# Announcements

## Assignments

- HW6
  - Due Mon, 11/2, 11:59 pm

## Midterm 2

- Mon, 11/9, during lecture
- See Piazza for details
- Forms for conflicts / tech issues due Fri, 10/30

## Fireside Chat about the CMU ML PhD Program

- Fri, 10/30, 8:00 pm
- See Piazza for details, including form to show interest

# Plan

## Last Time

- PAC Criteria and Learning Theorems
- Bias-Variance trade-off as we change $|\mathcal{H}|$ or $N$
- Started VC dimension for infinite $|\mathcal{H}|$

## Today

- VC dimensions
- Learning theory and regularization
- MLE
  - MLE for linear regression
- MAP (Maximum a posteriori) estimation
  - MAP for linear regression

# Wrap up Learning Theory

Learning theory slides

# Introduction to Machine Learning

# MLE & MAP

Instructor: Pat Virtue

# Reminder MLE

$Y=1$ Heads

$[1, 0, 1, 1]$

## Trick coin

| All T | 1/3 H | Fair | 2/3 H | All H |
| --- | --- | --- | --- | --- |
| $\phi^{(A)} = 0$ | $\phi^{(B)} = 1/3$ | $\phi^{(C)} = 1/2$ | $\phi^{(D)} = 2/3$ | $\phi^{(E)} = 1$ |

$$p(y^{(1)} \dots y^{(4)} \mid \phi^{A}) = \underline{\prod p(y^{(i)} \mid \phi^{(A)})}$$

$\phi$ heads

$(1-\phi)$ fail

$$= \phi^{A} \cdot (1-\phi^{A}) \cdot \phi^{A} \cdot \phi^{A}$$

$$= 0 \cdot 1 \cdot 0 \cdot 0 = 0$$

$$\hat{\phi}_{MLE} = \underset{\phi}{\operatorname{argmax}} \prod_{i}^{N} p(\, y^{(i)} \mid \phi \,)$$

# Previous Piazza Poll

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

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \ \ (Heads) \\ 1 - \phi, & y = 0 \ \ (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

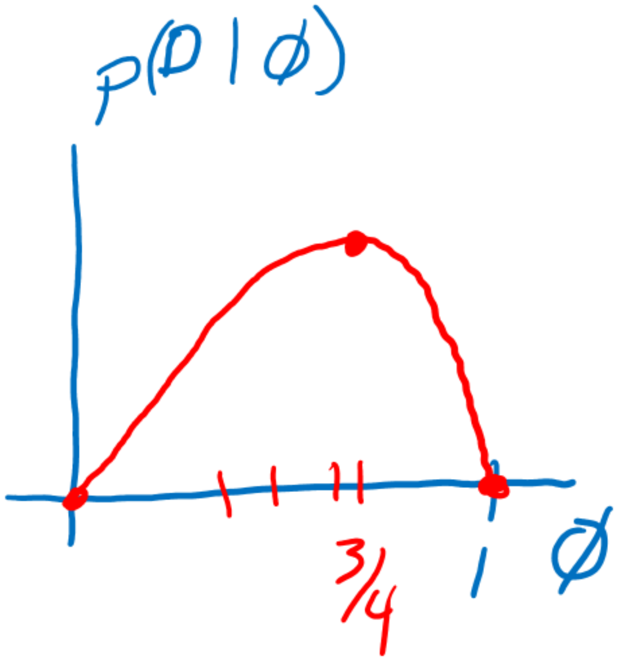
$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

A. 0.0   B. 1/8   C. 1/4   D. 1/2   E. 3/4   F. 3/8   G. 1.0

Why?

