



10-315

Introduction to ML

Probabilistic Models:  
MAP

Instructor: Pat Virtue

# Course Feedback

## Going well

- Recitation
- Pre-reading
- Homework
- Lecture Polls

## Room for improvement

- Lecture slide ink handwriting 
- Connecting lecture to other components
- Lecture detail
  - Not enough
  - Too much

# Plan

## Today

- MLE
  - Linear Regression 
-  ■ Probability Motivation
- MAP
  -  ■ ML Applications of Bayes Rule
-  ■ Linear Regression

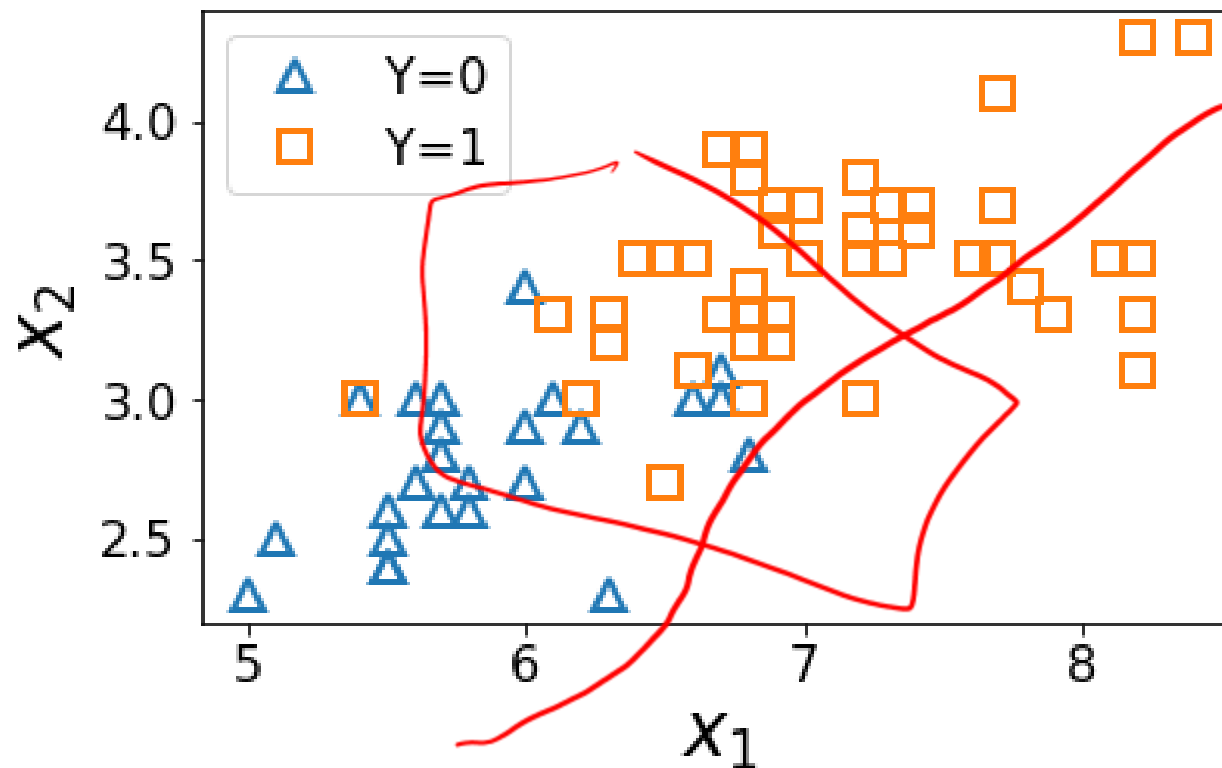
# Probability Motivation

# Empirical Risk Minimization vs MLE/MAP

We seem to be redoing a lot of work? ...well, we are. But there is a reason

# Why Probabilistic Models?

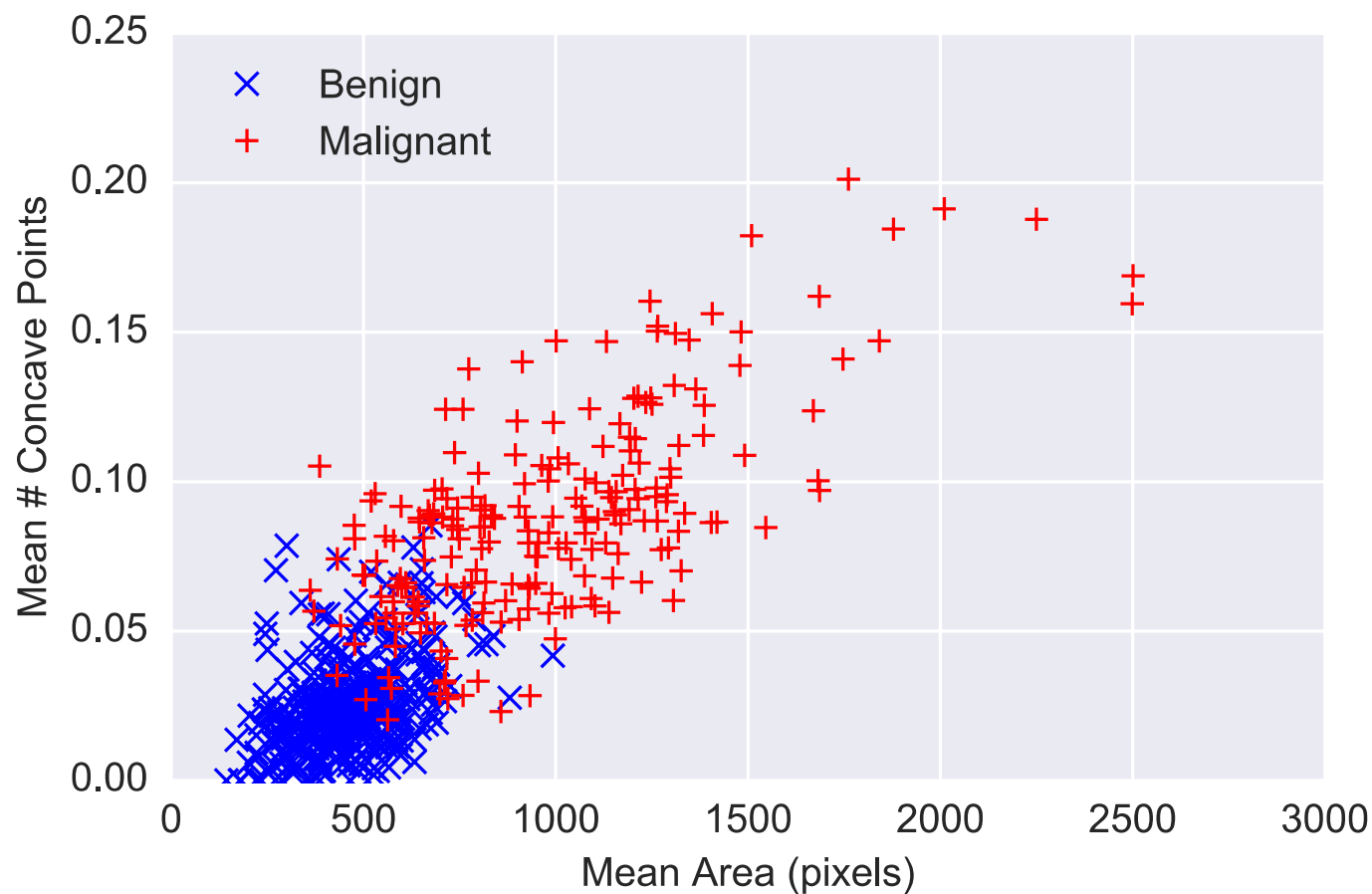
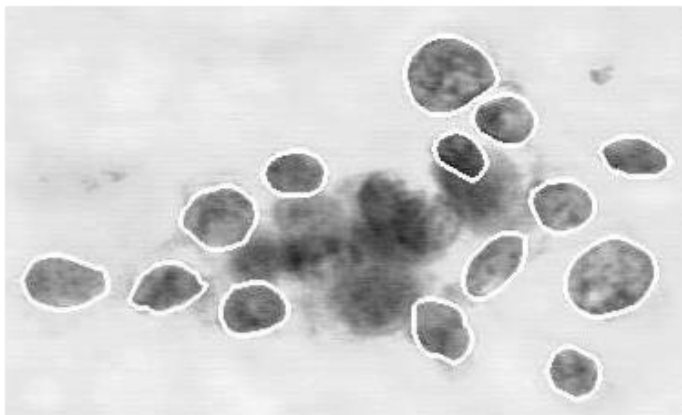
## Iris Data



$P(y|x, w, b)$   
cross-entropy

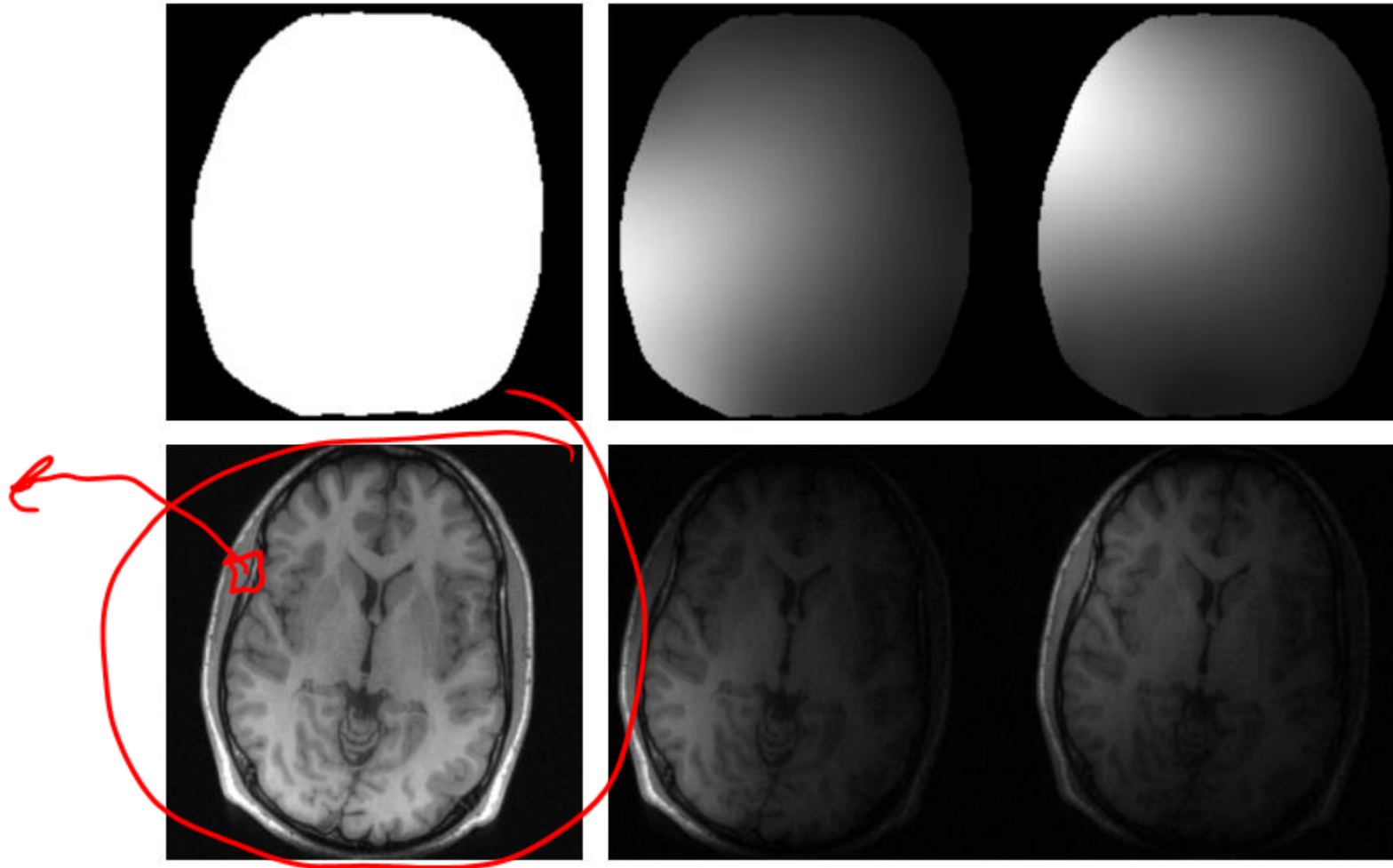
# Why Probabilistic Models?

