Maximum a-Posteriori Estimation

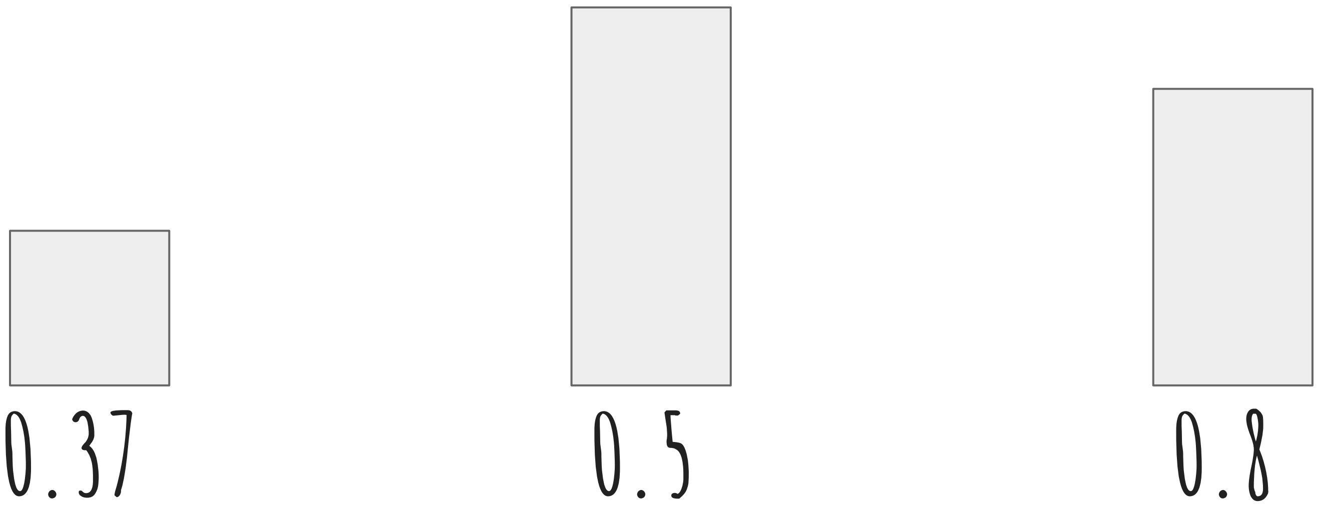
Aleks Jovcic
Slides by Joshua Fan & Alex Tsun

Agenda

- The Beta Random Variable
- MAP Estimation
- MAP Example

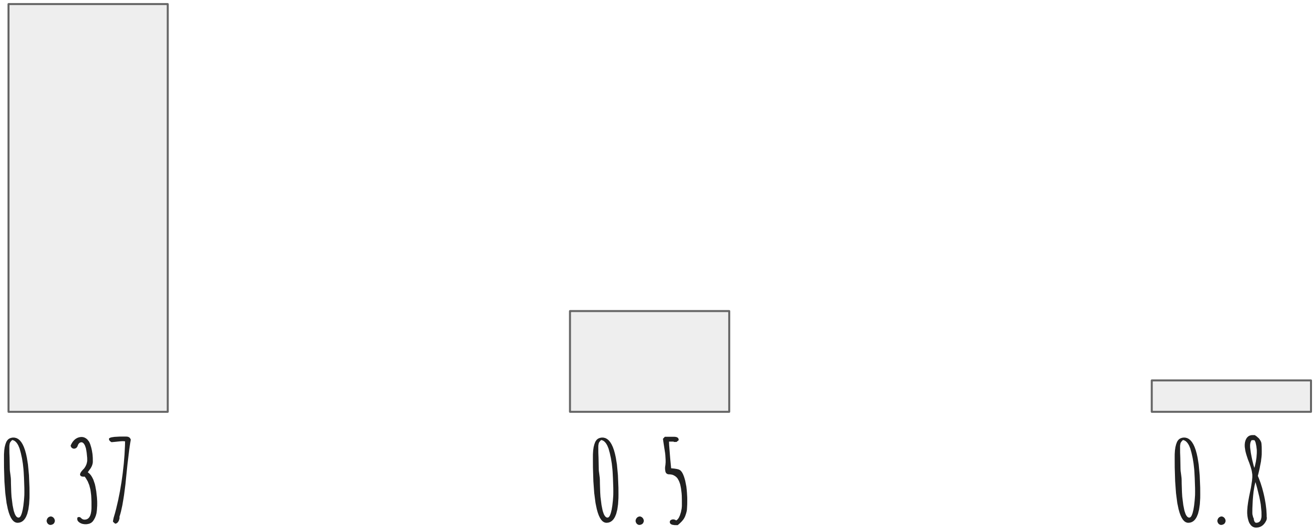
Beta Random Variable (Intuition)

Suppose you want to model your belief on the unknown probability X of heads. You could assign a probability distribution as follows:



Beta Random Variable (Intuition)

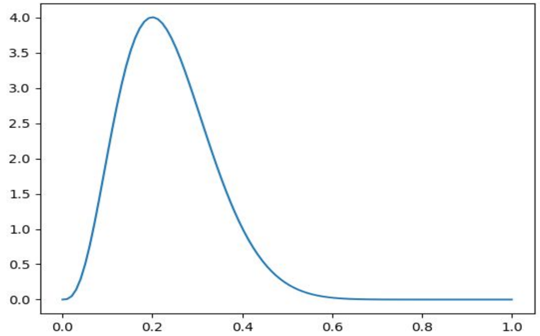
Suppose you want to model your belief on the unknown probability X of heads. You could assign a probability distribution as follows:

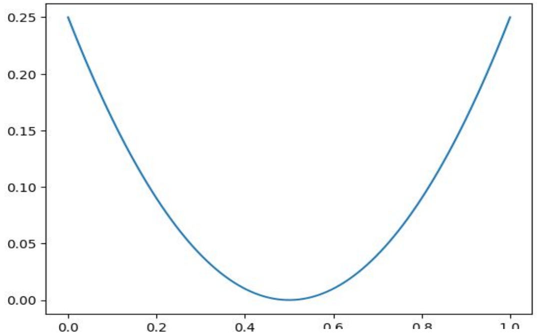


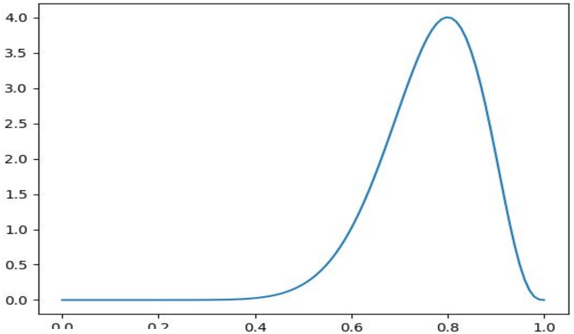
Beta Random Variable (Intuition)

