MLE/MAP for learning distributions

Aarti Singh

Machine Learning 10-701
Jan 25, 2023

Distribution of Inputs

Input $X \in \mathcal{X}$

Discrete Probability Distribution P(X) = P(X=x)

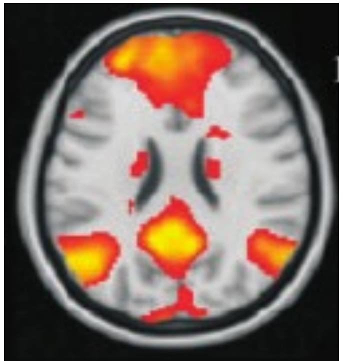
e.g. P(head) = ½, P(word x in text) = p_x

Probabilities in a distribution sum to 1

$\sum_x P(X=x) = 1$ P(tail) = 1 – p(head), $\sum_x p_x = 1$

Continuous Probability density p(x) $P(a <= X <= b) = \int_a^b p(x)dx$

e.g. p(brain activity)



Probability density integrate to 1

$\int p(x)dx = 1$

Distributions in Supervised tasks

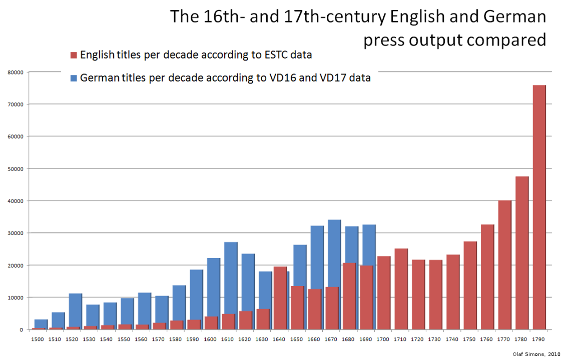
Input $X \in \mathcal{X}$

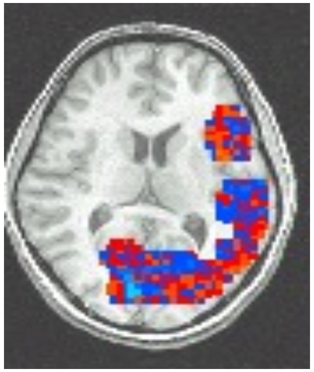
- Distribution learning also arises in supervised learning tasks e.g. classification

P(Y= y) Distribution of class labels

P(X = x | Y = y) Distribution of words in 'news' documents

Distribution of brain activity under 'stress'





Olaf simons'10

P(Y = y|X = x) Distribution of topics given document

How to learn parameters from data? MLE

(Discrete case)

Learning parameters in distributions

$Y \sim \text{Bernoulli}(\theta)$

$P(Y = \text{●}) = \theta$ $\qquad$ $P(Y = \text{●}) = 1 - \theta$

Learning θ is equivalent to learning probability of head in coin flip.

- How do you learn that?

Data =

Answer: 3/5

- Why??

Bernoulli distribution

Data, D = 

- Parameter θ : P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

Choose θ that maximizes the probability of observed data aka Likelihood

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

$D = \{X_1 \ldots X_n\}$

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$X_i \in \{H, T\}$
$\equiv \{0, 1\}$

$P(D|\theta) = P(X_1 \ldots X_n | \theta)$

$= \prod_{i=1}^{n} P(X_i|\theta)$ independent

$P(X_i|\theta) \sim \text{Ber}(\theta)$ identically distr

MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

$= \theta^{n_H} (1-\theta)^{n_T}$

$= 3/5$

$n_H = \sum_{i=1}^{n} 1_{\{X_i = 1\}}$ $n_T = \sum_{i=1}^{n} 1_{\{X_i = 0\}}$

"Frequency of heads"

