Maximum Likelihood and MAP Estimation

or

Doing Easy Things the Hard Way and Vice Versa

Overview

You've used MLE and MAP estimators, even if you didn't call them that. Today we will…

- … see some simple modeling examples in greater depth
- … tie together lots of concepts
- … see a method for building custom models

Part I:
Maximum Likelihood Estimation

Coin Flips

- Suppose three coin flips show (H,T,H)
- What does this tell us about the coin?

Setting up the Math

- This is the canonical example of a *binomial* random variable

- We'll stick to more generic variable names:

$$\begin{cases} n & = \text{Number of flips} \\ \theta & = P(\text{heads}) \\ x & = \text{Number of heads in } n \text{ flips} \end{cases}$$

- What do we know about $P(x|\theta)$?

Likelihood

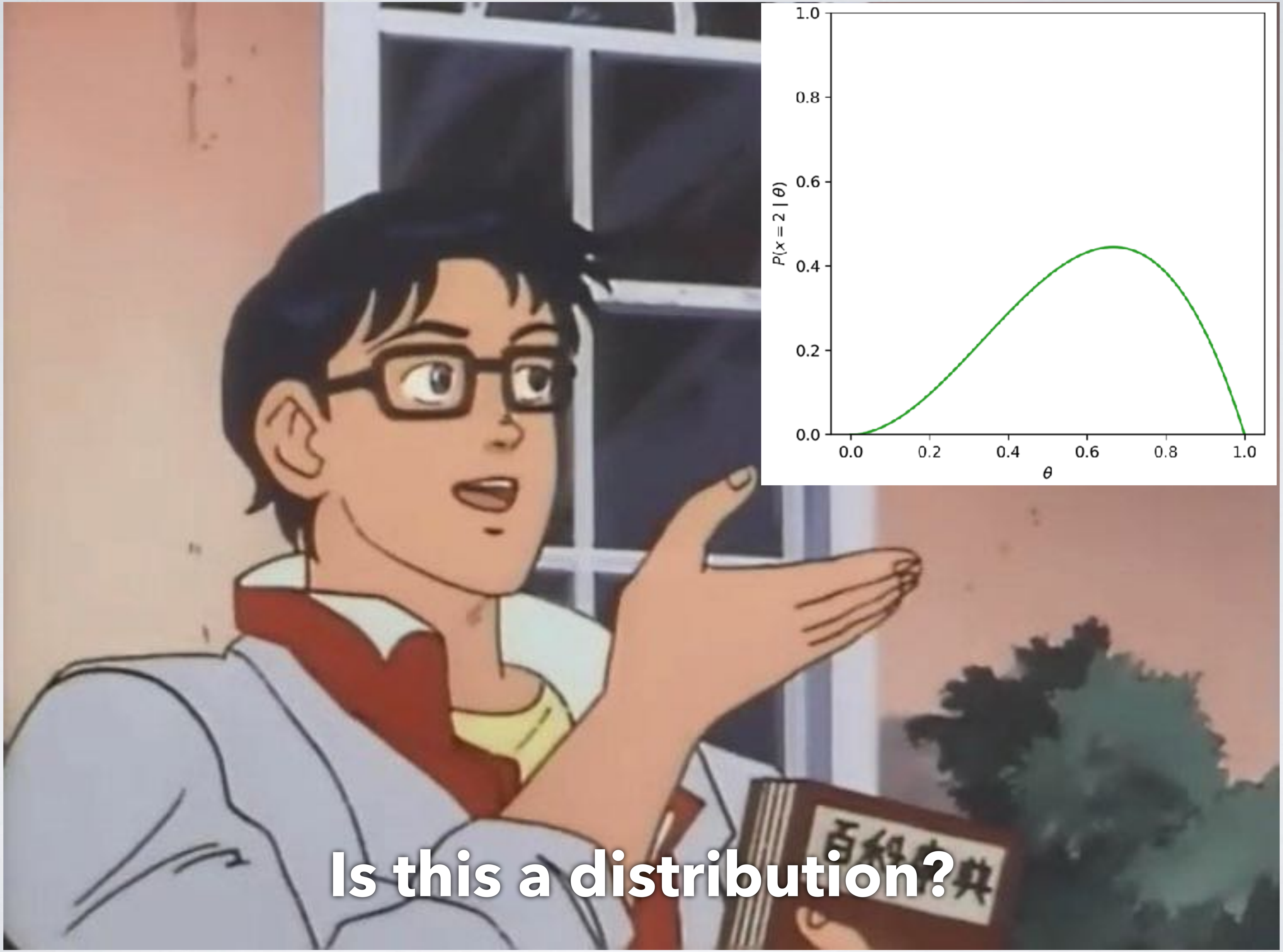
- $P(x|\theta)$ is called the *likelihood*