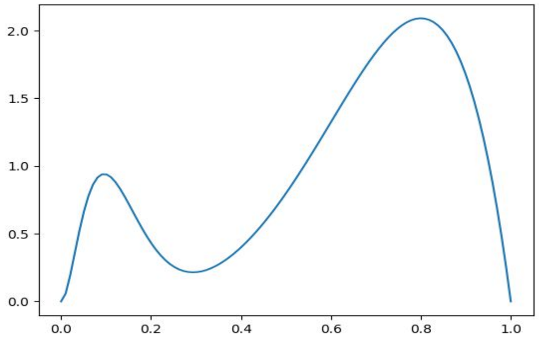
What if you wanted to be open to any value in [0,1]?

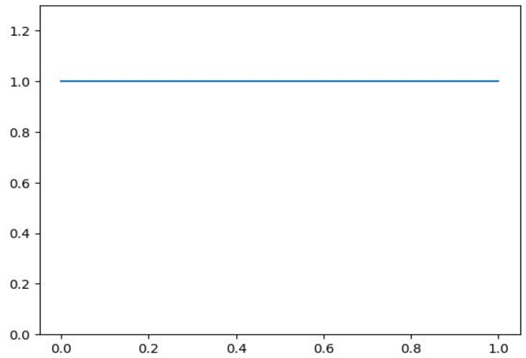
Need to use a *continuous* RV (with range [0,1])!







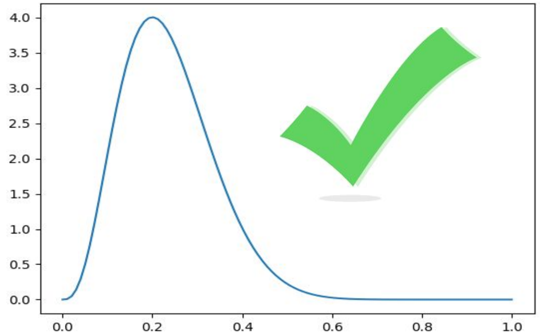


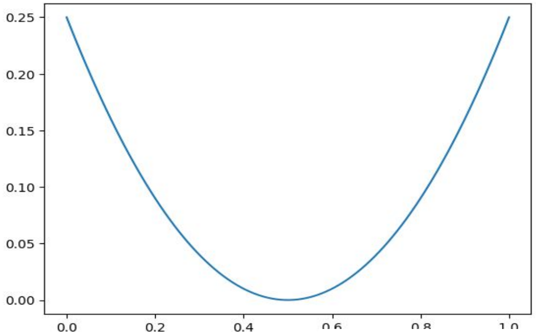


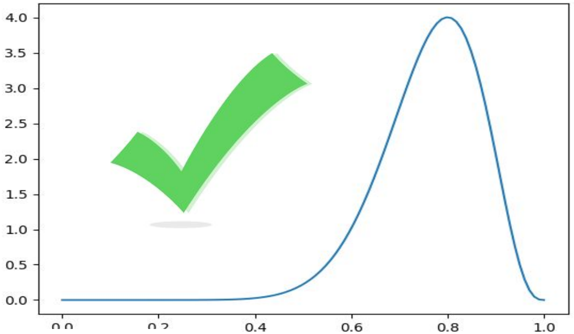
Beta Random Variable (Intuition)

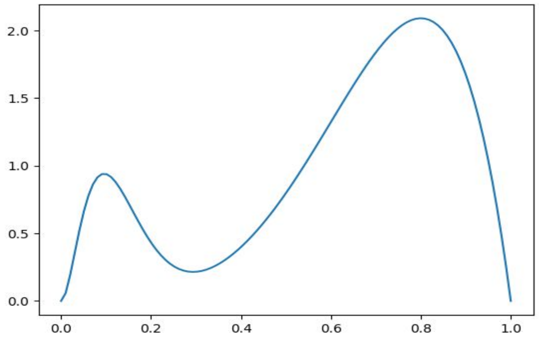
What if you wanted to be open to any value in [0,1]?

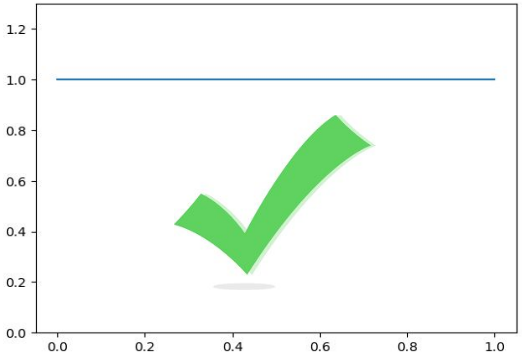
Need to use a *continuous* RV (with range [0,1])!







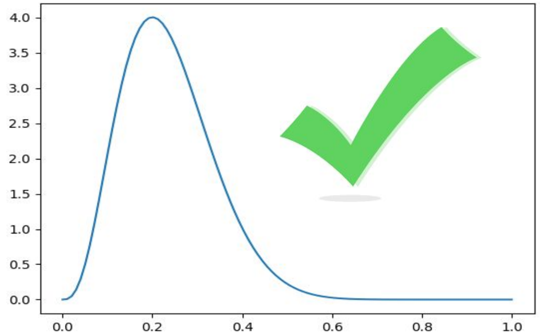


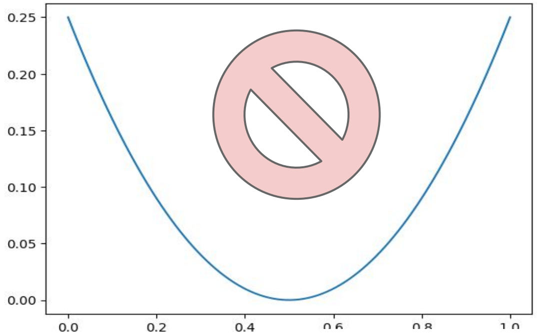


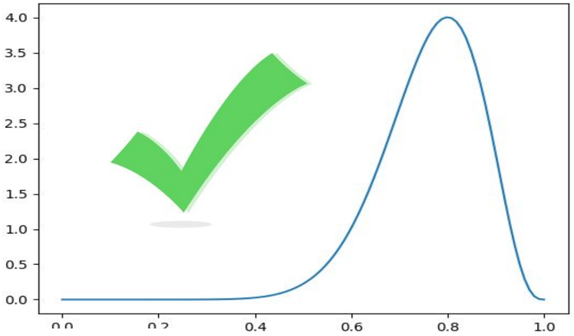
Beta Random Variable (Intuition)

