Bayesian Estimation

Machine Learning Techniques

Karthik Thiagarajan

References and Credits

- The content presented in these slides is derived from professor Arun Rajkumar's lectures and slides in the MLT course. This is the "ground truth" for almost all the content in these slides. As a result, these slides should be viewed as a presentation of the same content in the professor's lectures using a different medium. At the same time, these slides are not meant to be a replacement for the lectures.
- These slides were prepared using the tool mathcha.io.

Bayes' Theorem

Bayes' Theorem

Bayes' Theorem

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

Bayes' Theorem

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

Bayes' Theorem

Belief

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

1) Prior

Bayes' Theorem

Belief

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

1) Prior
2) Evidence (Data)

Bayes' Theorem

Belief

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

1) Prior
2) Evidence (Data)
3) Likelihood

Bayes' Theorem

Belief

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

Bayes' Theorem

Belief

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

Bayes' Theorem

Belief

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

Bayes' Theorem

Belief

Distribution

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

Bayes' Theorem

Belief

Distribution

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

Bayes' Theorem

Belief

Distribution

Function

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

Bayes' Theorem

Belief

Distribution

Function

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

Bayes' Theorem

Belief

Distribution Function

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

Scalar

1) Prior
2) Evidence (Data)
3) Likelihood
4) Posterior

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

Beta Distribution

$\mathsf{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$

Beta Distribution

$$\text{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

$$p \in (0, 1)$$

Beta Distribution

$$\text{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

$$p \in (0, 1)$$

$$f(p; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

Beta Distribution

$$\text{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

$$p \in (0, 1)$$

$$f(p; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du$$

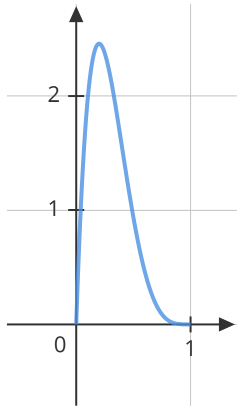
Beta Distribution

$$\mathrm{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

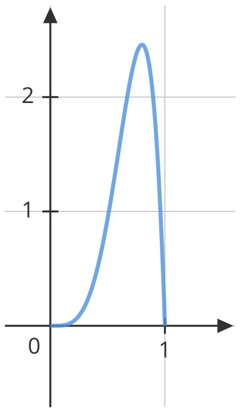
$$p \in (0, 1)$$

$$f(p; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du$$



$\mathrm{Beta}(2, 5)$



$\mathrm{Beta}(5, 2)$

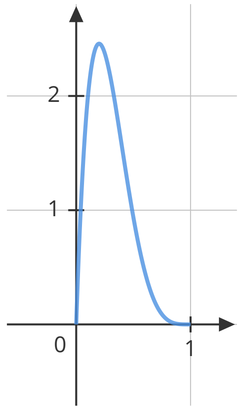
Beta Distribution

$$\text{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

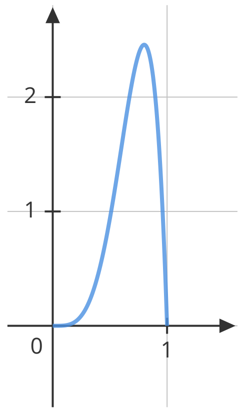
$$p \in (0, 1)$$

$$f(p; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

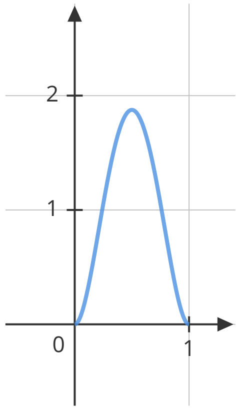
$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du$$



Beta$(2, 5)$



Beta$(5, 2)$



Beta$(3, 3)$

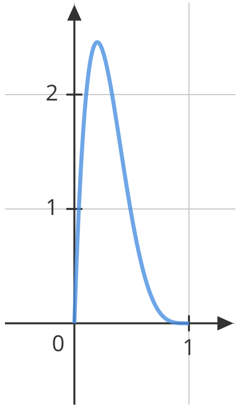
Beta Distribution

$$\text{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

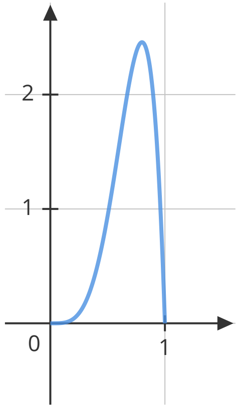
$$p \in (0, 1)$$

$$f(p; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

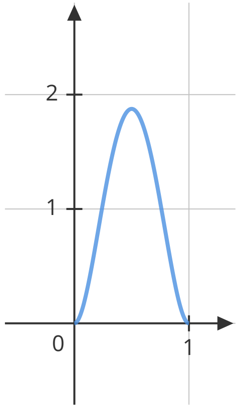
$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du$$



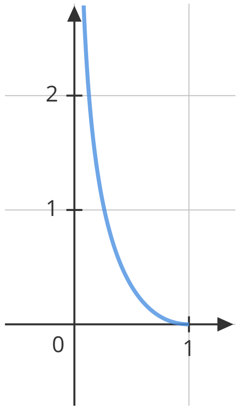
$\text{Beta}(2, 5)$



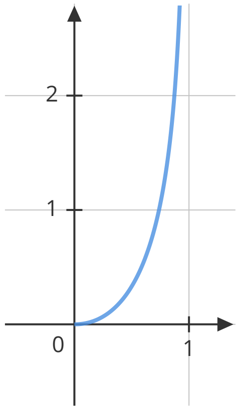
$\text{Beta}(5, 2)$



$\text{Beta}(3, 3)$



$\text{Beta}(0.5, 3)$



$\text{Beta}(3, 0.5)$

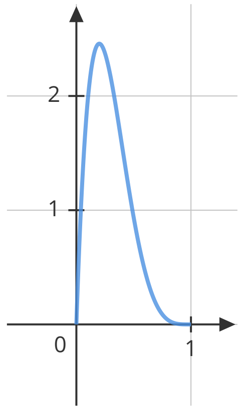
Beta Distribution

$$\text{Beta}(\alpha, \beta) \qquad \alpha > 0, \beta > 0$$

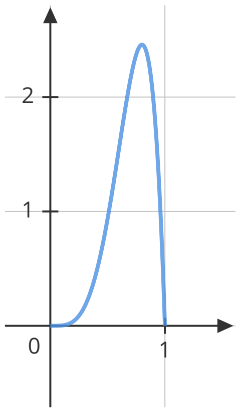
$$p \in (0, 1)$$

$$f(p; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

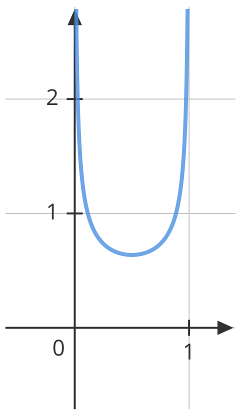
$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du$$



