6.7900: Machine Learning
Lecture 3

Lecture start: Tues/Thurs 2:35pm
Who's speaking today? Prof. Tamara Broderick
Course website: gradml.mit.edu
Questions? Ask here or on piazza.com/mit/fall2024/67900/
Materials: Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

Last Time

I. Empirical risk minimization
II. Modeling
III. Maximum likelihood estimate (MLE)
IV. Pros & cons of the MLE

Today's Plan

I. A Bayesian approach
II. Prior
III. Posterior
IV. Predictive

Recap

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification
- Recall example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification
- Recall example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - E.g. Conversion: ad click leads to purchase of item
  - E.g. Medical outcome after a medical event

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification
- Recall example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - E.g. Conversion: ad click leads to purchase of item
  - E.g. Medical outcome after a medical event
  - We saw MLE can overfit: $\hat{\theta} = N^{-1} \sum_{n=1}^{N} y^{(n)}$

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification
- Recall example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - E.g. Conversion: ad click leads to purchase of item
  - E.g. Medical outcome after a medical event
  - We saw MLE can overfit: $\hat{\theta} = N^{-1} \sum_{n=1}^{N} y^{(n)}$
  - Often have domain knowledge

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification
- Recall example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0, 1)$
  - E.g. Conversion: ad click leads to purchase of item
  - E.g. Medical outcome after a medical event
  - We saw MLE can overfit: $\hat{\theta} = N^{-1} \sum_{n=1}^{N} y^{(n)}$
  - Often have domain knowledge

not enough to restrict range

Recap

we'll get back to having features in the next lecture

- We've observed training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We'd like to approximate the distribution of the next data point we might see, $p(y)$, when training data and future data are iid
  - Idea: use the **empirical distribution** over the training data
  - Idea: introduce a (parametric) model $p(y|\theta)$
    - Sub-idea: Choose one $\hat{\theta}$ and use $p(y|\hat{\theta})$
      - Sub-idea: $\hat{\theta}$ = **maximum likelihood estimate (MLE)**
- Potential issues (with empirical distribution and MLE) include overfitting to training data and no uncertainty quantification
- Recall example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0, 1)$
  - E.g. Conversion: ad click leads to purchase of item
  - E.g. Medical outcome after a medical event
  - We saw MLE can overfit: $\hat{\theta} = N^{-1} \sum_{n=1}^{N} y^{(n)}$
  - Often have domain knowledge
- A **Bayesian approach** tries to address overfitting, quantify uncertainty, and incorporate domain knowledge

A Bayesian approach

A Bayesian approach

- **Bayes Theorem**.

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(y, \theta) = p(y|\theta)p(\theta)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$
  - **Posterior**: $p(\theta|y)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$
  - **Posterior**: $p(\theta|y)$
  - **Evidence**: $p(y)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$
  - **Posterior**: $p(\theta|y)$
  - **Evidence**: $p(y)$
- The data could be our full set of training data

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$
  - **Posterior**: $p(\theta|y)$
  - **Evidence**: $p(y)$
- The data could be our full set of training data

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

- Running example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0, 1)$

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$
  - **Posterior**: $p(\theta|y)$
  - **Evidence**: $p(y)$
- The data could be our full set of training data

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

- Running example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0, 1)$
  - We have a likelihood

A Bayesian approach

- **Bayes Theorem**. Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

- Just a fact of probability: $p(\theta|y)p(y) = p(y, \theta) = p(y|\theta)p(\theta)$
- **Bayesian inference**: interpret *y* as data and θ as parameter
  - **Likelihood**: $p(y|\theta)$
  - **Prior**: $p(\theta)$
  - **Posterior**: $p(\theta|y)$
  - **Evidence**: $p(y)$
- The data could be our full set of training data

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

- Running example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0, 1)$
  - We have a likelihood
  - A Bayesian model needs both a likelihood and a prior

Beta distribution

$$\mathrm{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

Beta distribution

hyperparameters

$$\mathrm{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

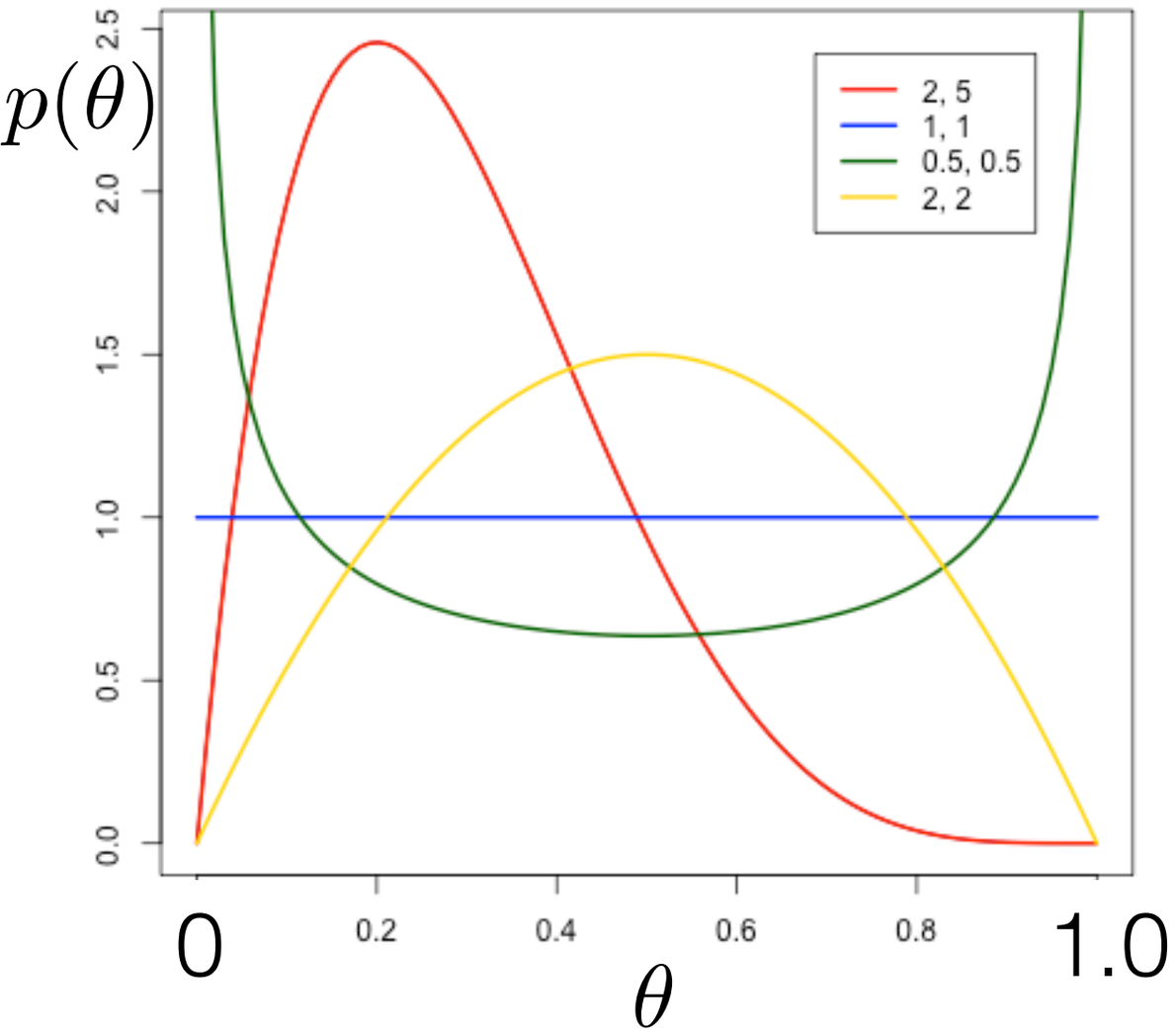
- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$