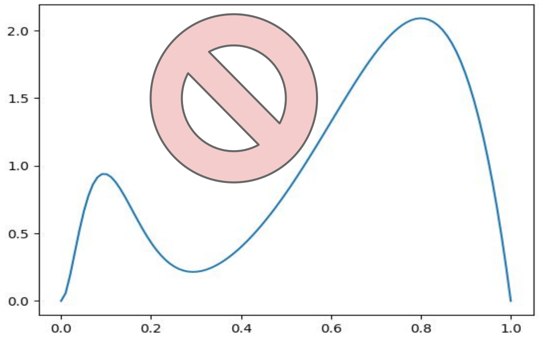
What if you wanted to be open to any value in [0,1]?

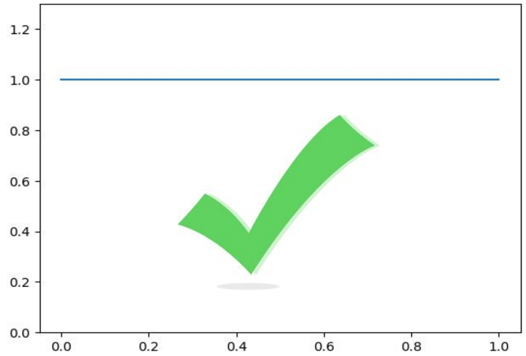
Need to use a *continuous* RV (with range [0,1])!







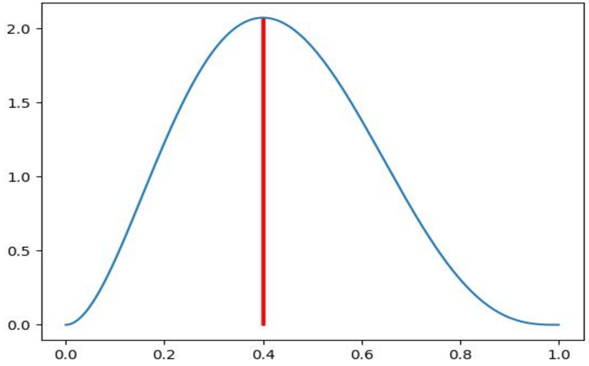


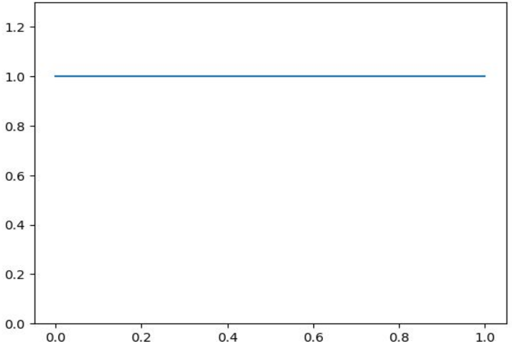


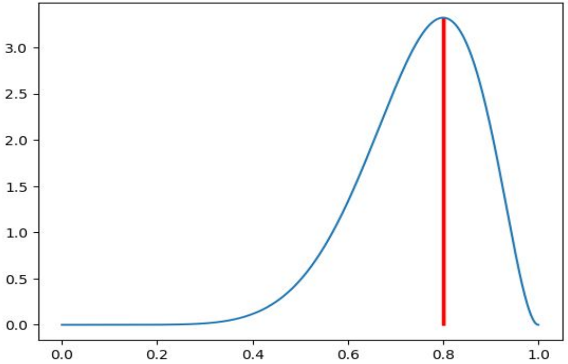
Beta Random Variable (Intuition)

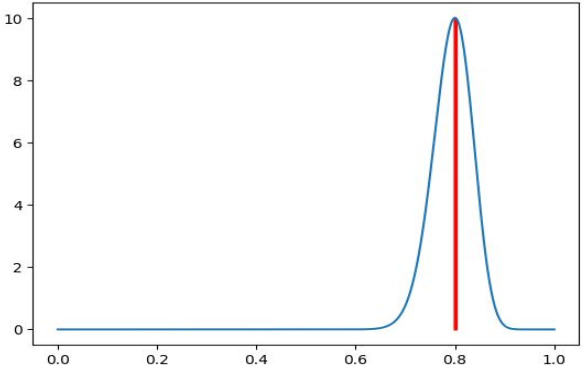
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





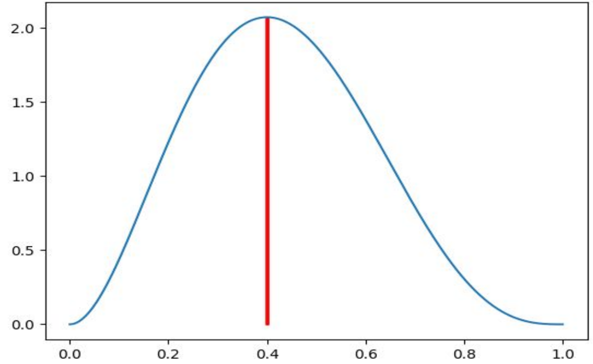


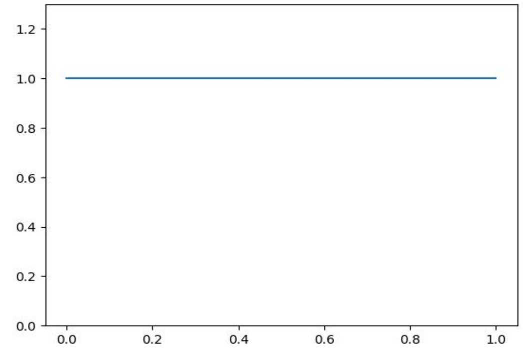


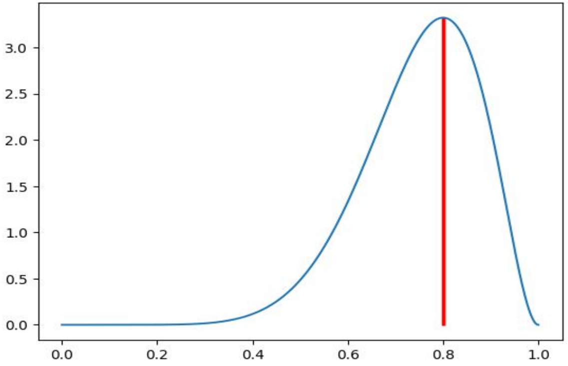
Beta Random Variable (Intuition)