## Breast Cancer Diagnosis



# Why Probabilistic Models?

## MRI Image Reconstruction



## Categorical Gaussian Generative Model

Estimating parameters

$$Y \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$$

$$\begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \\ y_3^{(1)} \end{bmatrix}$$

$$\frac{\partial l}{\partial \pi_i} = 0$$

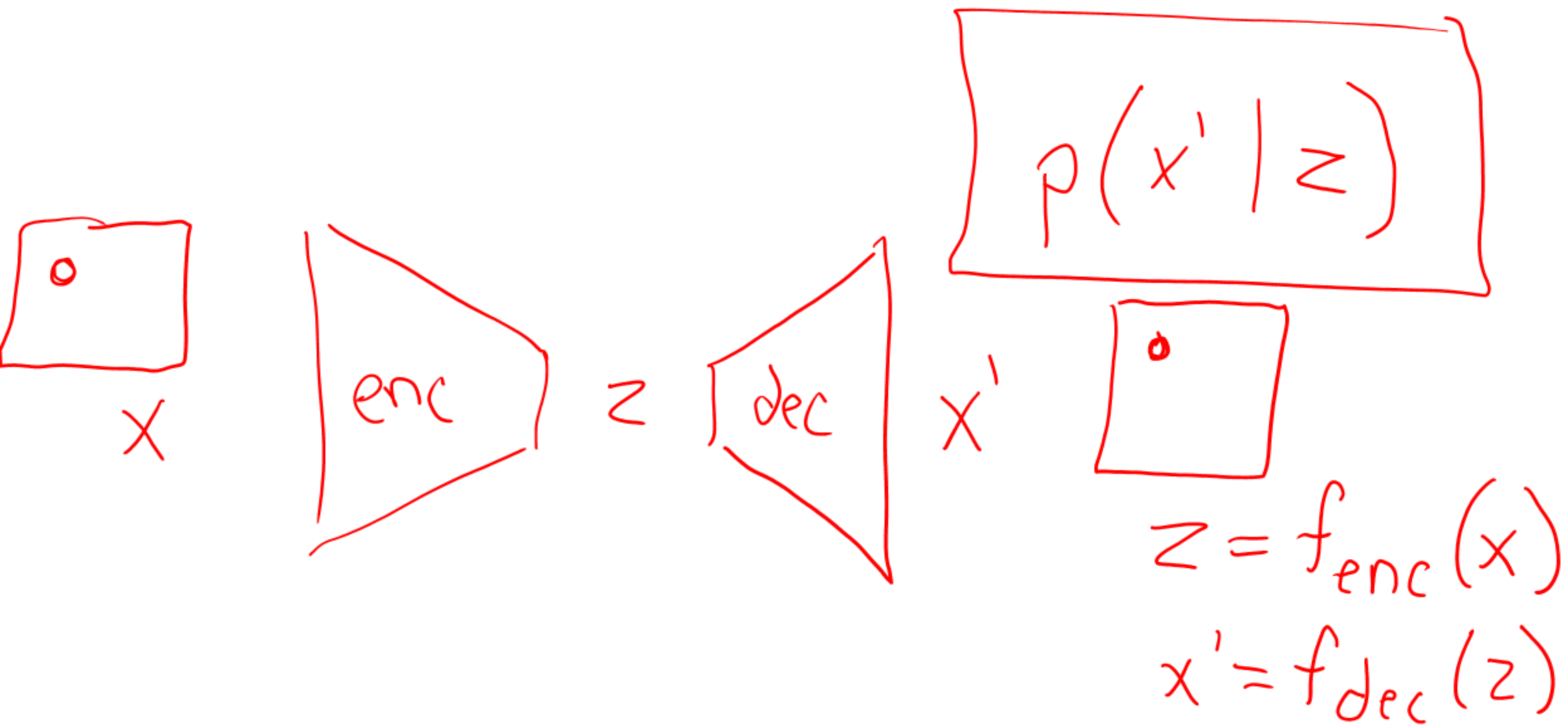
$$\frac{\partial l}{\partial \mu_3} = 0$$