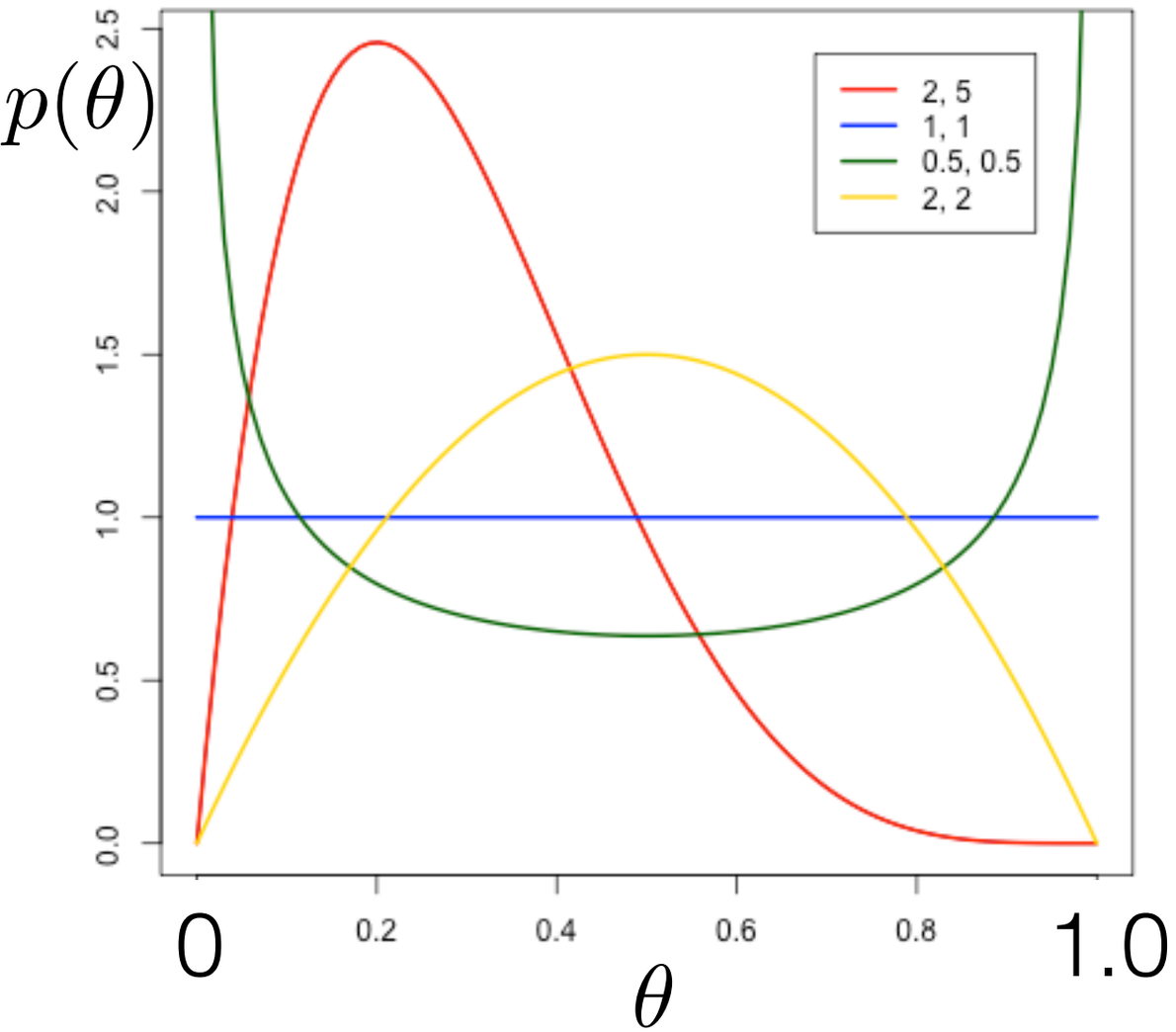


Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



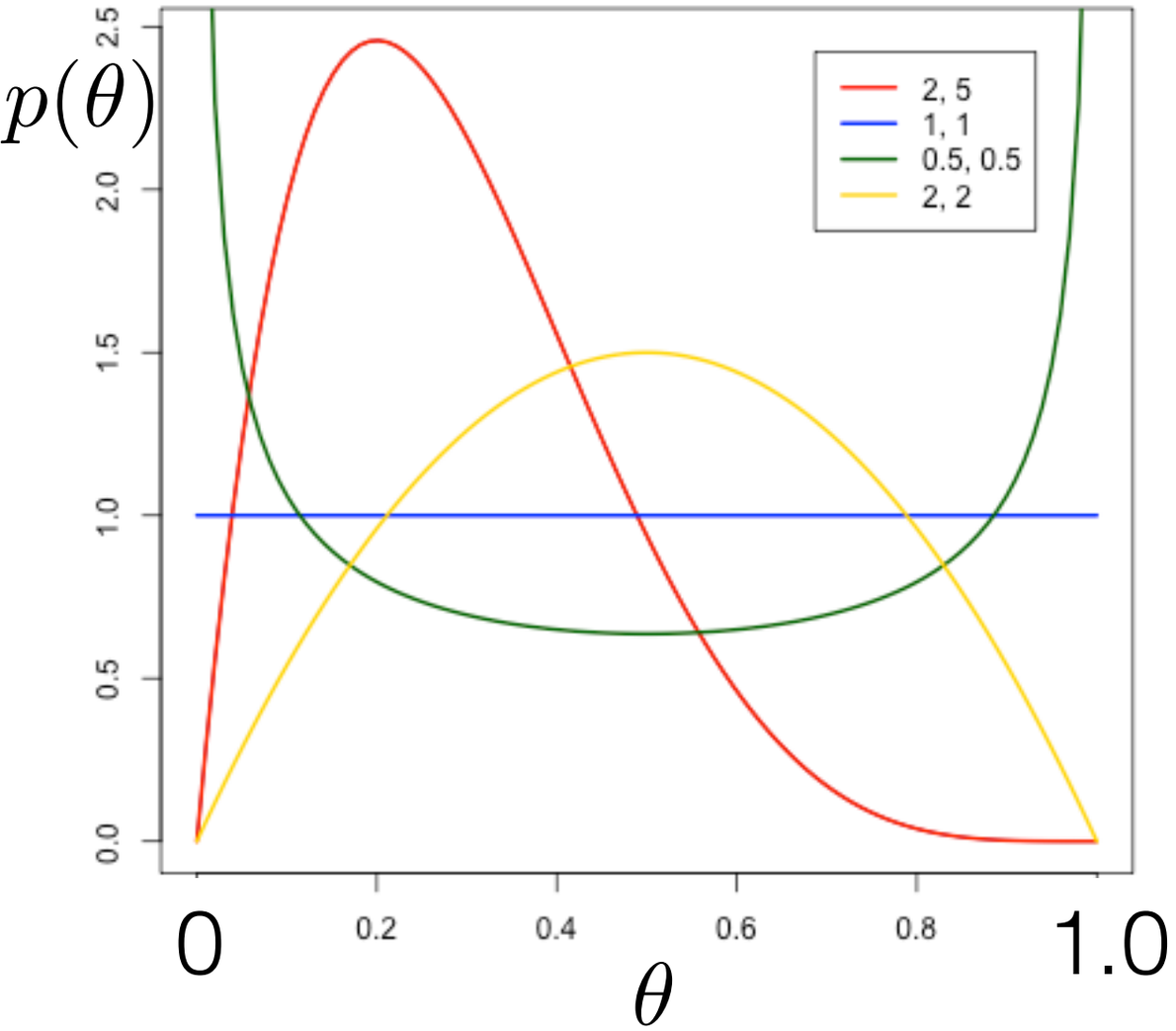
- What happens?