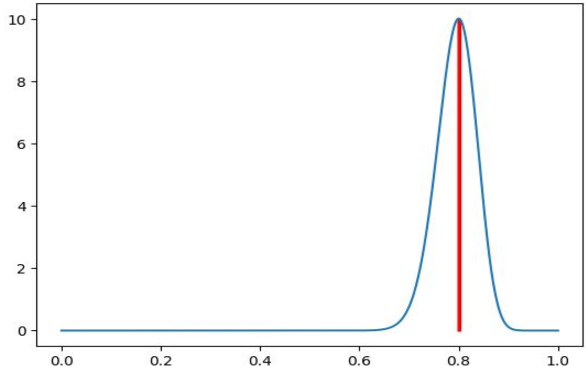
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





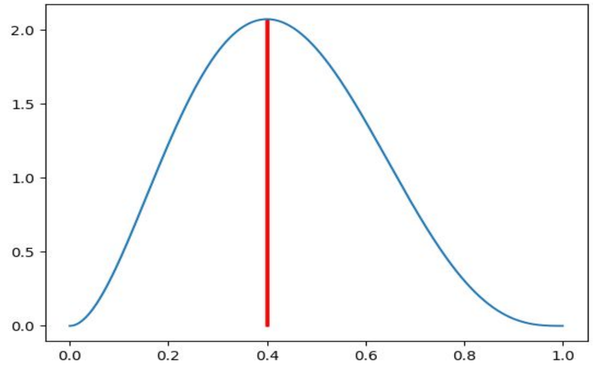


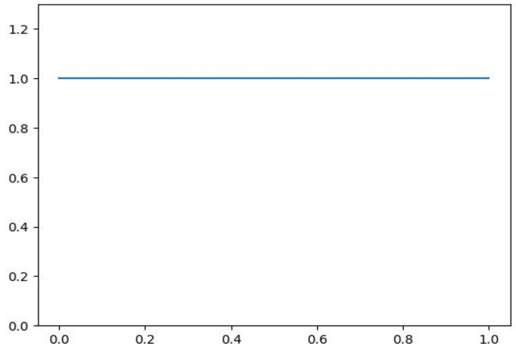


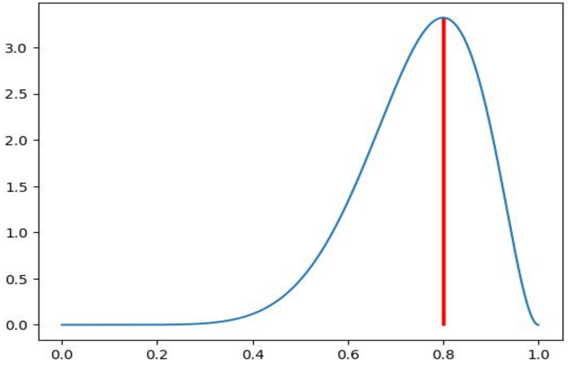
Beta Random Variable (Intuition)

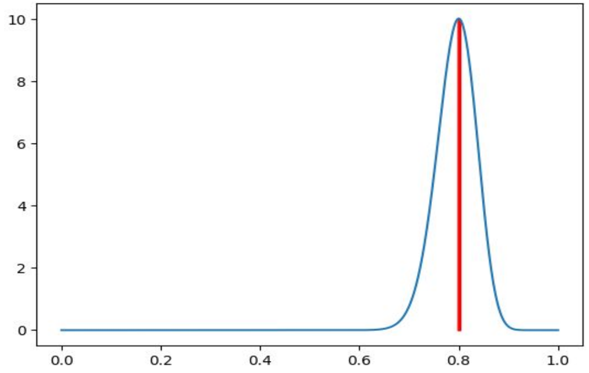
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





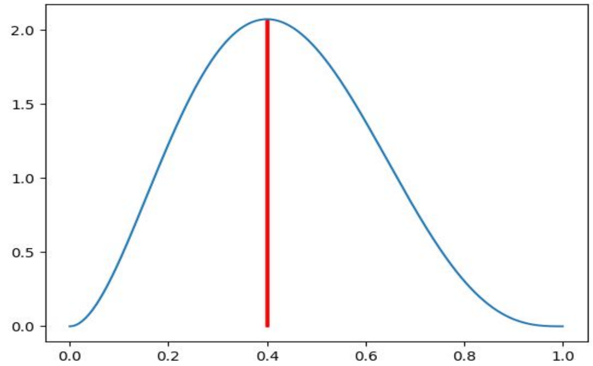


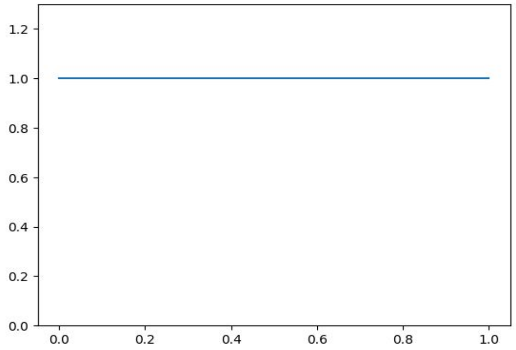


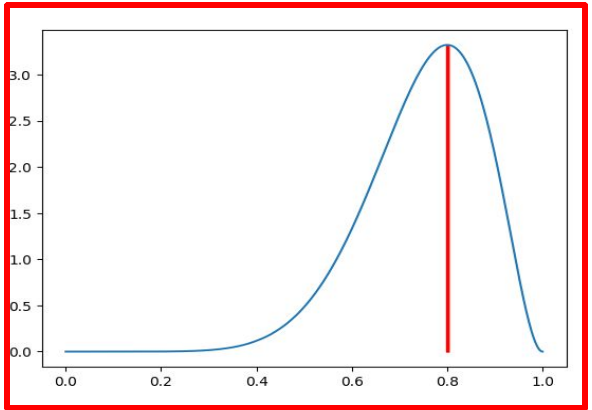
Beta Random Variable (Intuition)