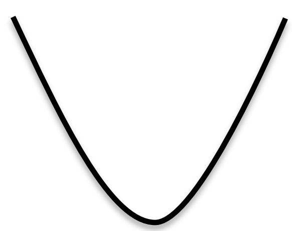
Derivation

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

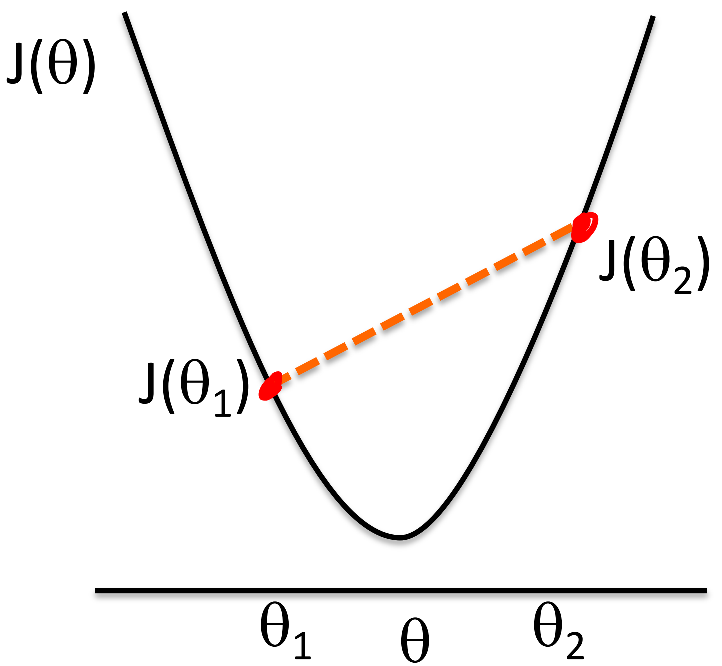
Short detour - Optimization

- Optimization objective J(θ) ← function of θ   e.g. $J(\theta) = \theta^{n_H}(1-\theta)^{n_T}$
- Minimum value J* = $\min_\theta$ J(θ)
- Minima (points at which minimum value is achieved) may not be unique

- If function is strictly convex, then minimum is unique

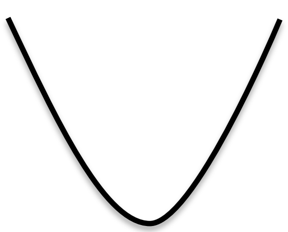


Convex functions

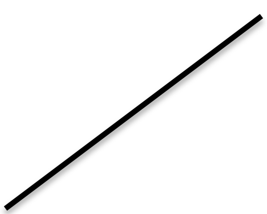


A function $J(\theta)$ is called **convex** if the line joining two points $J(\theta_1), J(\theta_2)$ on the function does not go below the function on the interval $[\theta_1, \theta_2]$

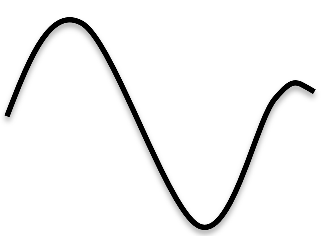
(Strictly) Convex functions have a unique minimum!