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}$$

$$\begin{aligned} \hat{\theta}_{MLE} &= \underset{\theta}{\operatorname{argmax}} \quad p(\mathcal{D} | \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \prod_{i=1}^N p(x^{(i)}, y^{(i)} | \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \log \prod_{i=1}^N p(x^{(i)}, y^{(i)} | \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \sum_{i=1}^N \log p(x^{(i)}, y^{(i)} | \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \sum_{i=1}^N \log p(y^{(i)} | \theta) p(x^{(i)} | y^{(i)} | \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \sum_{i=1}^N \log \left( \prod_{k=1}^K \pi_k^{y_k^{(i)}} f_{\mathcal{N}}(x^{(i)} | \mu_k, \sigma_k^2)^{y_k^{(i)}} \right) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \sum_{i=1}^N \sum_{k=1}^K y_k^{(i)} \log (\pi_k f_{\mathcal{N}}(x^{(i)} | \mu_k, \sigma_k^2)) \\ &= \underset{\theta}{\operatorname{argmax}} \quad \underbrace{\sum_{i=1}^N \sum_{k=1}^K y_k^{(i)} \log (\pi_k f_{\mathcal{N}}(x^{(i)} | \mu_k, \sigma_k^2))}_{l(\theta | \mathcal{D})} \end{aligned}$$

$$f_{\mathcal{N}}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# Autoencoders



# Language Models

# ML Applications of Bayes Rule

# Bayes Rule

$$p(a \mid b) = \frac{p(b \mid a) p(a)}{p(b)}$$

$$p(b \mid a) = \frac{p(a \mid b) p(b)}{p(a)}$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) p(\theta)}{p(\mathcal{D})}$$

$$p(\mathcal{D} \mid \theta) = \frac{p(\theta \mid \mathcal{D}) p(\mathcal{D})}{p(\theta)}$$

# Poll 1

Which of these terms is the likelihood?

Select all that apply

Weird



A →

B →

C →

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) p(\theta)}{p(\mathcal{D})}$$

D →

E →

F →

G →

$$p(\mathcal{D} | \theta) = \frac{p(\theta | \mathcal{D}) p(\mathcal{D})}{p(\theta)}$$

H →

# Poll

Which of these terms is the likelihood?

Select all that apply

A B C

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) p(\theta)}{p(\mathcal{D})}$$

D

x: input  
y: output

$$p(x, y | \theta)$$

E F G

$$\underline{p(y | x)} = \frac{p(x | y) p(y)}{p(x)}$$

condit. likelihood  $p(y | x, \theta)$

H

# Bayes Rule

## Terminology

Posterior

Likelihood

Prior


$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) p(\theta)}{p(\mathcal{D})}$$

# Two Applications of Bayes Rule

$$p(a | b) = \frac{p(b | a) p(a)}{p(b)}$$

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) p(\theta)}{p(\mathcal{D})}$$

---

$$p(y | x) = \frac{p(x | y) p(y)}{p(x)}$$

later

MLE and MAP

## Poll 2

pdf exp. dist

Where do we plug in the pdf that we used for MLE, e.g.,  $f(x) = \lambda e^{-\lambda x}$ ?

$\theta = \lambda$

$\mathcal{D} = \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, \dots\}$

A

B

C

D

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) p(\theta)}{p(\mathcal{D})} = \frac{\prod_{i=1}^N p(x^{(i)} | \lambda) p(\lambda)}{\sum}$$

# MLE and MAP

Maximum likelihood estimation

$$\begin{aligned}\theta_{MLE} &= \operatorname{argmax}_{\theta} p(\mathcal{D} \mid \theta) \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(y^{(i)} \mid \theta)\end{aligned}$$

Maximum *a posteriori* estimation

$$\begin{aligned}\theta_{MAP} &= \operatorname{argmax}_{\theta} p(\theta \mid \mathcal{D}) \\ &= \operatorname{argmax}_{\theta} \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \\ &= \operatorname{argmax}_{\theta} \frac{\prod_{i=1}^N p(y^{(i)} \mid \theta)p(\theta)}{p(\mathcal{D})} \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(y^{(i)} \mid \theta) \underbrace{p(\theta)}_{\begin{bmatrix} p(w) \\ p(b) \end{bmatrix}}\end{aligned}$$

# Recipe for Estimation

## MLE

1. Formulate the **likelihood**,  $p(\mathcal{D} \mid \theta)$
2. Set objective  $J(\theta)$  equal to negative log of the **likelihood**  
$$J(\theta) = -\log p(\mathcal{D} \mid \theta)$$
3. Compute derivative of objective,  $\partial J / \partial \theta$
4. Find  $\hat{\theta}$ , either
  - a. Set derivate equal to zero and solve for  $\theta$
  - b. Use (stochastic) gradient descent to step towards better  $\theta$

# Recipe for Estimation

## MAP

1. Formulate the likelihood times the prior,  $p(\mathcal{D} | \theta)p(\theta)$
2. Set objective  $J(\theta)$  equal to negative log of the likelihood times the prior
$$J(\theta) = -\log[p(\mathcal{D} | \theta)p(\theta)]$$
3. Compute derivative of objective,  $\partial J / \partial \theta$
4. Find  $\hat{\theta}$ , either
  - a. Set derivate equal to zero and solve for  $\theta$
  - b. Use (stochastic) gradient descent to step towards better  $\theta$

# Coin Flipping Example

## Trick coin from pre-reading

Initially: no information about the coin, so we just default to a uniform belief about the Bernoulli parameter  $\phi$

| Invoice: Weighted Coins |                 | Customer: Torch Tricks, Inc.        |                             |
|-------------------------|-----------------|-------------------------------------|-----------------------------|
| <u>Item</u>             | <u>Quantity</u> | <u>Coin type, <math>\phi</math></u> | <u><math>p(\phi)</math></u> |
| 0% Heads Coin           | 40 / 200        | 0.0                                 | 0.20                        |
| 20% Heads Coin          | 50 / 200        | 0.2                                 | 0.25                        |
| 50% Heads Coin          | 80 / 200        | 0.5                                 | 0.40                        |
| 80% Heads Coin          | 10 / 200        | 0.8                                 | 0.05                        |
| 100% Heads Coin         | 20 / 200        | 1.0                                 | 0.10                        |
| Total: 200              |                 |                                     | ✓ 1.00                      |