- We'll have fixed data and variable parameters

- New notation to emphasize this

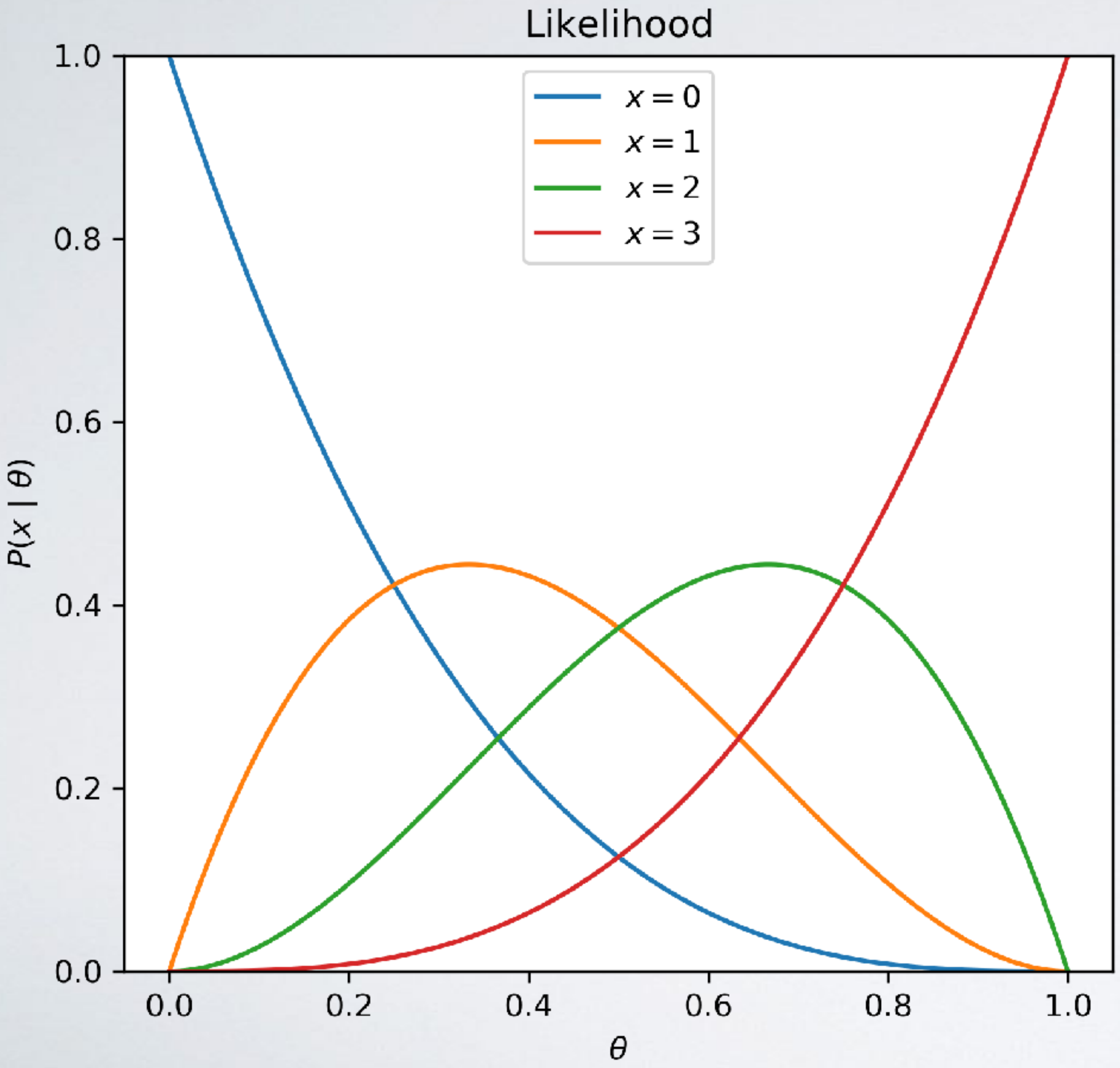
$$L(\theta|x) \equiv P(x|\theta)$$