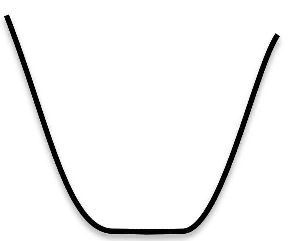
Convex



Both Concave & Convex



Neither

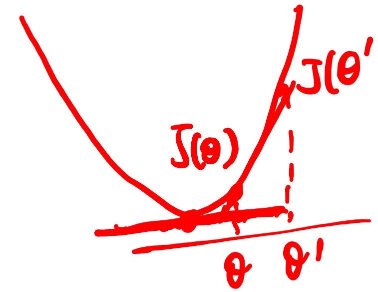


Convex but not strictly convex

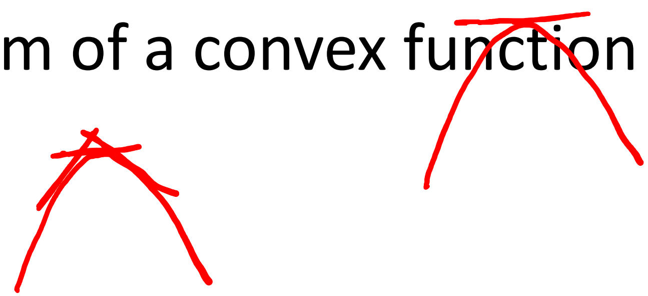
Optimizing convex (concave) functions

- Derivative of a function

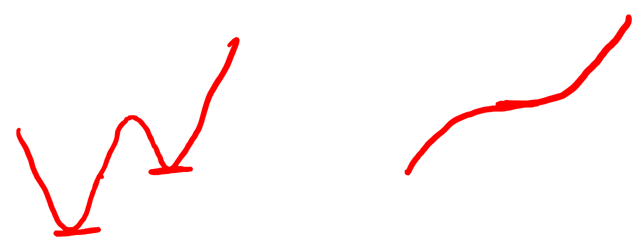
$$\left.\frac{d\,J(\theta)}{d\theta}\right|_{\theta} = \lim_{\theta' \to \theta} \frac{J(\theta) - J(\theta')}{\theta - \theta'}$$



- Derivative is zero at minimum of a convex function



- Second derivative is positive at minimum of a convex function



Optimizing convex (concave) functions

- What about

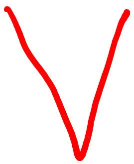
  concave functions?

  non-convex/non-concave functions?

  derivative = 0 may not have analytic solution?

  functions that are not differentiable?

  optimizing a function over a bounded domain aka constrained optimization?



Derivation

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

$$= \arg\max_{\theta} \theta^{n_H} (1-\theta)^{n_T}$$

$$= \arg\max_{\theta} n_H \log\theta + n_T \log(1-\theta)$$

$Ber(\theta)$

$$\frac{\partial}{\partial\theta} = \frac{n_H}{\theta} + \frac{n_T}{1-\theta}\cdot(-1) \Big|_{\hat{\theta}_{MLE}} = 0$$

$$\frac{n_H}{\theta} = \frac{n_T}{1-\theta}$$

$$(1-\theta) n_H = n_T \theta$$

$$n_H = (n_H + n_T)\theta$$

$$\Rightarrow \hat{\theta}_{MLE} = n_H / (n_H + n_T)$$

Categorical distribution

Data, D = rolls of a dice

- $P(1) = p_1$, $P(2) = p_2$, …, $P(6) = p_6$ $p_1 + \ldots + p_6 = 1$
- Rolls are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Categorical(θ) distribution where

$$\theta = \{p_1, p_2, \ldots, p_6\}$$

Choose θ that maximizes the probability of observed data aka "Likelihood"

$$P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6}$$

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

MLE of probability of rolls:

$$\widehat{\theta}_{MLE} = \hat{p}_{1,MLE}, \ldots, \hat{p}_{6,MLE}$$

$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y}$$

α_y ← Rolls that turn up y (no. of rolls that show y.)

$\sum_y \alpha_y$ ← Total number of rolls

"Frequency of roll y"

How to learn parameters from data? MLE

(Continuous case)

