Bayesian Methods

Rishabh Iyer
University of Texas at Dallas

based on the slides of Vibhav Gogate and Nick Rouzzi

Binary Variables

- Coin flipping: heads=1, tails=0 with bias μ

$$p(X = 1|\mu) = \mu$$

- Bernoulli Distribution

$$Bern(x|\mu) = \mu^x \cdot (1-\mu)^{1-x}$$

$$E[X] = \mu$$

$$var(X) = \mu \cdot (1-\mu)$$

$$E[x] = \sum_{x \in D} x\, p(x)$$

$$= 1 \cdot \mu + 0 \cdot (1-\mu) = \mu .$$

Binary Variables

m Heads , N − m tails.

- N coin flips: $X_1, \ldots, X_N$

$$p(\textstyle\sum_i X_i = m | N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Binomial Distribution

$$Bin(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$E\left[\sum_i X_i\right] = N\mu$$

$$var\left[\sum_i X_i\right] = N\mu(1-\mu)$$

N = # Coin Flips

m : Heads

N − m : Tails

$$E\left(\sum_{i=1}^{N} X_i\right) = \sum_i \underbrace{E(X_i)}_{\mu}$$

$$= N\mu.$$

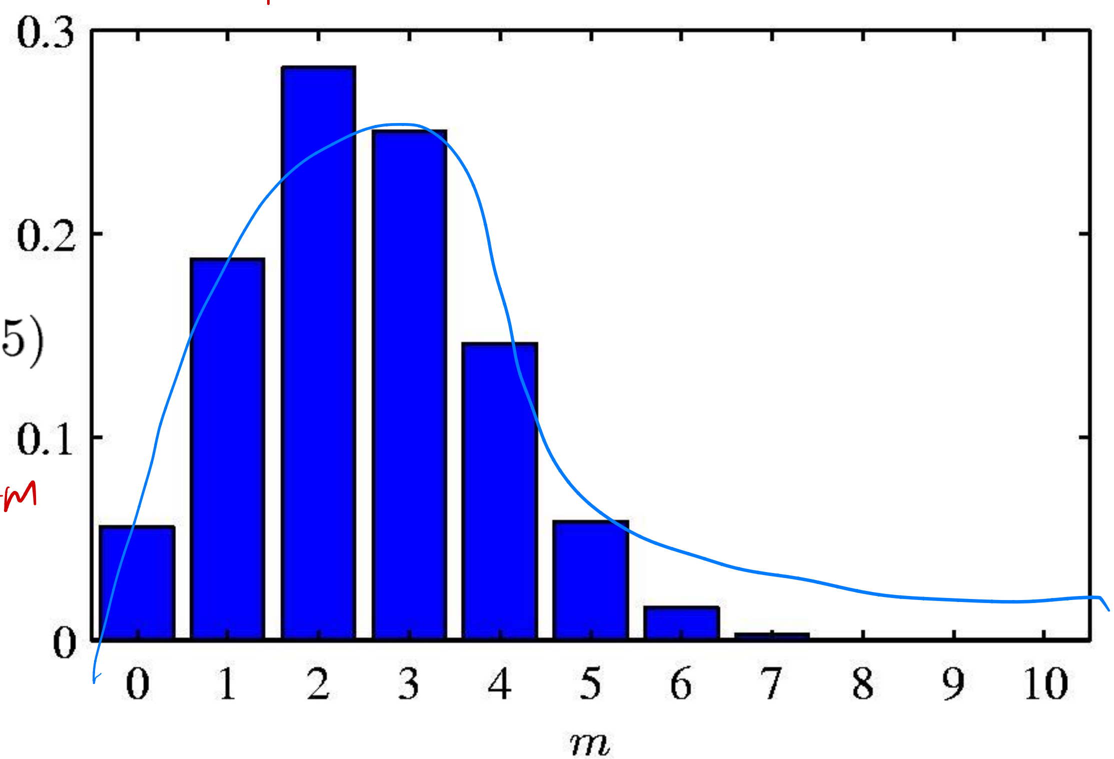
Binomial Distribution

$\mu = 0.25$

$N = 10$

$E(X) = N\mu = 2.5$

$\mathrm{Bin}(m|10, 0.25)$



$\binom{N}{m}\mu^{m}(1-\mu)^{N-m}$

Estimating the Bias of a Coin

- Suppose that we have a coin, and we would like to figure out what the probability is that it will flip up heads
  - How should we estimate the bias?

Estimating the Bias of a Coin

- Suppose that we have a coin, and we would like to figure out what the probability is that it will flip up heads
  - How should we estimate the bias?