- Which brings us to the first big question…



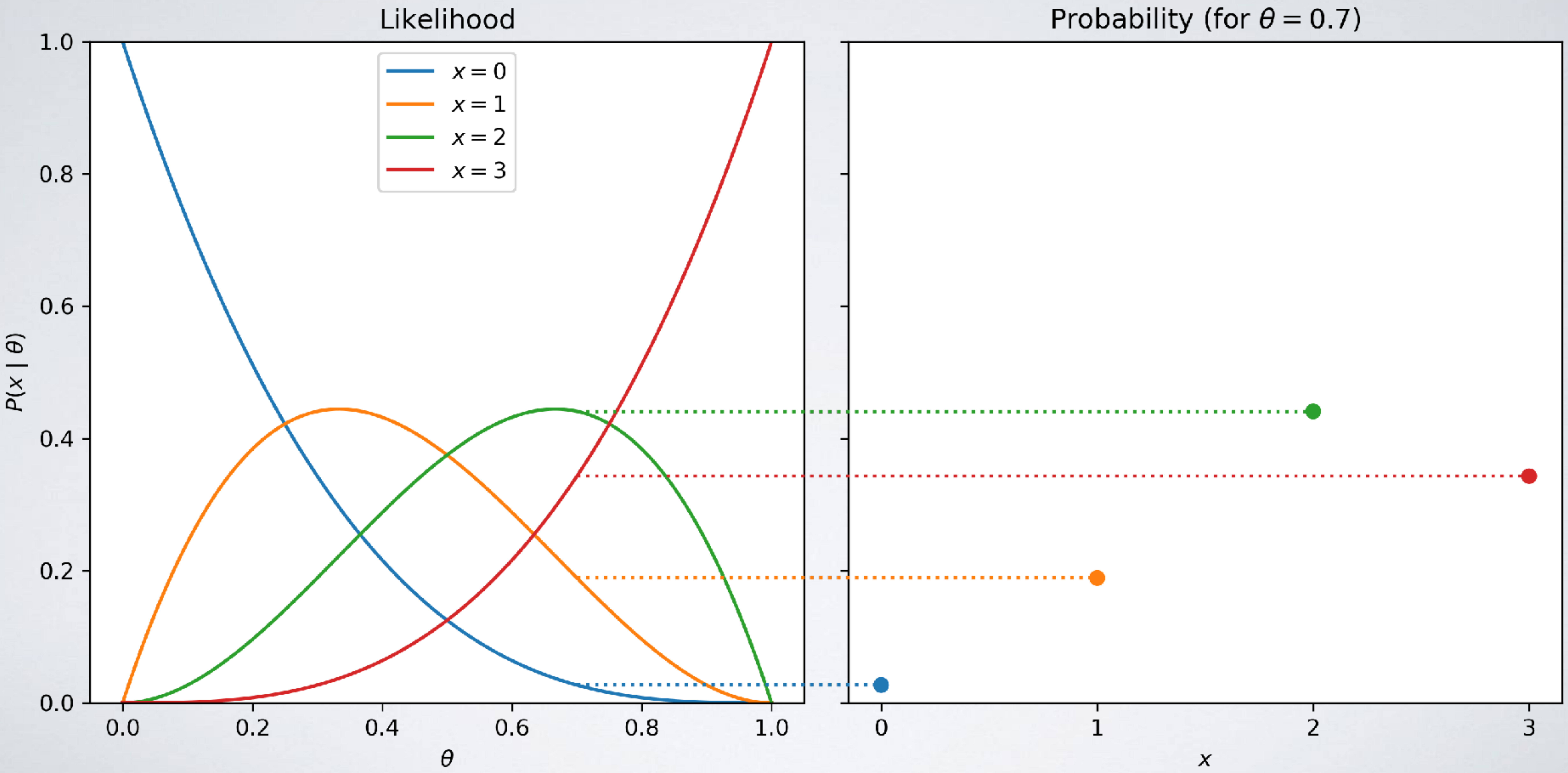
1.0
0.8
0.6
0.4
0.2
0.0
$P(x = 2 \mid \theta)$
0.0
0.2
0.4
0.6
0.8
1.0
θ
Is this a distribution?