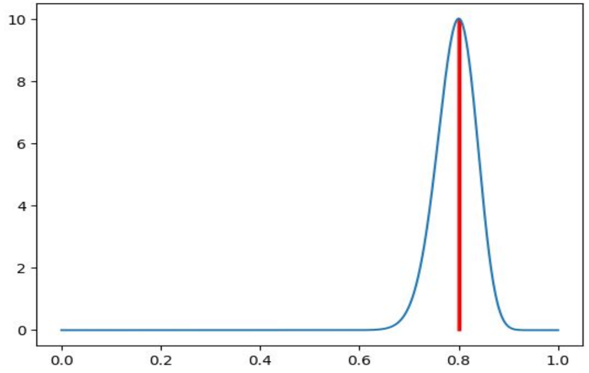
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





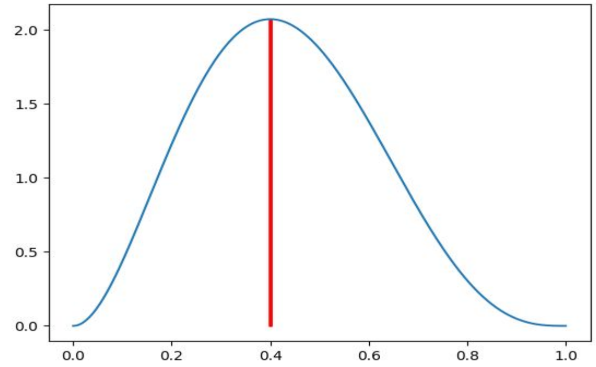


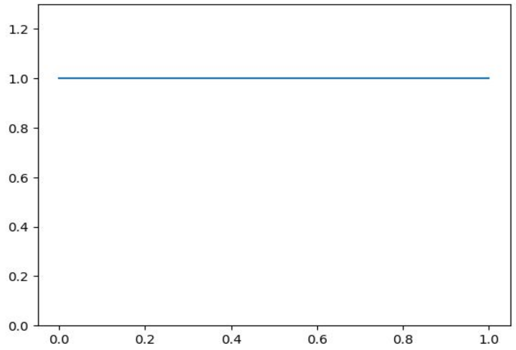


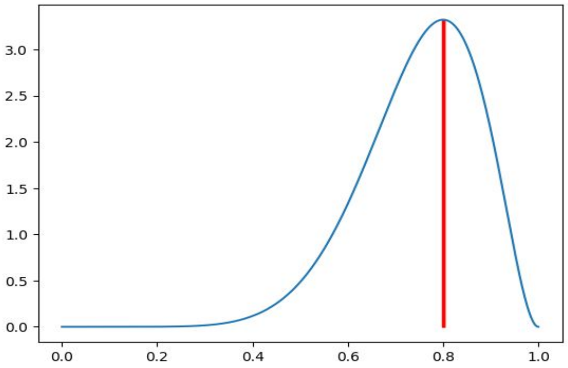
Beta Random Variable (Intuition)

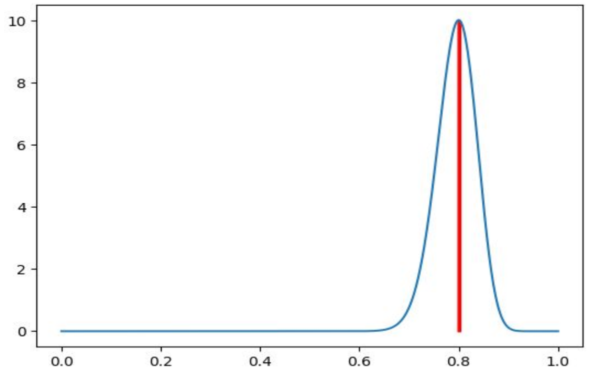
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





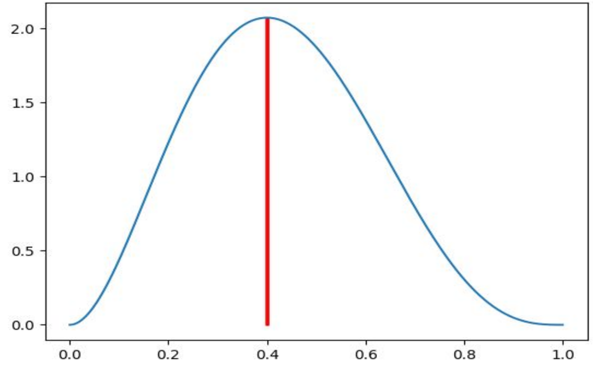


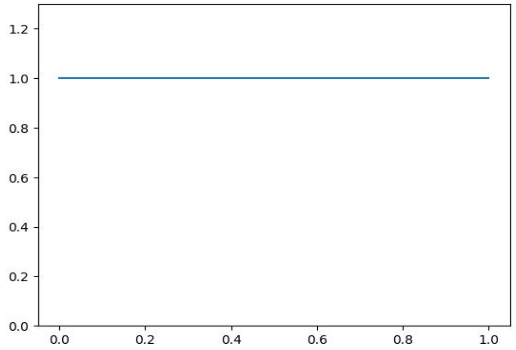


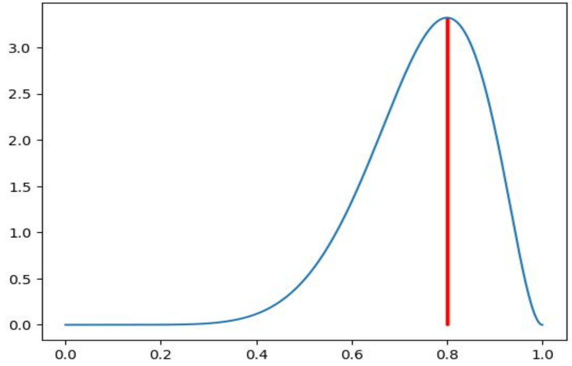
Beta Random Variable (Intuition)

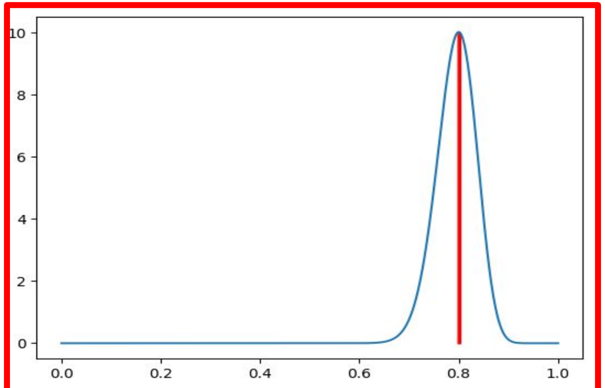
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





