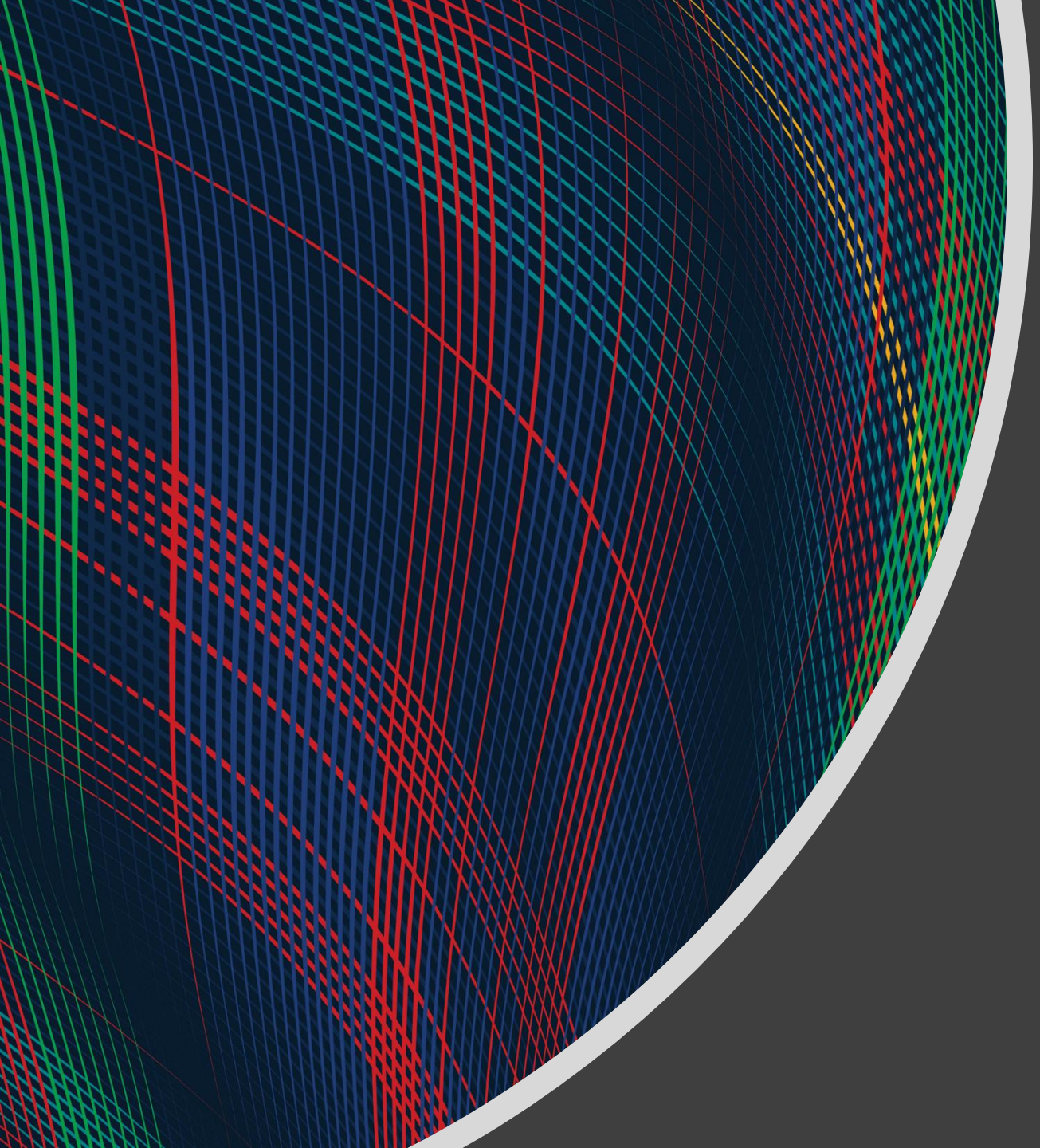


10-315 Introduction to ML

Probabilistic Models: MAP

Instructor: Pat Virtue

Course Feedback

Going well

- Recitation
- Pre-reading
- Homework
- Lecture Polls

Room for improvement

- Lecture slide ink handwritting
- 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