- When $a = b = 1$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



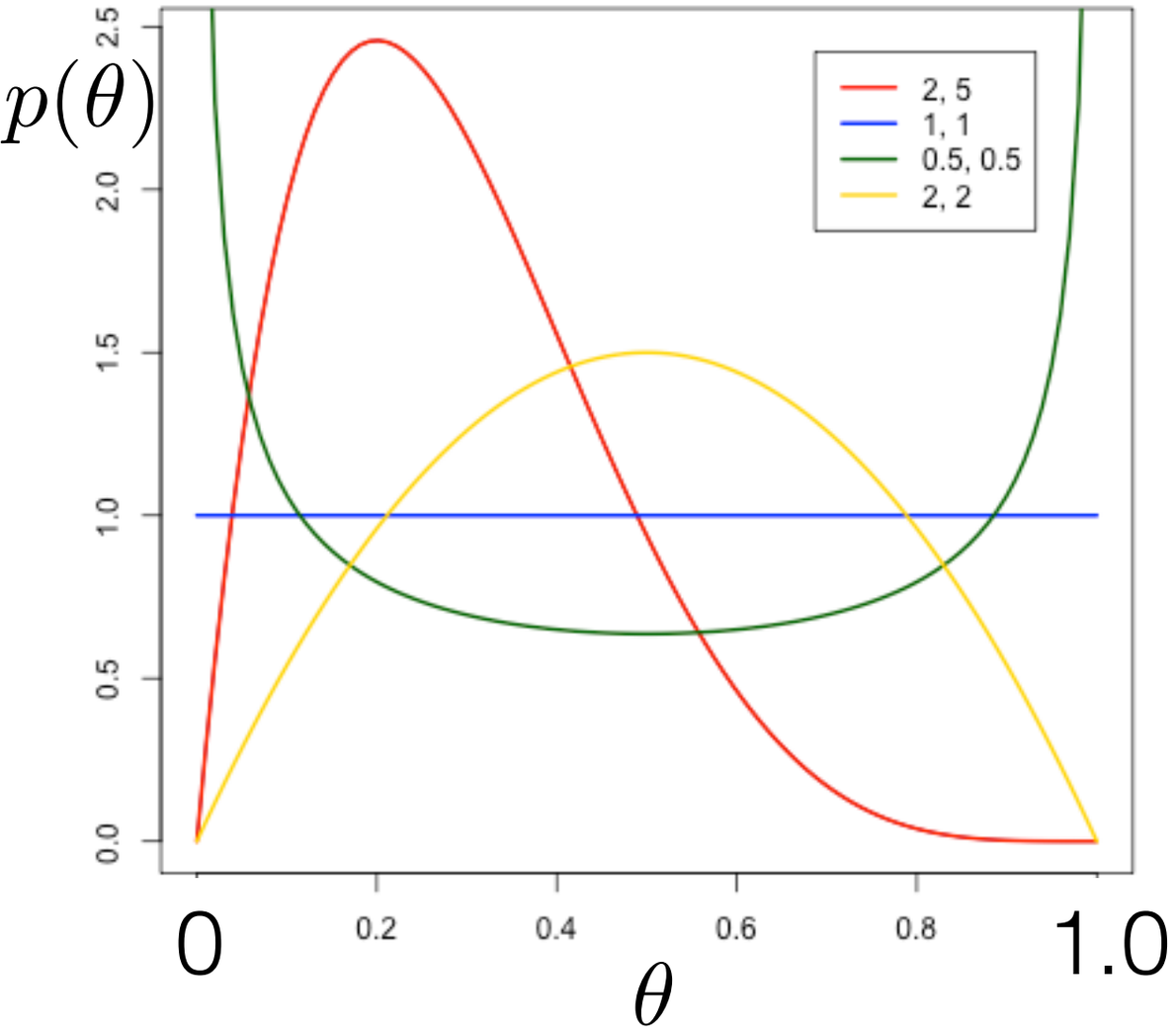
- What happens?
- When *a* = *b* = 1
- As *a*, *b* get very small

