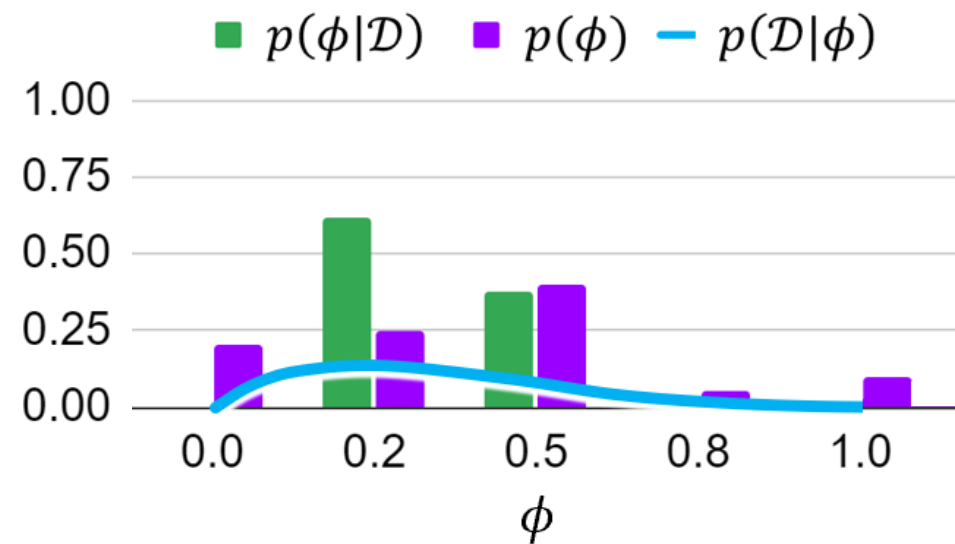
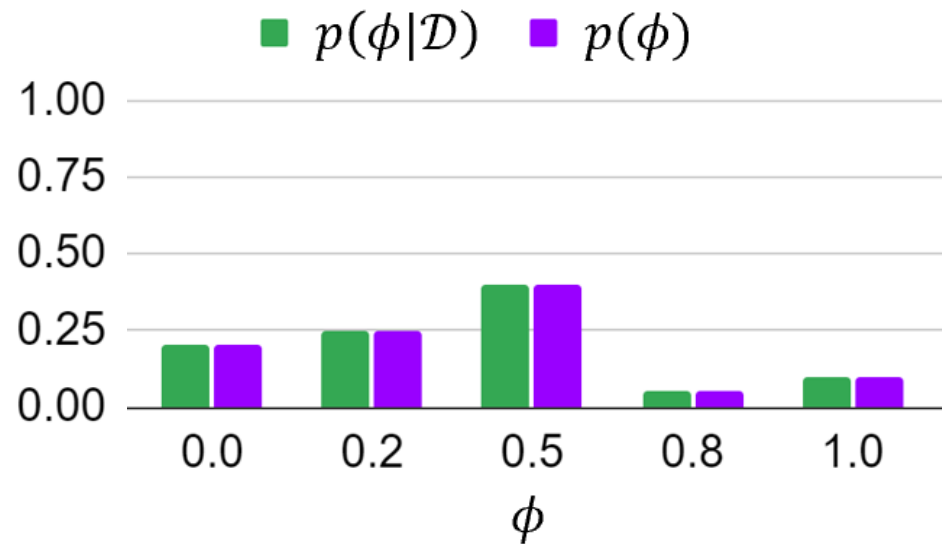


07-380 AI & ML II

Maximum a posteriori (MAP) Estimation



Instructors: Pat Virtue & Mohammad Salameh

ML Probability Notation

Language of Probability

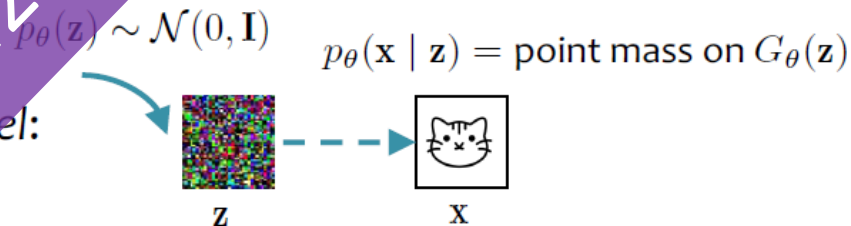
Like a magic key to open doors to learning SO much more...

GAN → VAE → Diffusion

(in 5 minutes)

GANs

Model:



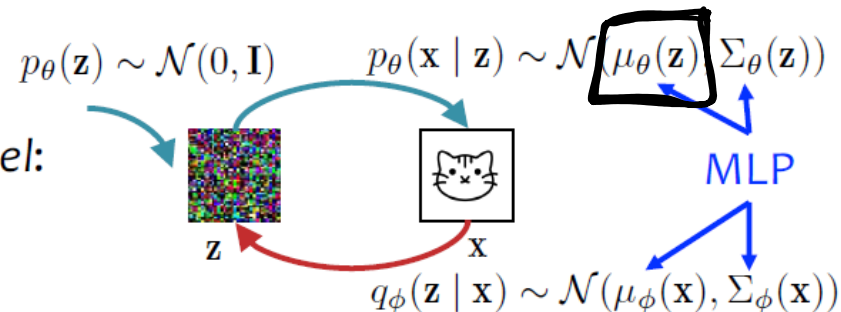
MLE?: $\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^N \log p(\mathbf{x}^{(i)})$ NOPE!