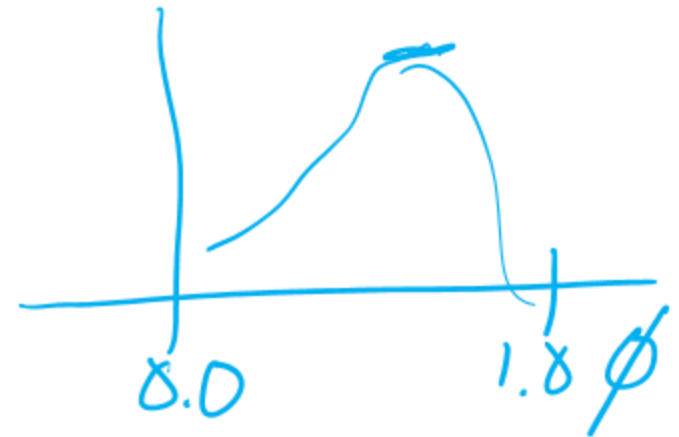
## Poll 3

As we collect more and more data (more coin flips), will the peak of the likelihood curve increase or decrease?

- A) Increase
- B) Decrease
- C) I have no idea

$$N \rightarrow \infty$$

$$p(\theta) \prod_{i=1}^N$$

# Coin Flipping Example

Trick coin from pre-reading

Suppose we discover information about the distribution of trick coin types?  
How can we use this information both before and after flipping coins?

| Invoice: Weighted Coins |                 | Customer: Torch Tricks, Inc.        |                             |
|-------------------------|-----------------|-------------------------------------|-----------------------------|
| <u>Item</u>             | <u>Quantity</u> | <u>Coin type, <math>\phi</math></u> | <u><math>p(\phi)</math></u> |
| 0% Heads Coin           | 40 / 200        | 0.0                                 | 0.20                        |
| 20% Heads Coin          | 50 / 200        | 0.2                                 | 0.25                        |
| 50% Heads Coin          | 80 / 200        | 0.5                                 | 0.40                        |
| 80% Heads Coin          | 10 / 200        | 0.8                                 | 0.05                        |
| 100% Heads Coin         | 20 / 200        | 1.0                                 | 0.10                        |
| Total: 200              |                 |                                     | ✓ 1.00                      |

# Coin Flipping Example

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} p(\theta) \prod_{i=1}^N p(y^{(i)} | \theta)$$

Trick coin from pre-reading

Suppose we discover information about the distribution of trick coin types?  
How can we use this information both before and after flipping coins?

| $\mathcal{D}$ | $p(\phi) \prod_{i=1}^N p(y^{(i)}   \phi)$ |
|---------------|-------------------------------------------|
| $\{\}$        | $p(\phi)$                                 |
| $\{H\}$       | $p(\phi) \phi$                            |
| $\{H, T\}$    | $p(\phi) \phi(1 - \phi)$                  |
| $\{H, T, T\}$ | $p(\phi) \phi(1 - \phi)(1 - \phi)$        |

## Poll 4

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta) p(\theta) \quad \text{posterior} \propto \text{likelihood} \cdot \text{prior}$$

$$p(\theta \mid \mathcal{D}) \propto \prod p(y^{(i)} \mid \theta) p(\theta)$$

As the number of data points increases, which of the following are true?

Select ALL that apply

- A. The MAP estimate approaches the MLE estimate
- B. The **posterior** distribution approaches the **prior** distribution
- C. The **likelihood** distribution approaches the **prior** distribution
- D. The **posterior** distribution approaches the **likelihood** distribution
- E. The **likelihood** has a lower impact on the **posterior**
- F. The **prior** has a lower impact on the **posterior**

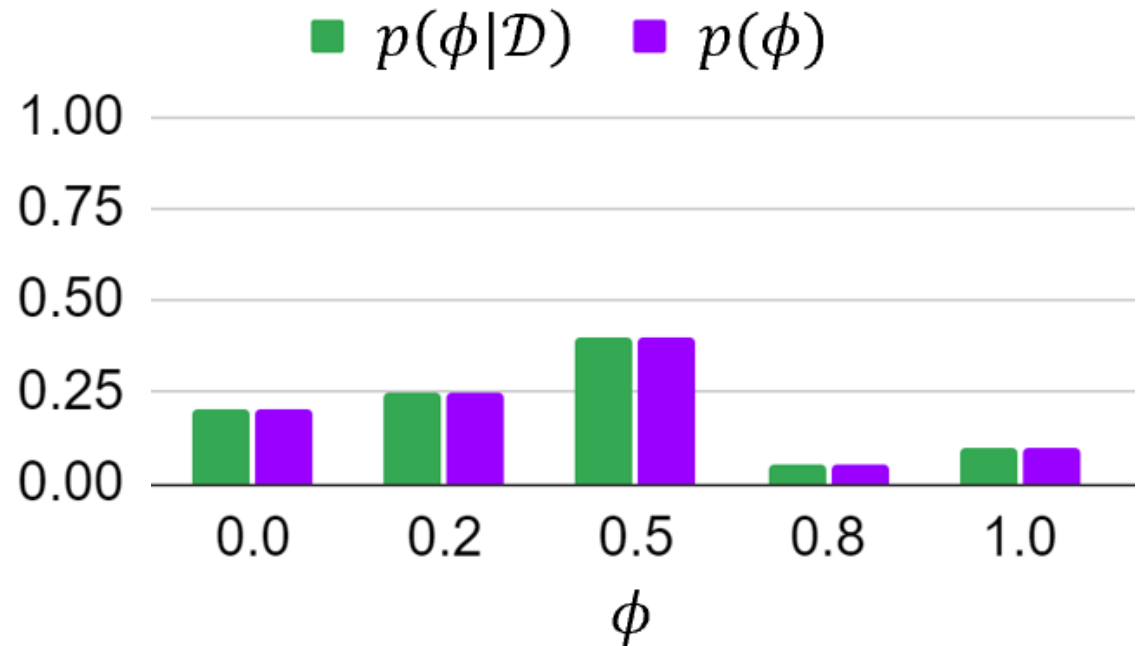
# MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$p(\mathcal{D} \mid \phi) p(\phi) = \prod_i^N p(y^{(i)} \mid \phi) p(\phi) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} p(\phi)$$

What happens as we flip more coins?