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



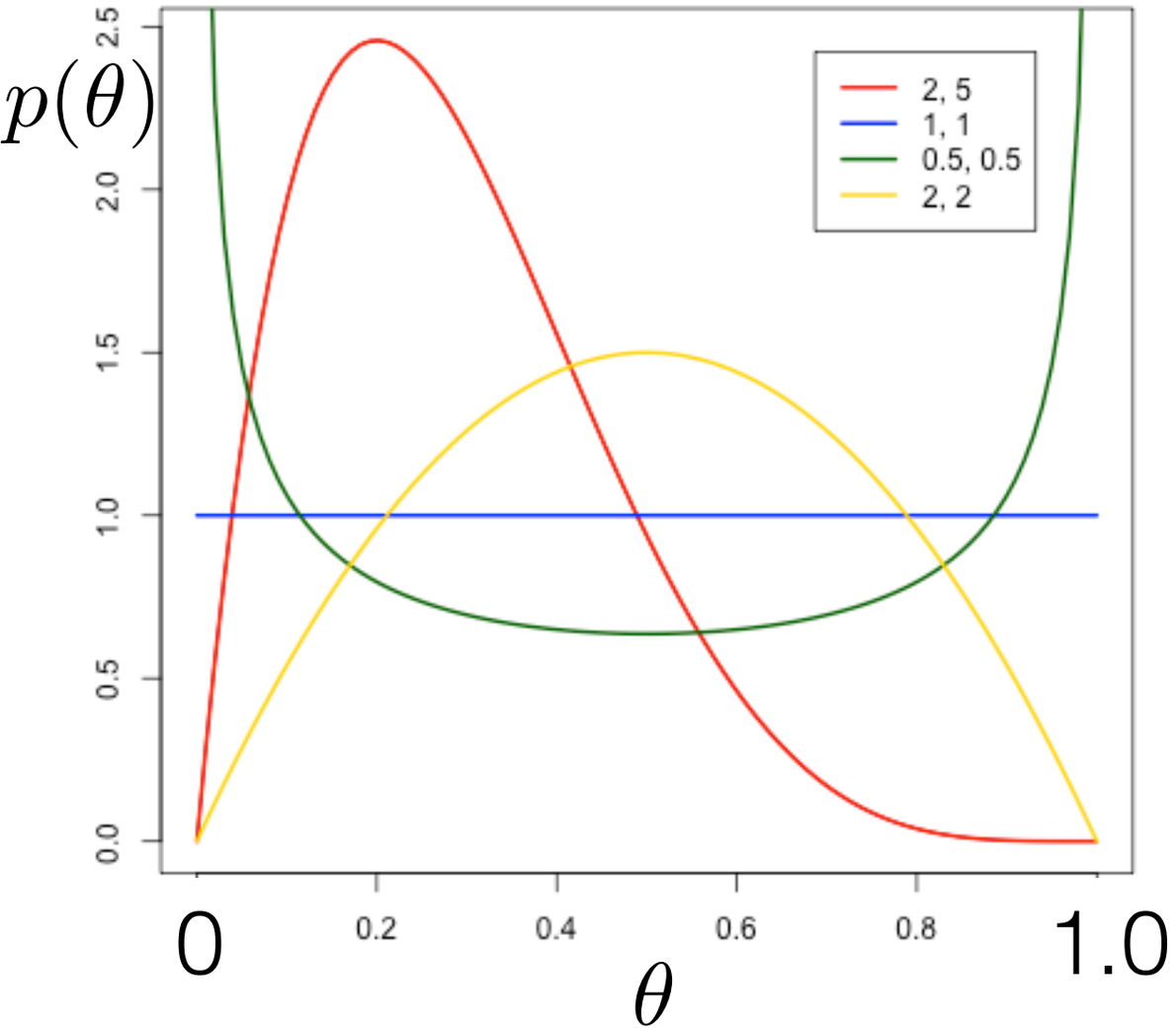
- What happens?
- When $a = b = 1$
- As a, b get very small
- As a, b get very large