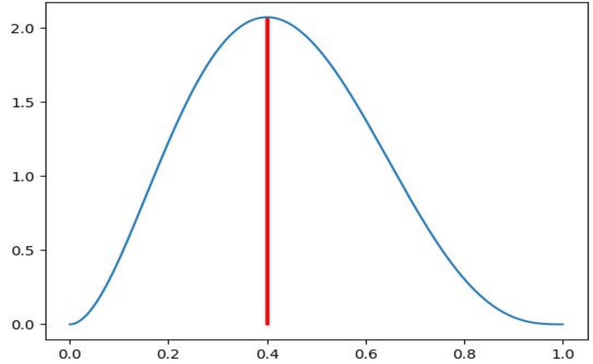


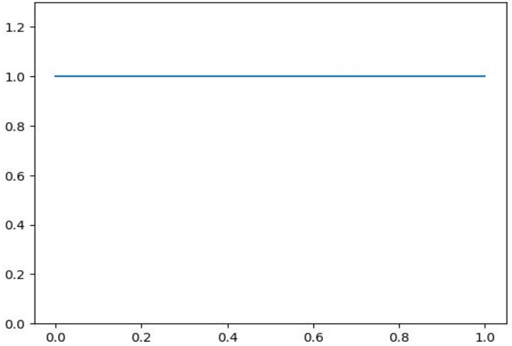


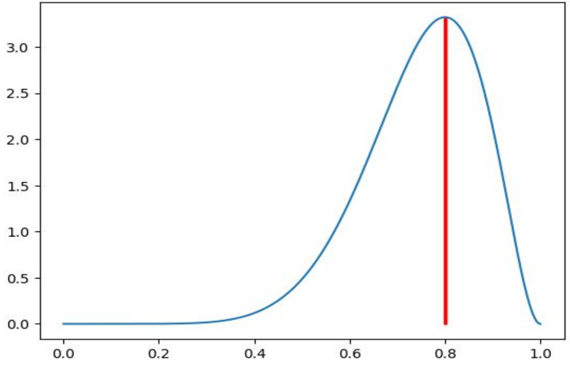
Beta Random Variable (Intuition)

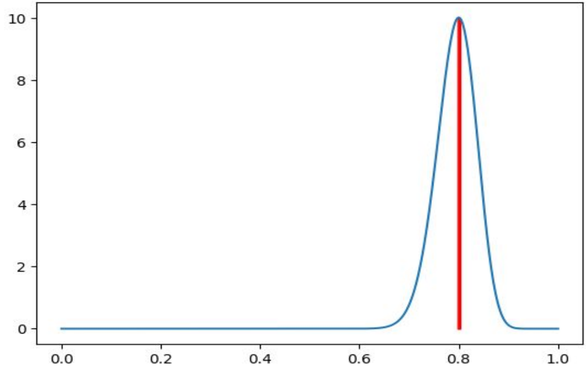
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?





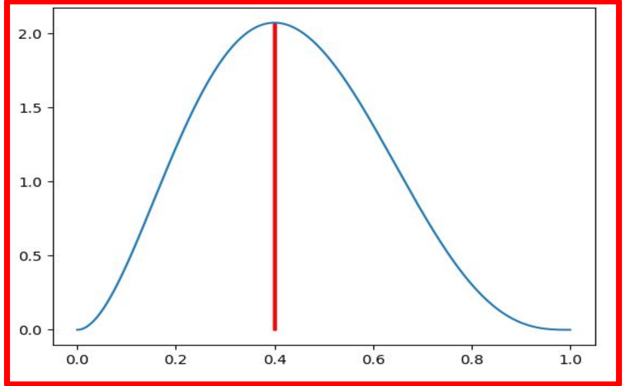


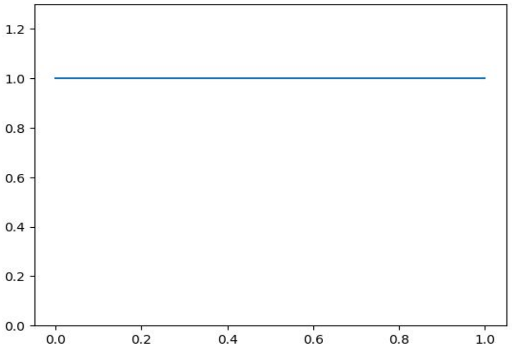


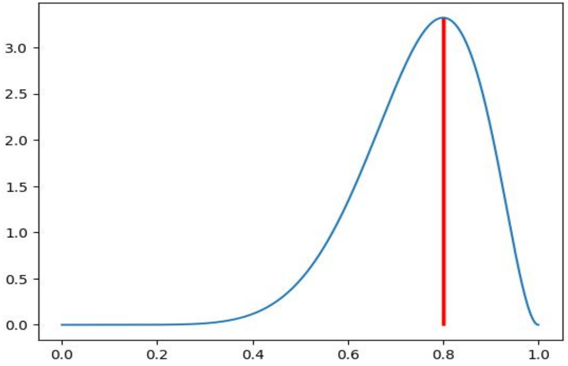
Beta Random Variable (Intuition)

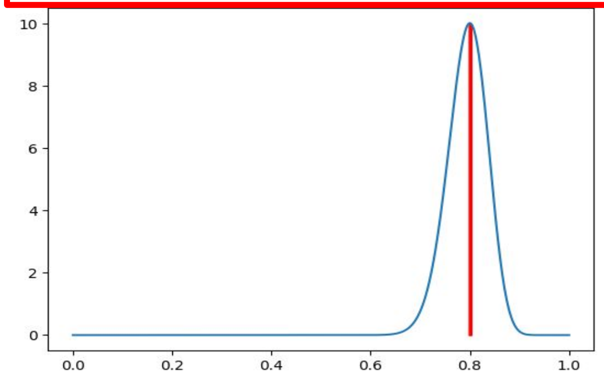
If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?









The Beta Random Variable

Beta RV: $X \sim Beta(\alpha, \beta)$, if and only if X has the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

The Beta Random Variable

Beta RV: $X \sim Beta(\alpha, \beta)$, if and only if X has the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

X is typically the belief distribution about some unknown probability of success, where we pretend we saw $\alpha - 1$ successes and $\beta - 1$ failures ahead of time. Hence, the mode, $\arg\max_{x \in [0,1]} f_X(x)$, is

$$\text{mode}[X] = \frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

$$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} = ???$$

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

  $$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} =???$$

- You observed 8 heads and 2 tails?

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

$$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} = ???$$

- You observed 8 heads and 2 tails?

$$Beta(8+1, 2+1) \equiv Beta(9,3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$$

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

$$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} = ???$$

- You observed 8 heads and 2 tails?

$$Beta(8+1, 2+1) \equiv Beta(9,3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$$

- You observed 80 heads and 20 tails?

