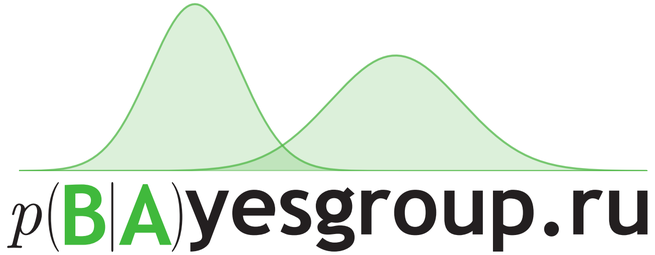
Introduction to Bayesian methods

Ekaterina Lobacheva

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

SAMSUNG
Research



Slides are partially based on lectures of Dmitry Vetrov, Dmitry Kropotov and Kirill Struminsky, deepbayes.ru/2018

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Practice
- Approximate Bayesian inference

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Practice
- Approximate Bayesian inference

How to work with distributions?

$$\text{Conditional} = \frac{\text{Joint}}{\text{Marginal}}, \quad p(x|y) = \frac{p(x,y)}{p(y)}$$

Product rule

any joint distribution can be expressed as a product of one-dimensional conditional distributions

$$p(x,y,z) = p(x|y,z)p(y|z)p(z)$$

Sum rule

any marginal distribution can be obtained from the joint distribution by integrating out

$$p(y) = \int p(x,y)dx$$

Example

- We have a joint distribution over three groups of variables $p(x, y, z)$
- We observe x and are interested in predicting y
- Values of z are unknown and irrelevant to us
- How to estimate $p(y|x)$ from $p(x, y, z)$?

Example

- We have a joint distribution over three groups of variables $p(x, y, z)$
- We observe x and are interested in predicting y
- Values of z are unknown and irrelevant to us
- How to estimate $p(y|x)$ from $p(x, y, z)$?

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{\int p(x, y, z)dz}{\int p(x, y, z)dzdy}$$

Sum rule and product rule allow to obtain arbitrary conditional distributions from the joint one

Bayes theorem

Bayes theorem (follows from product and sum rules):

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

Bayes theorem defines the rule for uncertainty conversion when new information arrives:

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

Statistical inference

Problem: given i.i.d. data $X = (x_1, ..., x_n)$ from distribution $p(x|\theta)$ one needs to estimate θ

Frequentist framework: use maximum likelihood estimation (MLE)

$$\theta_{ML} = \arg\max p(X|\theta) = \arg\max \prod_{i=1}^{n} p\left(x_i|\theta\right) = \arg\max \sum_{i=1}^{n} \log p\left(x_i|\theta\right)$$

Bayesian framework: encode uncertainty about θ in a prior $p(\theta)$ and apply Bayesian inference

$$p(\theta|X) = \frac{\prod_{i=1}^{n} p\left(x_i|\theta\right) p(\theta)}{\int \prod_{i=1}^{n} p\left(x_i|\theta\right) p(\theta) d\theta}$$

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: 2 tosses with a result (H,H)

Head (H) Tail (T)

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: 2 tosses with a result (H,H)

Head (H) Tail (T)

Frequentist framework:

In all experiments the coin landed heads up

$$\theta_{ML} = 1$$

→

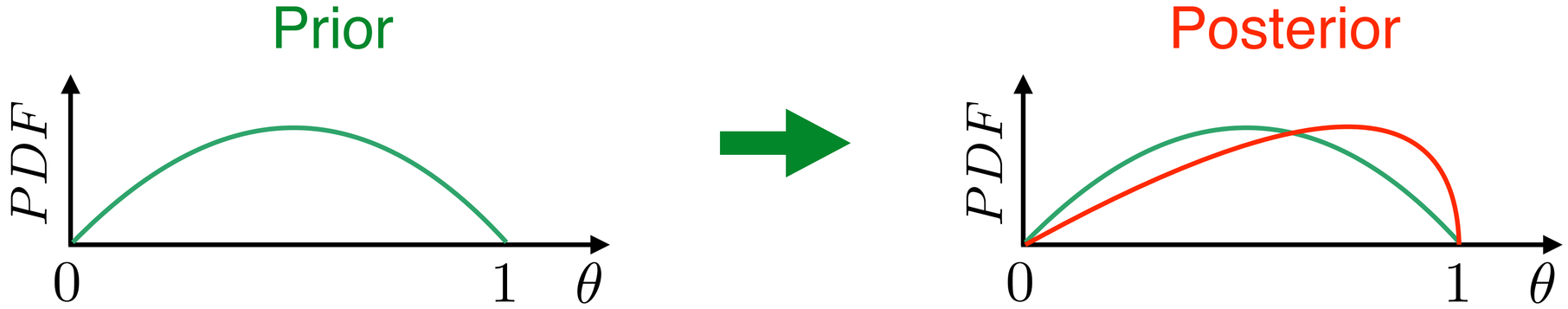
The coin is not fair and always lands heads up