$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z}$$

Learn: minimax game

VAEs

Model:



$$h = \tanh(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1)$$

$$\mu_{\theta}(\mathbf{z}) = \mathbf{W}_2 h + \mathbf{b}_2$$

$$\Sigma_{\theta}(\mathbf{z}) = (\mathbf{W}_3 h + \mathbf{b}_3) \mathbf{I}$$

MLE?: $\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^N \log p(\mathbf{x}^{(i)})$ NOPE!

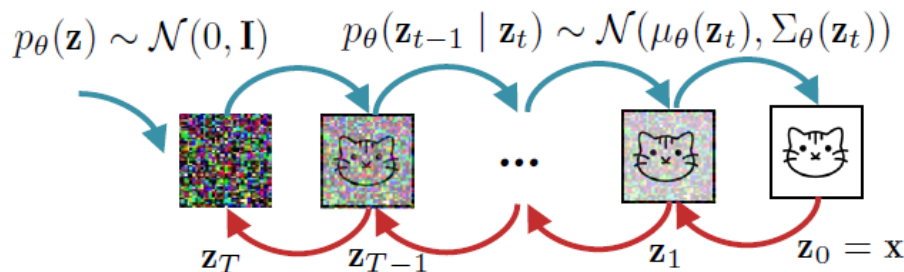
$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z}$$

Learn: $\theta, \phi = \arg \min_{\theta, \phi} \text{KL}(q_{\phi}(\mathbf{z} | \mathbf{x}) \| p_{\theta}(\mathbf{z} | \mathbf{x}))$

where $p_{\theta}(\mathbf{z} | \mathbf{x})$ is true posterior of $p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})$

Diffusion

Model:



Learn: $\theta = \arg \min_{\theta} \text{KL}(q(\mathbf{z}_{1:T} | \mathbf{x}) \| p_{\theta}(\mathbf{z}_{1:T} | \mathbf{x}))$

where $p_{\theta}(\mathbf{z}_{1:T} | \mathbf{x})$ is true posterior of $p(\mathbf{x} | \mathbf{z}_1) \cdots p(\mathbf{z}_{T-1} | \mathbf{z}_T) p(\mathbf{z}_T)$

MLE?: NOPE!

Links to Notes

Notation and Probability Notes

07-380: AI & ML II		Staff	OH	Schedule	Recitation	Assignments	Notes	Quizzes	Project	Policies	FAQ
9/21 Mon	Reasoning Under Uncertainty		8. MAP		Desmos: Bernoulli Likelihood and Posterior		MAP.pdf Checkpoint due 9/20 Sun, 11:59 pm			07-280 Notation Guide 07-280 Probability Background 07-280 MLE Pre-reading MML 9.2.3-4 Mitchell MLE and MAP	

$p(x; \theta)$ or $p(x | \theta)$ or $p(x)$ or $p_\theta(x)$: (marginal) distribution (pmf/pdf) defined by parameter(s) θ . The $|$ is required when we are considering the variable that is conditioned upon to also be a random variable value (Bayesian reasoning), e.g., if we also have prior distribution over θ , $p(\theta)$. The colon notation represents a **Frequentist's** view of probability (as opposed to Bayesian). The colon and $|$ are often used interchangeably when we don't explicitly model distributions for the given variable.

Probability Notation

Let X be a random variable modeled by a (univariate) Gaussian distribution.

We have a dataset of N i.i.d. samples of X , $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

How do we write this?

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

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

$$p(x; \mu, \sigma^2) \leftarrow \mu, \sigma^2 \text{ are } \underline{\text{not}} \text{ R.V.s}$$

$$p(x | \underline{\mu}, \sigma^2) \leftarrow \mu, \sigma^2 \text{ may be } \underline{\text{R.V.s}}$$

$p(\mu; \tau, \varsigma^2) \quad p(\sigma^2; \dots)$

Probability Notation

Density estimation

Coin flips

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

~~$p(x) = \phi$~~ $p(x | \phi) = \phi$

Exam grades

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x | \mu, \sigma^2)$$

$$p(x^{(1)}, x^{(2)}, x^{(3)}, \dots | \phi) = P(\mathcal{D} | \phi)$$

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

X : input features
 Y : output

Linear regression

$$p(y | x) = \mathcal{N}(w^T x + b; \sigma^2)$$

Logistic regression

$$p(y | x) = \text{Bernoulli}(\phi)$$

$$\phi = \frac{1}{1 + e^{-(w^T x + b)}}$$

Neural networks?