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

$$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} = ???$$

- You observed 8 heads and 2 tails?

$$Beta(8+1, 2+1) \equiv Beta(9,3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$$

- You observed 80 heads and 20 tails?

$$Beta(80+1, 20+1) \equiv Beta(81,21) \rightarrow \text{mode} = \frac{(81-1)}{(81-1)+(21-1)} = \frac{80}{100}$$

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

$$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} =???$$

- You observed 8 heads and 2 tails?

$$Beta(8+1, 2+1) \equiv Beta(9,3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$$

- You observed 80 heads and 20 tails?

$$Beta(80+1, 20+1) \equiv Beta(81,21) \rightarrow \text{mode} = \frac{(81-1)}{(81-1)+(21-1)} = \frac{80}{100}$$

- You observed 2 heads and 3 tails?

Beta Random Variable Examples

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

- You didn't observe anything?

$$Beta(0+1, 0+1) \equiv Beta(1,1) \equiv Unif(0,1) \rightarrow \text{mode} = ???$$

- You observed 8 heads and 2 tails?

$$Beta(8+1, 2+1) \equiv Beta(9,3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$$

- You observed 80 heads and 20 tails?

$$Beta(80+1, 20+1) \equiv Beta(81,21) \rightarrow \text{mode} = \frac{(81-1)}{(81-1)+(21-1)} = \frac{80}{100}$$

- You observed 2 heads and 3 tails?

$$Beta(2+1, 3+1) \equiv Beta(3,4) \rightarrow \text{mode} = \frac{(3-1)}{(3-1)+(4-1)} = \frac{2}{5}$$

Agenda

- The Beta Random Variable
- MAP Estimation
- MAP Example

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Shouldn't we maximize "$P(\theta \mid \boldsymbol{x})$" instead? Doesn't make sense unless Θ is a rv.

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Shouldn't we maximize "$P(\theta \mid \boldsymbol{x})$" instead? Doesn't make sense unless Θ is a rv.

Maximum a Posteriori (MAP) Estimation Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data) $\pi_\Theta(\theta \mid \boldsymbol{x})$.

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Shouldn't we maximize "$P(\theta \mid \boldsymbol{x})$" instead? Doesn't make sense unless Θ is a rv.

Maximum a Posteriori (MAP) Estimation Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data) $\pi_\Theta(\theta \mid \boldsymbol{x})$. By Bayes' Theorem,

$$\pi_\Theta(\theta \mid \boldsymbol{x}) = \frac{L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)}{P(\boldsymbol{x})}$$

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Shouldn't we maximize "$P(\theta \mid \boldsymbol{x})$" instead? Doesn't make sense unless Θ is a rv.

Maximum a Posteriori (MAP) Estimation Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data) $\pi_\Theta(\theta \mid \boldsymbol{x})$. By Bayes' Theorem,

$$\pi_\Theta(\theta \mid \boldsymbol{x}) = \frac{L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)}{P(\boldsymbol{x})} \propto L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)$$

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Shouldn't we maximize "$P(\theta \mid \boldsymbol{x})$" instead? Doesn't make sense unless Θ is a rv.

Maximum a Posteriori (MAP) Estimation Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data) $\pi_\Theta(\theta \mid \boldsymbol{x})$. By Bayes' Theorem,

$$\pi_\Theta(\theta \mid \boldsymbol{x}) = \frac{L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)}{P(\boldsymbol{x})} \propto L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi_\Theta(\theta \mid \boldsymbol{x}) =$$

Maximum a Posteriori (Intuition)

In maximum likelihood estimation, we use iid samples $\boldsymbol{x} = (x_1, \dots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i; \theta)$$

Shouldn't we maximize "$P(\theta \mid \boldsymbol{x})$" instead? Doesn't make sense unless Θ is a rv.

Maximum a Posteriori (MAP) Estimation Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data) $\pi_\Theta(\theta \mid \boldsymbol{x})$. By Bayes' Theorem,

$$\pi_\Theta(\theta \mid \boldsymbol{x}) = \frac{L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)}{P(\boldsymbol{x})} \propto L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi_\Theta(\theta \mid \boldsymbol{x}) = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta)\pi_\Theta(\theta)$$

Maximum A Posteriori (MAP) Estimation

<u>Maximum A Posteriori (MAP) Estimation:</u> Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,; \Theta = \theta)$ (if X discrete), or from density $f_X(t\,; \Theta = \theta)$ (if X continuous), where Θ is the random variable representing the parameter (or vector of parameters). We define the maximum a posteriori (MAP) estimator $\hat{\theta}_{MAP}$ of Θ to be the parameter which maximizes the posterior distribution of Θ given the data.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi_{\Theta}(\theta \mid \boldsymbol{x}) = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta)\pi_{\Theta}(\theta)$$

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \frac{2}{5}$$

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \boxed{\frac{2}{5}}$$

b. Suppose we impose $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \boxed{\frac{2}{5}}$$

b. Suppose we impose $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \{0.2,0.5,0.7\}} L(x|\theta)$$

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in[0,1]} \theta^2(1-\theta)^3 = \boxed{\frac{2}{5}}$$

b. Suppose we impose $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\{0.2,0.5,0.7\}} L(x|\theta)$$

$$L(x|0.2) = (0.2^2 0.8^3)$$

$$L(x|0.5) = (0.5^2 0.5^3)$$

$$L(x|0.7) = (0.7^2 0.3^3)$$

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \boxed{\frac{2}{5}}$$

b. Suppose we impose $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \{0.2,0.5,0.7\}} L(x|\theta)$$

$$L(x|0.2) = (0.2^2 0.8^3) = 0.02048$$

$$L(x|0.5) = (0.5^2 0.5^3) = 0.03125$$

$$L(x|0.7) = (0.7^2 0.3^3) = 0.01323$$

Maximum a Posteriori (Example)

a. Suppose our samples are $x = (0,0,1,1,0)$, from $Bernoulli(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0,1)$. What is the MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \boxed{\frac{2}{5}}$$

b. Suppose we impose $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \{0.2,0.5,0.7\}} L(x|\theta) = \boxed{0.5}$$

$$L(x|0.2) = (0.2^2 0.8^3) = 0.02048$$

$$L(x|0.5) = (0.5^2 0.5^3) = 0.03125$$

$$L(x|0.7) = (0.7^2 0.3^3) = 0.01323$$

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta)\pi_\Theta(\theta)$$

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

$$\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta)\pi_\Theta(\theta)$$

$$\pi_\Theta(0.2|x) = L(x|0.2)\pi_\Theta(0.2)$$
$$\pi_\Theta(0.5|x) = L(x|0.5)\pi_\Theta(0.5)$$
$$\pi_\Theta(0.7|x) = L(x|0.7)\pi_\Theta(0.7)$$