Likelihood vs Probability



- None of these add to one!
- Is there a distribution hiding here somewhere?

Likelihood vs Probability



Some Questions

Some Questions

- How should we interpret the maximum of the likelihood?

Some Questions

- How should we interpret the maximum of the likelihood?
- The MLE is the parameter value for which the data has the highest probability

Some Questions

- How should we interpret the maximum of the likelihood?
- The MLE is the parameter value for which the data has the highest probability
- Can you think of real-world cases where this gives strange results?

Maximizing the Likelihood

- The likelihood for our coin problem is

$$L(\theta | x) = P(x | \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

- Maximizing a function usually involves working with the derivative, but products and exponentials are a mess.
- What can we do to make this easier?

The Log-Likelihood

- Maximizing L is the same as maximizing $\log L$

$$\begin{aligned}\ell(\theta \mid x) &= \log L(\theta \mid x) \\ &= \log \binom{n}{x} + x \log \theta + (n - x)\log(1 - \theta)\end{aligned}$$

- Much better! What's next?

Differentiate!

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)\right]$$

$$= 0 + \frac{x}{\theta} + \frac{n-x}{1-\theta}(-1)$$

- Solving $\frac{\partial \ell}{\partial \theta} = 0$ gives the maximum likelihood estimate, $\hat{\theta} = \frac{x}{n}$
- Stats trivia: $\frac{\partial \ell}{\partial \theta}$ is called the *score function*

More Questions

More Questions

- We found a nice formula for the MLE. Is that always possible?

More Questions

- We found a nice formula for the MLE. Is that always possible?
- No! What do we do then?

More Questions

- We found a nice formula for the MLE. Is that always possible?
- No! What do we do then?
- If there's no *closed-form solution*, we need an iterative, numeric method

Some Terminology

- An *estimate* is a parameter value
- An *estimator* is a function that returns an estimate
- *Estimation* is the process of finding or using an estimator

More Questions

More Questions

- What if there's more than one variable?