$$p(y^{(1)}, y^{(2)}, \dots | x^{(1)}, x^{(2)}, \dots, \vec{w}, b)$$

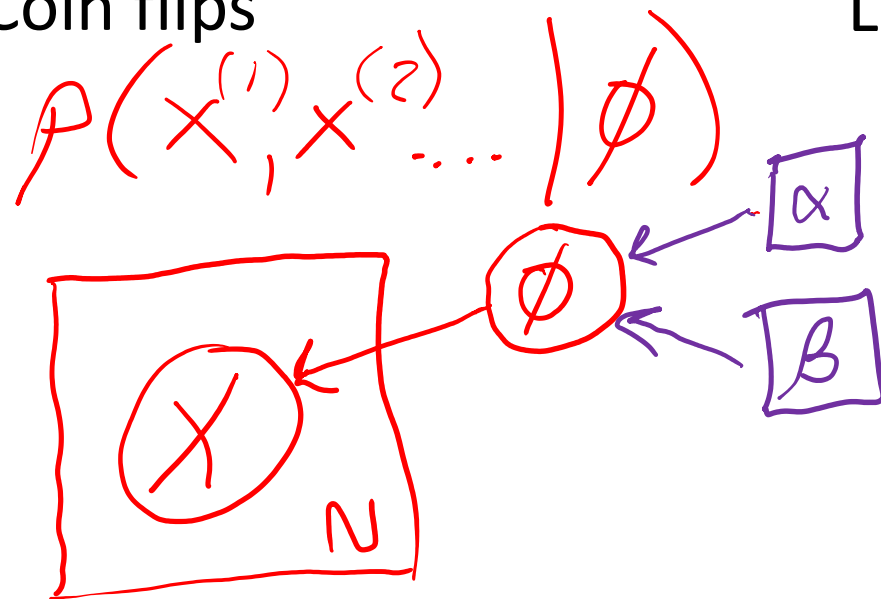
Classification Net

$$p(y | x, \theta) = \frac{1}{1 + e^{-f_{\theta}(x)}}$$

Graphical Model Notation (quick intro)

Density estimation

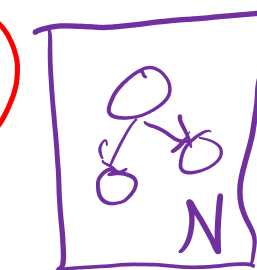
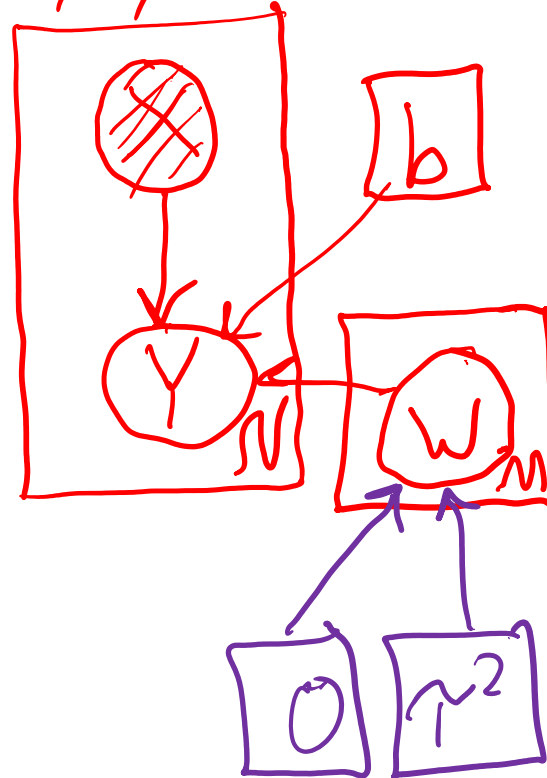
Coin flips



Supervised learning

Linear regression

$p(y^{(1)}, y^{(2)}, \dots | x^{(1)}, x^{(2)}, \dots, \vec{w}, b)$



Legend

Frame:
repeated
 N times

$\bigcirc \times$ (unobserved)
R.V.

$\bigcirc \times \times$ (observed)
R.V.

$\square \times$ another non-RV
factor

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

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

A $\rightarrow$ $p(\theta | \mathcal{D})$ (green box)

B $\rightarrow$ $p(\mathcal{D} | \theta)$ (blue box)

C $\rightarrow$ $p(\theta)$ (purple box)

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

D $\rightarrow$ $p(\mathcal{D})$ (red underline)

Normalization term

E $\rightarrow$ $p(\mathcal{D} | \theta)$ (blue box)

F $\rightarrow$ $p(\theta | \mathcal{D})$

G $\rightarrow$ $p(\mathcal{D})$

H $\rightarrow$ $p(\theta)$

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

Weird

Poll

Which of these terms is the likelihood?

Select all that apply

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

A blue arrow points from 'A' to $p(\theta | \mathcal{D})$. A blue arrow points from 'B' to $p(\mathcal{D} | \theta)$. A blue arrow points from 'C' to $p(\theta)$. A blue arrow points from 'D' to $p(\mathcal{D})$.

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

A blue arrow points from 'E' to $p(y | x)$. A blue arrow points from 'F' to $p(x | y)$. A blue arrow points from 'G' to $p(y)$. A blue arrow points from 'H' to $p(x)$. Red handwritten annotations include θ_1 with an arrow pointing to $p(x | y)$, θ_1, θ_2 with an arrow pointing to $p(y)$, and θ_2 with an arrow pointing to $p(x)$.