$$\mathrm{p}(\mathcal{D} \mid \phi) = \phi^3 (1 - \phi)^1$$



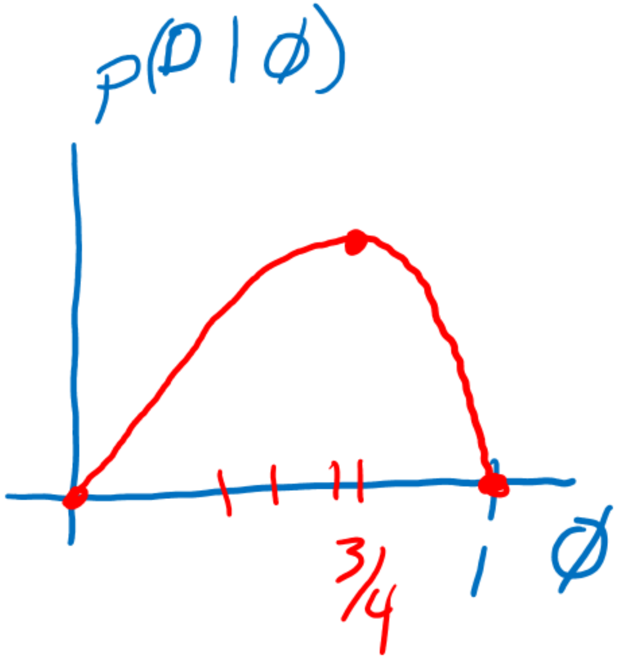
# MLE as Data Increases

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

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

What happens as we flip more coins?



# MLE for Gaussian

Gaussian distribution:

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

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

What is the log likelihood for three i.i.d. samples, given parameters $\mu, \sigma^2$?

$$\mathcal{D} = \{y^{(1)} = 65, y^{(2)} = 95, y^{(3)} = 85\}$$

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

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

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

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i}^{N} \log p\left( y^{(i)} \mid \boldsymbol{\theta} \right)$$

# Recipe for Estimation

## MLE

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

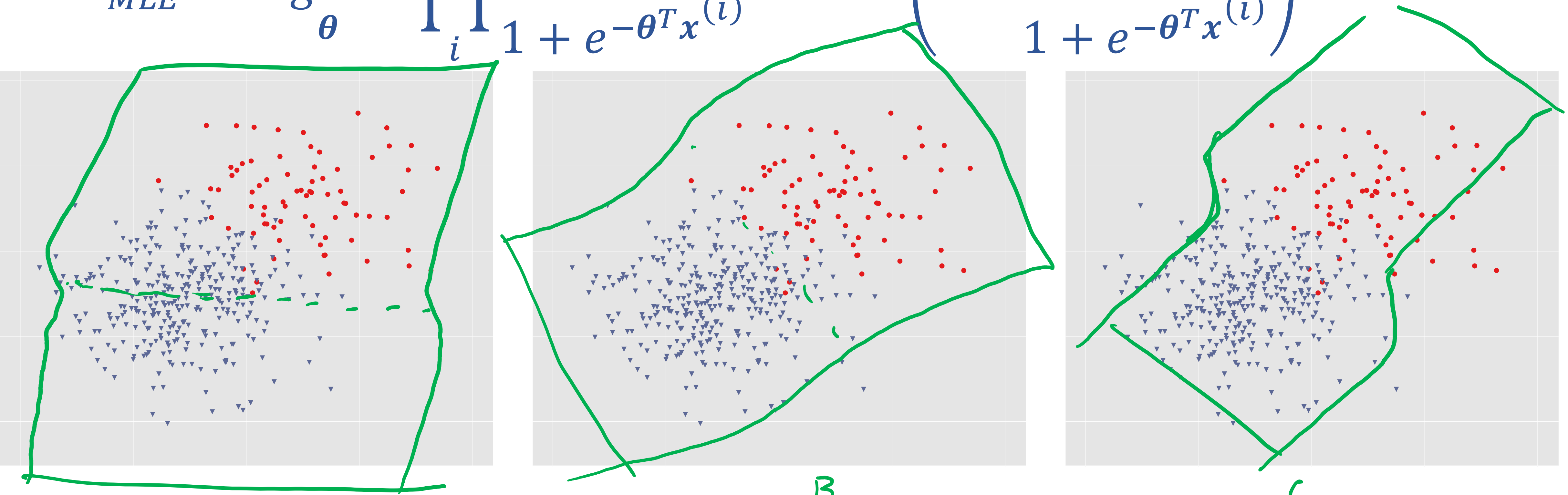
$$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$

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

# M(C)LE for Logistic Regression

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, $X_A$.