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



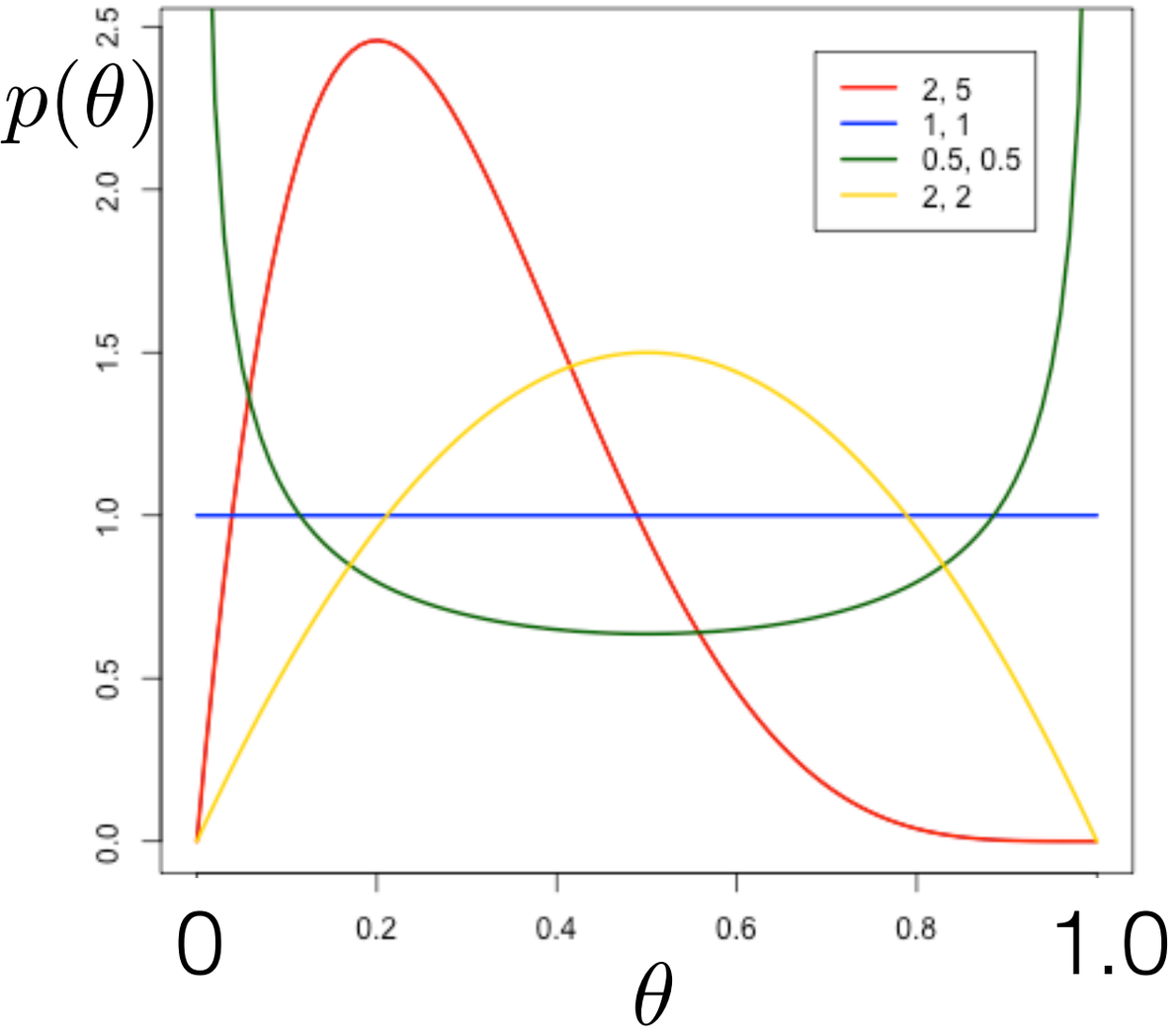
- What happens?
- When $a = b = 1$
- As a, b get very small
- As a, b get very large
- When $a > b$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



- What happens?
- When $a = b = 1$
- As a, b get very small
- As a, b get very large
- When $a > b$

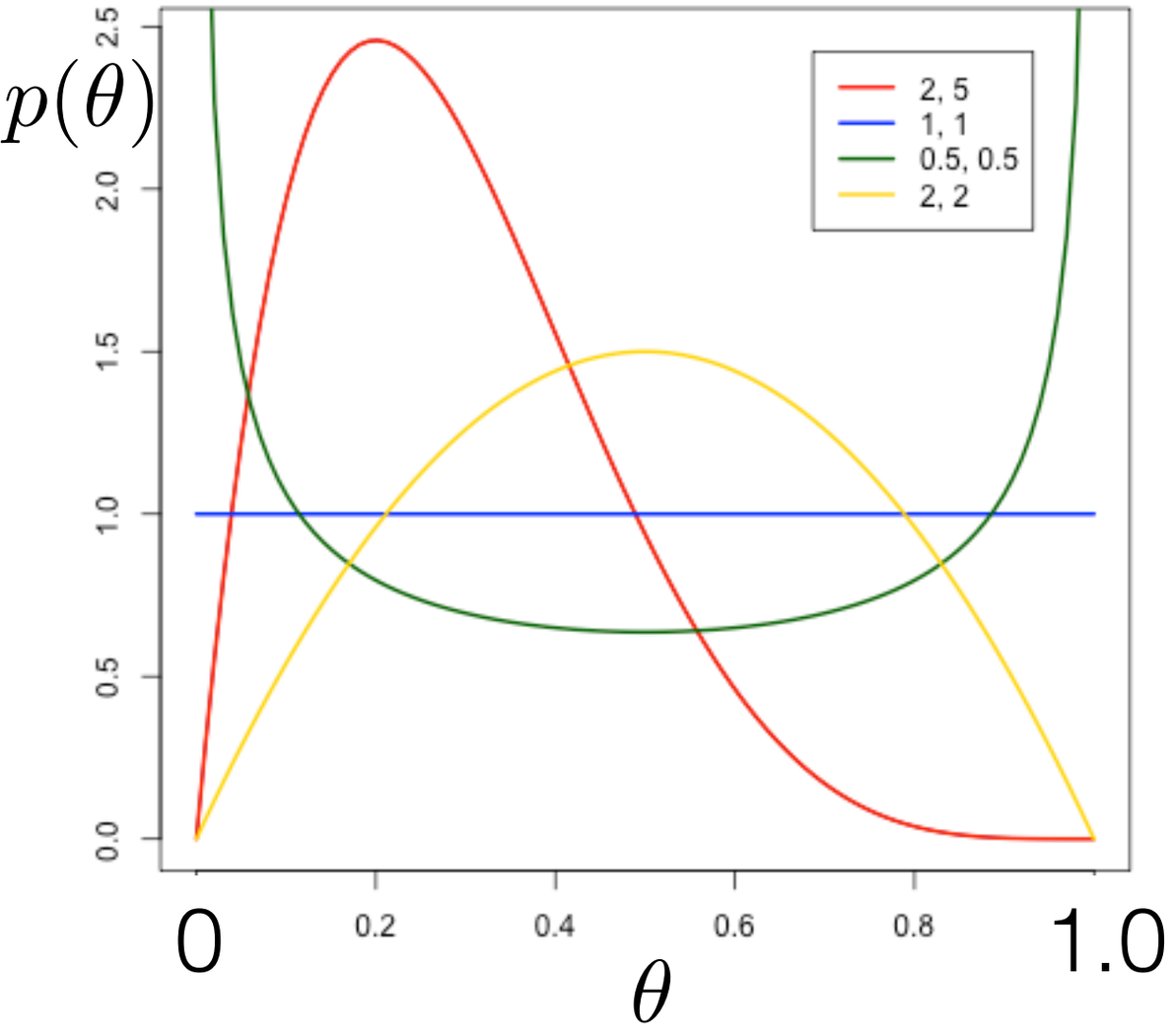
[demo]