Bayes Rule

Terminology

Posterior

Likelihood

Prior

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

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

$p(\theta)$

$p(\mathcal{D})$

Two Applications of Bayes Rule

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

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

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

MLE and MAP

Poll 2

For a dataset D drawn from the following distribution pdf with parameter λ , where do we plug in the pdf that we used for MLE?

$$\underline{f(x) = \lambda e^{-\lambda x} ?}$$

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

A

B

C

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

D

MLE and MAP

Maximum likelihood estimation

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

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

Maximum *a posteriori* estimation

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

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

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

$$= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(y^{(i)} \mid \theta) p(\theta)$$
$$\sum_{i=1}^N \log p(y^{(i)} \mid \theta)$$
$$+ \log p(\theta)$$

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

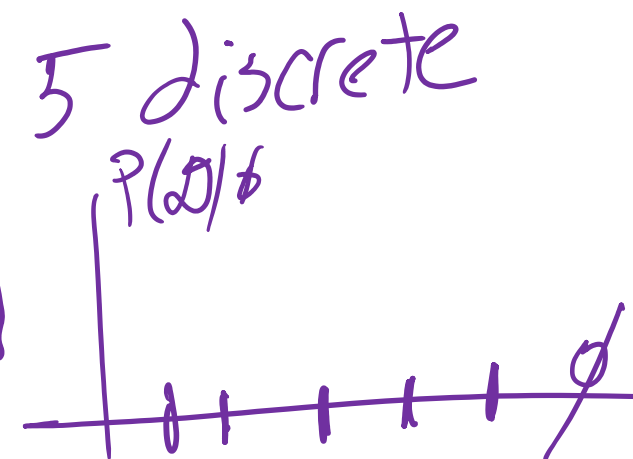
1. Formulate the likelihood times the prior, $p(\mathcal{D} | \theta)p(\theta)$
2. Set objective $J(\theta)$ equal to negative log of the likelihood times the prior
$$J(\theta) = -\log[p(\mathcal{D} | \theta)p(\theta)]$$
3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$,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

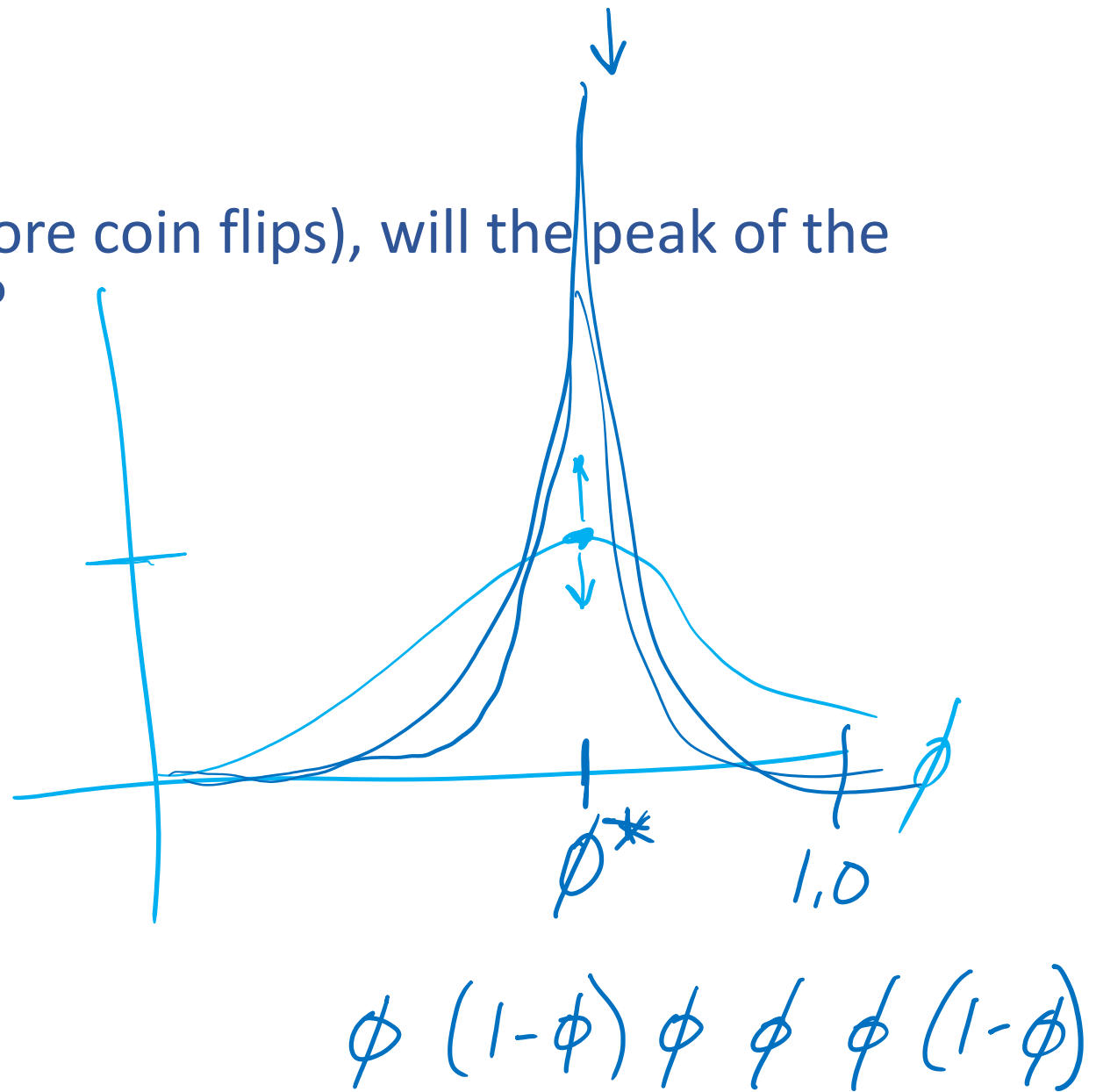
Invoice: Weighted Coins			Customer: Torch Tricks, Inc.	
<u>Item</u>	<u>Quantity</u>	<u>Coin type, ϕ</u>	<u>$p(\phi)$</u>	
0% Heads Coin	40 / 200	0.0	0.20	5 discrete $p(\phi)$
20% Heads Coin	50 / 200	0.2	0.25	
50% Heads Coin	80 / 200	0.5	0.40	
80% Heads Coin	10 / 200	0.8	0.05	
100% Heads Coin	20 / 200	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.	
<u>Item</u>	<u>Quantity</u>	<u>Coin type, ϕ</u>	<u>$p(\phi)$</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