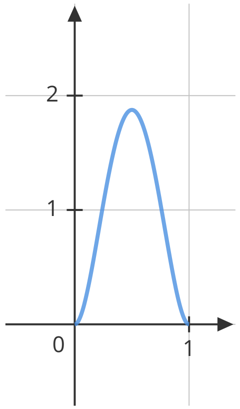
$\text{Beta}(2, 5)$



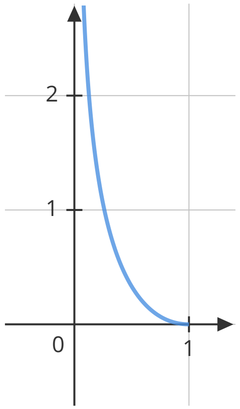
$\text{Beta}(5, 2)$



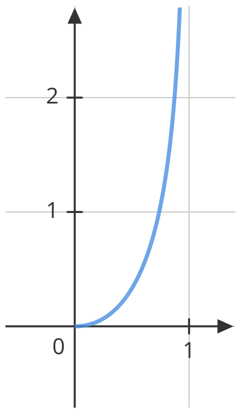
$\text{Beta}(0.5, 0.5)$



$\text{Beta}(3, 3)$



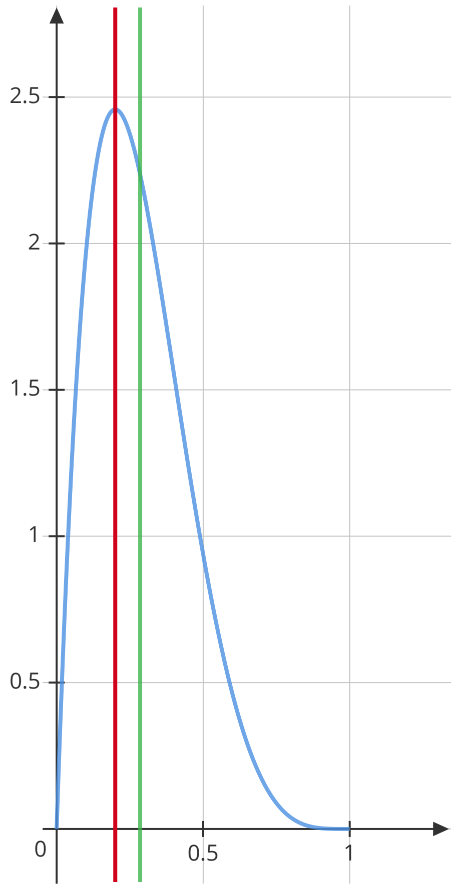
$\text{Beta}(0.5, 3)$



$\text{Beta}(3, 0.5)$

Beta Distribution

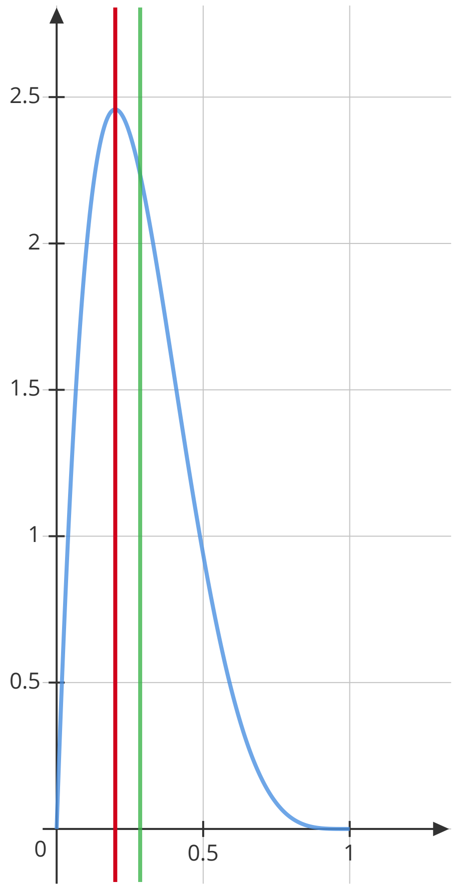
Beta$(2, 5)$



Beta Distribution

$$\text{Mean} = \frac{\alpha}{\alpha + \beta}$$

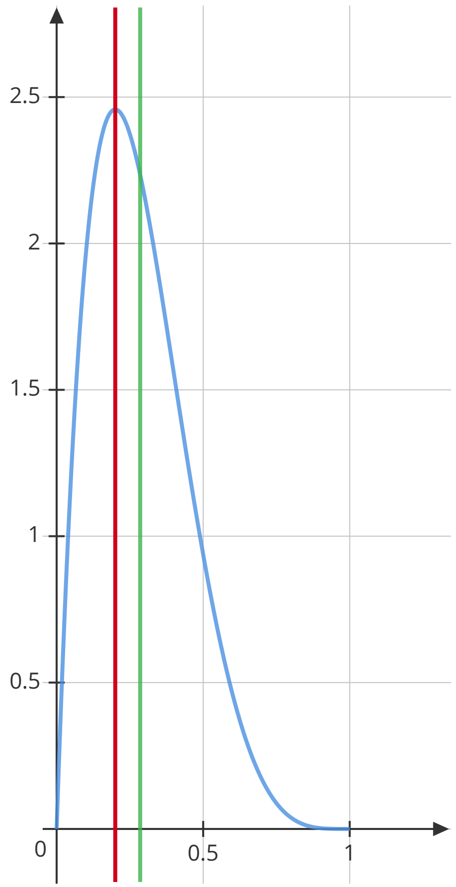
$\text{Beta}(2, 5)$



Beta Distribution

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}$$

Beta$(2, 5)$



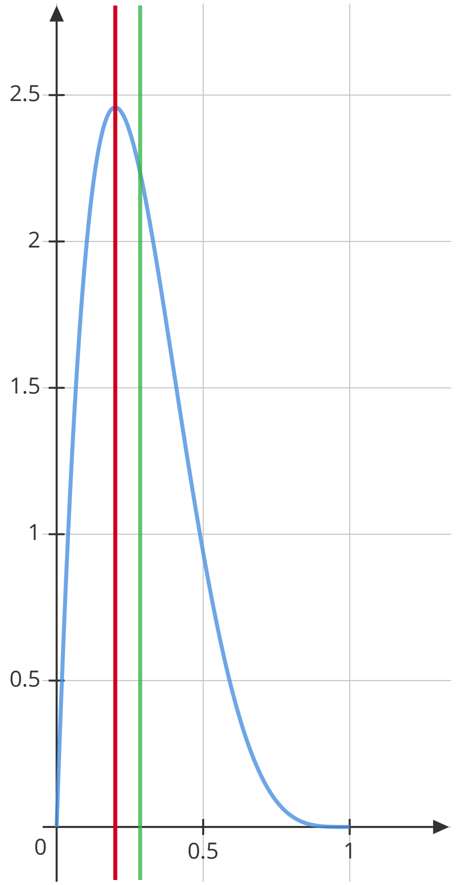
$$\text{Mode} = \begin{cases} \dfrac{\alpha-1}{\alpha+\beta-2} & \alpha, \beta > 1 \\ 0 & \alpha \leqslant 1, \beta > 1 \\ 1 & \alpha > 1, \beta \leqslant 1 \\ (0, 1) & \alpha = \beta = 1 \\ \{0, 1\} & \alpha, \beta < 1 \end{cases}$$

Beta Distribution

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}$$

$$\text{Mode} = \begin{cases} \dfrac{\alpha-1}{\alpha+\beta-2} & \alpha, \beta > 1 \\ 0 & \alpha \leqslant 1, \beta > 1 \\ 1 & \alpha > 1, \beta \leqslant 1 \\ (0,1) & \alpha = \beta = 1 \\ \{0,1\} & \alpha, \beta < 1 \end{cases}$$