  - With these coin flips, our estimate of the bias is: ?

Estimating the Bias of a Coin

- Suppose that we have a coin, and we would like to figure out what the probability is that it will flip up heads
  - How should we estimate the bias?

  - With these coin flips, our estimate of the bias is: 3/5
    - Why is this a good estimate?

Coin Flipping – Binomial Distribution

D1 D2 D3 D4 D5

- $P(Heads) = \theta,\ P(Tails) = 1 - \theta$
- Flips are i.i.d. (Independent & Identically Distributed)
  - Independent events

    $N = \alpha_H + \alpha_T$
  - Identically distributed according to Binomial distribution
- Our training data consists of α_H heads and α_T tails

$$p(D|\theta) = \theta^{\alpha_H} \cdot (1-\theta)^{\alpha_T}$$

α_H Heads

α_T Tails

Maximum Likelihood Estimation (MLE)

- **Data:** Observed set of α_H heads and α_T tails
- **Hypothesis:** Coin flips follow a Bernoulli distribution
- **Learning:** Find the "best" θ
- **MLE:** Choose θ to maximize probability of D given θ

$f(x_1) \geq f(x_2)$

$\log f(x_1) \geq \log f(x_2)$

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

Set derivative to zero, and solve!

$$\begin{aligned}\frac{d}{d\theta}\ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right] \\ &= \frac{d}{d\theta}\left[\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\right] \\ &= \alpha_H \frac{d}{d\theta}\ln\theta + \alpha_T \frac{d}{d\theta}\ln(1-\theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \quad = 0\end{aligned}$$

First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

Set derivative to zero, and solve!

$$\begin{aligned}\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right] \\ &= \frac{d}{d\theta}\left[\alpha_H \ln \theta + \alpha_T \ln(1-\theta)\right] \\ &= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1-\theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \quad = 0\end{aligned}$$

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

Dice Rolls

$$(x_1, \dots, x_M)$$

$$p(x_i|\theta) = \theta_1^{\mathbb{1}(x_i=1)} \theta_2^{\mathbb{1}(x_i=2)} \cdots \theta_5^{\mathbb{1}(x_i=5)} (1-\theta_1-\dots-\theta_5)^{\mathbb{1}(x_i=6)}$$

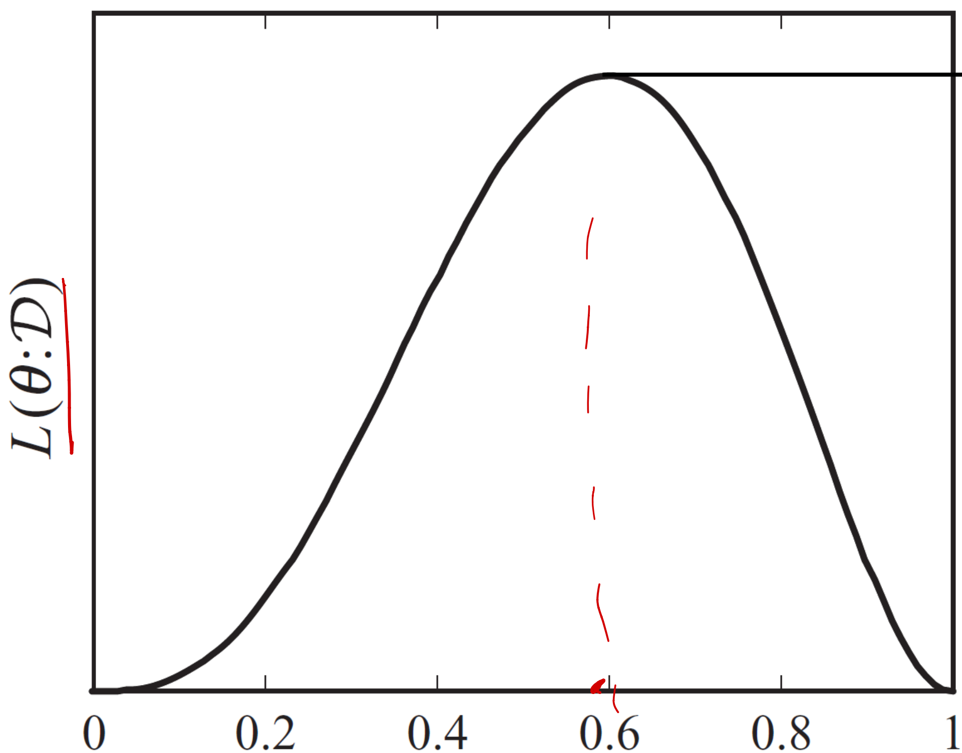
$$p(x_1 \dots x_M|\theta) = \theta_1^{\#1(x)} \theta_2^{\#2(x)} \cdots$$

$$p(\theta) \propto \theta_1^{\beta_1} \theta_2^{\beta_2} \cdots \theta_5^{\beta_5} (1-\theta_1-\dots-\theta_5)^{\beta_6}$$