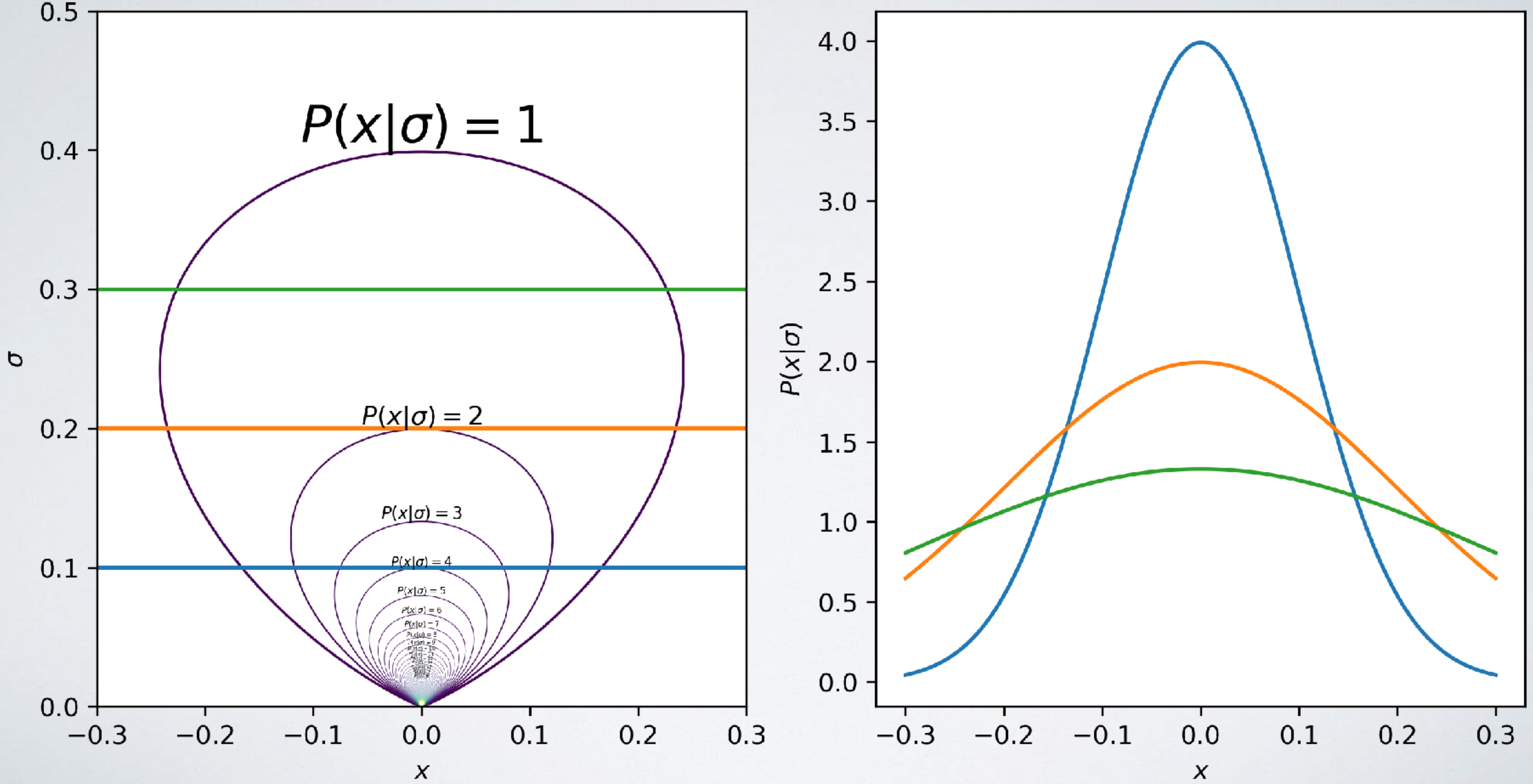
More Questions

- What if there's more than one variable?
- Instead of the partial derivative, we need to use the *gradient*