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: 2 tosses with a result (H,H)

Head (H) Tail (T)

Bayesian framework:



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: 1000 tosses with a result (H,H,T,…) — 489 tails and 511 heads

Head (H)

Tail (T)

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: 1000 tosses with a result (H,H,T,…) — 489 tails and 511 heads

Head (H) Tail (T)

Both frameworks:

Sufficient amount of data matches our expectations → The coin is fair

Frequentist vs. Bayesian frameworks

	Frequentist	Bayesian
Variables	random and deterministic	everything is random
Applicability	$n >> d$	$\forall n$

- The number of tunable parameters in modern ML models is comparable with the sizes of training data
- Frequentist framework is a limit case of Bayesian one:

$$\lim_{n/d \to \infty} p\left(\theta | x_1, \ldots, x_n\right) = \delta\left(\theta - \theta_{ML}\right)$$

Advantages of Bayesian framework

- We can encode our prior knowledge or desired properties of the final solution into a prior distribution
- Prior is a form of regularization
- Additionally to the point estimate of θ posterior contains information about the uncertainty of the estimate

Bayesian framework just provides an alternative point of view, it DOES NOT contradict or deny frequentist framework

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Practice
- Approximate Bayesian inference

Probabilistic ML model

For each object in the data:

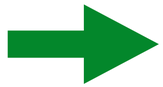
- x — set of observed variables (features)
- y — set of hidden / latent variables (class label / hidden representation, etc.)

Model:

- θ — model parameters (e.g. weights of the linear model)

Discriminative probabilistic ML model

Models $p(y, \theta \mid x)$



Cannot generate new objects — needs x as an input

Usually assumes that prior over θ does not depend on x:

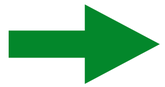
$$p(y, \theta \mid x) = p(y \mid x, \theta)p(\theta)$$

Examples:

- Classification or regression task (hidden space is much easier than the observed one)
- Machine translation (hidden and observed spaces have the same complexity)

Generative probabilistic ML model

Models joint distribution
$p(x, y, \theta) = p(x, y \mid \theta)p(\theta)$



Can generate new objects,
i.e. pairs (x, y)

May be quite difficult to train since the observed space is usually much more complicated than the hidden one

Examples:

- Generation of text, speech, images, etc.

Training Bayesian ML models

We are given training data (X_{tr}, Y_{tr}) and a discriminative model $p(y, \theta \mid x)$

Training stage — Bayesian inference over θ:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}$$

Result: ensemble of algorithms rather than a single one θ_{ML}

- Ensemble usually outperforms single best model
- Posterior capture all dependencies from the training data that the model could extract and may be used as a new prior later

Predictions of Bayesian ML models

Testing stage:

- From training we have a posterior distribution $p\left(\theta \mid X_{tr}, Y_{tr}\right)$
- New data point x arrives
- We need to compute the predictive distribution on its hidden value y

Ensembling w.r.t. posterior over the parameters θ:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta$$

Bayesian ML models

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}$$

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta$$

Bayesian ML models

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}$$

May be intractable

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta$$

When are the integrals tractable?
What can we do if they are intractable?

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Practice
- Approximate Bayesian inference

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \qquad \longrightarrow \qquad p(\theta \mid x) \in \mathcal{A}\,(\alpha')$$

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \qquad \longrightarrow \qquad p(\theta \mid x) \in \mathcal{A}\,(\alpha')$$

Intuition:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta}$$

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \quad \longrightarrow \quad p(\theta \mid x) \in \mathcal{A}(\alpha')$$

Intuition:

$$p(\theta \mid x) = \frac{\boxed{p(x \mid \theta)p(\theta)}}{\int p(x \mid \theta)p(\theta)d\theta} \leftarrow \text{conjugate}$$