MLE: $\max_{\theta} \log p(x_1 \dots x_M|\theta)$

Coin Flip MLE

$\frac{3}{5} = \hat{\theta}_{MLE}$



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

$p(D|\theta)$

Priors

- Suppose we have 5 coin flips all of which are heads
  - Our estimate of the bias is?

Priors

- Suppose we have 5 coin flips all of which are heads $\hat{\theta} = \frac{5}{5+0} = 1$
  - MLE would give $\theta_{MLE} = 1$
  - This event occurs with probability $\frac{1}{2^5} = \frac{1}{32}$ for a fair coin
  - Are we willing to commit to such a strong conclusion with such little evidence?

Priors

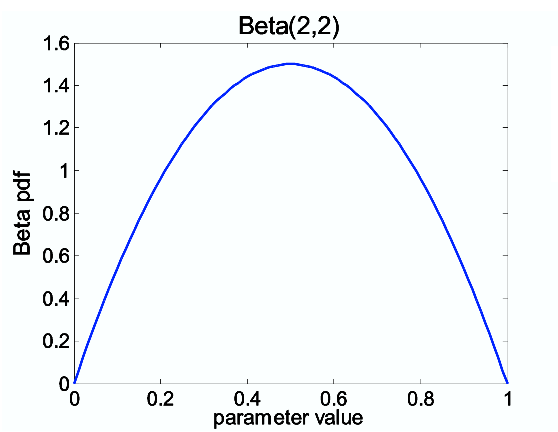
- Priors are a Bayesian mechanism that allow us to take into account "prior" knowledge about our belief in the outcome

- Rather than estimating a single θ, consider a distribution over possible values of θ given the data
  - Update our prior after seeing data

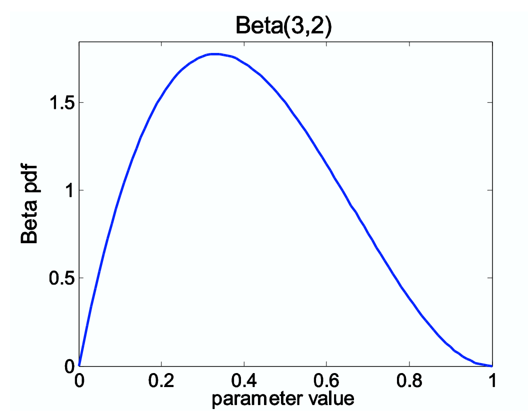
Our best guess in the absence of any data

Our estimate after we see some data

$\beta(\theta)$



Observe flips
e.g.: {tails, tails}



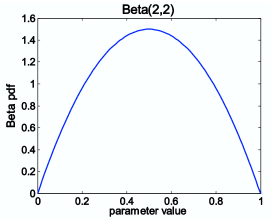
Bayesian Learning

Apply Bayes rule:

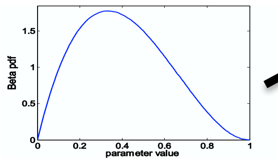
MLE: $\max_{\theta} p(\theta|D)$

Data Likelihood

Prior



Posterior



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

$p(D) = p(X_1, \ldots X_M)$

Normalization

(Ind of θ)

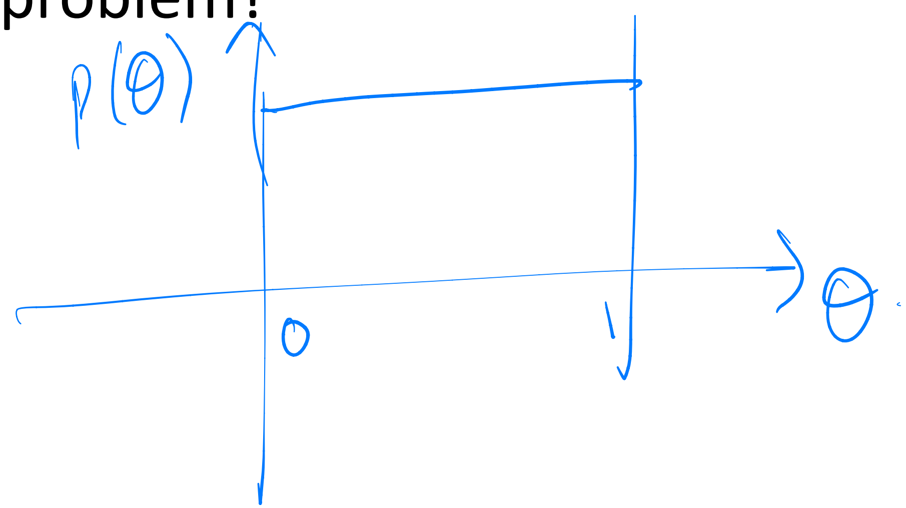
- Or equivalently: $p(\theta|D) \propto p(D|\theta)p(\theta)$
- For uniform priors this reduces to the MLE objective