Beta$(2, 5)$

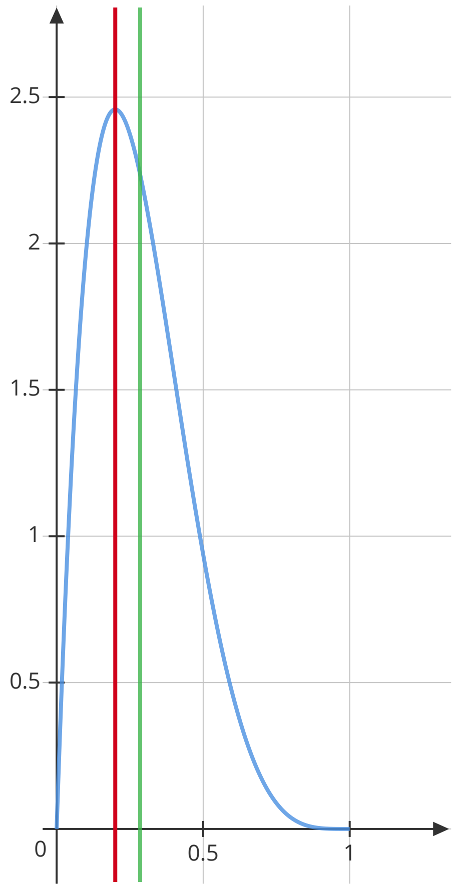


$$\frac{d \log(f(p))}{dp} = 0$$

Beta Distribution

$$\text{Mean} = \frac{\alpha}{\alpha + \beta}$$

$\text{Beta}(2, 5)$



$$\text{Mode} = \begin{cases} \dfrac{\alpha - 1}{\alpha + \beta - 2} & \alpha, \beta > 1 \\ 0 & \alpha \leqslant 1, \beta > 1 \\ 1 & \alpha > 1, \beta \leqslant 1 \\ (0, 1) & \alpha = \beta = 1 \\ \{0, 1\} & \alpha, \beta < 1 \end{cases}$$

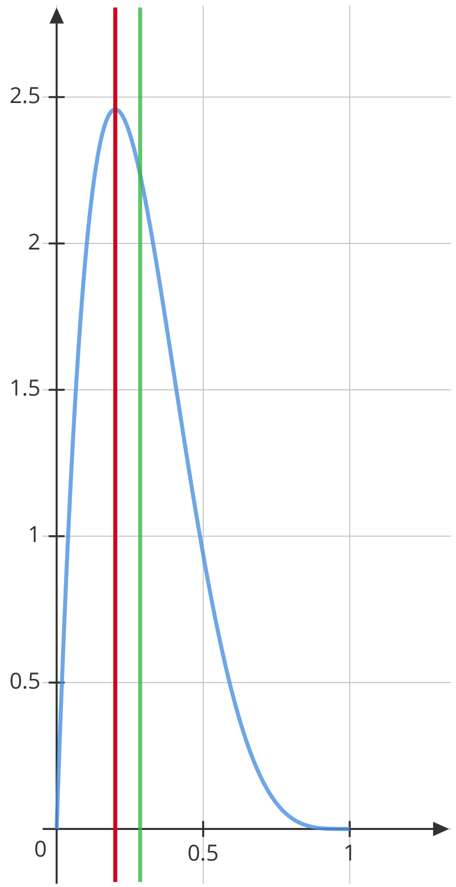
$$\frac{d \log(f(p))}{dp} = 0$$

$$\frac{(\alpha - 1)}{p} - \frac{(\beta - 1)}{1 - p} = 0$$

Beta Distribution

$$\text{Mean} = \frac{\alpha}{\alpha + \beta}$$

Beta$(2, 5)$



$$\text{Mode} = \begin{cases} \dfrac{\alpha - 1}{\alpha + \beta - 2} & \alpha, \beta > 1 \\ 0 & \alpha \leqslant 1, \beta > 1 \\ 1 & \alpha > 1, \beta \leqslant 1 \\ (0, 1) & \alpha = \beta = 1 \\ \{0, 1\} & \alpha, \beta < 1 \end{cases}$$

$$\frac{d \log(f(p))}{dp} = 0$$

$$\frac{(\alpha - 1)}{p} - \frac{(\beta - 1)}{1 - p} = 0$$

$$p = \frac{\alpha - 1}{\alpha + \beta - 2}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$X_i \sim Br(p)$