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?

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

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

Poll 4

dist over θ
↓

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

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

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**

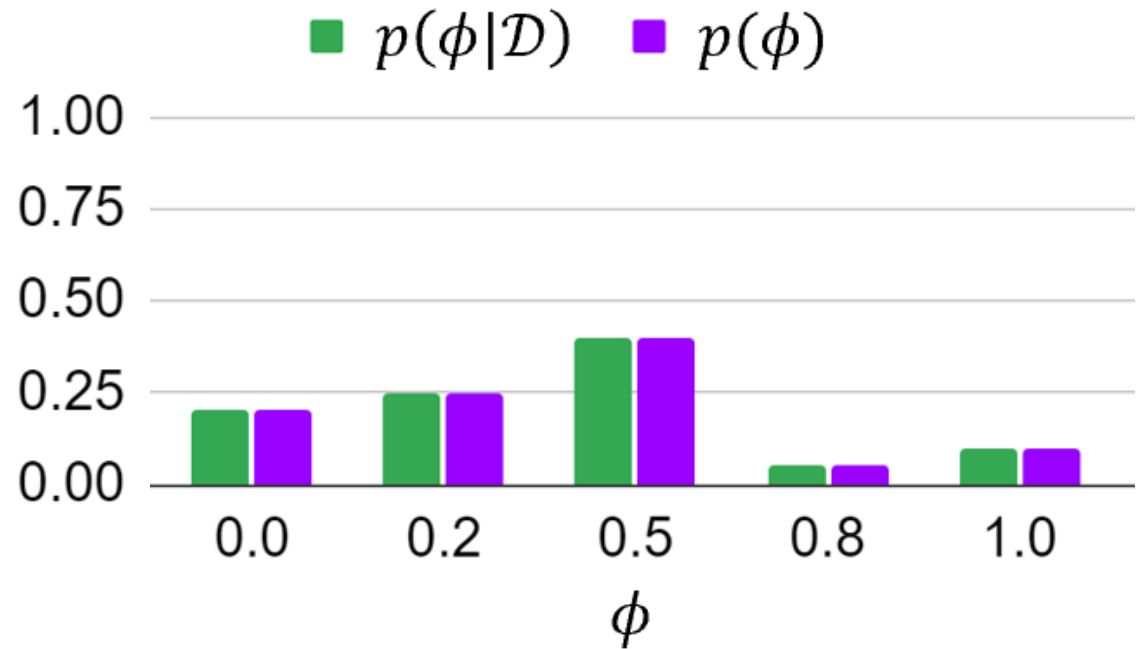
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

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

What happens as we flip more coins?

