



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

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

$$E[X] = \mu$$

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

$$\sum_{x \in \alpha} x p(x=x) \quad / \quad \int_{\alpha} x p(x)$$

Binary Variables

- N coin flips: X_1, \dots, X_N

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

- Binomial Distribution

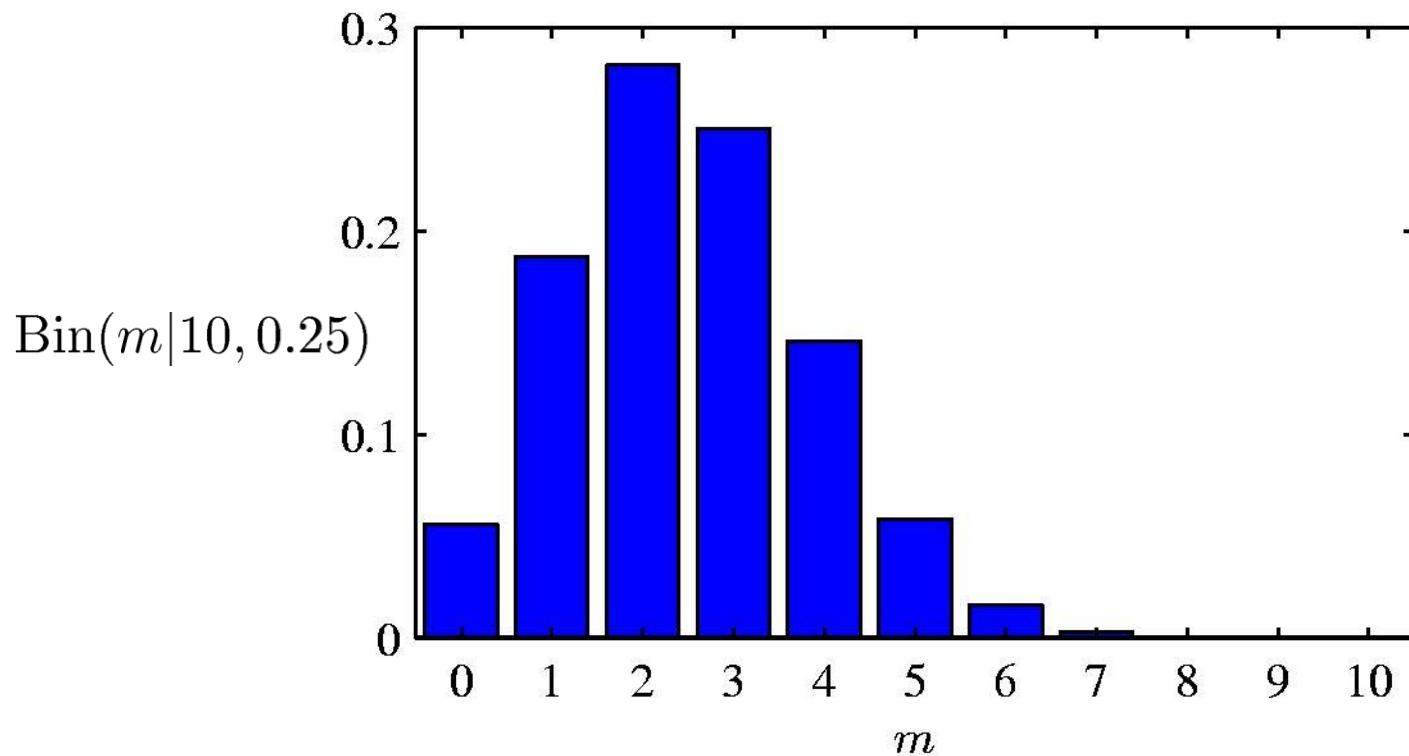
$$Bin(m | N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