- Denominator is tractable since any distribution in $\mathcal{A}$ is normalized

Conjugate distributions

Distribution $p(\theta)$ and $p(x \mid \theta)$ are conjugate iff $p(\theta \mid x)$ belongs to the same parametric family as $p(\theta)$:

$$p(\theta) \in \mathcal{A}(\alpha), \quad p(x \mid \theta) \in \mathcal{B}(\theta) \qquad \longrightarrow \qquad p(\theta \mid x) \in \mathcal{A}\,(\alpha')$$

Intuition:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta} \propto p(x \mid \theta)p(\theta)$$

- Denominator is tractable since any distribution in $\mathcal{A}$ is normalized
- All we need is to compute α'

Full Bayesian inference

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}$$

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta$$

Integrals are tractable if prior and likelihood are conjugate

Full Bayesian inference

- Easy to use - analytical formulas for training and testing stages
- Strong assumptions on the model - conjugacy of prior and likelihood

  → Choose conjugate prior

  → Only simple models (not flexible enough for most of the cases)

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:

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

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 **(purple prior)**
- 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

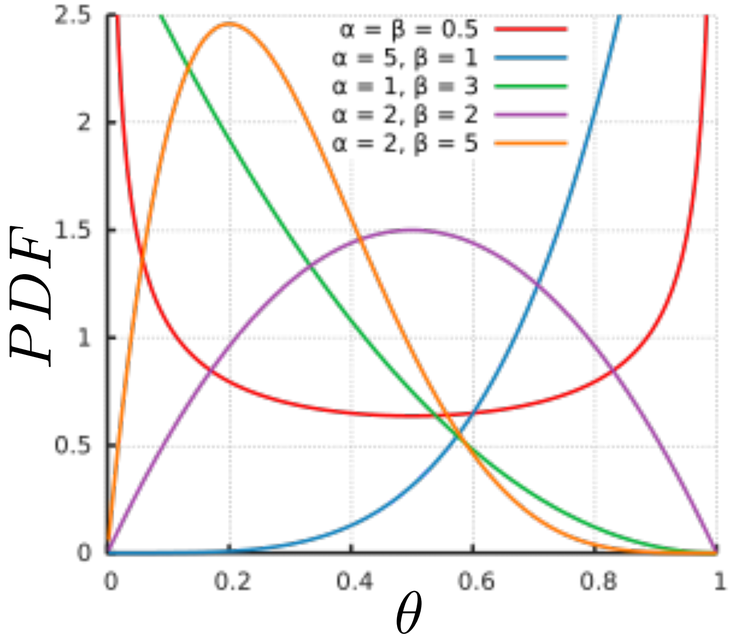


Image: https://en.wikipedia.org/wiki/Beta_distribution

Example: coin tossing

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

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

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:

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

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:

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

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:

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

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

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

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

$$= \boxed{C\theta^C(1-\theta)^C}$$ 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: $a' = a + \sum_{i=1}^{n} x_i \qquad b' = b + n - \sum_{i=1}^{n} x_i$

Example: coin tossing

How to encode different prior knowledge?

- a coin is most likely fair — **purple prior**

Beta distribution

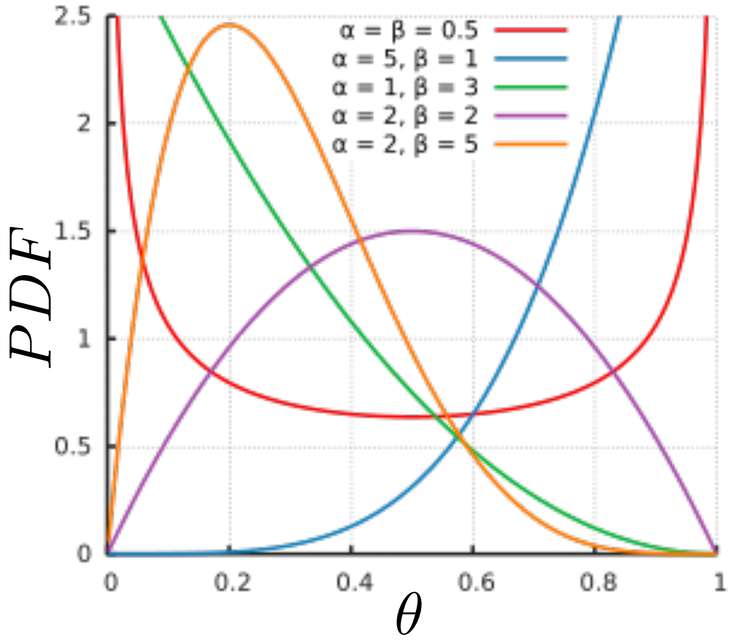


Image: https://en.wikipedia.org/wiki/Beta_distribution

Example: coin tossing

How to encode different prior knowledge?