$N = 0: \mathcal{D} = \{\}$



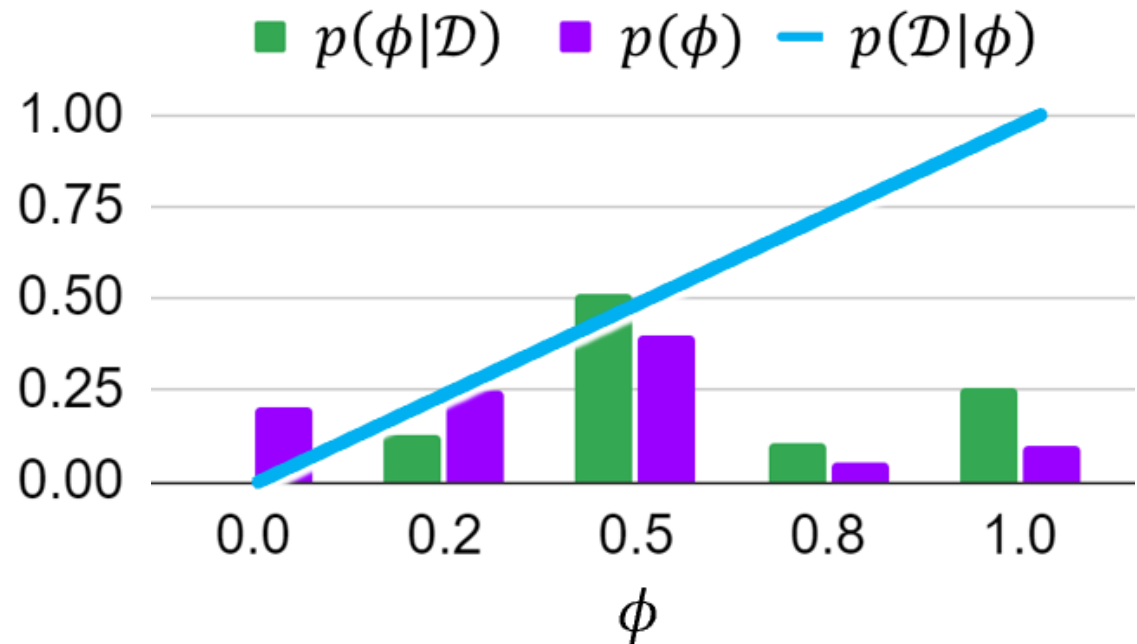
# MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$p(\mathcal{D} \mid \phi) p(\phi) = \prod_i^N p(y^{(i)} \mid \phi) p(\phi) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} p(\phi)$$

What happens as we flip more coins?

$N = 0: \mathcal{D} = \{H\}$



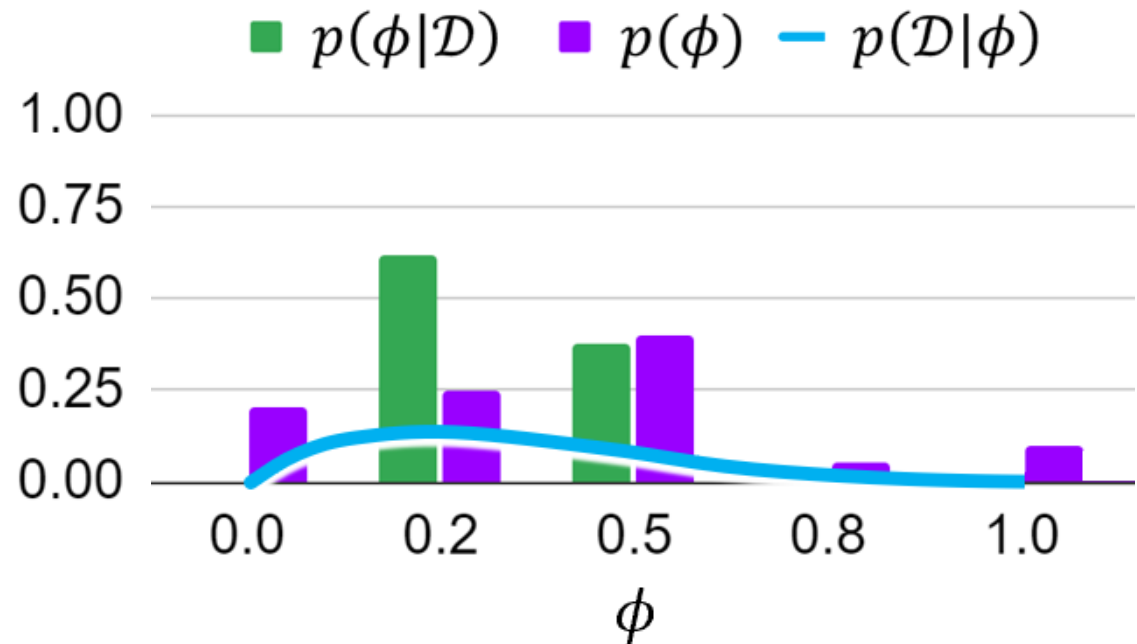
# MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$p(\mathcal{D} \mid \phi) p(\phi) = \prod_i^N p(y^{(i)} \mid \phi) p(\phi) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} p(\phi)$$

What happens as we flip more coins?