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



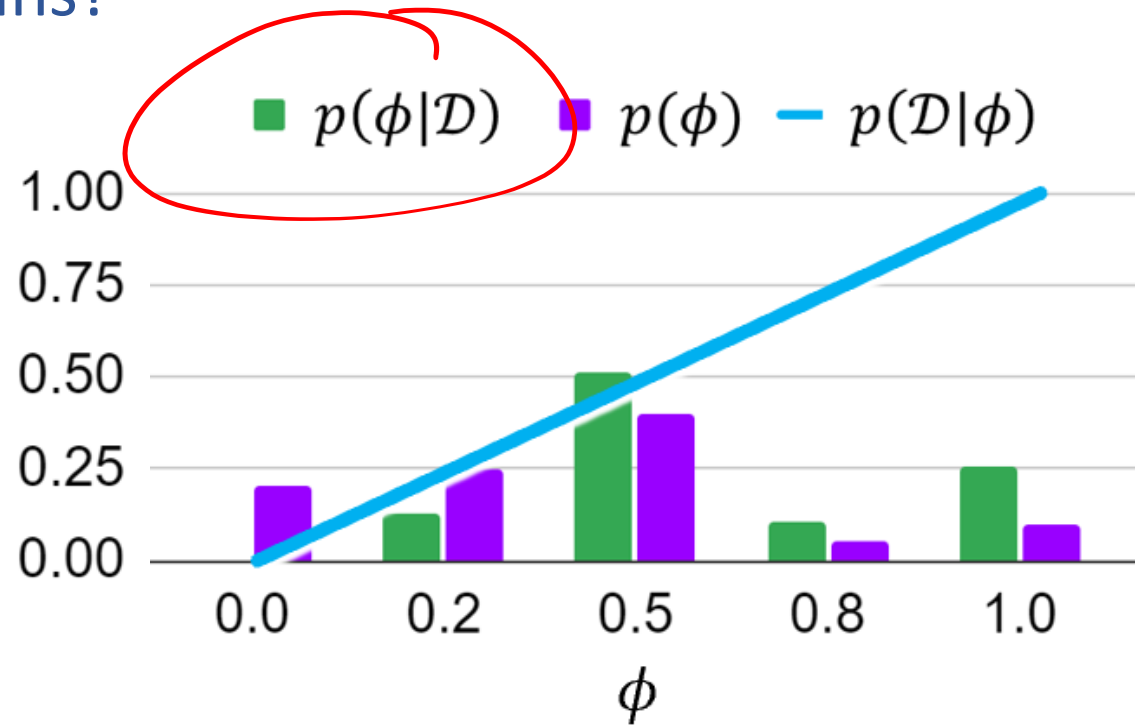
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

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

What happens as we flip more coins?