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



- What happens?
- When $a = b = 1$
- As a, b get very small
- As a, b get very large
- When $a > b$

[demo]

- Can check, if $\Theta \sim \text{Beta}(a,b)$,

$$\mathbb{E}[\Theta] = \frac{a}{a+b}$$

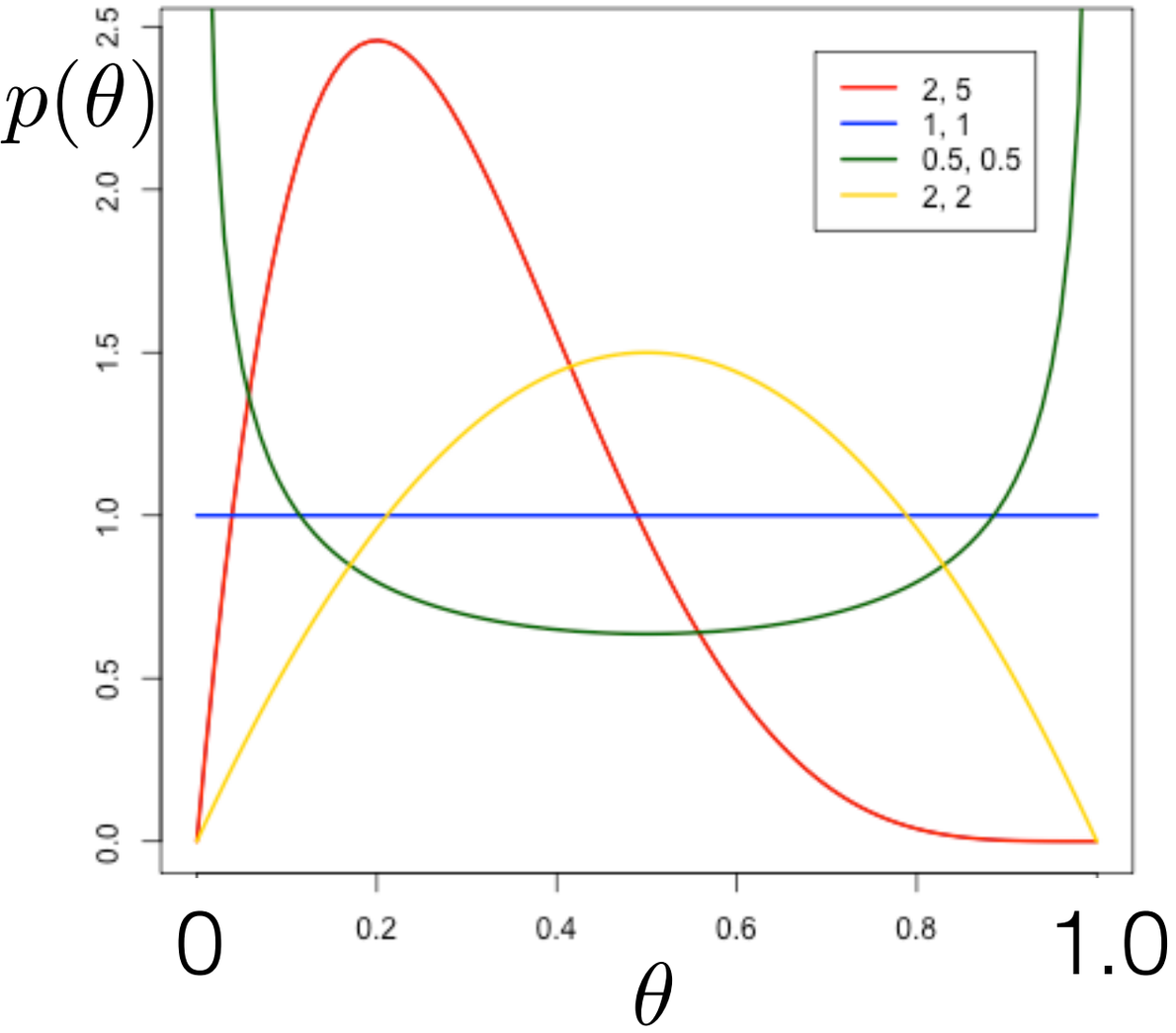
$$\text{Var}[\Theta] = \frac{ab}{(a+b)^2(a+b+1)}$$

Beta distribution

hyperparameters

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1), \quad a,b>0$$

- For positive integers: $\Gamma(n) = (n-1)!$
- For $t > 0$, $\Gamma(t+1) = t\Gamma(t)$



- What happens?
- When $a = b = 1$
- As a, b get very small
- As a, b get very large
- When $a > b$

[demo]