d-dim Gaussian distribution

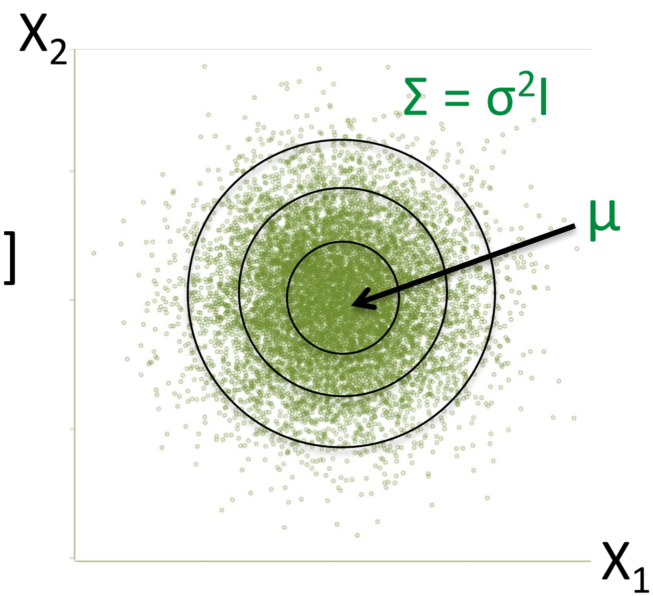
X is Gaussian N(μ, Σ)

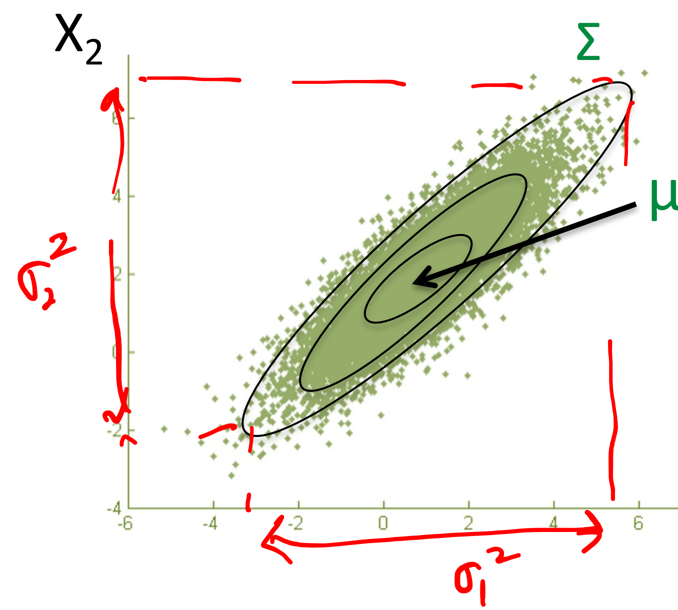
μ is d-dim vector, Σ is dxd dim matrix

$$P(X = x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

d=2

$X = [X_1; X_2]$



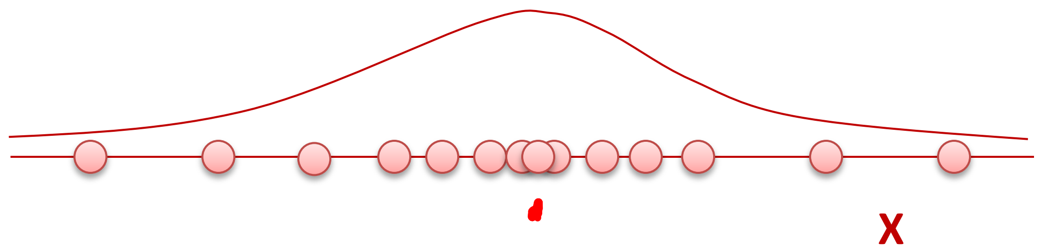


How to learn parameters from data? MLE

(Continuous case)

Gaussian distribution

Data, D =



- Parameters: μ – mean, σ^2 - variance

- Data are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Maximum Likelihood Estimation (MLE)

Choose θ= (μ,σ²) that maximizes the probability of observed data

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

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

Maximum Likelihood Estimation (MLE)

Choose θ= (μ,σ²) that maximizes the probability of observed data

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

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}$$ Identically distributed

Maximum Likelihood Estimation (MLE)

Choose θ= (μ,σ²) that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}$$ Identically distributed

$$= \arg\max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

MLE for Gaussian mean

➢ Poll

$$P(D|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

A. $\max_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$

B. $\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$

C. $\max_{\mu} \mu^2 - 2\mu \sum_{i=1}^{n} X_i$

D. $\max_{\mu} n\mu^2 - 2\mu \sum_{i=1}^{n} X_i$

$\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$

$\frac{\partial}{\partial\theta}$ $2\sum_{i=1}^{n}(X_i-\mu) = 0 \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} X_i$

MLE for Gaussian mean and variance

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$E[\hat{\mu}_{MLE}] = \mu$

unbiased

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Self exercise:

Derive MLE of variance?