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



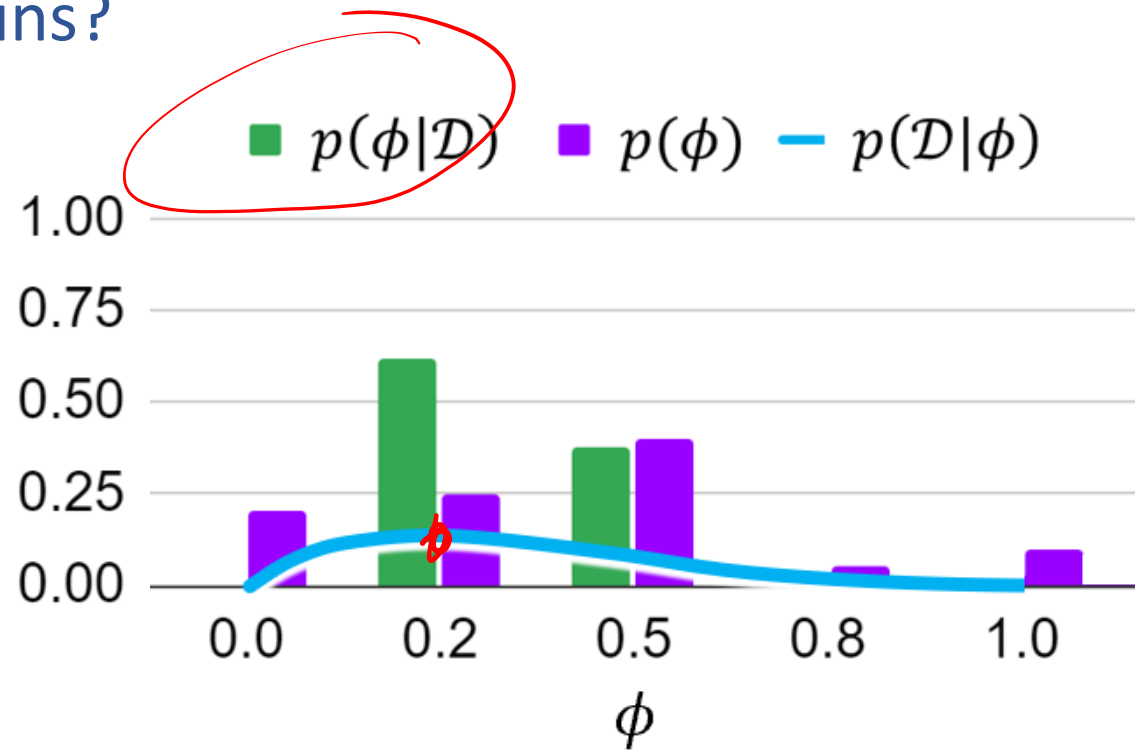
MAP as Data Increases

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

Beta distribution PDF

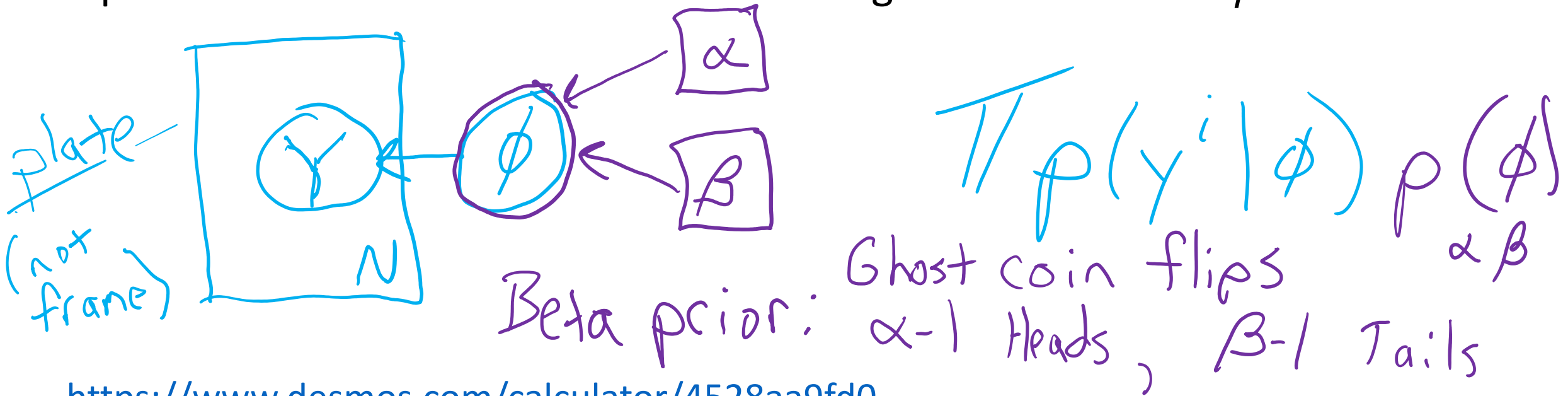
$$p(\phi; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \phi^{\alpha-1} (1 - \phi)^{\beta-1}$$

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/4528aa9fd0>

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

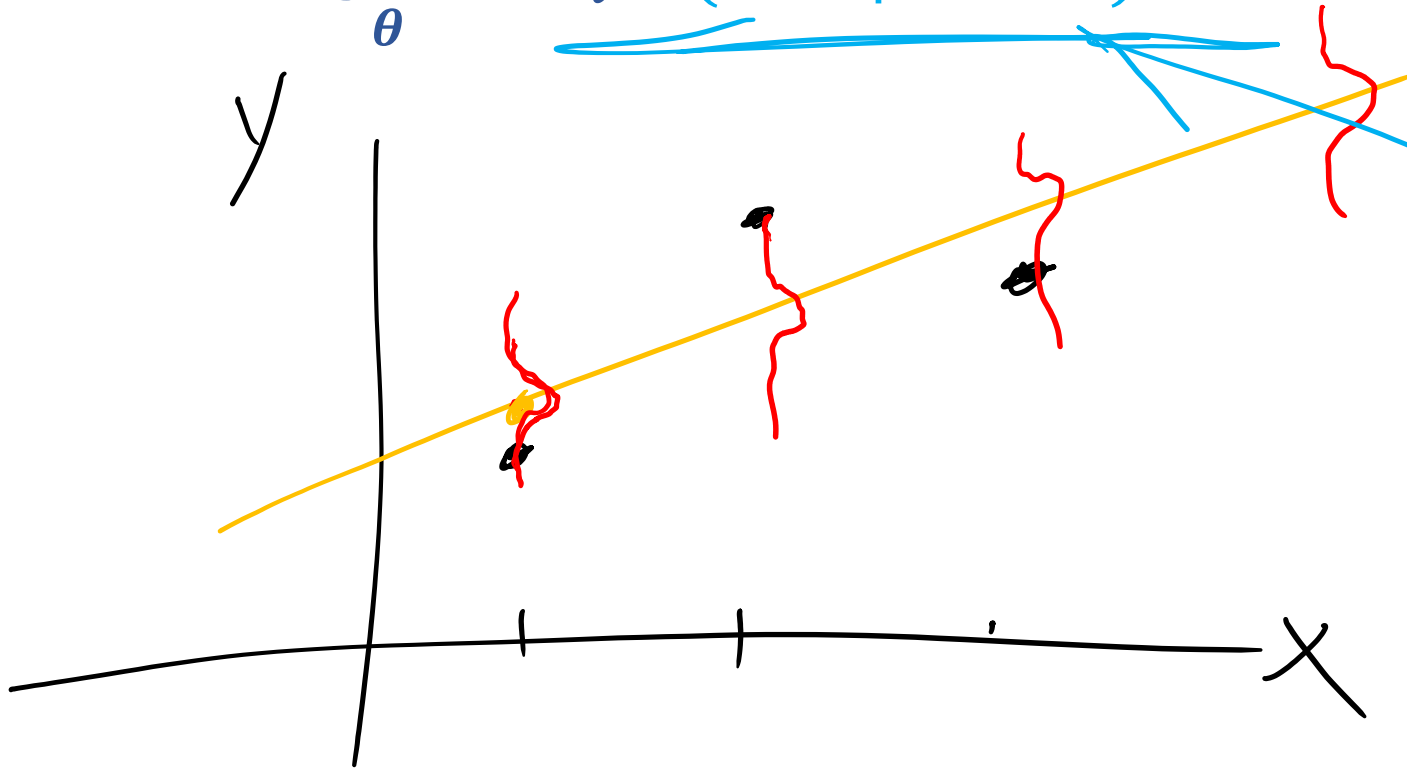
MLE & MAP

Linear Regression

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

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



$$y = \underline{wx + b}$$
$$= \underline{\vec{\theta}^T \vec{x}}$$

$$y = \underline{\theta^T x} + \underline{\epsilon}$$
$$\underline{\epsilon \sim \mathcal{N}(0, \sigma^2)}$$