- Can check, if $\Theta \sim \text{Beta}(a,b)$,

$$\mathbb{E}[\Theta] = \frac{a}{a+b}$$

$$\text{Var}[\Theta] = \frac{ab}{(a+b)^2(a+b+1)}$$

revisit questions above

From prior to posterior

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta)$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$p(\theta|\mathcal{D})$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$p(\theta|\mathcal{D})$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$p(\theta|\mathcal{D}) \propto_\theta \left[\prod_{n=1}^{N} \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a, b)$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^N p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^N \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\ &\propto_\theta \left[\prod_{n=1}^N \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^{N} \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\ &\propto_\theta \left[\prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^{N} \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\ &\propto_\theta \left[\prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^N p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^N \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\ &\propto_\theta \left[\prod_{n=1}^N \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1} \end{aligned}$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^N p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned}
p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^N \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\
&\propto_\theta \left[\prod_{n=1}^N \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \\
&= \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1} \\
&\propto_\theta \text{Beta}(\theta|\sum_{n=1}^N y^{(n)} + a, \sum_{n=1}^N (1-y^{(n)}) + b)
\end{aligned}$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^{N} p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^{N} \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\ &\propto_\theta \left[\prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{\sum_{n=1}^{N} y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^{N}(1-y^{(n)})+b-1} \\ &\propto_\theta \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1-y^{(n)}) + b) \end{aligned}$$

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^N p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned}
p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^N \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\
&\propto_\theta \left[\prod_{n=1}^N \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \\
&= \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1} \\
&\propto_\theta \text{Beta}(\theta|\sum_{n=1}^N y^{(n)} + a, \sum_{n=1}^N (1-y^{(n)}) + b)
\end{aligned}$$

- What happens when we have no data?

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^N p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned}
p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^N \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\
&\propto_\theta \left[\prod_{n=1}^N \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \\
&= \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1} \\
&\propto_\theta \text{Beta}(\theta|\sum_{n=1}^N y^{(n)} + a, \sum_{n=1}^N (1-y^{(n)}) + b)
\end{aligned}$$

- What happens when we have no data?
- What happens as we get more and more data?

From prior to posterior

- We have our training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$
- We have our model (likelihood and prior): $p(y|\theta), p(\theta)$
- We can compute the posterior:

$$p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \propto_\theta p(\mathcal{D}|\theta)p(\theta) = \left[\prod_{n=1}^N p(y^{(n)}|\theta)\right] p(\theta)$$

  - $f(\theta) \propto_\theta g(\theta)$: $f(\theta) = cg(\theta)$ for some $c \neq 0$, constant in θ

- Running example:

$$\begin{aligned}
p(\theta|\mathcal{D}) &\propto_\theta \left[\prod_{n=1}^N \text{Bernoulli}(y^{(n)}|\theta)\right] \text{Beta}(\theta|a,b) \\
&\propto_\theta \left[\prod_{n=1}^N \theta^{y^{(n)}}(1-\theta)^{(1-y^{(n)})}\right] \theta^{a-1}(1-\theta)^{b-1} \\
&= \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1} \\
&\propto_\theta \text{Beta}(\theta|\sum_{n=1}^N y^{(n)} + a, \sum_{n=1}^N (1-y^{(n)}) + b)
\end{aligned}$$

- What happens when we have no data?
- What happens as we get more and more data?
- **Conjugate prior** for a likelihood: posterior same form as prior

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

can think of *a* and *b* as "pseudocounts"

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

can think of *a* and *b* as "pseudocounts"

$$= \left(\frac{N}{N+a+b}\right) \frac{\sum_{n=1}^{N} y^{(n)}}{N} + \left(\frac{a+b}{N+a+b}\right) \frac{a}{a+b}$$

MLE

prior mean

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

can think of *a* and *b* as "pseudocounts"

$$= \left(\frac{N}{N+a+b}\right) \frac{\sum_{n=1}^{N} y^{(n)}}{N} + \left(\frac{a+b}{N+a+b}\right) \frac{a}{a+b}$$

MLE — prior mean

- As long as *a*,*b*>0, this estimate can't be 0 or 1

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

can think of *a* and *b* as "pseudocounts"

$$= \left(\frac{N}{N+a+b}\right) \frac{\sum_{n=1}^{N} y^{(n)}}{N} + \left(\frac{a+b}{N+a+b}\right) \frac{a}{a+b}$$

MLE — prior mean

- As long as *a*,*b*>0, this estimate can't be 0 or 1
- In this case, avoiding the overfitting of the MLE

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

can think of *a* and *b* as "pseudocounts"

$$= \left(\frac{N}{N+a+b}\right) \frac{\sum_{n=1}^{N} y^{(n)}}{N} + \left(\frac{a+b}{N+a+b}\right) \frac{a}{a+b}$$

MLE — prior mean

- As long as *a*,*b*>0, this estimate can't be 0 or 1
- In this case, avoiding the overfitting of the MLE
- But same behavior as the MLE when *N* is very large

Interpreting the posterior

- Running example:
  - Likelihood: $y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$
  - Prior: $\theta \sim \text{Beta}(a, b)$
  - We computed the posterior:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

    - Posterior standard deviation: one way to quantify uncertainty (check: gets smaller with more data)
    - Posterior mean: one particular estimate of the parameter

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

can think of *a* and *b* as "pseudocounts"

domain knowledge

$$= \left(\frac{N}{N+a+b}\right) \frac{\sum_{n=1}^{N} y^{(n)}}{N} + \left(\frac{a+b}{N+a+b}\right) \frac{a}{a+b}$$

MLE

prior mean

- As long as *a*,*b*>0, this estimate can't be 0 or 1
- In this case, avoiding the overfitting of the MLE
- But same behavior as the MLE when *N* is very large

From posterior to predictive

From posterior to predictive

- We want to approximate the distribution of a future data point

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D})$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\, d\theta$$

law of total probability

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta$$

law of total probability

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta$$

law of total probability

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D})$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta$$

law of total probability

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D})$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D})$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta$$

law of total probability

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D}) = \int_\theta \theta\, p(\theta|\mathcal{D})\ d\theta$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D}) = \int_\theta \theta\ p(\theta|\mathcal{D})\ d\theta = \mathbb{E}(\theta|\mathcal{D})$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D}) = \int_\theta \theta\ p(\theta|\mathcal{D})\ d\theta = \mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D}) = \int_\theta \theta\ p(\theta|\mathcal{D})\ d\theta = \mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

- The posterior and posterior predictive depend on the likelihood and prior (form & hyperparameters). How do we choose those?