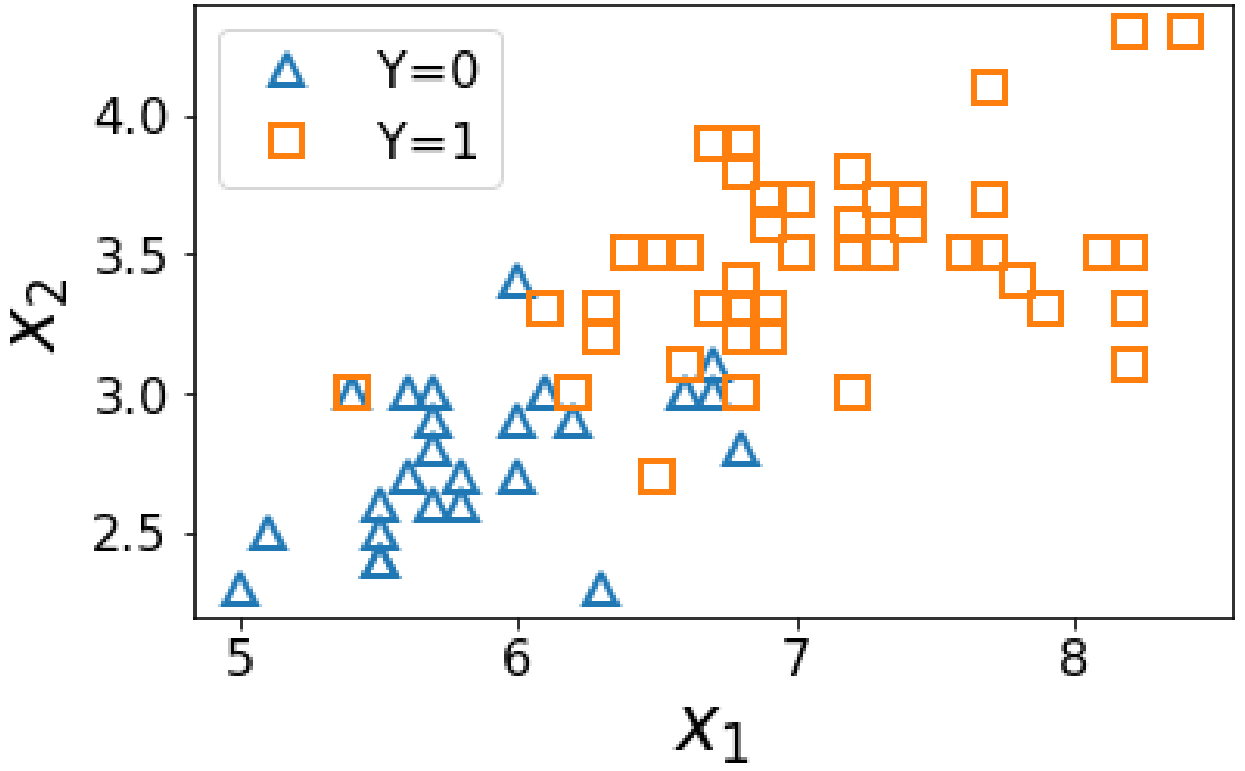
Probability Motivation

Empirical Risk Minimization vs MLE/MAP

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

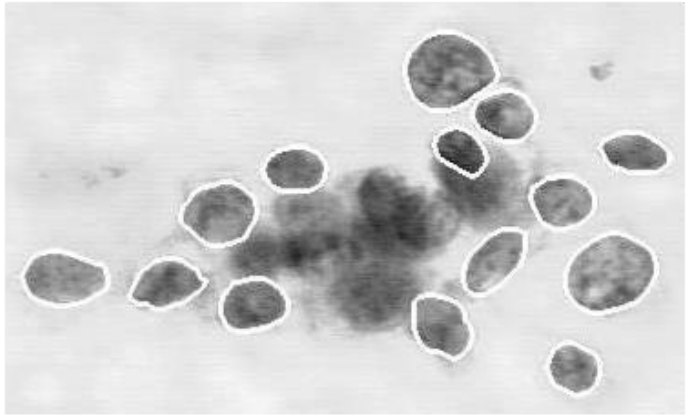
Why Probabilistic Models?

