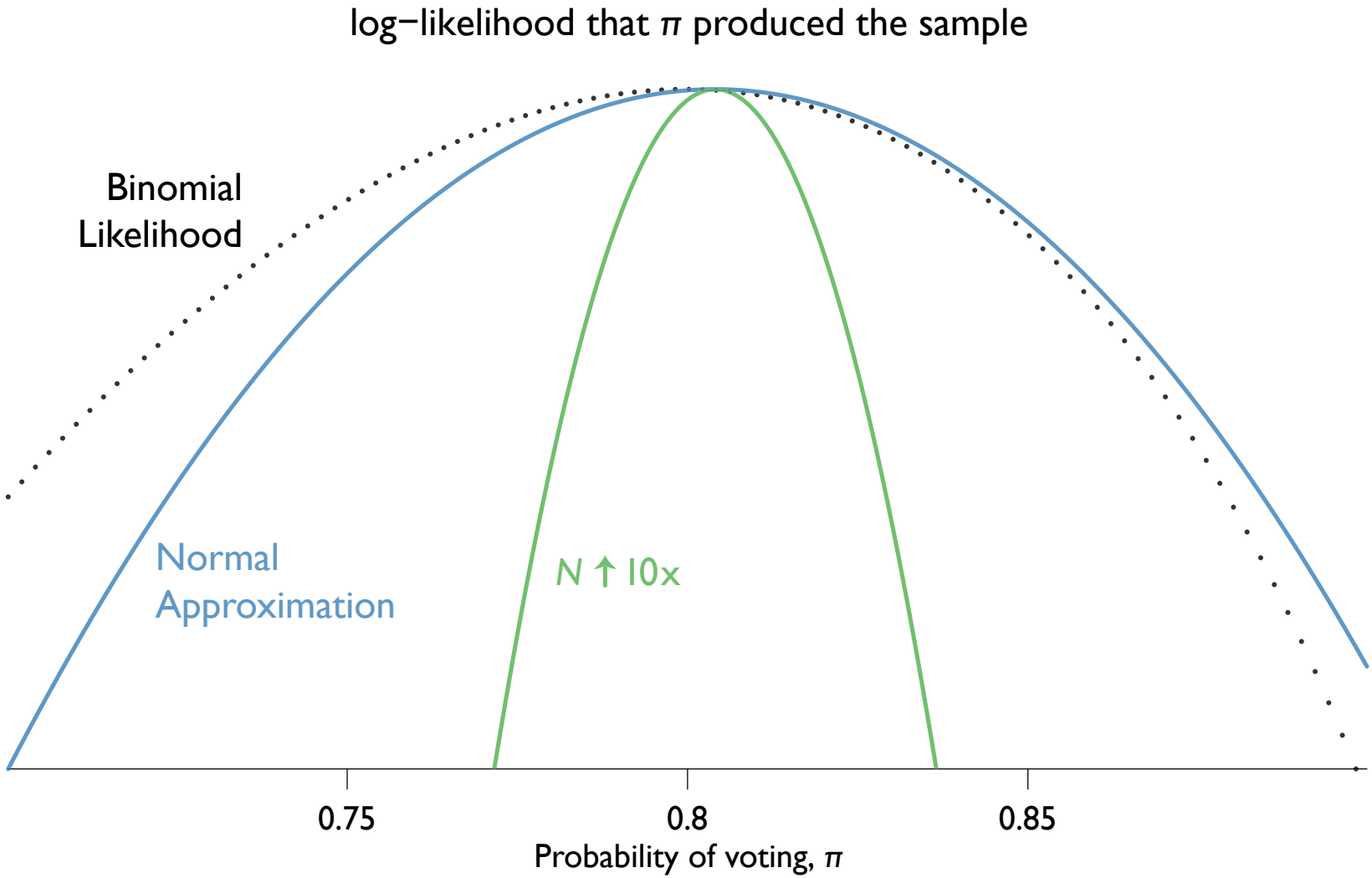
We wouldn't want to use the Normal in place of the Binomial here

But the parameterization of the Normal helps reveal
how the likelihood gets sharper when there is more information in the data



Suppose we shrunk the variance of the outcome by 10 times, so that each observation had more signal and less noise

The likelihood gets much steeper (and se's much smaller)



Suppose instead we had 10x more observations with the original variance

Same benefit – steeper curve; smaller standard errors

Can we formalize this for the general case?

The score vector and the Fisher information matrix

Recall that at the MLE, we expect the derivative of the likelihood with respect to each parameter θ (the score) to be zero:

$$\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Now calculate the variance of the score, recalling it has expectation 0, and assuming $\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is twice differentiable with respect to $\boldsymbol{\theta}$:

$$\operatorname{var}\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right) = \mathbb{E}\left(\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} - \mathbb{E}\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right)\right)^2\right)$$

The score vector and the Fisher information matrix

Recall that at the MLE, we expect the derivative of the likelihood with respect to each parameter $\boldsymbol{\theta}$ (the score) to be zero:

$$\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Now calculate the variance of the score, recalling it has expectation 0, and assuming $\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is twice differentiable with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\operatorname{var}\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right) &= \mathbb{E}\left(\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} - \mathbb{E}\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right)\right)^2\right) \\
&= \mathbb{E}\left(\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} - 0\right)^2\right)
\end{aligned}$$