$$p(\theta) \propto 1 \qquad \Rightarrow \qquad p(\theta|D) \propto p(D|\theta)$$

$p(D) = \int_{\theta=0}^{1} p(X, \theta)$

Picking Priors

- How do we pick a good prior distribution?
  - Could represent expert domain knowledge (makes sense)
  - Statisticians choose them to make the posterior distribution "nice" (conjugate priors) (looks nice)
- What is a good prior for the bias in the coin flipping problem?



Picking Priors

- How do we pick a good prior distribution?
  - Could represent expert domain knowledge
  - Statisticians choose them to make the posterior distribution “nice” (conjugate priors)
- What is a good prior for the bias in the coin flipping problem?
  - Truncated Gaussian (tough to work with)
  - Beta distribution (works well for binary random variables)

Coin Flips with Beta Distribution

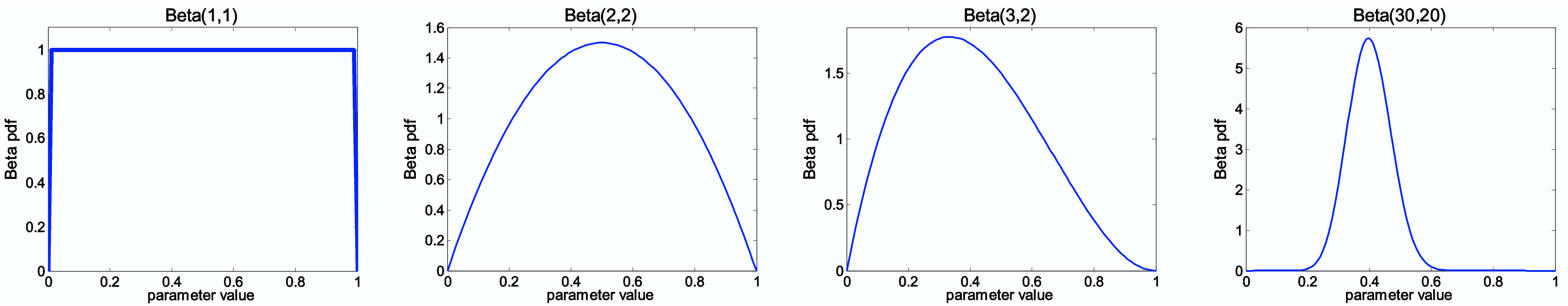
Likelihood function:

\# Heads   \# Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Prior:

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

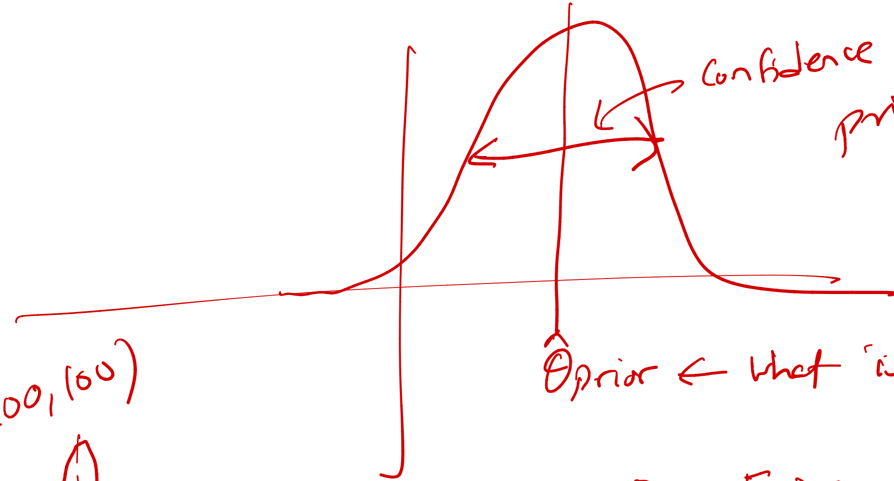


$$P(\theta \mid \mathcal{D}) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\ \theta^{\beta_H-1}(1-\theta)^{\beta_T-1}$$