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} \frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}^{\mathbb{I}(y^{(i)}=1)} \left(1-\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}\right)^{\mathbb{I}(y^{(i)}=0)}$$

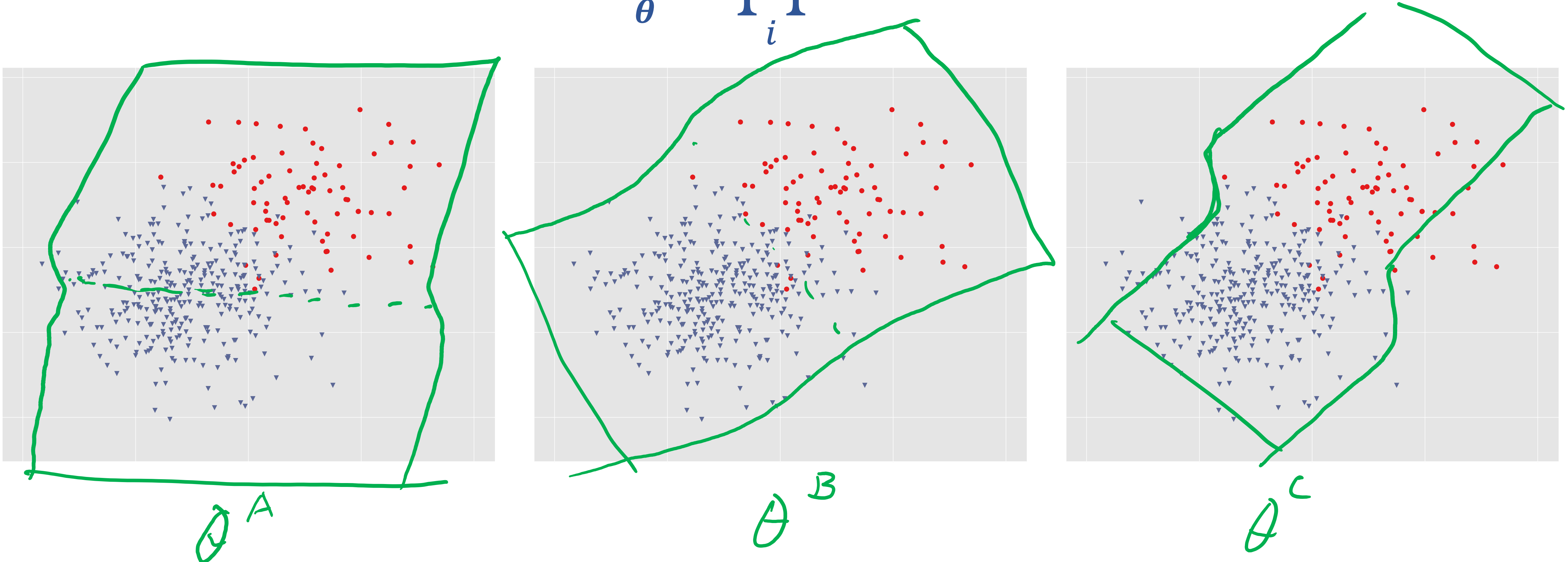


$\theta^A$ $\theta^B$ $\theta^C$

# M(C)LE for Logistic Regression

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, $X_A$.

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



# M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

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

# FROM MLE TO MAP

# Product Rule

Construct the joint by multiplying the conditional by the appropriate marginal

$$P(A, B) = P(B \mid A)P(A)$$

$$P(A, B) = P(A \mid B)P(B)$$

Also works when something is given everywhere

$$P(A, B \mid C) = P(A \mid B, C)P(B \mid C)$$

$$P(A, B \mid C, D, E) = P(A \mid B, C, D, E)P(B \mid C, D, E)$$

# Coin Flipping Example

Trick coin: Suppose I know how many coins are in each container in the store. How can I use this information both before and after flipping coins?

| All T | 1/3 H | Fair | 2/3 H | All H |
|---|---|---|---|---|
| $\phi^{(A)} = 0$ | $\phi^{(B)} = 1/3$ | $\phi^{(C)} = 1/2$ | $\phi^{(D)} = 2/3$ | $\phi^{(E)} = 1$ |