$N = 0$ :  $\mathcal{D} = \{H, T, T, T, T\}$



# Prior Distributions for MAP

If the prior  $p(\theta)$  is uniform, then MLE and MAP are the same!

$$p(\mathcal{D} \mid \phi) p(\phi) = \prod_i^N p(y^{(i)} \mid \phi) p(\phi) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} p(\phi)$$

# Prior Distributions for MAP

If the prior  $p(\theta)$  is uniform, then MLE and MAP are the same!

$$p(\mathcal{D} \mid \phi) p(\phi) = \prod_i^N p(y^{(i)} \mid \phi) p(\phi) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} p(\phi)$$

**Conjugate priors:** when the prior and the posterior distributions are in the same family

Bernoulli likelihood with a Beta prior has Beta posterior

Categorical likelihood with a Dirichlet prior has Dirichlet posterior

Gaussian likelihood with a Gaussian prior has Gaussian posterior

<https://www.desmos.com/calculator/kr7m2m6cf7>

# Prior Distributions for MAP

If the prior  $p(\theta)$  is uniform, then MLE and MAP are the same!

$$p(\mathcal{D} \mid \phi) p(\phi) = \prod_i^N p(y^{(i)} \mid \phi) p(\phi) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} p(\phi)$$

**Conjugate priors:** when the prior and the posterior distributions are in the same family

Bernoulli likelihood with a Beta prior has Beta posterior

$$\phi^{N_{y=1}} (1 - \phi)^{N_{y=0}} \text{Beta}(\alpha, \beta) = \text{Beta}(\alpha + N_{y=1}, \beta + N_{y=0})$$

Tip: Think of the Beta distribution as having  $\alpha - 1$  heads and  $\beta - 1$  tails

<https://www.desmos.com/calculator/kr7m2m6cf7>

# M(C)LE for Linear Regression

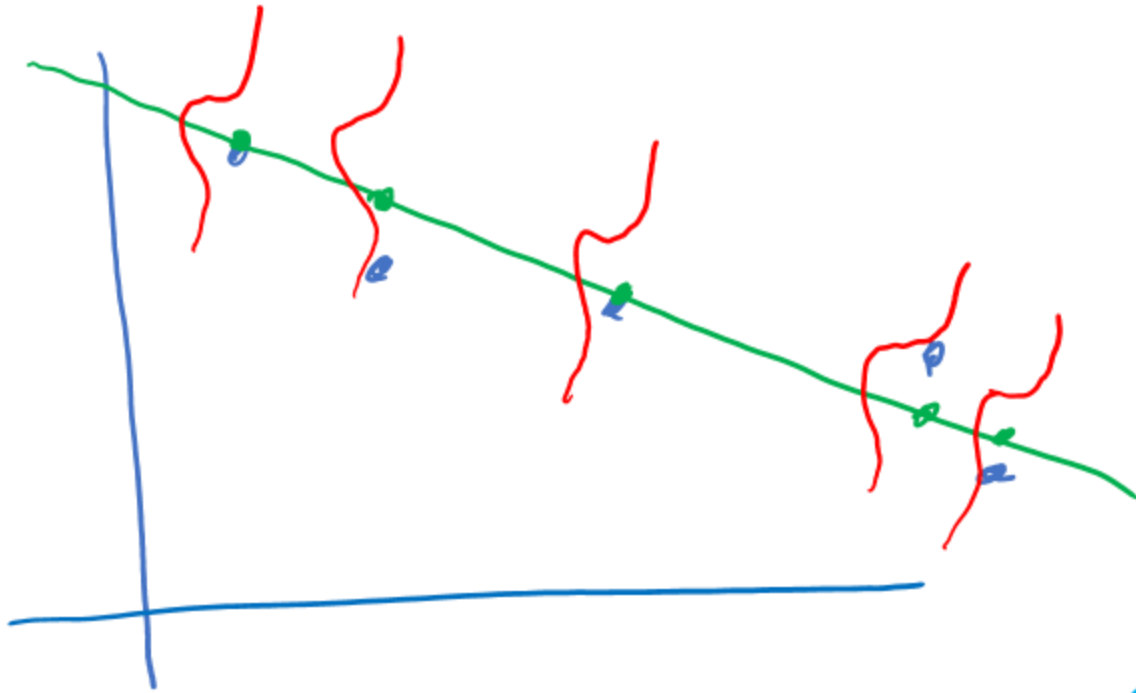
Probabilistic interpretation of linear regression

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_i^N p(y^{(i)} | \mathbf{x}^{(i)}, \theta)$$

$p(w)$   
 $p(b)$

$$\begin{cases} y = w^T x + b + \epsilon \\ \epsilon \sim \mathcal{N}(0, \sigma) \end{cases}$$