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

\sum_m Bin(m), m

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

Binomial Distribution



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



- $P(\text{Heads}) = \theta, P(\text{Tails}) = 1 - \theta$
- Flips are i.i.d. (Independent & Identically Distributed)
 - Independent events
 - Identically distributed according to Binomial distribution
- Our training data consists of α_H heads and α_T tails

$$p(D|\theta) = \underbrace{\theta^{\alpha_H}}_{\text{heads}} \cdot \underbrace{(1-\theta)^{\alpha_T}}_{\text{tails}}$$
$$p(D|\theta) = \prod_{i=1}^{|\mathcal{D}|} p(D_i|\theta)$$

$$P(D|\theta) = P(x_1, x_2, x_3, \dots, x_N | \theta)$$
$$= \prod_{i=1}^n P(x_i | \theta) \quad \text{--- Independent.}$$

$$= \prod_{i=1}^n \text{Bern}(x_i | \theta) \quad \text{--- Inden. Dist.}$$

$$= \prod_{i=1}^n \mu^{1(x_i=H)} (1-\mu)^{1(x_i=T)}$$

$$= \mu^{d_H} (1-\mu)^{d_T}$$

$$d_H = \# \text{Heads}, \quad d_T = \# \text{Tails} \quad | \quad d_H + d_T = N$$

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

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\mathcal{D} | \theta) \\ &= \arg \max_{\theta} \underbrace{\ln P(\mathcal{D} | \theta)}_{\text{Log-Likelihood}}\end{aligned}$$

First Parameter Learning Algorithm



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

Set derivative to zero, and solve!

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

$$P_1 \geq P_2$$

$$\log P_1 \geq \log P_2$$

First Parameter Learning Algorithm



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

Set derivative to zero, and solve!

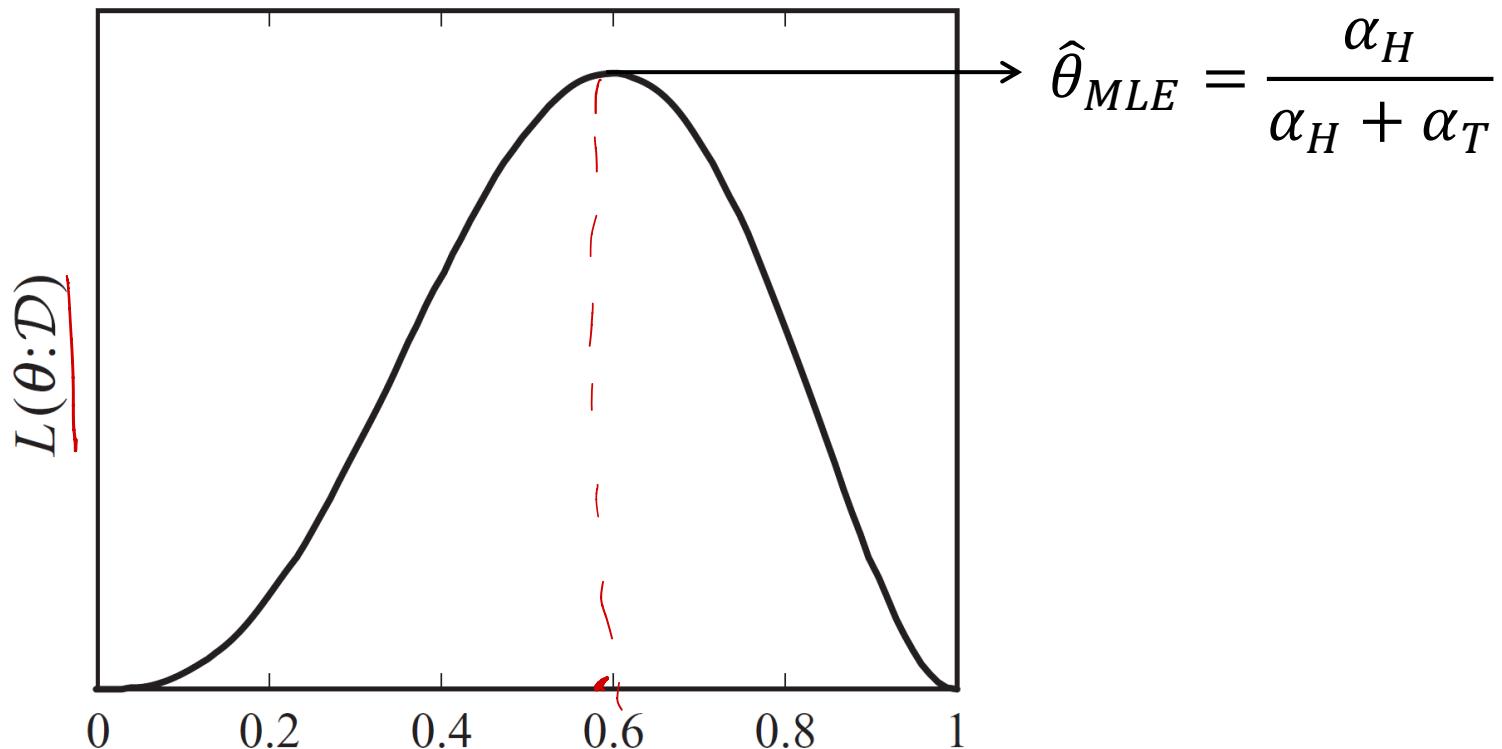
$$\begin{aligned}\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) &= \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}] \\ &= \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1 - \theta)] \\ &= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1 - \theta) \\ &= \frac{\alpha_H}{\theta}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