# Likelihood, Prior, and Posterior

Likelihood: $p(\mathcal{D} \mid \theta)$

Joint: $p(\mathcal{D}, \theta)$

Prior: $p(\theta)$

Posterior: $p(\theta \mid \mathcal{D})$

Relating these with Bayes rule

# MLE and MAP

Likelihood: $p(\mathcal{D} \mid \theta)$

Joint: $p(\mathcal{D}, \theta)$

Prior: $p(\theta)$

Posterior: $p(\theta \mid \mathcal{D})$

$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta) p(\theta)$

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

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

Maximum *a posteriori* estimation

# Coin Flipping Example

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

Trick coin: Suppose I know how many coins are in each container in the store. How can I use this information both before and after flipping coins?

| All T | 1/3 H | Fair | 2/3 H | All H |
|---|---|---|---|---|
| $\phi^{(A)} = 0$ | $\phi^{(B)} = 1/3$ | $\phi^{(C)} = 1/2$ | $\phi^{(D)} = 2/3$ | $\phi^{(E)} = 1$ |

# Piazza Poll 1:

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

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

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

# Piazza Poll 1:

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

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

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:

$$\mathcal{D} = [1, 0, 1, 1]$$

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

What happens as we flip more coins?

# Recipe for Estimation

## MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log of 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 $\theta$
   b. Use (stochastic) gradient descent to step towards better $\theta$

# Recipe for Estimation

MAP

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

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

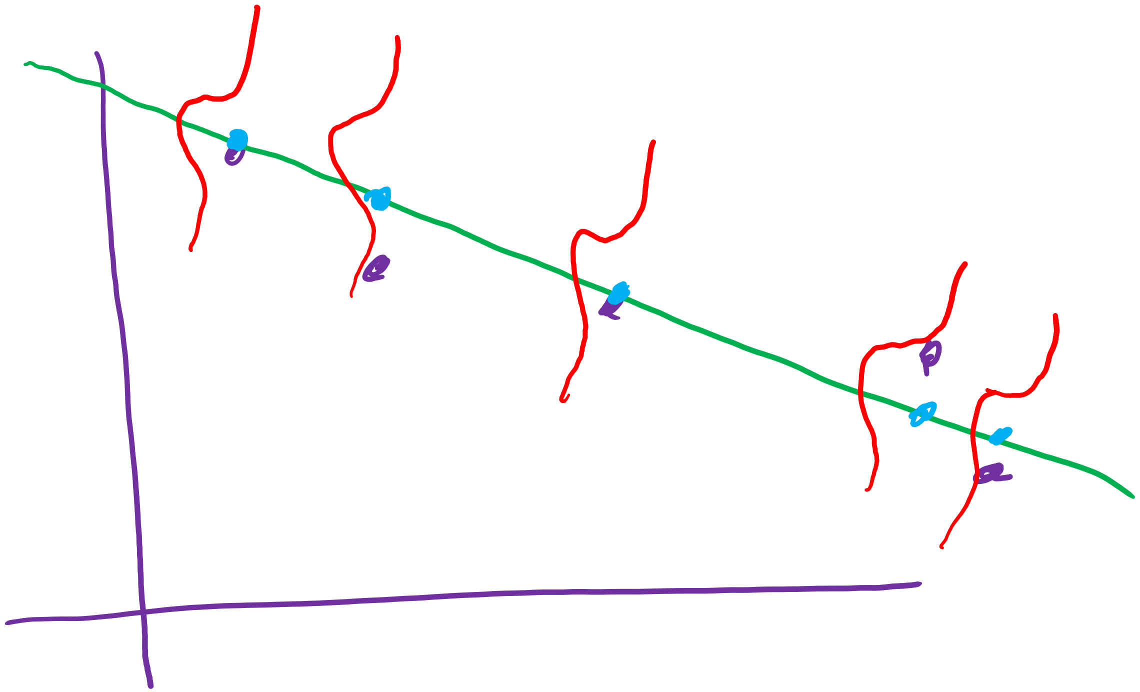
3. Compute derivative of objective, $\partial J/\partial\theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for $\theta$
   b. Use (stochastic) gradient descent to step towards better $\theta$

# M(C)LE for Linear Regression

## Probabilistic interpretation of linear regression

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

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



$$y \sim \underline{\qquad\qquad\qquad}$$

$$p(y \mid x, \vec{w}) = \underline{\qquad\qquad\qquad}$$

# MAP for Linear Regression

What assumptions are we making about our parameters?