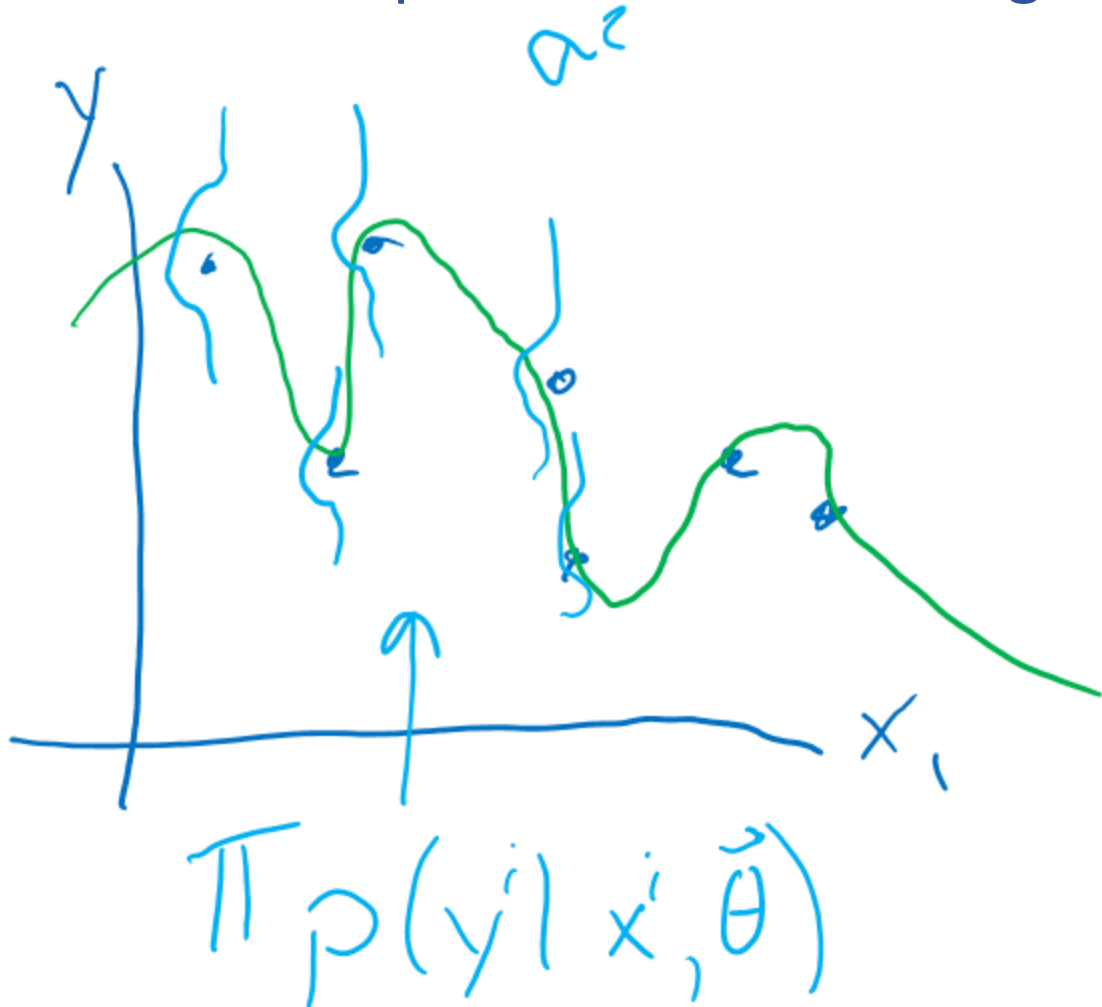
$$y \sim \underbrace{\mathcal{N}(w^T x + b, \sigma)}_{\uparrow}$$



$$p(y | x, w, b) = \underbrace{\hspace{10em}}$$

# MAP for Linear Regression

What assumptions are we making about our parameters?



$$\begin{bmatrix} 1 \\ x_i \end{bmatrix} \xrightarrow{\phi} \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ x_{i,4} \end{bmatrix} \begin{matrix} b \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{matrix} \leftarrow$$

$p(w_j) \sim \mathcal{N}(0, \tau^2)$

# MAP for Linear Regression

Recall prereading example of Gaussian prior for Gaussian likelihood

$$\begin{aligned} p(\mu \mid \mathcal{D}) &\propto \underline{p(\mu)} \prod_{i=1}^4 p(x^{(i)} \mid \mu) \\ &= \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mu-\nu)^2} \prod_{i=1}^4 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2} \end{aligned}$$

Linear Regression with Gaussian prior on weights

# M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_i p(y^{(i)} | \mathbf{x}^{(i)}, \theta)$$

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^N -\log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)} - \mu^{(i)})^2}{2\sigma^2}}$$

$$-\ell(\theta; \mathcal{D}) = N \log \sqrt{2\pi\sigma^2} + \frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - \theta^T x^{(i)})^2$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y^{(i)} - \theta^T x^{(i)})^2$$

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

$$\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z^{(i)} - \mu)^2}{2\sigma^2}}$$

$$\sum_{i=1}^N -\log \sqrt{2\pi\sigma^2} - \frac{(z^{(i)} - \mu)^2}{2\sigma^2}$$

$$\mu^{(i)} = \theta^T x^{(i)}$$

$$\sigma^2 = 1$$

# Regularization and MAP

## Linear Regression

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

$w_j \sim \mathcal{N}(0, \tau)$

$\uparrow$   
 $\uparrow$

argmax  $\vec{w}, b, \tau, \sigma^2$

$$\prod_{j=1}^M p(w_j) \prod_{i=1}^N p(y^{(i)} | x^{(i)}; \vec{w}, b)$$

$$\prod_{j=1}^M \frac{1}{\sqrt{\tau}} e^{-\frac{(w_j - 0)^2}{2\tau^2}} \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)} - \mu^{(i)})^2}{2\sigma^2}}$$

$$\left( + \sum_{j=1}^M \frac{1}{2} w_j^2 + N \log \sqrt{2\pi\sigma^2} + \frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - \theta^T x^{(i)})^2 \right)$$

$$\lambda \|\vec{w}\|_2^2 + \text{MSE}$$

# Regularization and MAP

Linear Regression

|                         | (MSE)<br>ERM    | MLE/Map                   |
|-------------------------|-----------------|---------------------------|
| OLS                     | MSE no reg      | MLE                       |
| Ridge Regr.<br>Tychonov | MSE $\ w\ _2^2$ | MAP $N(0, \tau)$<br>prior |
| Lasso                   | MSE $\ w\ _1$   | MAP Laplace<br>prior      |