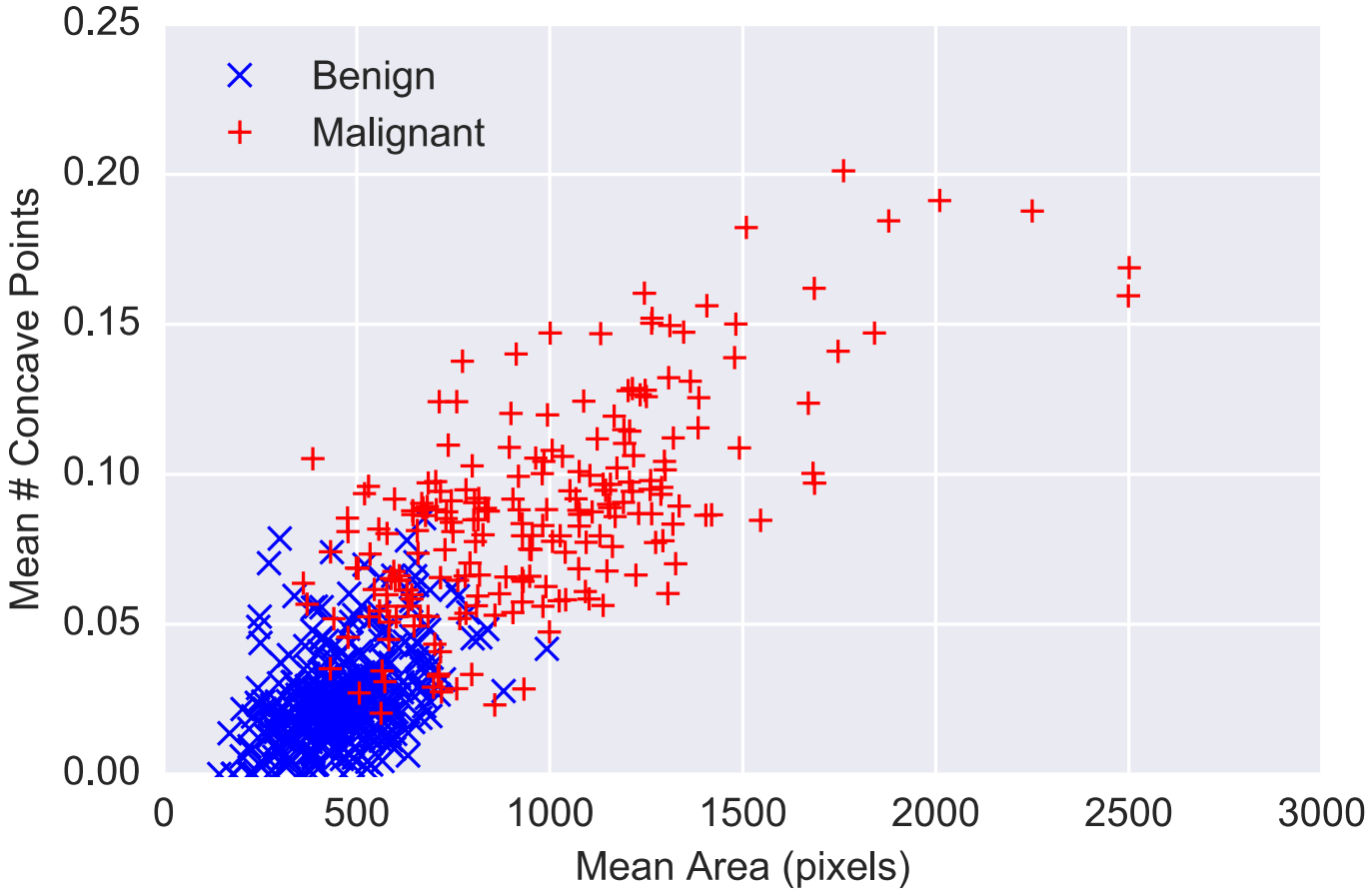
Iris Data



Why Probabilistic Models?