$p \to$ parameter

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

$\alpha, \beta \to$ hyperparameters

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$X_i \sim Br(p)$ $p \sim \text{Beta}(\alpha, \beta)$

$p \to$ parameter $\alpha, \beta \to$ hyperparameters

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha - 1}(1-p)^{\beta - 1}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\text{Likelihood}: \; p^{n_h}(1-p)^{n_t}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \ f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\text{Posterior} \propto$$

$$\text{Likelihood}: \ p^{n_h}(1-p)^{n_t}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \ f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$\text{Posterior} \propto \text{Prior} \ \times \ \text{Likelihood}$$

$$\text{Likelihood}: \ p^{n_h} (1-p)^{n_t}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \ f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\text{Posterior} \propto \text{Prior} \ \times \ \text{Likelihood}$$

$$\propto p^{n_h+\alpha-1} \cdot (1-p)^{n_t+\beta-1}$$

$$\text{Likelihood}: \ p^{n_h}(1-p)^{n_t}$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$\text{Likelihood}: \; p^{n_h} (1-p)^{n_t}$$

$$\text{Posterior} \propto \text{Prior} \; \times \; \text{Likelihood}$$

$$\propto p^{n_h+\alpha-1} \cdot (1-p)^{n_t+\beta-1}$$

$$\propto \text{Beta}(n_h+\alpha, n_t+\beta)$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

$$\propto p^{n_h+\alpha-1} \cdot (1-p)^{n_t+\beta-1}$$

$$\text{Likelihood}: \; p^{n_h}(1-p)^{n_t}$$

$$\propto \text{Beta}(n_h+\alpha, n_t+\beta)$$

$$\text{Posterior} = \text{Beta}(n_h+\alpha, n_t+\beta)$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$X_i \sim Br(p)$

$p \rightarrow$ parameter

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

$\alpha, \beta \rightarrow$ hyperparameters

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\text{Likelihood}: \; p^{n_h}(1-p)^{n_t}$$

$$\text{Posterior} = \text{Beta}(n_h+\alpha, n_t+\beta)$$

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

$$\propto p^{n_h+\alpha-1} \cdot (1-p)^{n_t+\beta-1}$$

$$\propto \text{Beta}(n_h+\alpha, n_t+\beta)$$

Beta distribution is a conjugate prior for the Bernoulli distribution

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \to \text{parameter} \qquad \alpha, \beta \to \text{hyperparameters}$$

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$\text{Posterior} \propto \text{Prior} \; \times \; \text{Likelihood}$$

$$\propto p^{n_h+\alpha-1} \cdot (1-p)^{n_t+\beta-1}$$

$$\text{Likelihood}: \; p^{n_h} (1-p)^{n_t}$$

$$\propto \text{Beta}(n_h + \alpha, n_t + \beta)$$

$$\text{Posterior} = \text{Beta}(n_h + \alpha, n_t + \beta)$$

Beta distribution is a conjugate
prior for the Bernoulli distribution

$$\text{Hard to compute} \longleftarrow P(D) = \int_\theta P(\theta) \cdot P(D \mid \theta)$$

$$\text{Prior} \xrightarrow{\text{Likelihood}} \text{Posterior}$$

$$X_i \sim Br(p) \qquad p \sim \text{Beta}(\alpha, \beta)$$

$$p \rightarrow \text{parameter} \qquad \alpha, \beta \rightarrow \text{hyperparameters}$$

$$\text{Prior}: \; f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

$$\propto p^{n_h+\alpha-1} \cdot (1-p)^{n_t+\beta-1}$$

$$\text{Likelihood}: \; p^{n_h}(1-p)^{n_t}$$

$$\propto \text{Beta}(n_h + \alpha, n_t + \beta)$$

$$\text{Posterior} = \text{Beta}(n_h + \alpha, n_t + \beta)$$

Beta distribution is a conjugate prior for the Bernoulli distribution

α, β: Pseudo-observations

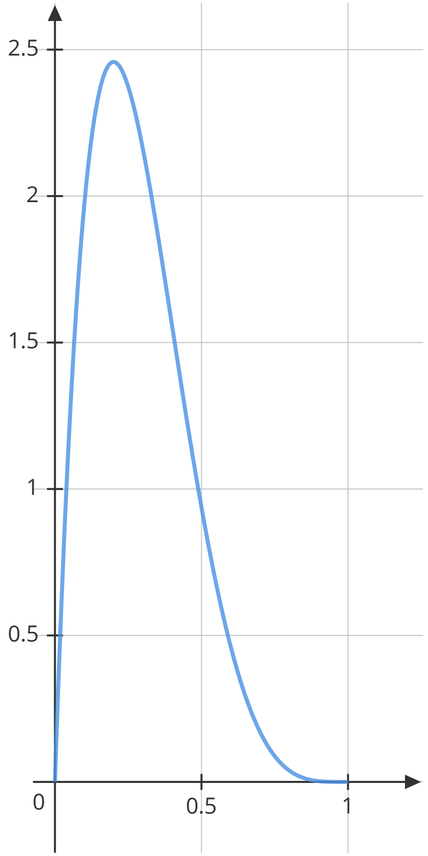
$$\text{Hard to compute} \longleftarrow P(D) = \int_\theta P(\theta) \cdot P(D \mid \theta)$$