$$= \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

$$= Beta(\alpha_H+\beta_H, \alpha_T+\beta_T)$$

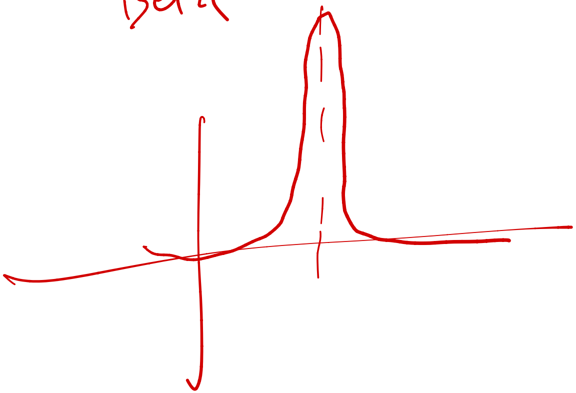
Prior knowledge



confidence on my prior

$\hat{\theta}_{prior}$ ← what is my prior belief

Beta(100, 100)



e.g.: Fair coin.

$\hat{\theta}_{prior} = 0.5$

lower value of $\alpha, \beta \Rightarrow$ Flat / Less peak.

MAP Estimation

- Choosing θ to maximize the posterior distribution is called maximum a posteriori (MAP) estimation

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D)$$

- The only difference between θ_{MLE} and θ_{MAP} is that one assumes a uniform prior (MLE) and the other allows an arbitrary prior

Priors

- Suppose we have 5 coin flips all of which are heads
  - MLE would give $\theta_{MLE} = 1$
  - MLE with a $Beta(2,2)$ prior gives $\theta_{MAP} = \frac{6}{7} \approx .857$
  - As we see more data, the effect of the prior diminishes
    - $\theta_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \approx \frac{\alpha_H}{\alpha_H + \alpha_T}$ for large # of observations

MAP for Coin Flipping

Dataset: α_H heads, α_T tails (D)

$$p(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

$D = \{x_1, x_2 \ldots x_N\}$ $\quad \alpha_H + \alpha_T = N.$

$x_i \in \{0,1\}$

MLE

$$\max_\theta \log p(D|\theta).$$

$$\max_\theta \log\left[\theta^{\alpha_H}(1-\theta)^{\alpha_T}\right]$$

$$\alpha_H \log\theta + \alpha_T \log(1-\theta).$$

$$\nabla LL_\theta = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0 \Rightarrow \theta = \frac{\alpha_H}{\alpha_H+\alpha_T}$$

MAP. $[\beta_H \sim 2, \beta_T \sim 2]$

$$\max_\theta \log\left[p(D|\theta)\,p(\theta)\right]$$

$$p(\theta) \propto \theta^{\beta_H-1}(1-\theta)^{\beta_T-1}$$

$$p(D|\theta)p(\theta) = \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

$$\theta_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

$$\alpha_H = 5, \ \alpha_T = 0$$

$$\theta_{MLE} = 1.$$

$$\beta_H = 2, \ \beta_T = 2. \ / \ \beta_H = 3, \ \beta_T = 3$$

$$\theta_{MAP} = \frac{5+2-1}{5+2+2-2} = \frac{6}{7}$$

$$\frac{5+3-1}{5+3+3-2} = \frac{7}{9}.$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound

  - Suppose $Y_1, \dots, Y_N$ are i.i.d. random variables taking values in $\{0, 1\}$ such that $E_p[Y_i] = y$. For $\epsilon > 0$,

$$p\left(\left|y - \frac{1}{N}\sum_i Y_i\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound

  - For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p\left(\left|\theta_{true} - \frac{1}{N}\sum_i X_i\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound

  - For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p(|\theta_{true} - \theta_{MLE}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

$\epsilon = 10^{-2}$

$\delta = 10^{-4}$

- Can use Chernoff bound
    - For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

δ

$\epsilon = 10^{-3}$

$\frac{1}{\epsilon^2} = 10^6$

$$p(|\theta_{true} - \theta_{MLE}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

$$\underline{\delta \geq 2e^{-2N\epsilon^2}} \Rightarrow \boxed{N \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}}$$

$\epsilon = 10^{-5}$

$= 10^{10}$