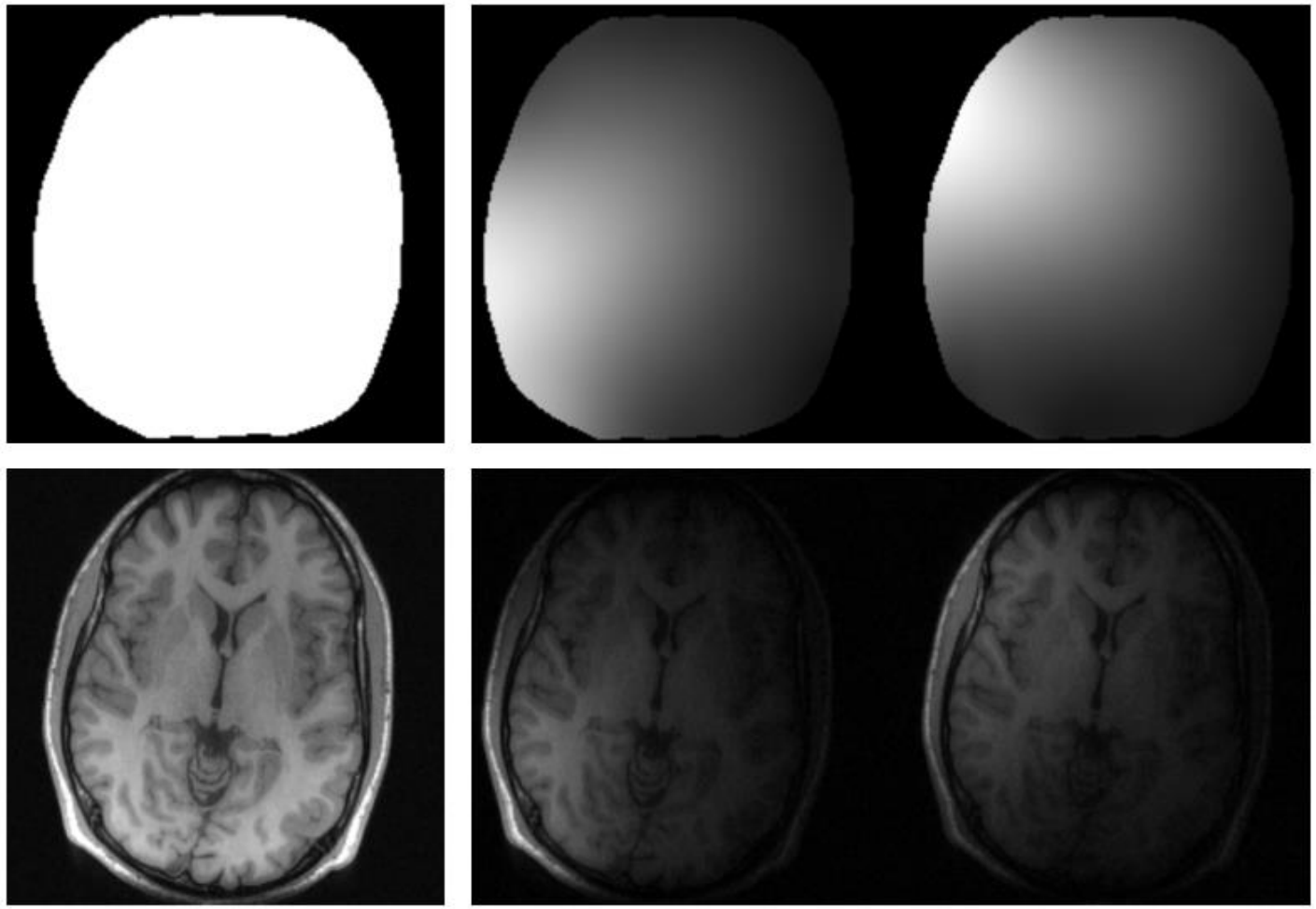
Breast Cancer Diagnosis





Why Probabilistic Models?

MRI Image Reconstruction



Categorical Gaussian Generative Model

Estimating parameters

$$Y \sim Categorical(\pi_1, \pi_2, \pi_3)$$

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

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

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

$$\frac{\partial l}{\partial \hat{\pi}_1} = 0$$

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

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

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(\mathcal{D} \mid \theta)$$

$$= \prod_{i=1}^{N} p(x^i, y^i \mid \theta)$$

$$= \log \prod p(x^i, y^i \mid \theta)$$

$$= \sum_{i=1}^{N} \log p(x^i, y^i \mid \theta)$$

$$= \arg\max_{\theta} \sum \log p(y^{(i)} \mid \theta)\, p(x^i \mid y^i, \theta)$$

$$\sum_{i=1}^{N} \log \prod_{k=1}^{K} \hat{\pi}_k^{y_k^{(i)}} f_{\mathcal{N}}(x^i \mid \mu_k, \sigma_k^2)^{y_k^{(i)}}$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \log \left(\hat{\pi}_k f_{\mathcal{N}}(x^i \mid \mu_k, \sigma_k^2)\right)}_{l(\theta \mid \mathcal{D})}$$

Autoencoders

Language Models

ML Applications of Bayes Rule

Bayes Rule

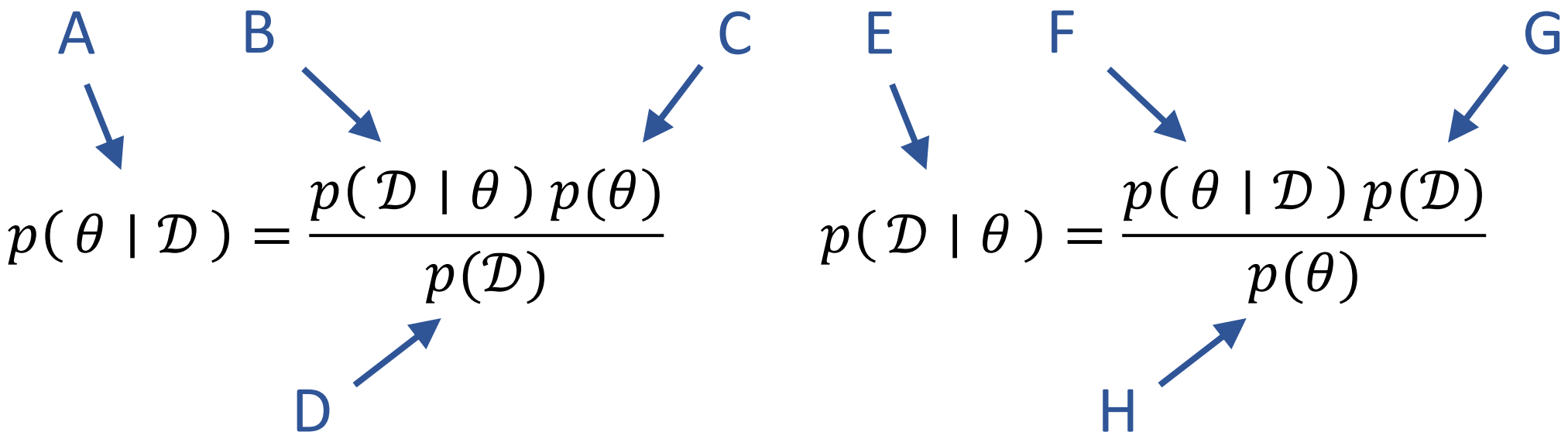
$$p(a \mid b) = \frac{p(b \mid a)\, p(a)}{p(b)} \qquad p(b \mid a) = \frac{p(a \mid b)\, p(b)}{p(a)}$$

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

Poll 1

Which of these terms is the likelihood?

