Introduction to Bayesian methods

Ekaterina Lobacheva

Higher School of Economics, Samsung-HSE Laboratory
Moscow, Russia

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

SAMSUNG Research

p(B|A)yesgroup.ru

Problem set

The problem set is
available here:

tiny.cc/ASGM_bayes_problems

Problem 1: Bayesian reasoning

Setting

During medical checkup, one of the tests indicates a serious disease. The test has high accuracy 99% (probability of true positive is 99%, probability of true negative is 99%). However, the disease is quite rare, and only one person in 10000 is affected.

Question

Calculate the probability that the examined person has the disease.

Problem 1: Bayesian reasoning

- $d \in \{0, 1\}$ — disease (1 means that the person has a disease)
- $t \in \{0, 1\}$ — test (1 means that test says that the person has a disease)

Setting: $p(t = 1 \mid d = 1) = p(t = 0 \mid d = 0) = 0.99, \quad p(d = 1) = 10^{-4}$

Question: $p(d = 1 \mid t = 1) =?$

Problem 1: Bayesian reasoning

- $d \in \{0,1\}$ — disease (1 means that the person has a disease)
- $t \in \{0,1\}$ — test (1 means that test says that the person has a disease)

Setting: $p(t=1 \mid d=1) = p(t=0 \mid d=0) = 0.99, \quad p(d=1) = 10^{-4}$

Question: $p(d=1 \mid t=1) = ?$

$$p(d=1 \mid t=1) = \frac{p(t=1 \mid d=1)p(d=1)}{p(t=1 \mid d=1)p(d=1) + p(t=1 \mid d=0)p(d=0)} =$$

$$= \frac{0.99 \cdot 10^{-4}}{0.99 \cdot 10^{-4} + 0.01 \cdot (1 - 10^{-4})} \approx 1\%$$

Example: coin tossing

- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: $X = (x_1, \ldots, x_n)\,, \quad x \in \{0, 1\}$

Head (H)

Tail (T)

Probabilistic model:

$$p(x, \theta) = p(x \mid \theta) p(\theta)$$

Example: coin tossing

- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: $X = (x_1, \ldots, x_n)\,, \quad x \in \{0, 1\}$


Head (H)


Tail (T)

Probabilistic model:

$$p(x, \theta) = p(x \mid \theta) p(\theta)$$

$$\boxed{\textbf{Likelihood: } Bern(x \mid \theta) = \theta^x (1 - \theta)^{1-x}}$$

Example: coin tossing

- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: $X = (x_1, \ldots, x_n)\,, \quad x \in \{0, 1\}$


Head (H)


Tail (T)

Probabilistic model:

$$p(x, \theta) = p(x \mid \theta) p(\theta)$$

Likelihood: $Bern(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$

Prior: ???

Example: coin tossing

How to choose a prior?

- Correct domain: $\theta \in [0, 1]$
- Include prior knowledge: a coin is most likely fair
- Inference complexity: use conjugate prior