$$N = \alpha_H + \alpha_T$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{N}$$

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

Coin Flip MLE



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

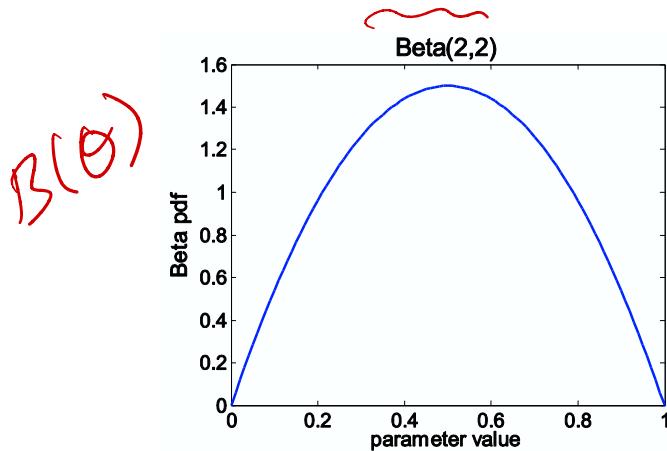
$$\text{Beta}(\alpha, \beta)$$

$$p(\theta)$$



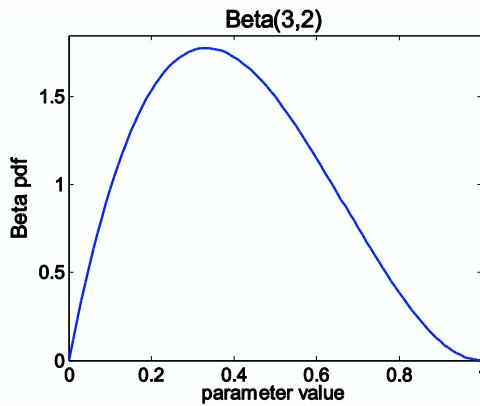
- Priors are a Bayesian mechanism that allow us to take into account “prior” knowledge about our belief in the outcome
- Rather than estimating a single θ , consider a distribution over possible values of θ given the data
 - Update our prior after seeing data

Our best guess in the absence of any data



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

Our estimate after we see some data



Bayesian Learning

$$D = \{(x_1, y_1), \dots, (x_m, y_m)\}$$

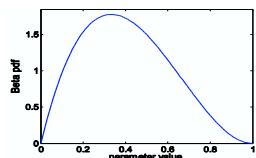

Apply Bayes rule:

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

Data Likelihood



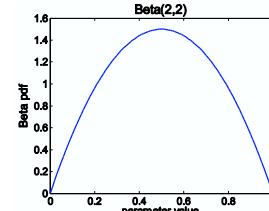
Posterior



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

Bayes Rule

Prior



$$p(D) = P(X_1, \dots, X_m)$$

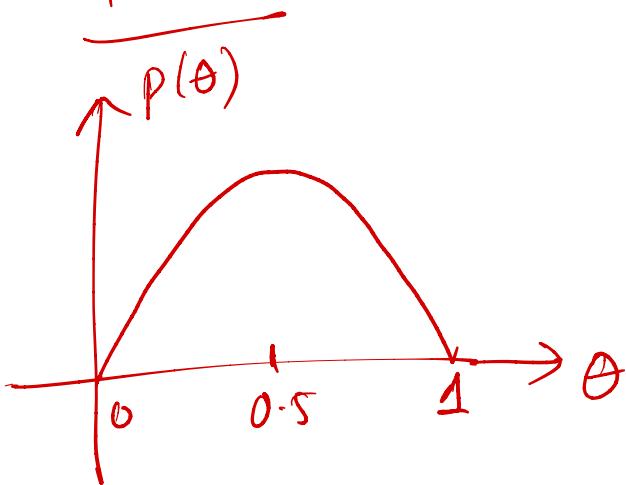
Normalization
(Ind of θ)