$$\underline{y \sim \mathcal{N}(\theta^T x, \sigma^2)}$$

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$\begin{aligned}\hat{\theta}_{MLE} &= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(y^{(i)} | \mathbf{x}^{(i)}, \theta) \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)} - \mu^{(i)})^2}{2\sigma^2}}\end{aligned}$$

$$= \operatorname{argmin}_{\theta} \frac{1}{N} \sum (y^{(i)} - \mu^{(i)})^2$$

RECALL

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

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

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^N p(y^{(i)} | \mathbf{x}^{(i)}, \theta)$$

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

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

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

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

$$\sigma^2 = 1$$

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

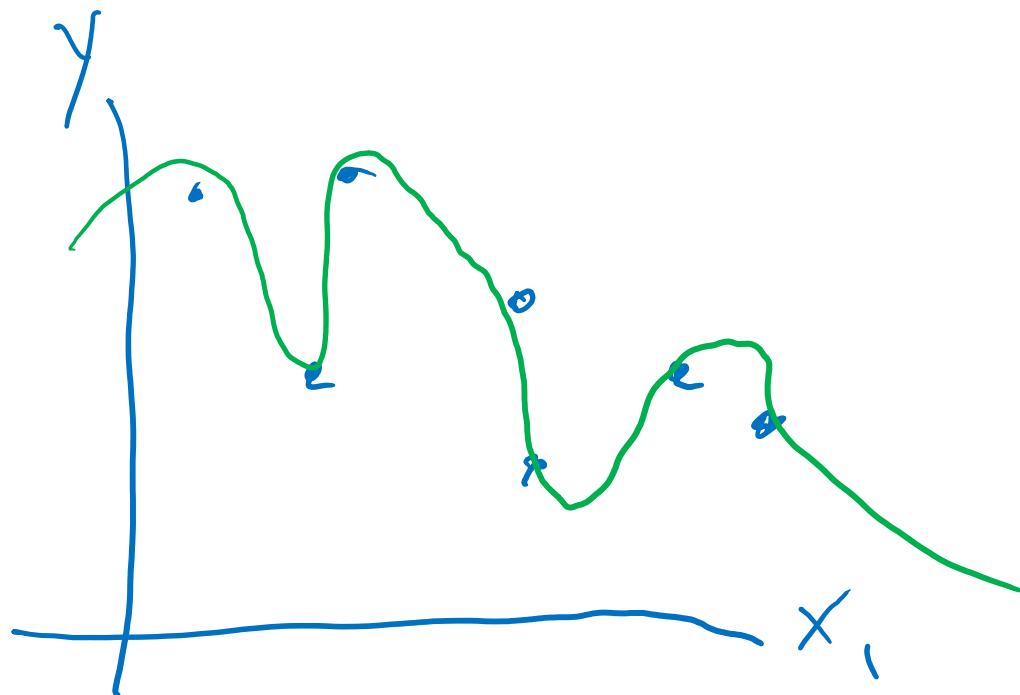
$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y^{(i)} - \theta^T \mathbf{x}^{(i)})^2$$

RECALL

MAP for Linear Regression

Example: Linear regression with polynomial feature maps

What do we want to make about our parameters?



$$\prod p(y^i | x^i, \vec{\theta})$$

$$\begin{bmatrix} 1 \\ x_1 \end{bmatrix} \xrightarrow{\phi} \begin{bmatrix} 1 \\ x_{1,2} \\ x_{1,3} \\ x_{1,4} \end{bmatrix} \quad \begin{matrix} b \\ w \\ w_2 \\ w_3 \\ w_4 \end{matrix}$$
$$p(w_j) \sim \mathcal{N}(0, \tau^2)$$

MAP for Linear Regression

Recall prereading example of Gaussian 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

$$a e^{-b-c} = a e^{-b-c}$$

$$y \sim \mathcal{N}(\underline{wx+b}, \sigma^2)$$

$$\underline{w} \sim \mathcal{N}(\underline{0}, \tau^2)$$

Regularization and MAP

$$y \sim \mathcal{N}(\vec{w}^\top x + b, \sigma^2)$$

Linear Regression

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(y^{(i)} | x^{(i)}, \theta) && p(w_1) p(w_2) p(w_3) \dots \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)} - \mu^{(i)})^2}{2\sigma^2}} && \cdot \prod_{j=1}^n \mathcal{N}(0, \tau^2)\end{aligned}$$

Regularization and MAP

Linear Regression

	ERM	
DLS	$MSE + \text{no reg}$	MLE
Ridge regression Tychonov regression	$MSE + \ \vec{w}\ _2$	MAP w/ Gaussian prior w_j
Lasso	$MSE + \ \vec{w}\ _1$	MAP w/ Laplace prior w_j