Example: coin tossing

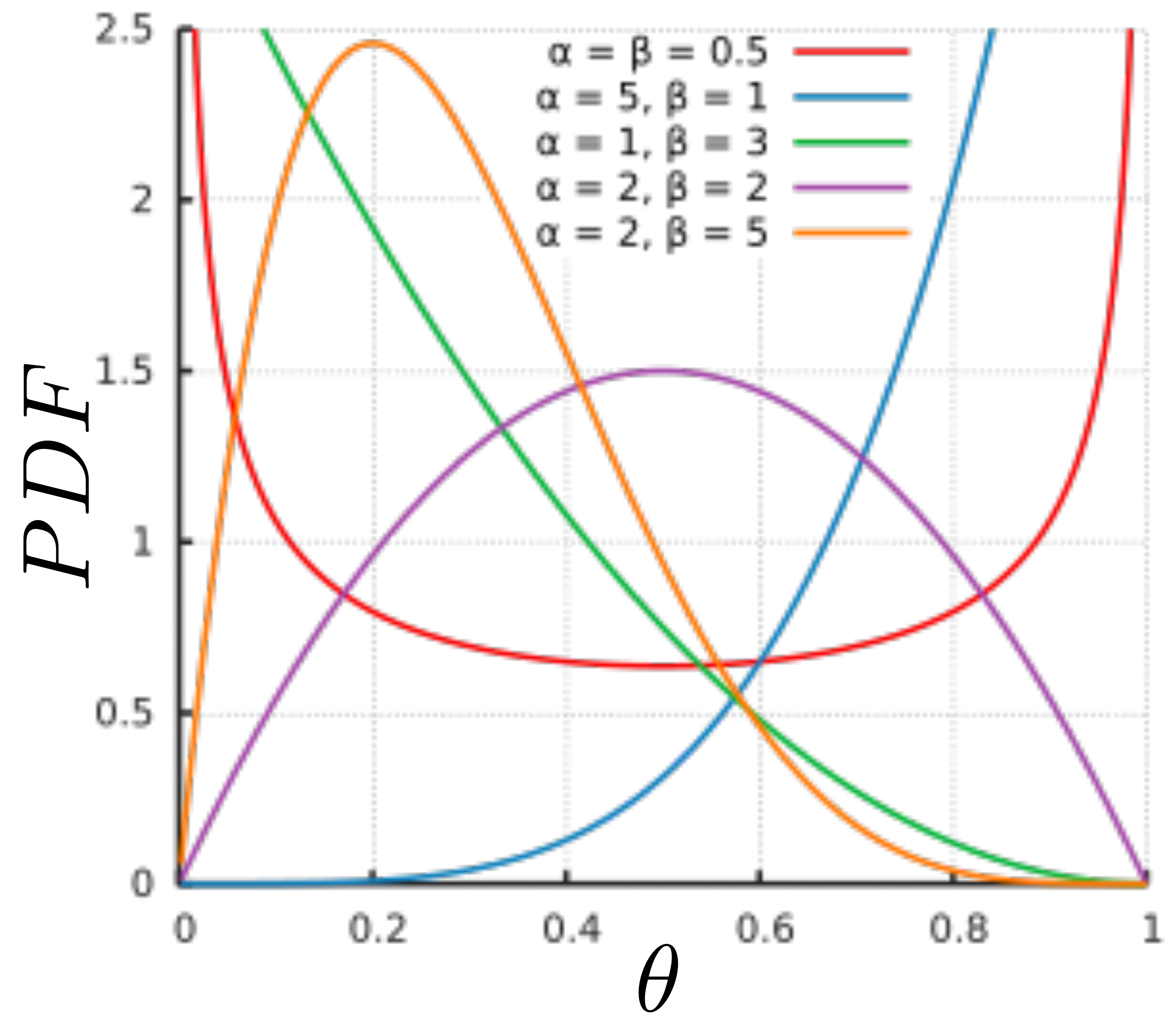
How to choose a prior?

- Correct domain: $\theta \in [0, 1]$
- Include prior knowledge: a coin is most likely fair
- Inference complexity: use conjugate prior

Beta distribution matches all requirements:

$$Beta(\theta \mid a, b) = \frac{1}{\mathrm{B}(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

Beta distribution



Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1}(1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1}(1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

$$p(\theta) = C\theta^C(1-\theta)^C$$

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

$$p(\theta) = C\theta^C (1-\theta)^C$$

$$p(\theta \mid x) = \frac{1}{C} p(x \mid \theta) p(\theta) = \frac{1}{C} \theta^x (1-\theta)^{1-x} \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} =$$

$$= C\theta^C (1-\theta)^C$$

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Let's check that our likelihood and prior are conjugate:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta) = \frac{1}{\mathrm{B}(a,b)} \theta^{a-1}(1-\theta)^{b-1}$$

Idea — check that prior and posterior lay in the same parametric family:

$$p(\theta) = \boxed{C\theta^C (1-\theta)^C} \text{ conjugacy}$$

$$p(\theta \mid x) = \frac{1}{C} p(x \mid \theta) p(\theta) = \frac{1}{C} \theta^x (1-\theta)^{1-x} \frac{1}{\mathrm{B}(a,b)} \theta^{a-1}(1-\theta)^{b-1} =$$

$$= \boxed{C\theta^C (1-\theta)^C} \text{ conjugacy}$$

Here different constants are denoted with
the same letter C for demonstration reasons.