The score vector and the Fisher information matrix

Recall that at the MLE, we expect the derivative of the likelihood with respect to each parameter $\boldsymbol{\theta}$ (the score) to be zero:

$$\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Now calculate the variance of the score, recalling it has expectation 0, and assuming $\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is twice differentiable with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\operatorname{var}\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right) &= \mathbb{E}\left(\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} - \mathbb{E}\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right)\right)^2\right) \\
&= \mathbb{E}\left(\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}} - 0\right)^2\right) \\
&= \mathbb{E}\left(\left(\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta}}\right)^2\right) = -\mathbb{E}\left(\frac{\partial^2 \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)
\end{aligned}$$

yielding a $k \times k$ matrix known as the *Fisher information*, $\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y})$

MLE standard errors

Define the *Fisher information* of the likelihood as

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

i.e., a $k \times k$ matrix of 2nd derivatives of the likelihood with respect to the k parameters $\boldsymbol{\theta}$

MLE standard errors

Define the *Fisher information* of the likelihood as

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

i.e., a $k \times k$ matrix of 2nd derivatives of the likelihood
with respect to the k parameters $\boldsymbol{\theta}$

Asymptotically, the variance covariance matrix is related to the information:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\left[\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}$$

MLE standard errors

Define the *Fisher information* of the likelihood as

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

i.e., a $k \times k$ matrix of 2nd derivatives of the likelihood
with respect to the k parameters $\boldsymbol{\theta}$

Asymptotically, the variance covariance matrix is related to the information:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\left[\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}$$

MLE standard errors are given by the square roots of the diagonal
of the inverse of the matrix of second derivatives (the Hessian matrix)

MLE standard errors

Define the *Fisher information* of the likelihood as

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

i.e., a $k \times k$ matrix of 2nd derivatives of the likelihood
with respect to the k parameters $\boldsymbol{\theta}$

Asymptotically, the variance covariance matrix is related to the information:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\left[\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}$$

MLE standard errors are given by the square roots of the diagonal
of the inverse of the matrix of second derivatives (the Hessian matrix)

Note we can only invert the Hessian if it is positive definite!

guaranteed in theory

may fail computationally for complex or low-information likelihoods

MLE standard errors: Turnout Example

Because there's just one parameter ($\hat{\pi}_{\text{MLE}} = 0.7998$),
the information matrix for the turnout example is a 1×1:

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = 21911001$$

MLE standard errors: Turnout Example

Because there's just one parameter ($\hat{\pi}_{\text{MLE}} = 0.7998$),
the information matrix for the turnout example is a 1×1:

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = 21911001$$

The variance-covariance matrix is also 1×1

$$\text{Var}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\left[\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} = 4.56 \times 10^{-8}$$

MLE standard errors: Turnout Example

Because there's just one parameter ($\hat{\pi}_{\text{MLE}} = 0.7998$),
the information matrix for the turnout example is a 1×1:

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = 21911001$$

The variance-covariance matrix is also 1×1

$$\text{Var}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = -\left[\frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} = 4.56 \times 10^{-8}$$

The square root gives the standard error of $\hat{\pi}_{\text{MLE}} = 0.0002136$

Assuming $\hat{\pi}$ is asymptotically normal implies a 95% CI of $[0.7994, 0.8002]$

MLE standard errors: Heteroskedasticity Example

With four parameters in the heteroskedasticity example $(\beta_0, \beta_1, \gamma_0, \gamma_1)$, the information matrix is a symmetric 4×4

$$\mathcal{I}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = \begin{bmatrix} 179.3 & 48.66 & 0.0001460 & 0.00004229 \\ 48.66 & 22.85 & 0.00004206 & 0.3980 \\ 0.0001460 & 0.00004206 & 750.0 & 377.7 \\ 0.00004229 & 0.3980 & 377.7 & 254.0 \end{bmatrix}$$

The variance-covariance matrix is also symmetric 4×4

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = \begin{bmatrix} 0.01320 & -0.02811 & -0.00008838 & 0.0001755 \\ -0.02811 & 0.1036 & 0.0003258 & -0.0006469 \\ -0.00008838 & 0.0003258 & 0.005313 & -0.007902 \\ 0.0001755 & -0.0006469 & -0.007902 & 0.01569 \end{bmatrix}$$

The standard errors of our parameter estimates are the square roots of the diagonal

So for example, $\hat{\gamma}_1 = 3.247$ and $\mathrm{se}(\gamma_1) = 0.1253$, for a 95% CI of $[3.007, 3.497]$

Practically speaking. . .

So how do we do this in R?

Practically speaking. . .

So how do we do this in `R`?

Overview:

1. Use a generic optimizer to maximize $\log \mathcal{L}$. `optim()` is good

Practically speaking. . .

So how do we do this in R?

Overview:

1. Use a generic optimizer to maximize $\log \mathcal{L}$. `optim()` is good

2. Get from `optim()` the value of $\log \mathcal{L}$ at its maximum,
   the corresponding parameter point estimates,
   and the variance-covariance matrix of the parameter estimates

Practically speaking. . .

So how do we do this in R?

Overview:

1. Use a generic optimizer to maximize $\log \mathcal{L}$. `optim()` is good

2. Get from `optim()` the value of $\log \mathcal{L}$ at its maximum,
   the corresponding parameter point estimates,
   and the variance-covariance matrix of the parameter estimates

3. Construct any desired summary, statistic, or goodness of fit test from these

An example: MLE for normal data

We're going to write `R` code to estimate the Normal linear model by ML

This code, with small changes, will suffice for many other MLEs

Where to start?

It helps when writing a program to first think through what the program needs to accomplish, in the order it needs to be done.

Writing out a plain English version of the algorithm is called "pseudo-code."

An example: MLE for normal data

Pseudo-code for Normal MLE:

1. Load needed libraries
2. Create an artificial dataset (with known properties)
3. Fit the data with LS
4. Fit the data with ML
5. Simulate quantities of interest (Qol's) from the MLE
6. Plot the simulated Qol's with confidence intervals