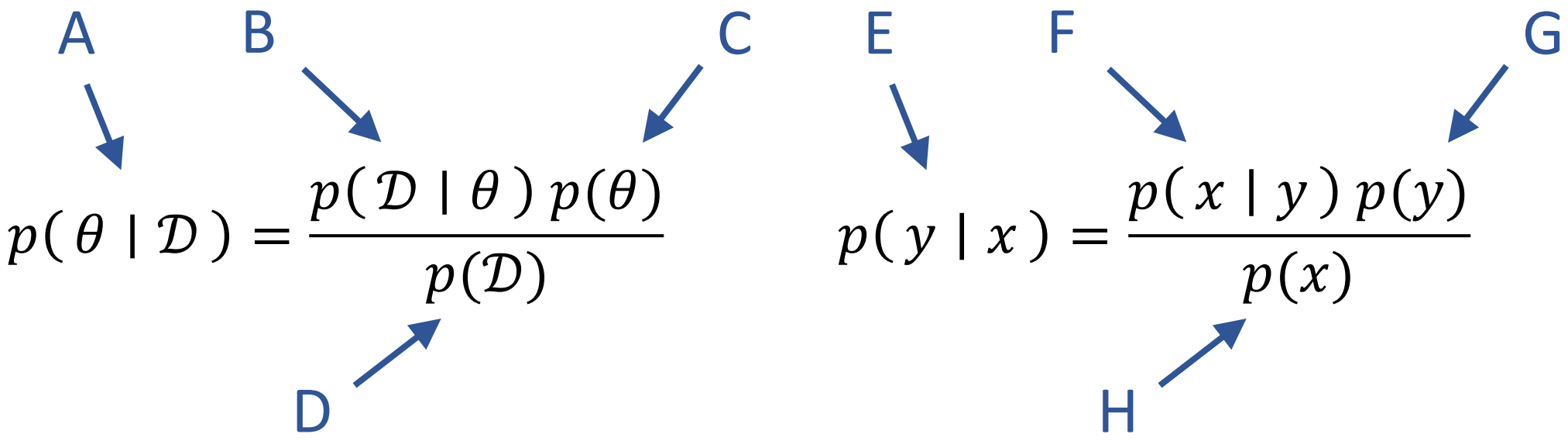
Select all that apply



Poll

Which of these terms is the likelihood?

Select all that apply



Bayes Rule

Terminology

Posterior Likelihood Prior

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

Two Applications of Bayes Rule

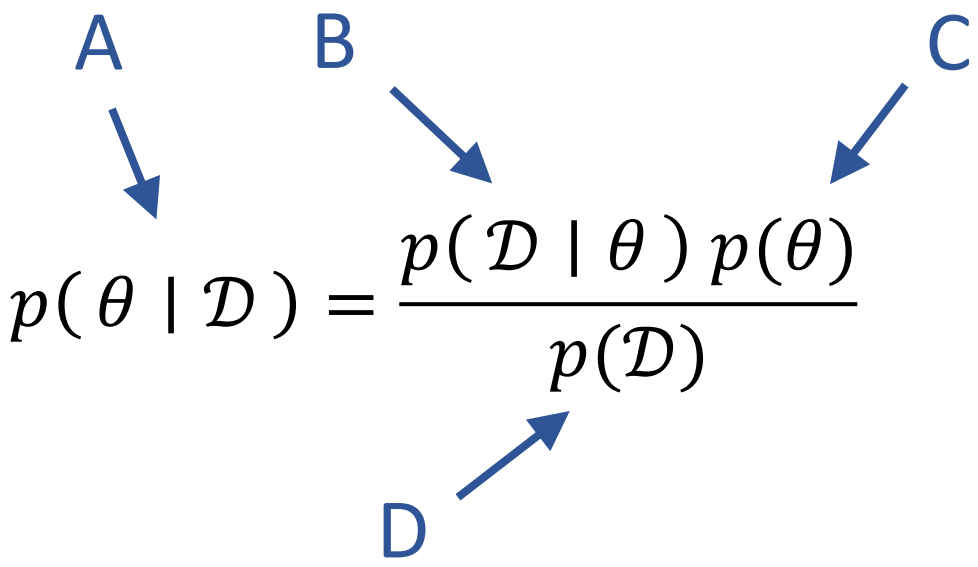
$$p(a \mid b) = \frac{p(b \mid a)\, p(a)}{p(b)}$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})} \qquad p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

MLE and MAP

Poll 2

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



MLE and MAP

Maximum likelihood estimation

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(y^{(i)} \mid \theta)$$

Maximum *a prosteriori* estimation

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D})$$

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

$$= \underset{\theta}{\operatorname{argmax}} \frac{\prod_{i=1}^{N} p(y^{(i)} \mid \theta) p(\theta)}{p(\mathcal{D})}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(y^{(i)} \mid \theta)\, p(\theta)$$

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

   $$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

Recipe for Estimation

MAP

1. Formulate the likelihood times the prior, $p(\mathcal{D} \mid \theta)p(\theta)$
2. Set objective $J(\theta)$ equal to negative log of the likelihood times the prior

   $$\mathrm{J}(\theta) = -\log[p(\mathcal{D} \mid \theta)p(\theta)]$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

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 ϕ

Invoice: Weighted Coins Customer: Torch Tricks, Inc.