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$p(\theta \mid X) = \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z} \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right) \right] p(\theta) =$$

Example: coin tossing

Bayesian inference after receiving data $X=(x_1,\ldots,x_n)$:

$$p(\theta \mid X)=\frac{1}{Z}p(X \mid \theta)p(\theta)=\frac{1}{Z}\left[\prod_{i=1}^{n}p\left(x_i \mid \theta\right)\right]p(\theta)=$$

$$=\frac{1}{Z}\left[\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}\right]\frac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}=$$

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$p(\theta \mid X) = \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z} \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right) \right] p(\theta) =$$

$$= \frac{1}{Z} \left[\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right] \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} =$$

$$= \frac{1}{Z'} \theta^{a + \sum_{i=1}^{n} x_i - 1} (1-\theta)^{b + n - \sum_{i=1}^{n} x_i - 1}$$

Example: coin tossing

Bayesian inference after receiving data $X = (x_1, \ldots, x_n)$:

$$p(\theta \mid X) = \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z} \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right) \right] p(\theta) =$$

$$= \frac{1}{Z} \left[\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right] \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} =$$

$$= \frac{1}{Z'} \theta^{a + \sum_{i=1}^{n} x_i - 1} (1-\theta)^{b + n - \sum_{i=1}^{n} x_i - 1} = Beta\left(\theta \mid a', b'\right)$$

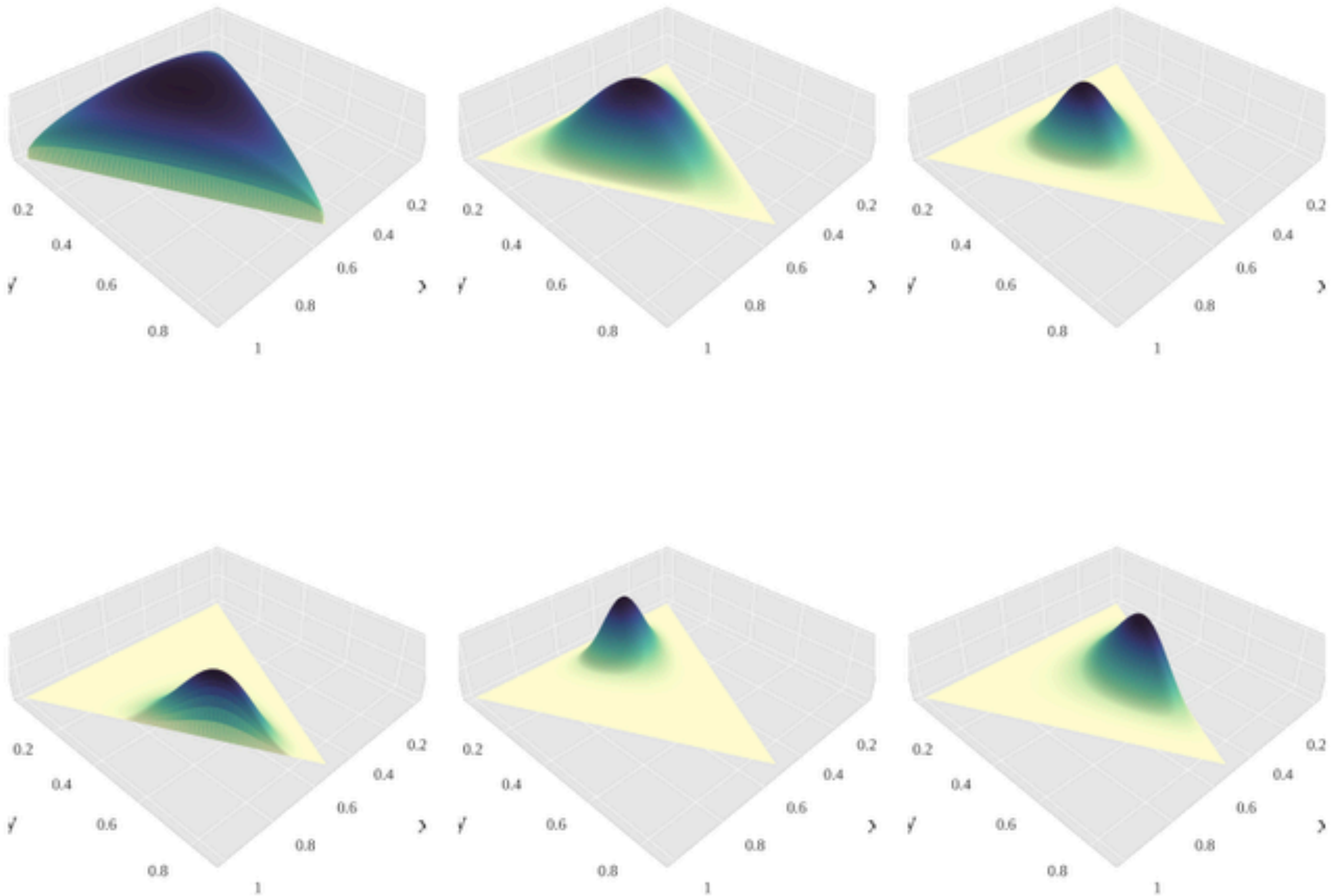
New parameters: $\quad a' = a + \sum_{i=1}^{n} x_i \qquad b' = b + n - \sum_{i=1}^{n} x_i$