Is the MLE of mean unbiased?
Is the MLE of variance unbiased?
How can you make it unbiased?

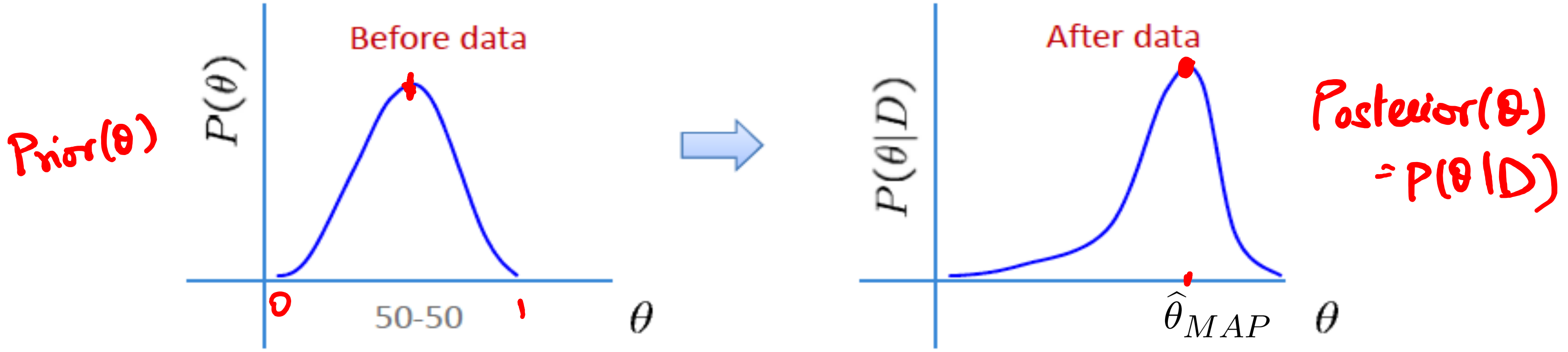
d-dimensional versions?

MLE for uniform or exponential distribution?

Max A Posteriori (MAP) estimation

Can we bring in prior knowledge if data is not enough?

- Assume a prior (before seeing data D) distribution P(θ) for parameters θ



- Choose value that maximizes a posterior distribution P(θ|D) of parameters θ

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$
$$= \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

How to choose prior distribution?

- $P(\theta)$
  - Prior knowledge about domain e.g. unbiased coin $P(\theta) = 1/2$
  - A mathematically convenient form e.g. "conjugate" prior

    If $P(\theta)$ is conjugate prior for $P(D|\theta)$, then Posterior has same form as prior

    Posterior ∝ Likelihood x Prior

    $P(\theta|D) \propto P(D|\theta) \times P(\theta)$

e.g.	Beta	Bernoulli	Beta	θ = bias
	Dirichlet	Categorical	Dirichlet	θ = bias vector
	Gaussian	Gaussian	Gaussian	θ = mean μ (known Σ)
	inv-Wishart	Gaussian	inv-Wishart	θ = cov matrix Σ (known μ)