Example-1

p is closer to 0 than it is to 1

Example-1

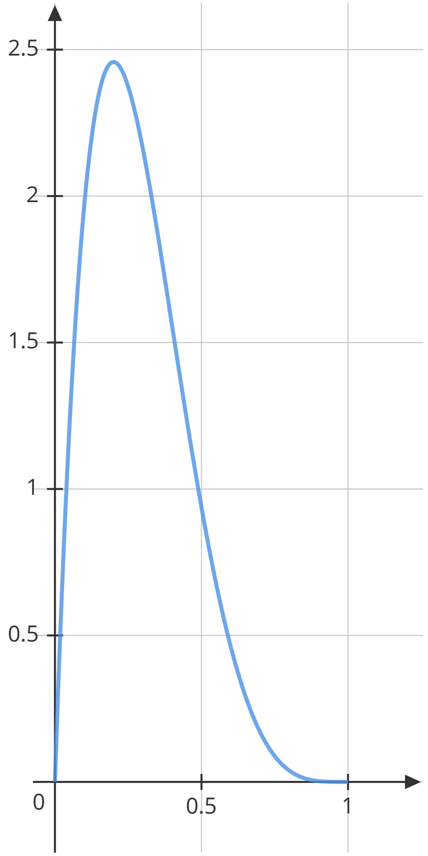
p is closer to 0 than it is to 1



Example-1

p is closer to 0 than it is to 1

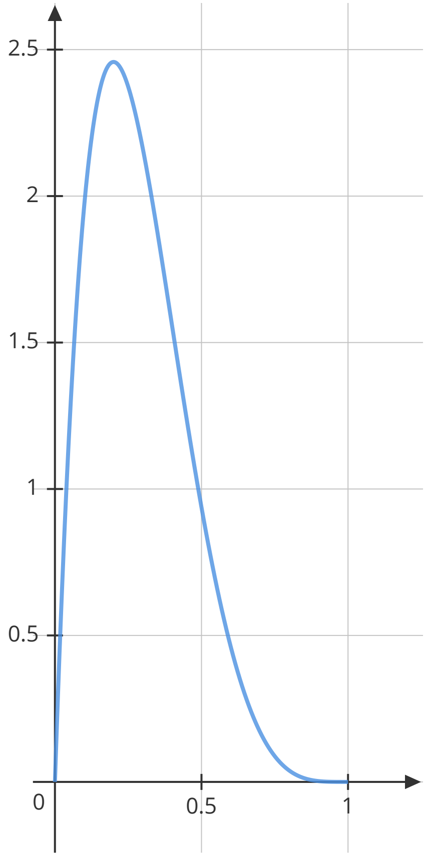
Beta$(2, 5)$



Example-1

p is closer to 0 than it is to 1

Beta$(2, 5)$



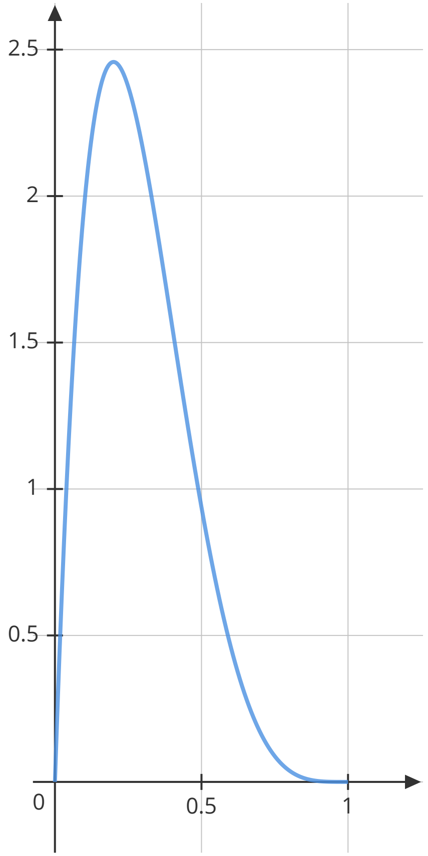
$$D = \{1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1\}$$