Problem 2: Bayesian framework

Setting

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior:
$$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Dirichlet distribution



Beta distribution is a special case of Dirichlet distribution:

$$\mathrm{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$$Beta(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

https://en.wikipedia.org/wiki/Dirichlet distribution

Problem 2: Bayesian framework

Setting

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior: $$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(\theta | X, \alpha)$

✳ Compare $\mathbb{E}_{p(\theta|X,\alpha)}\theta$ and θ_{ML}

✳ Compute the predictive posterior $p(x_{N+1} = j | X, \alpha)$

Problem 2: Bayesian framework

Setting

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior: $$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(\theta|X, \alpha)$

✳ Compare $\mathbb{E}_{p(\theta|X,\alpha)}\theta$ and θ_{ML}

✳ Compute the predictive posterior $p(x_{N+1} = j|X, \alpha)$

Problem 2: Bayesian framework

Probabilistic model: $p(X, \theta) = p(X \mid \theta) p(\theta) = p(X \mid \theta) Dir(\theta \mid \alpha)$

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior:

$$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Here different constants are denoted with
the same letter C for demonstration reasons.

Problem 2: Bayesian framework

Probabilistic model: $p(X, \theta) = p(X \mid \theta)p(\theta) = p(X \mid \theta)Dir(\theta \mid \alpha)$

Prior: $p(\theta) = \dfrac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = C \prod_{k=1}^{K} \theta_k^{C}$

Here different constants are denoted with
the same letter C for demonstration reasons.

Problem 2: Bayesian framework

Probabilistic model: $p(X, \theta) = p(X \mid \theta) p(\theta) = p(X \mid \theta) Dir(\theta \mid \alpha)$

Prior: $$p(\theta) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = C \prod_{k=1}^{K} \theta_k^{C}$$

Posterior: $$p(\theta \mid X) \propto p(X \mid \theta) p(\theta) = \prod_{k=1}^{K} \theta_k^{N_k} \cdot \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} =$$

$$= C \prod_{k=1}^{K} \theta_k^{C}$$

Here different constants are denoted with the same letter C for demonstration reasons.

Problem 2: Bayesian framework

Probabilistic model: $p(X, \theta) = p(X \mid \theta)p(\theta) = p(X \mid \theta)Dir(\theta \mid \alpha)$

Prior: $$p(\theta) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \boxed{C \prod_{k=1}^{K} \theta_k^{C}}$$

Posterior: $$p(\theta \mid X) \propto p(X \mid \theta)p(\theta) = \prod_{k=1}^{K} \theta_k^{N_k} \cdot \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} =$$

$$= \boxed{C \prod_{k=1}^{K} \theta_k^{C}}$$

conjugate

Here different constants are denoted with the same letter C for demonstration reasons.

Problem 2: Bayesian framework

Setting

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior:

$$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Questions

- Check that likelihood and prior are conjugate
- $\boxed{\text{Compute the posterior } p(\theta|X, \alpha)}$