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

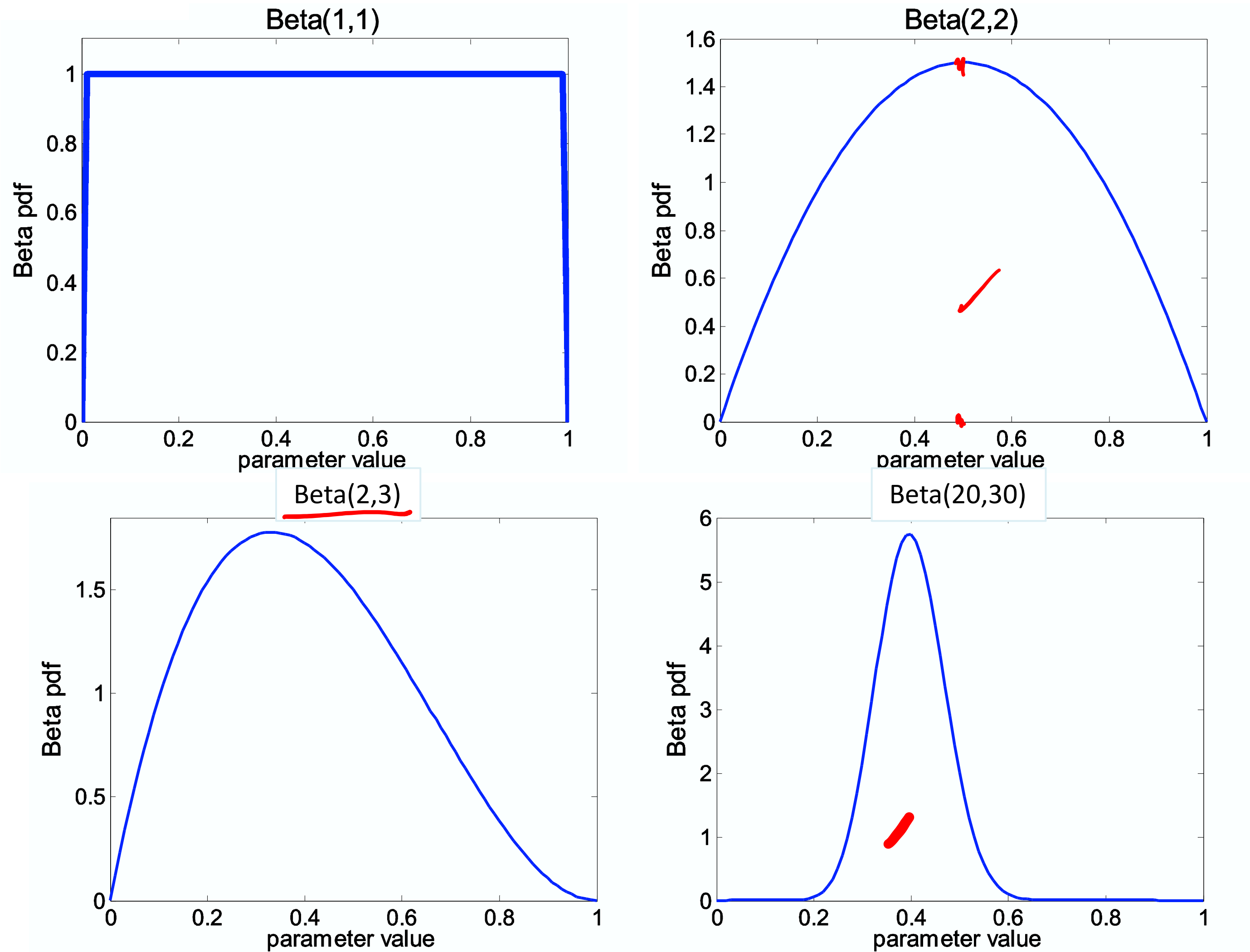
$$\widehat{\theta}_{MAP} = \arg\max_{\theta} \; P(\theta \mid D)$$
$$= \arg\max_{\theta} \; P(D \mid \theta) P(\theta)$$

MAP estimate of probability of head (using Beta conjugate prior):

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

Beta distribution

$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase



MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$
$$= \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

MAP estimate of probability of head (using Beta conjugate prior):