Item	Quantity	Coin type, ϕ	$p(\phi)$
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

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.

Item	Quantity	Coin type, ϕ	$p(\phi)$
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

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta) \prod_{i=1}^{N} p\left(y^{(i)} \mid \theta \right)$$

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?

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

Poll 4

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

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

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

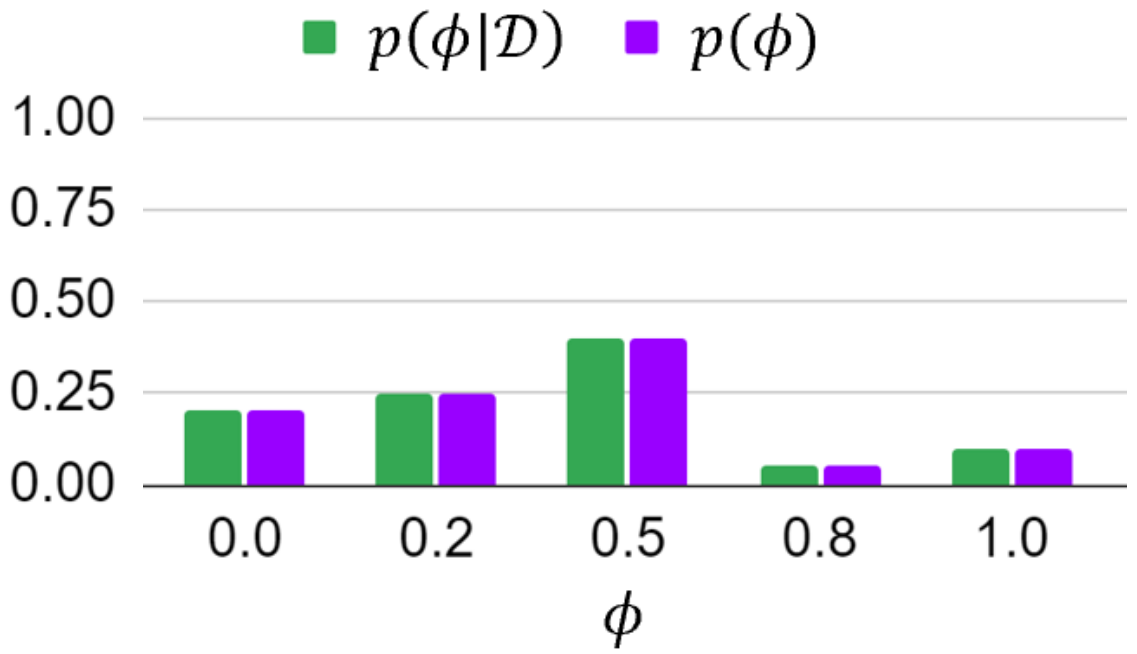
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) = \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

What happens as we flip more coins?

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



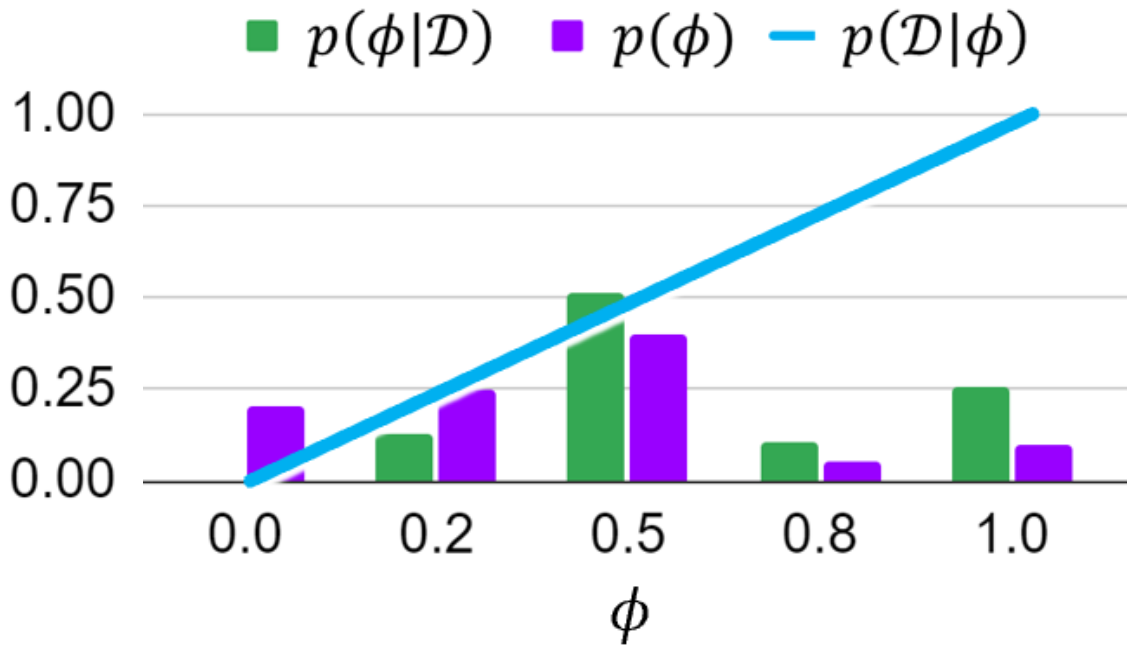
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) \;=\; \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