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D}) = \int_\theta \theta\ p(\theta|\mathcal{D})\ d\theta = \mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

- The posterior and posterior predictive depend on the likelihood and prior (form & hyperparameters). How do we choose those?
  - More engineering than math

From posterior to predictive

- We want to approximate the distribution of a future data point
- **Posterior predictive distribution**:

$$p(y^{(N+1)}|\mathcal{D}) = \int_\theta p(y^{(N+1)}, \theta|\mathcal{D})\ d\theta \qquad \text{law of total probability}$$

$$= \int_\theta p(y^{(N+1)}|\theta, \mathcal{D}) p(\theta|\mathcal{D})\ d\theta$$

$$= \int_\theta p(y^{(N+1)}|\theta) p(\theta|\mathcal{D})\ d\theta$$

- Running example:

$$p(y^{(N+1)} = 1|\mathcal{D}) = \int_\theta \theta\ p(\theta|\mathcal{D})\ d\theta = \mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b}$$

- The posterior and posterior predictive depend on the likelihood and prior (form & hyperparameters). How do we choose those?
  - More engineering than math
  - Can use standard/convenient options until something doesn't work

Streaming data

- The concept of a Bayesian update works conveniently with streaming data

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1)$$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$
$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$
$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$

$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$
$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$
$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$

$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$

$$= p(\mathcal{D}_2, \mathcal{D}_1|\theta)p(\theta)$$

likelihood of all data original prior

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$
$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$
$$= p(\mathcal{D}_2, \mathcal{D}_1|\theta)p(\theta)$$

likelihood of all data original prior

- Running example:

Streaming data

- The concept of a Bayesian update works conveniently with streaming data
  - Suppose I get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$
  - Then I get another batch of data $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$
  - I can treat my posterior after the first batch as my prior for the new batch. And it's the same as updating using all the data.

$$p(\theta|\mathcal{D}_2, \mathcal{D}_1) \propto_\theta p(\mathcal{D}_2|\theta, \mathcal{D}_1)p(\theta|\mathcal{D}_1)$$
$$= p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1)$$

likelihood of new data old posterior / new prior

$$\propto_\theta \left[\prod_{n=1}^{N_2} p(y^{(n+N_1)}|\theta)\right] \left[\prod_{n=1}^{N_1} p(y^{(n)}|\theta)\right] p(\theta)$$
$$= p(\mathcal{D}_2, \mathcal{D}_1|\theta)p(\theta)$$

likelihood of all data original prior

- Running example:

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\sum_{n=1}^{N} y^{(n)} + a, \sum_{n=1}^{N}(1 - y^{(n)}) + b)$$

MAP estimation and prediction

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D})$$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D})$ vs. $\theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) \qquad \text{vs. } \theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\begin{aligned}\theta_{\mathrm{MAP}} &= \arg\max_\theta p(\theta|\mathcal{D}) \\ &= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \\ &= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

vs. $\theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\begin{aligned}\theta_{\mathrm{MAP}} &= \arg\max_\theta p(\theta|\mathcal{D}) \\ &= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \\ &= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

vs. $\theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$

as with MLE, in practice you want to optimize the log posterior

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\begin{aligned}\theta_{\mathrm{MAP}} &= \arg\max_\theta p(\theta|\mathcal{D}) \\ &= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D}) \\ &= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

vs. $\theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$

as with MLE, in practice you want to optimize the log posterior

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) \qquad \text{vs. } \theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$$

as with MLE, in practice you want to optimize the log posterior

- Example: $p(\theta|\mathcal{D}) \propto_\theta \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1}$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) \qquad \text{vs. } \theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$$

as with MLE, in practice you want to optimize the log posterior

- Example: $p(\theta|\mathcal{D}) \propto_\theta \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1}$

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\sum_{n=1}^N y^{(n)}+a-1}{N+(a-1)+(b-1)}$$

if posterior exponents are positive, by the derivation in lecture 2

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) \quad \text{vs.} \quad \theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$$

as with MLE, in practice you want to optimize the log posterior

- Example: $p(\theta|\mathcal{D}) \propto_\theta \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1}$

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\sum_{n=1}^N y^{(n)}+a-1}{N+(a-1)+(b-1)}$$

if posterior exponents are positive, by the derivation in lecture 2

vs. $\hat{\theta}_{\mathrm{MLE}} = \frac{\sum_{n=1}^N y^{(n)}}{N}$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) \quad \text{vs.} \quad \theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$$

as with MLE, in practice you want to optimize the log posterior

- Example: $p(\theta|\mathcal{D}) \propto_\theta \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1}$

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\sum_{n=1}^N y^{(n)}+a-1}{N+(a-1)+(b-1)}$$

if posterior exponents are positive, by the derivation in lecture 2

vs. $\hat{\theta}_{\mathrm{MLE}} = \frac{\sum_{n=1}^N y^{(n)}}{N}$ vs. $p(\theta|\mathcal{D})$ with $\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^N y^{(n)}+a}{N+a+b}$

MAP estimation and prediction

- More complex models → Bayesian posteriors/predictives not typically available in closed form
  - Requires approximate integration, which is difficult and computationally expensive to do well
  - Recall: with the MLE, it's easy to use fast optimizers
- A possible intermediate choice: using the **maximum a posteriori** (MAP) estimate

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) \quad \text{vs. } \theta_{\mathrm{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$
$$= \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$$

as with MLE, in practice you want to optimize the log posterior

- Example: $p(\theta|\mathcal{D}) \propto_\theta \theta^{\sum_{n=1}^N y^{(n)}+a-1}(1-\theta)^{\sum_{n=1}^N (1-y^{(n)})+b-1}$

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\sum_{n=1}^N y^{(n)}+a-1}{N+(a-1)+(b-1)}$$

if posterior exponents are positive, by the derivation in lecture 2

vs. $\hat{\theta}_{\mathrm{MLE}} = \frac{\sum_{n=1}^N y^{(n)}}{N}$ vs. $p(\theta|\mathcal{D})$ with $\mathbb{E}(\theta|\mathcal{D}) = \frac{\sum_{n=1}^N y^{(n)}+a}{N+a+b}$

- Predictive: $p(y|\hat{\theta}_{\mathrm{MAP}})$ vs. $p(y|\hat{\theta}_{\mathrm{MLE}})$ vs. $p(y^{(n+1)} = y|\mathcal{D})$