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

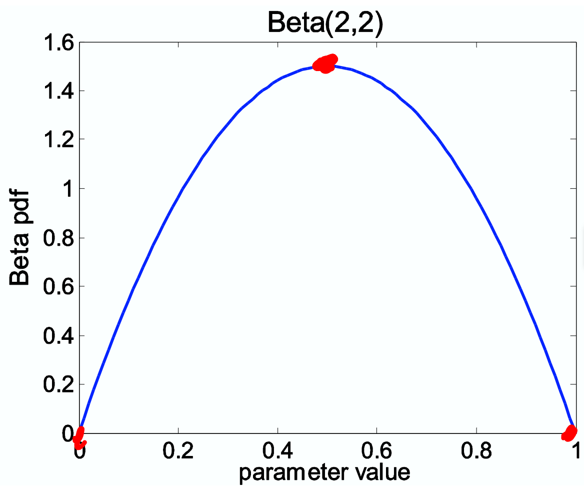
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

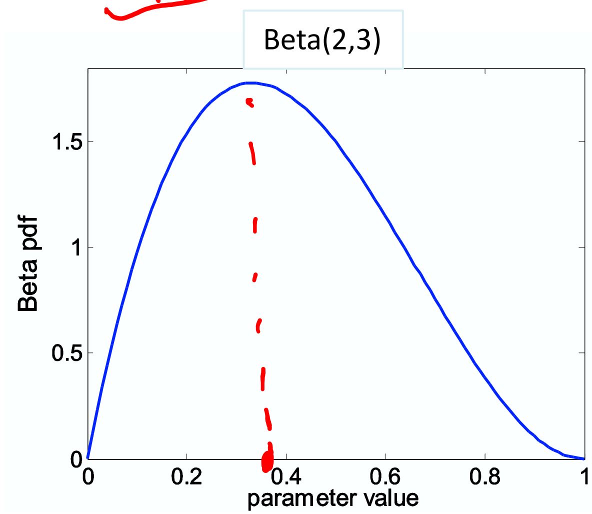
Count of H/T simply get added to parameters

$\alpha_H = n_H$ $\alpha_T = n_T$

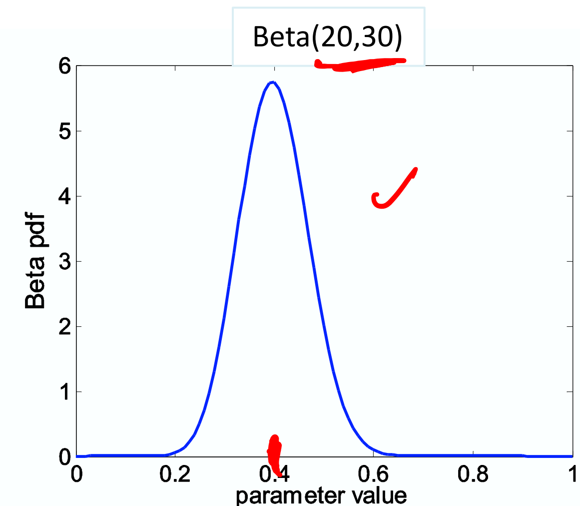
Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T) \qquad P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$





After observing 1 Tail



After observing 18 Heads and 28 Tails

As $n = \alpha_H + \alpha_T$ increases, posterior distribution becomes more concentrated

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$
$$= \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

MAP estimate of probability of head:

Mode $\frac{\beta_H - 1}{\beta_H - 1 + \beta_T - 1}$

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

Count of H/T simply get added to parameters

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

Equivalent to adding extra coin flips (β_H - 1 heads, β_T - 1 tails)

As we get more data, effect of prior is "washed out"

MAP estimation for Gaussian r.v.

Parameters $\theta = (\mu, \sigma^2)$

- Mean μ (known σ^2): Gaussian prior $P(\mu) = N(\eta, \lambda^2)$

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

$$\widehat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}} \qquad \widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

As we get more data, effect of prior is "washed out"

- Variance σ^2 (known μ): inv-Wishart Distribution

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When is MAP same as MLE?