Example-1

p is closer to 0 than it is to 1

Beta$(4,\ 12)$

Beta$(2, 5)$

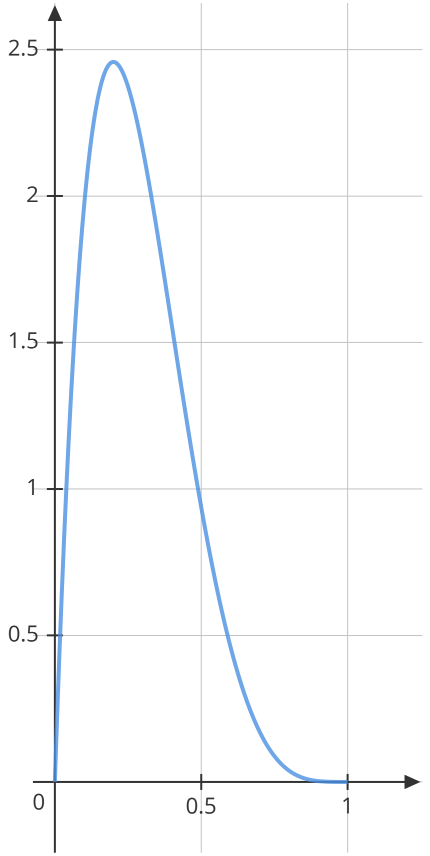


$$D = \{1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1\}$$

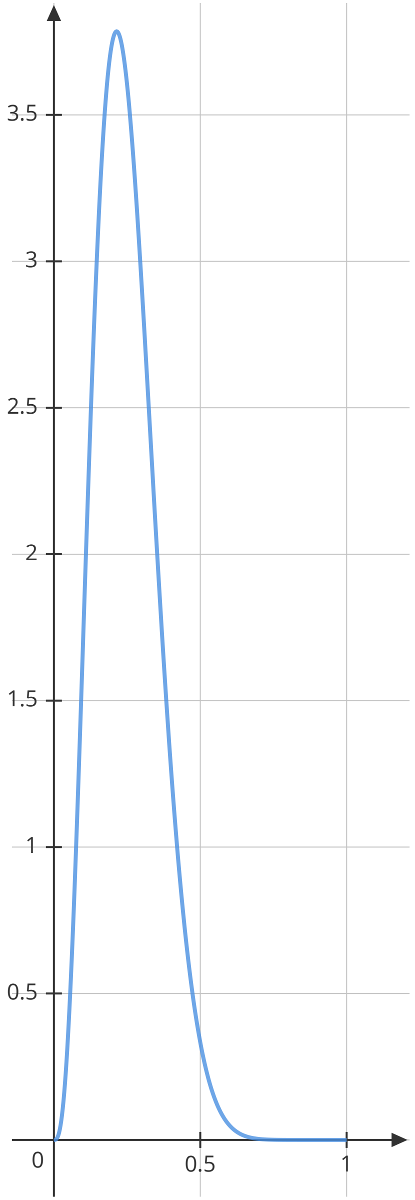
Example-1

p is closer to 0 than it is to 1

Beta$(2, 5)$



Beta$(4,\ 12)$

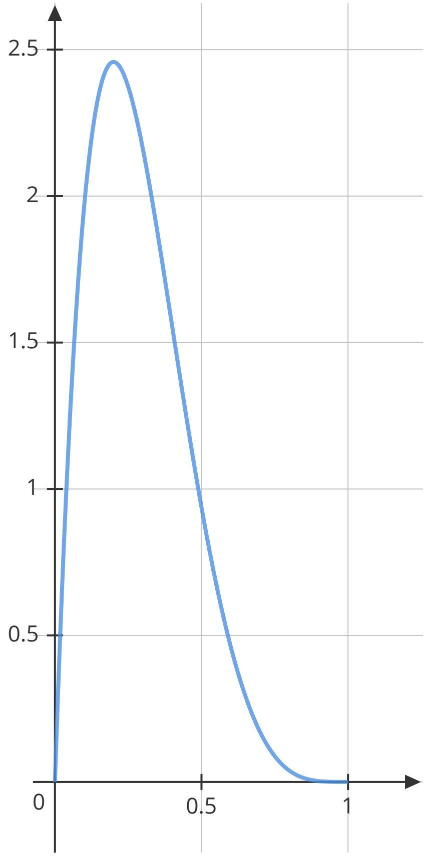


$$D = \{1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1\}$$

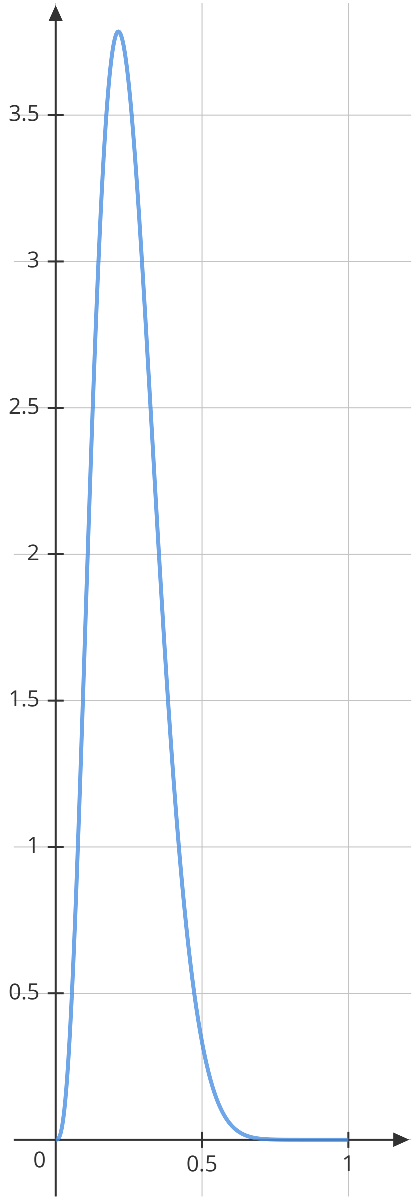
Example-1

p is closer to 0 than it is to 1

Beta$(2, 5)$



Beta$(4,\ 12)$



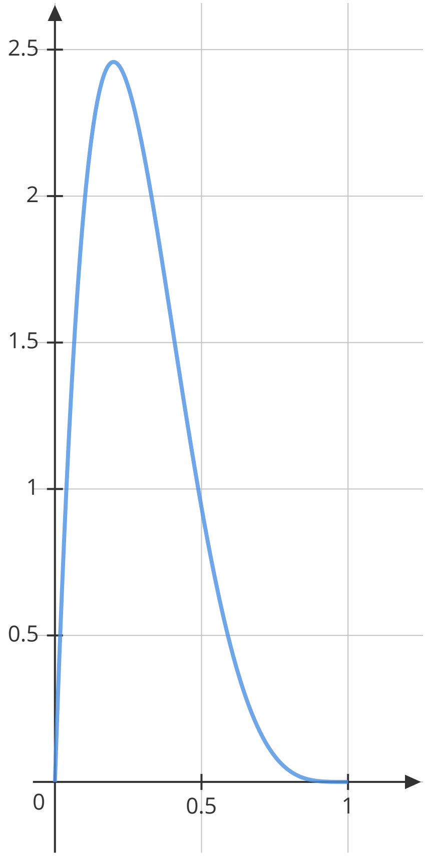
$$D_{\text{pseudo}} = \{1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0\}$$