What happens as we flip more coins?

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



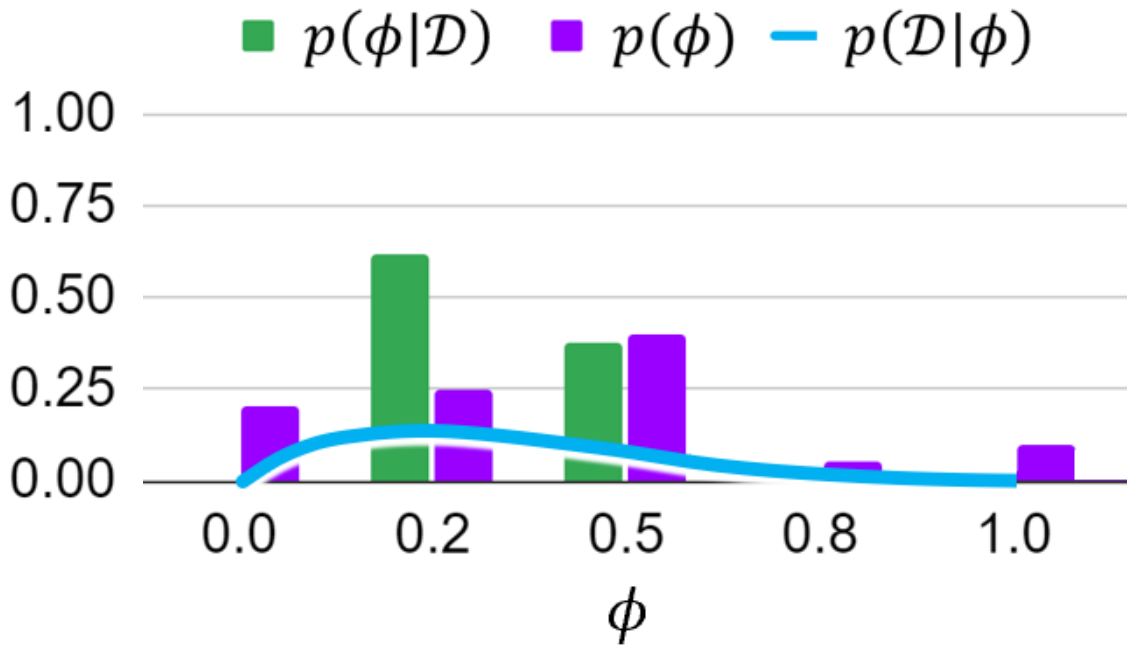
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) = \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

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!

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) = \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

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

Prior Distributions for MAP

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

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) = \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

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!

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) = \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