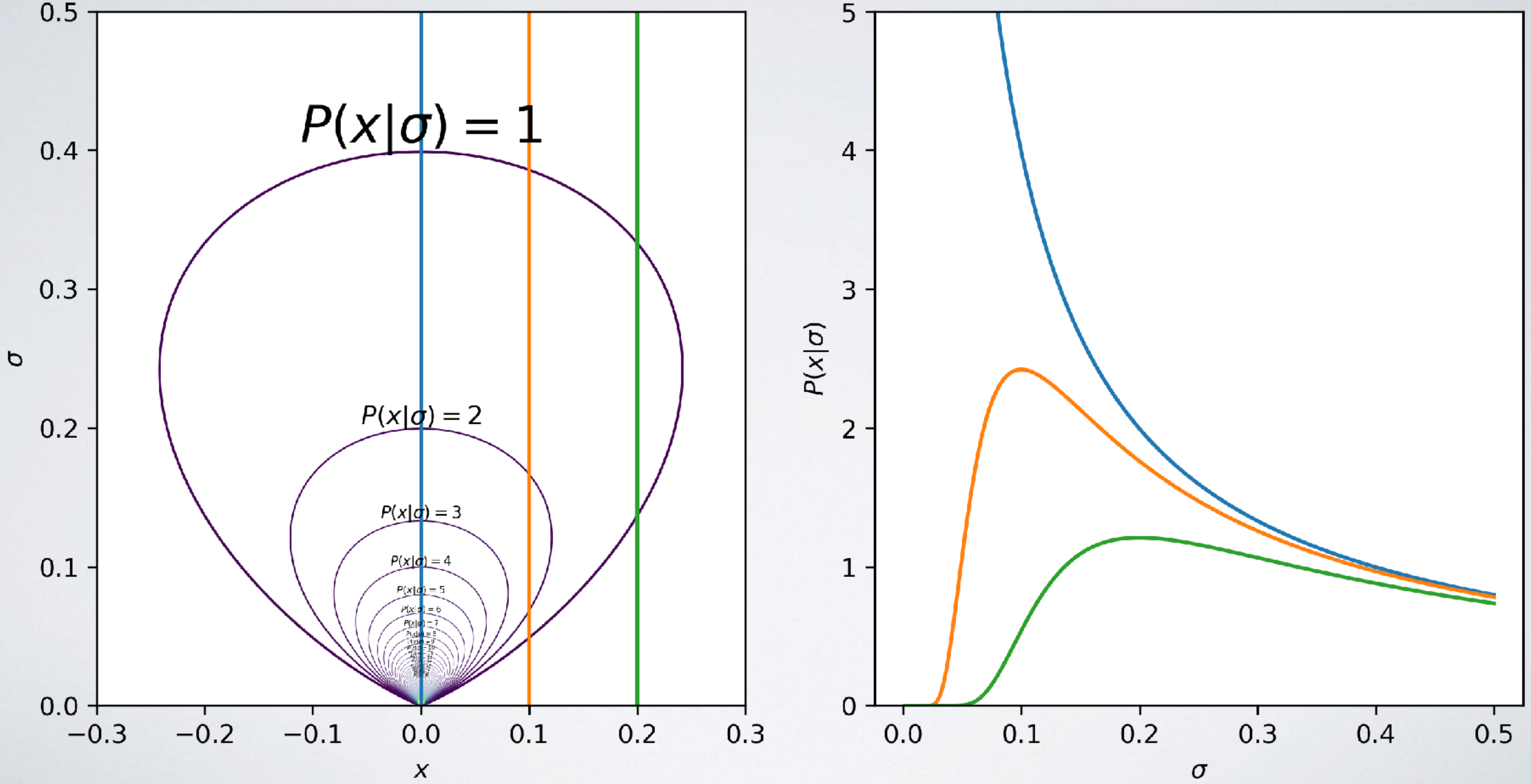
Another Example

- What if we only have one data point, $x \sim$ Normal$(0,\sigma)$
- If we plug any x and σ into $P(x\,|\,\sigma)$, we'll have a surface
- What will this look like?
- If we fix x or σ, what will the slices look like?

Probability Slices



Likelihood Slices



Origin of Least Squares (Simplified)

- Say we have independent $x_j \sim \text{Normal}(\mu, 1)$
- Then $$L = \prod_j \text{Normal}(x_j \mid \mu, 1) = \prod_j C_1 e^{-\frac{1}{2}(x_j - \mu)^2}$$
- So $$\ell = C_2 - \frac{1}{2} \sum_j (x_j - \mu)^2$$
- You'll also see references to -2ℓ, which in this case relates to the sum of squared residuals

Part II:
MAP Estimation

L_2 Redux

- We've seen connections between "sum of squares" and normal distributions
- L_2 regularization uses a sum of squares
- Is there something Gaussian about L_2?

Back to Bayes (~~icks~~)

- If we start with Bayes

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}$$

- And take the log, we get

$$\log P(\theta|x) = \log P(x|\theta) + \log P(\theta) - \log P(x)$$

- Do you see the connection?

Deconstructing L_2

- Remember the objective function for Ridge regression?

$$\hat{\beta} = \arg\min_\beta \left[||y - X\beta||^2 + \lambda ||\beta||^2 \right]$$

- Now we can see where these terms come from, since

$$\text{Log-likelihood} = \log P(y \mid X, \beta)$$

$$\text{Log-prior} = \log P(\beta \mid \lambda)$$

- The great thing about this is that **we can change either or both**!!

- What other examples have you seen?

MAP Estimation

- Choose a likelihood $P(y \mid \theta)$
- Choose a prior $P(\theta \mid \lambda)$ (hyperparameter λ optional)
- Find θ to **maximize** $\log P(y \mid \theta) + \log P(\theta \mid \lambda)$
- Cross-validate to tune λ

Final Thoughts

- For lots of models, inference is optimization
- Often, the objective function is a likelihood or posterior
- This approach can be used to build custom models specific to a given domain, or even to a particular data set