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

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

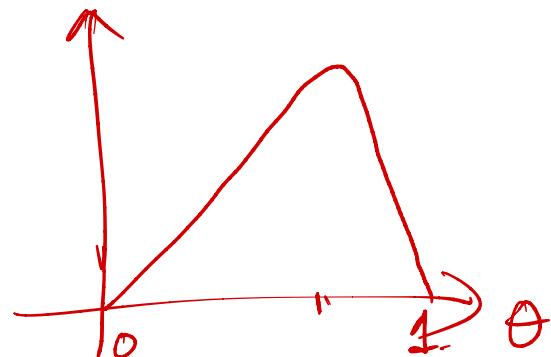
$$P(D) = \int_{\theta=0}^{\infty} p(X_1, \theta) d\theta$$

Prior



Posterior

$$p(\theta | D)$$



Dice Rolls

(x_1, \dots, x_m)

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

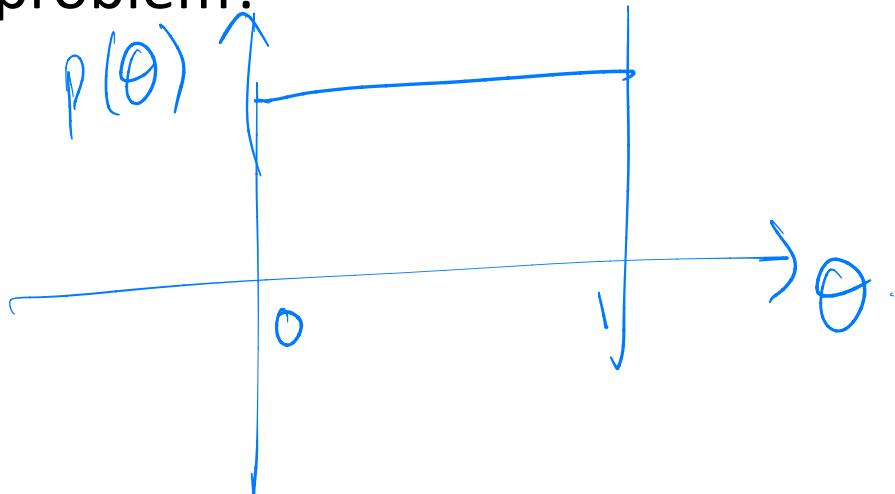
$$p(x_1 \dots x_m | \theta) = \theta_1^{\#1(x)} \theta_2^{\#2(x)} \cdots$$

$$p(\theta) \propto \theta_1^{\#1} \theta_2^{\#2} \cdots \theta_5^{\#5} (1 - \theta_1 - \cdots - \theta_5)^{\#6}$$

$$\text{MLE: } \underset{\theta}{\operatorname{max}} \log p(x_1 \dots x_m | \theta)$$

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

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

$\beta_H = 100$

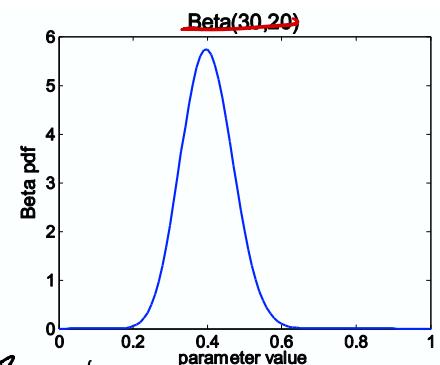
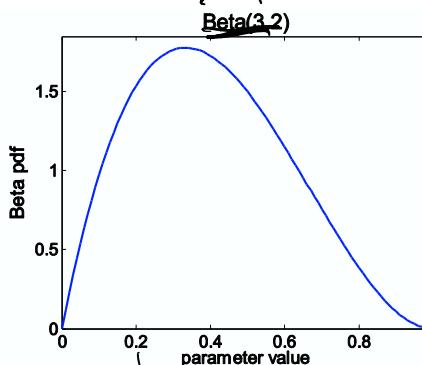
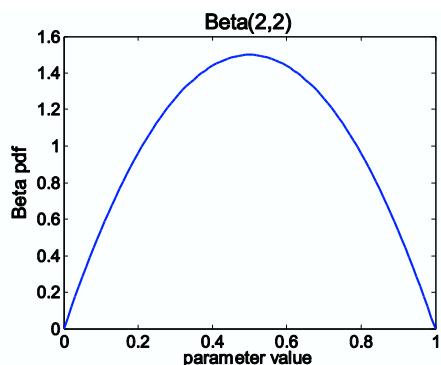
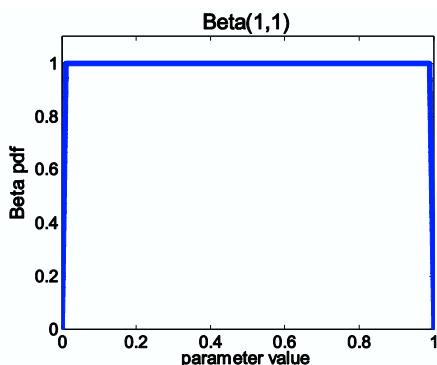
Prior:

Beta(2, 3)

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

Beta(2, 3)

Beta(20, 20)



$$B(\beta_H, \beta_T) = \int_0^1 \theta^{\beta_H-1} (1 - \theta)^{\beta_T-1}$$

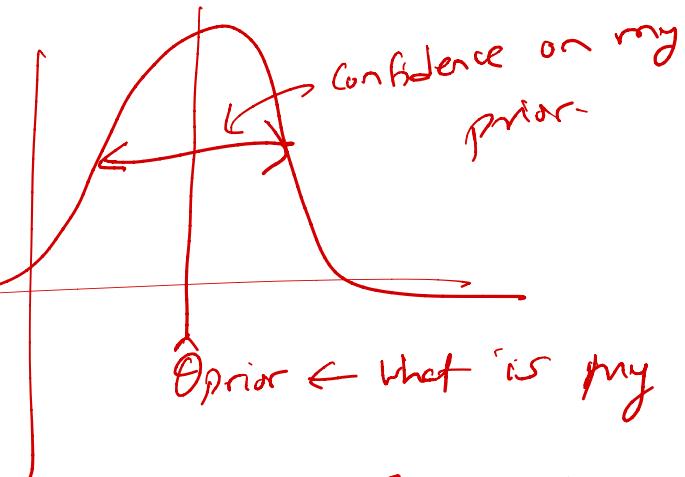
$$P(\theta | \mathcal{D}) \propto \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \theta^{\beta_H-1} (1 - \theta)^{\beta_T-1}$$

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

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

Prior knowledge

Beta(100, 50)



e.g.: Fair coin.

$$\stackrel{\wedge}{\text{Prior}} = 0.5$$

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

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 $\underline{\theta_{MLE}} = 1$
 - MLE with a $Beta(2,2)$ prior gives $\underline{\theta_{MAP}} = \frac{6}{7} \approx .857$
 - As we see more data, the effect of the prior diminishes

$$p(\theta | D) \propto p(D|\theta) p(\theta)$$
$$\text{Beta}(\alpha_H, \alpha_T) \quad \text{Beta}(\beta_H, \beta_T)$$

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

for large # of observations

MAP for Coin Flipping

Dataset: α_H heads, α_T tails (D)

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

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

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

MLE

$$\max_{\theta} \log P(D|\theta).$$

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

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

$$\partial L \partial \theta = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0 \Rightarrow \theta = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

$$\text{MAP. } [p_H \sim 2, p_T \sim 2]$$

$$\max_{\theta} \log [P(D|\theta) P(\theta)]$$

$$P(\theta) \propto \theta^{p_H-1} (1-\theta)^{p_T-1}$$

$$P(D|\theta) P(\theta) = \theta^{\alpha_H + p_H - 1} (1-\theta)^{\alpha_T + p_T - 1}$$

$$\theta_{\text{MAP}} = \frac{\alpha_H + p_H - 1}{\alpha_H + p_H + \alpha_T + p_T - 2}$$

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

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

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

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

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

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, \dots, X_n iid coin flips and $\epsilon > 0$,

$$\epsilon \approx 10^{-3}$$

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

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, \dots, 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, \dots, X_n iid coin flips and $\epsilon > 0$,

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

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