$$D = \{1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1\}$$

Example-2

p is closer to 0 than it is to 1

Beta$(2, 5)$

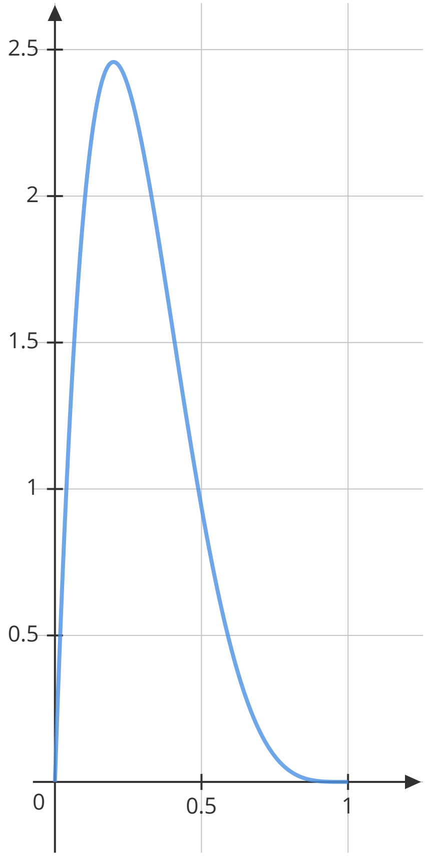


$$D = \{0,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0\}$$

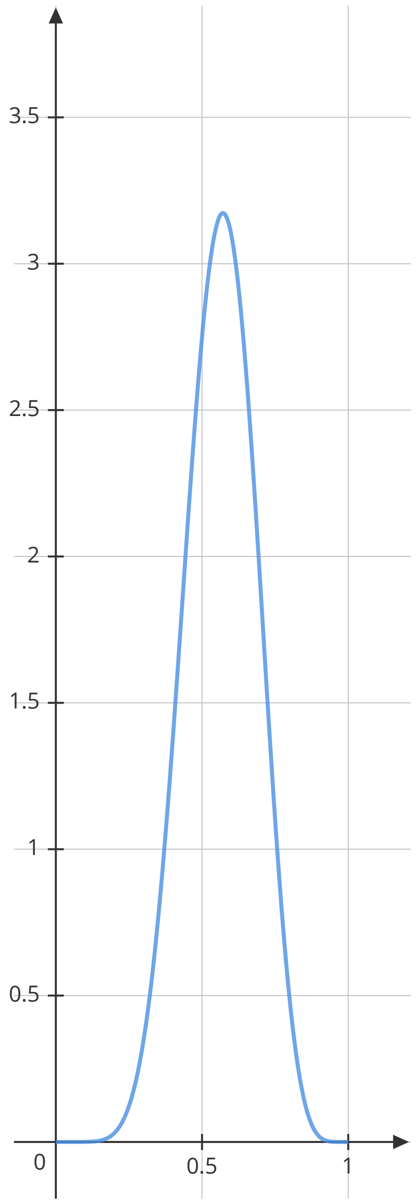
Example-2

p is closer to 0 than it is to 1

Beta$(2, 5)$



Beta$(9,\ 7)$



$$D_{\text{pseudo}} = \{1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0\}$$

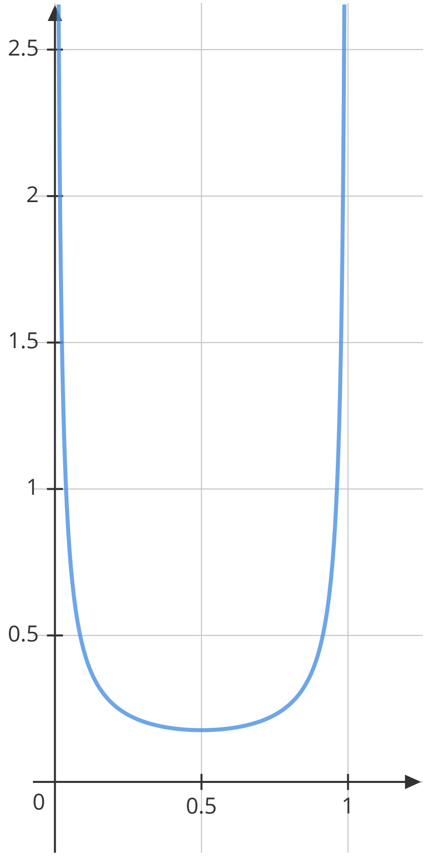
$$D = \{0,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0\}$$