- a coin is most likely fair — **purple prior**
- two sides of a coin are most likely the same — ?

Beta distribution

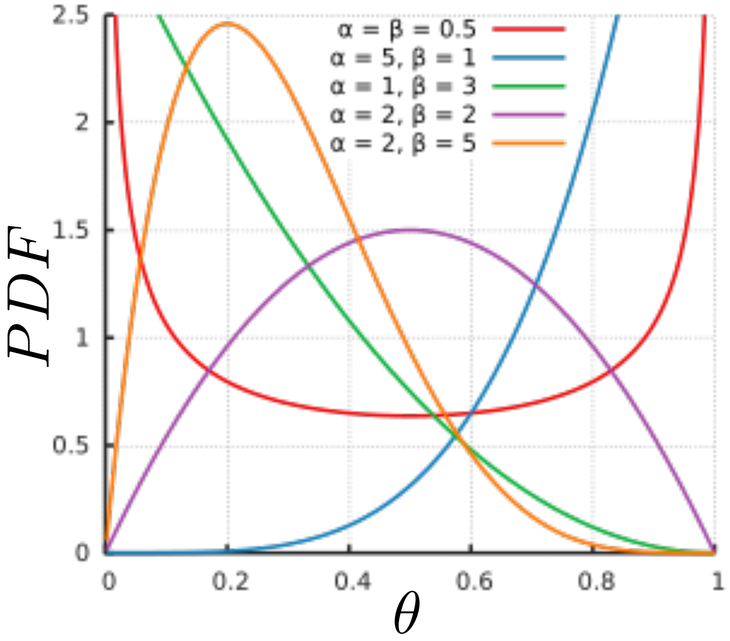


Image: https://en.wikipedia.org/wiki/Beta_distribution

Example: coin tossing

How to encode different prior knowledge?

- a coin is most likely fair — **purple prior**
- two sides of a coin are most likely the same — **red prior**
- a coin center of mass is biased —?

Beta distribution

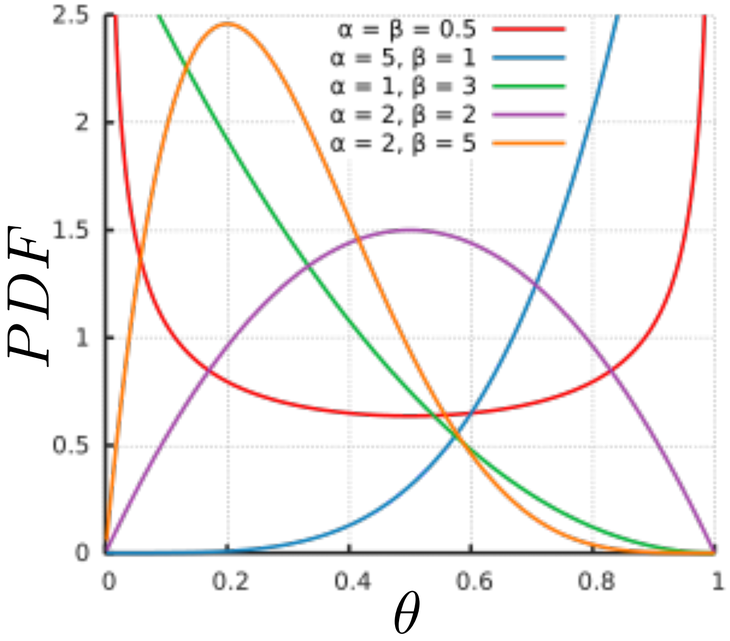


Image: https://en.wikipedia.org/wiki/Beta_distribution

Example: coin tossing

How to encode different prior knowledge?