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

Bernoulli likelihood with a Beta prior has Beta posterior

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

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

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

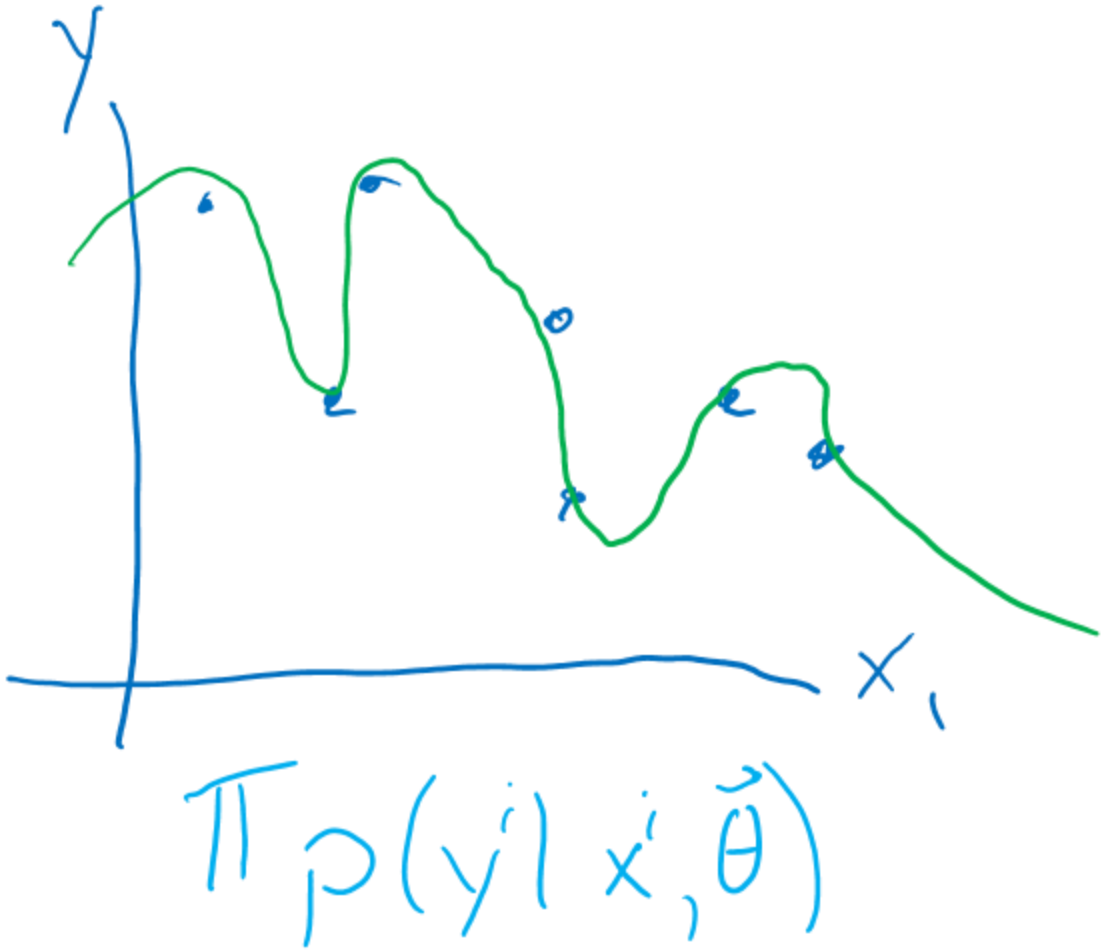
M(C)LE for Linear Regression

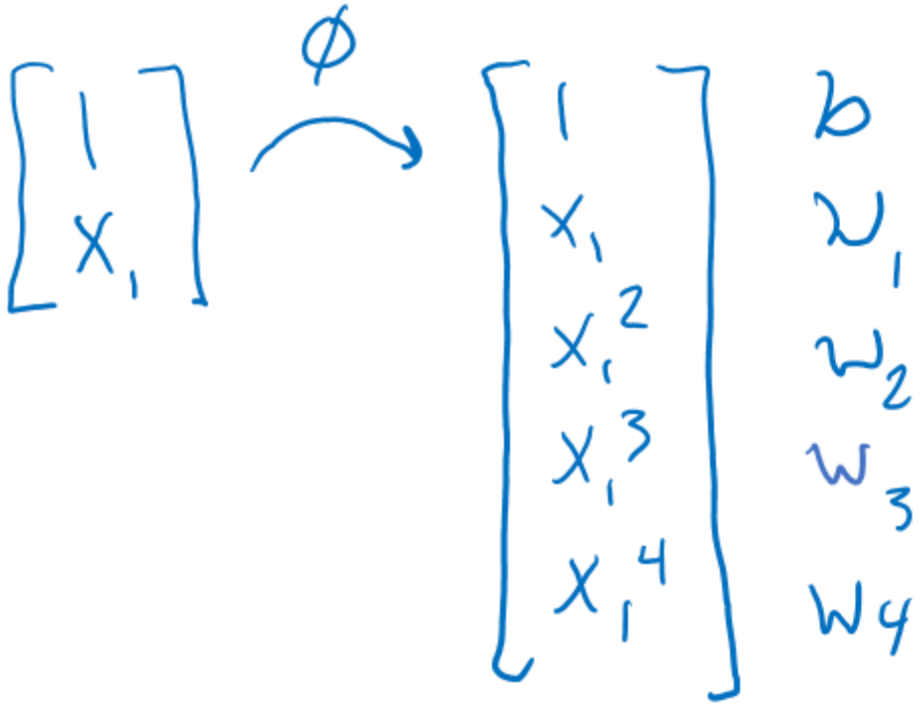
Probabilistic interpretation of linear regression

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta} \right)$$

MAP for Linear Regression

What assumptions are we making about our parameters?





MAP for Linear Regression

Recall prereading example of Gausssian prior for Gaussian likelihood

$$p(\mu \mid \mathcal{D}) \propto p(\mu) \prod_{i=1}^{4} p(x^{(i)} \mid \mu)$$

$$= \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mu-\nu)^2} \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2}$$

Linear Regression with Gaussian prior on weights

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i}^{N} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta} \right)$$

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

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

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

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

$$\mu^{(i)} = \theta^T x^{(i)} \qquad \sigma^2 = 1$$

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

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

RECALL

Regularization and MAP

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

Linear Regression

Regularization and MAP

Linear Regression