∗ Compare $\mathbb{E}_{p(\theta|X,\alpha)}\theta$ and θ_{ML}

∗ Compute the predictive posterior $p(x_{N+1} = j|X, \alpha)$

Problem 2: Bayesian framework

Likelihood and prior are conjugate $\longrightarrow$ posterior is Dirichlet

$$p(\theta \mid X) \propto p(X \mid \theta) p(\theta) = \prod_{k=1}^{K} \theta_k^{N_k} \cdot \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \propto$$

$$\propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

Problem 2: Bayesian framework

Likelihood and prior are conjugate $\longrightarrow$ posterior is Dirichlet

$$p(\theta \mid X) \propto p(X \mid \theta)p(\theta) = \prod_{k=1}^{K} \theta_k^{N_k} \cdot \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \propto$$

$$\propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

$$p(\theta \mid X) = Dir(\theta \mid \alpha'), \quad \alpha' = (\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

Problem 2: Bayesian framework

Setting

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior: $$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(\theta|X, \alpha)$

∗ Compare $\mathbb{E}_{p(\theta|X,\alpha)}\theta$ and θ_{ML}

∗ Compute the predictive posterior $p(x_{N+1} = j|X, \alpha)$

Problem 2: frequentist framework

θ is restricted to simplex. To omit the inequality restrictions change parameterization to $\mu_k = \log \theta_k, \quad \mu_k \in \mathbb{R}$

The Lagrangian has the form:

$$\mathcal{L}(\mu, \lambda) = \log p(X \mid \exp \mu) - \lambda(\textstyle\sum_{k=1}^K \exp \mu_k - 1) =$$
$$= \textstyle\sum_{k=1}^K (N_k \mu_k - \lambda \exp \mu_k) + \lambda$$

Differentiation:

$$0 = \frac{\partial \mathcal{L}(\boldsymbol{\mu}, \lambda)}{\partial \mu_k} = N_k - \lambda \exp \mu_k \Rightarrow \theta_k = \exp \mu_k = \frac{N_k}{\lambda}$$
$$0 = \frac{\partial \mathcal{L}(\boldsymbol{\mu}, \lambda)}{\partial \lambda} = -\textstyle\sum_{k=1}^K \exp \mu_k + 1 \Rightarrow \lambda = \textstyle\sum_{k=1}^K N_k$$

$$\theta_k = \frac{N_k}{\sum_{l=1}^K N_l}$$

Problem 2: Bayesian framework

Maximum likelihood estimate: $$\theta_k = \frac{N_k}{\sum_{l=1}^{K} N_l}$$

Expectation of the posterior: $$\mathbb{E}_{p(\theta|X)}\theta_k = \frac{\alpha_k + N_k}{\sum_{l=1}^{K} \alpha_l + N_l}$$

Small K $\longrightarrow$ Bayesian estimate is mostly based on prior

Large K $\longrightarrow$ Bayesian estimate is very similar to ML estimate

Problem 2: Bayesian framework

Setting

- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$
- Dirichlet prior: $$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(\theta|X, \alpha)$

∗ Compare $\mathbb{E}_{p(\theta|X,\alpha)}\theta$ and θ_{ML}

∗ Compute the predictive posterior $p(x_{N+1} = j|X, \alpha)$

Problem 2: Bayesian framework

$$p(x_{N+1} = j \mid X, \alpha) = \int_{S_K} p(x_{N+1} = j \mid \theta) p(\theta \mid X, \alpha) \mathrm{d}\theta =$$

Problem 2: Bayesian framework

$$p(x_{N+1} = j \mid X, \alpha) = \int_{S_K} p(x_{N+1} = j \mid \theta) p(\theta \mid X, \alpha) \mathrm{d}\theta =$$

$$= \frac{\int_{S_K} \theta_j \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \mathrm{d}\theta}{B(\alpha_1 + N_1, \ldots, \alpha_K + N_K)} = \frac{B(\alpha_1 + N_1, \ldots, \alpha_j + N_j + 1, \ldots, \alpha_K + N_K)}{B(\alpha_1 + N_1, \ldots, \alpha_j + N_j, \ldots, \alpha_K + N_K)} =$$

$$= \frac{\Gamma(\alpha_1 + N_1) \ldots \Gamma(\alpha_j + N_j + 1) \ldots \Gamma(\alpha_K + N_K)}{\Gamma(\alpha_1 + N_1) \ldots \Gamma(\alpha_j + N_j) \ldots \Gamma(\alpha_K + N_K)} \cdot \frac{\Gamma(\sum_l (\alpha_l + N_l))}{\Gamma(\sum_l (\alpha_l + N_l) + 1)} =$$

$$= \frac{\alpha_j + N_j}{\sum_k \alpha_k + N}$$