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

$$\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta)\pi_\Theta(\theta)$$

$$\pi_\Theta(0.2|x) = L(x|0.2)\pi_\Theta(0.2) = (0.2^2 0.8^3)(0.1) = 0.0020480$$
$$\pi_\Theta(0.5|x) = L(x|0.5)\pi_\Theta(0.5) = (0.5^2 0.5^3)(0.01) = 0.0003125$$
$$\pi_\Theta(0.7|x) = L(x|0.7)\pi_\Theta(0.7) = (0.7^2 0.3^3)(0.89) = 0.0117747$$

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

$$\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta)\pi_\Theta(\theta) = \boxed{0.7}$$

$$\pi_\Theta(0.2|x) = L(x|0.2)\pi_\Theta(0.2) = (0.2^2 0.8^3)(0.1) = 0.0020480$$
$$\pi_\Theta(0.5|x) = L(x|0.5)\pi_\Theta(0.5) = (0.5^2 0.5^3)(0.01) = 0.0003125$$
$$\pi_\Theta(0.7|x) = L(x|0.7)\pi_\Theta(0.7) = (0.7^2 0.3^3)(0.89) = 0.0117747$$

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

$$\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta)\pi_\Theta(\theta) = \boxed{0.7}$$

$$\pi_\Theta(0.2|x) = L(x|0.2)\pi_\Theta(0.2) = (0.2^2 0.8^3)(0.1) = 0.0020480$$
$$\pi_\Theta(0.5|x) = L(x|0.5)\pi_\Theta(0.5) = (0.5^2 0.5^3)(0.01) = 0.0003125$$
$$\pi_\Theta(0.7|x) = L(x|0.7)\pi_\Theta(0.7) = (0.7^2 0.3^3)(0.89) = 0.0117747$$

d. Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

Maximum a Posteriori (Example)

c. Assume Θ is restricted as in part b (now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

$$\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2,0.5,0.7\}} L(x|\theta)\pi_\Theta(\theta) = \boxed{0.7}$$

$$\pi_\Theta(0.2|x) = L(x|0.2)\pi_\Theta(0.2) = (0.2^2 0.8^3)(0.1) = 0.0020480$$
$$\pi_\Theta(0.5|x) = L(x|0.5)\pi_\Theta(0.5) = (0.5^2 0.5^3)(0.01) = 0.0003125$$
$$\pi_\Theta(0.7|x) = L(x|0.7)\pi_\Theta(0.7) = (0.7^2 0.3^3)(0.89) = 0.0117747$$

d. Show that we can make the MAP whatever we like, by finding a prior over $\{0.2,0.5,0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

Choose $\pi_\Theta(\theta) = 1$ for the θ you want.

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_{\Theta}(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_{w} f_W(w)$).

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k} \theta^k (1-\theta)^{n-k}\right) \cdot$$

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k} \theta^k (1-\theta)^{n-k} \right) \cdot \left(\frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right)$$

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \dots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta | x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\cancel{\binom{n}{k}} \theta^k (1-\theta)^{n-k}\right) \cdot \left(\cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2,0.5,0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \dots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k} \theta^k (1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

Hence our posterior is $Beta(k+\alpha, n-k+\beta)$. The mode of this beta is given:

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

Hence our posterior is $Beta(k+\alpha, n-k+\beta)$. The mode of this beta is given:

$$\frac{k+\alpha-1}{(k+\alpha-1)+(n-k+\beta-1)}$$

Maximum a Posteriori (Example)

e. Typically, for the Bernoulli/Binomial distribution, we want to be able to get any value $\in (0,1)$, (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\Theta \sim Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0,1)$. Recall the mode of a $W \sim Beta(\alpha, \beta)$ rv is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \dots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is k/n, where $k = \sum x_i$. Show that the posterior $\pi_\Theta(\theta|x)$ is $Beta(k+\alpha, n-k+\beta)$, and find the MAP. (Hint: use the mode given).

$$\pi_\Theta(\theta|x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$
$$= \left(\binom{n}{k} \theta^k (1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$
$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

Hence our posterior is $Beta(k+\alpha, n-k+\beta)$. The mode of this beta is given:

$$\frac{k+\alpha-1}{(k+\alpha-1)+(n-k+\beta-1)} = \boxed{\frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}}$$

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)}$$

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} =$$

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

\# True Heads

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

\# True Trials

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

Fake Heads

True Heads

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

True Trials

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

Fake Heads

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

True Heads

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

True Trials

Fake Trials

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

It means: pretend we saw $\alpha - 1$ heads ahead of time, and $\beta - 1$ tails ahead of time. Then our total heads is $k + \alpha - 1$ and our total trials is $n + \alpha + \beta - 2$.

Maximum a Posteriori (Example)

f. Notice that $Beta(1,1) \equiv Unif(0,1)$. If we used this as the prior, how would the MLE and MAP compare?

From previous slide, if $\alpha = \beta = 1$, then our MAP estimate is the same as our ML estimate!

$$\hat{\theta}_{MAP} = \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

g. Since the posterior is also a Beta distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution. Interpret α, β.

It means: pretend we saw $\alpha - 1$ heads ahead of time, and $\beta - 1$ tails ahead of time. Then our total heads is $k + \alpha - 1$ and our total trials is $n + \alpha + \beta - 2$.

$$\pi_\Theta(\theta | x) \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

"Lucky" Us!!

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

"Lucky" Us!!

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior.

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior.

i. Which do you think is "better", MLE or MAP?

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior.

i. Which do you think is "better", MLE or MAP?

- There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior.

i. Which do you think is "better", MLE or MAP?

- There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.
- Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a fixed quantity.

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior.

i. Which do you think is "better", MLE or MAP?

- There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.
- Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a fixed quantity.
- On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a random variable, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?

Maximum a Posteriori (Example)

h. As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our prior when n is small, or n is large?

They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior.

i. Which do you think is "better", MLE or MAP?

- There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.
- Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a fixed quantity.
- On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a random variable, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?
- Anyway, in the long run, the prior ``washes out", and the only thing that matters is the likelihood; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate. However, as the number of samples goes to infinity, the MAP and MLE are equal.