Example-3

p is extremely close to either 0 or 1

Example-3

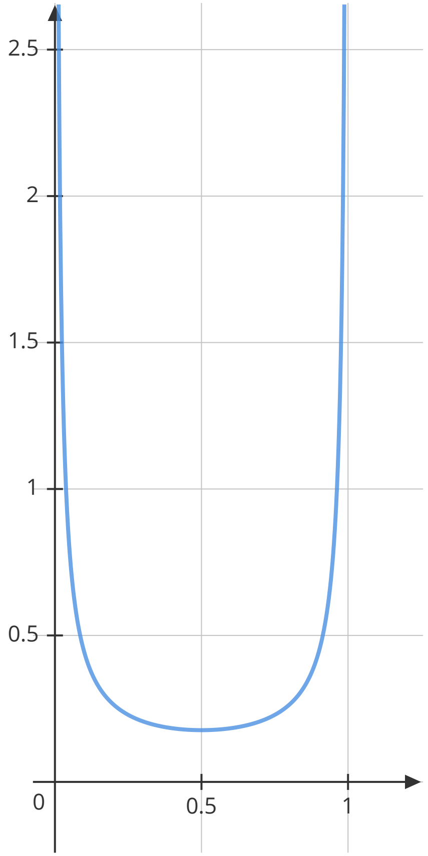
p is extremely close to either 0 or 1



Example-3

p is extremely close to either 0 or 1

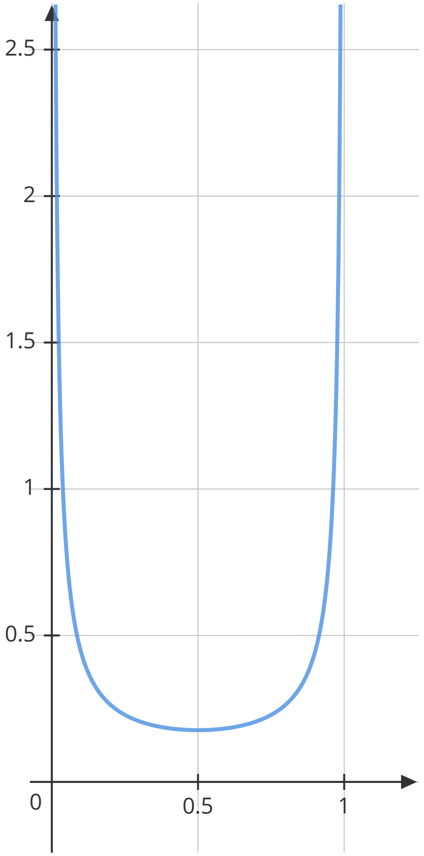
Beta$(0.1, 0.1)$



Example-3

p is extremely close to either 0 or 1

Beta$(0.1, 0.1)$



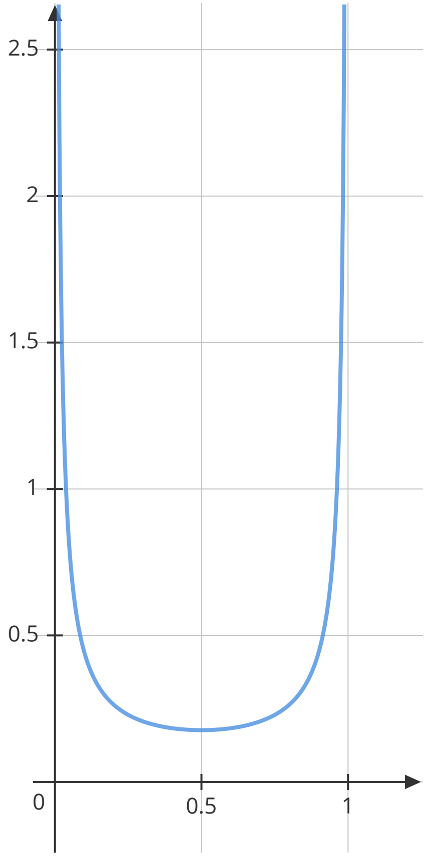
$$D = \{1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0\}$$

Example-3

p is extremely close to either 0 or 1

Beta$(8.1,\ 1.1)$

Beta$(0.1, 0.1)$

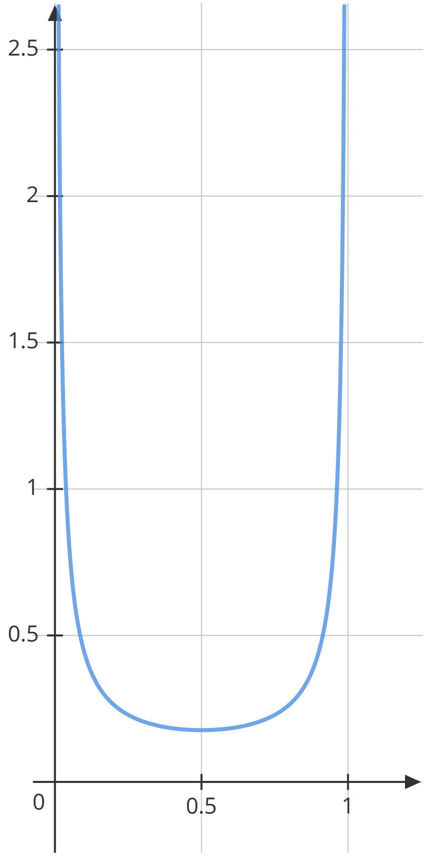


$$D = \{1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0\}$$

Example-3

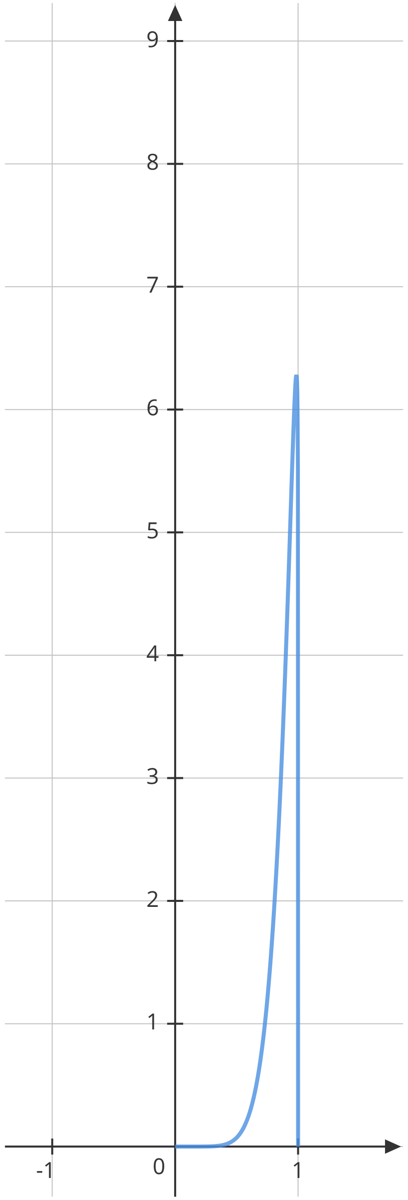
p is extremely close to either 0 or 1

Beta$(0.1, 0.1)$



Beta$(8.1,\ 1.1)$

$$D = \{1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0\}$$



Point Estimates

$$P(\theta \mid D)$$

Point Estimates

Posterior: $\text{Beta}(\alpha + n_h, \beta + n_t)$

$$P(\theta \mid D)$$

Point Estimates

Posterior: Beta$(\alpha + n_h, \beta + n_t)$

$$P(\theta \mid D) \longrightarrow \text{Mean}$$

$$\text{Mean} = \frac{\alpha + n_h}{\alpha + \beta + n}$$

Point Estimates

$$\begin{tikzcd} & \text{Mean} \\ P(\theta \mid D) \arrow[ur] \arrow[dr] & \\ & \text{Mode (MAP)} \end{tikzcd}$$

Posterior: $\text{Beta}(\alpha + n_h, \beta + n_t)$

$$\text{Mean} = \frac{\alpha + n_h}{\alpha + \beta + n}$$

$$\text{Mode} = \frac{\alpha + n_h - 1}{\alpha + \beta + n - 2}$$

$$\alpha + n_h > 1$$
$$\beta + n_t > 1$$

Point Estimates

Posterior: Beta$(\alpha + n_h, \beta + n_t)$

$$
\begin{tikzcd}
 & \text{Mean} \\
P(\theta \mid D) \arrow[ur] \arrow[dr] & \\
 & \text{Mode (MAP)}
\end{tikzcd}
$$

$$\text{Mean} = \frac{\alpha + n_h}{\alpha + \beta + n}$$

$$\text{Mode} = \frac{\alpha + n_h - 1}{\alpha + \beta + n - 2}$$

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

Maximum A Posteriori estimate

$$\alpha + n_h > 1$$
$$\beta + n_t > 1$$