- a coin is most likely fair — **purple prior**
- two sides of a coin are most likely the same — **red prior**
- a coin center of mass is biased — mixture of two: **orange prior** and its reflection

Beta distribution

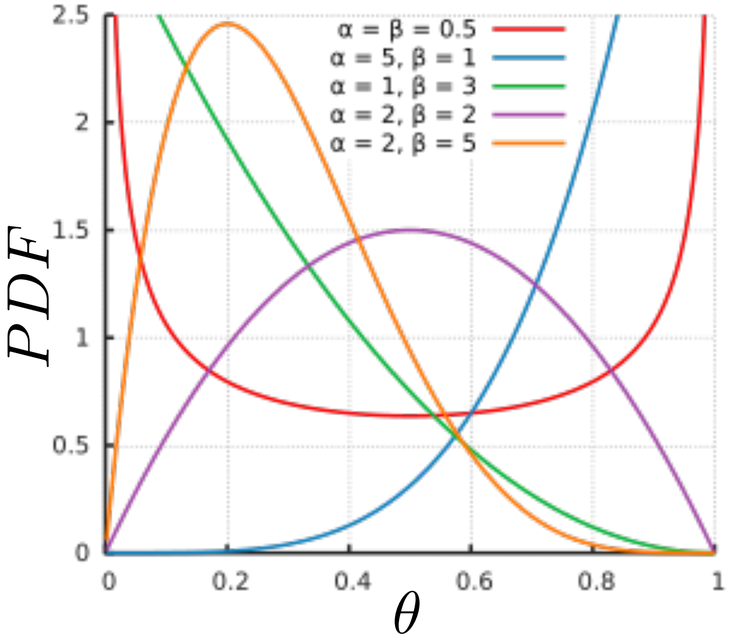


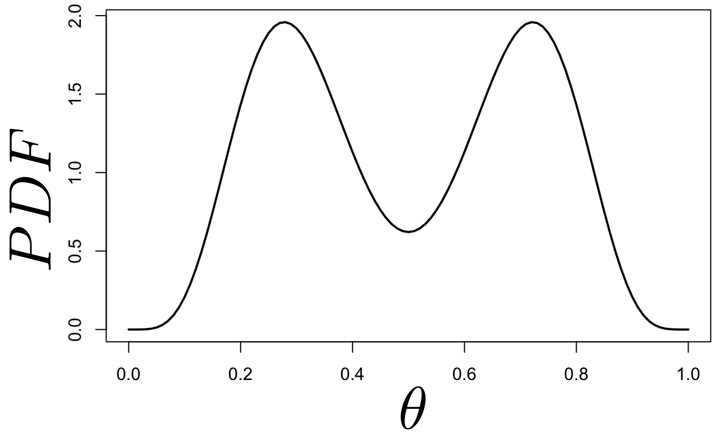
Image: https://en.wikipedia.org/wiki/Beta_distribution

Example: coin tossing

How to encode different prior knowledge?

- a coin is most likely fair — **purple prior**
- two sides of a coin are most likely the same — **red prior**
- a coin center of mass is biased — mixture of two: **orange prior** and its reflection

Mixture of two Beta distributions



Mixture of two Beta distributions:

$$p(\theta) = wBeta(\theta \mid a_1, b_1) + (1-w)Beta(\theta \mid a_2, b_2), \quad w \in [0,1]$$

Image: http://websites.uwlax.edu/storibio/stat762_spr19/r_codes_day8_mix.html

What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$

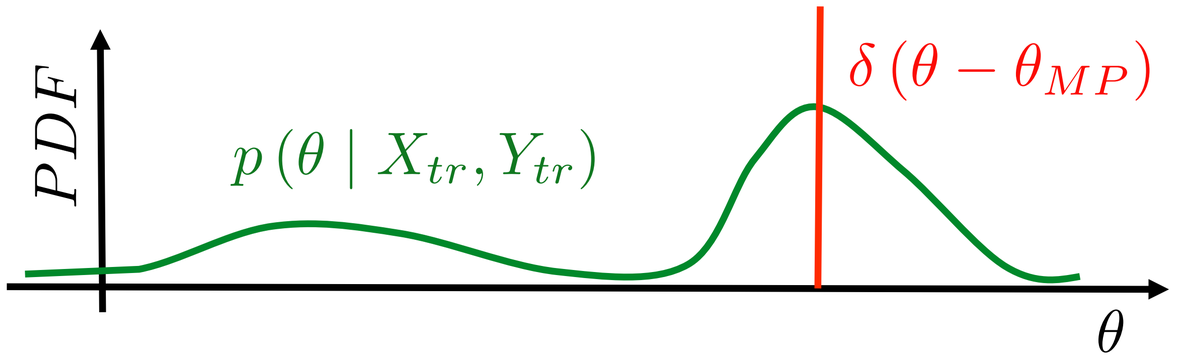
What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$



What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$

✳ Not the same as θ_{ML} — here we use prior

What to do if there is no conjugacy?

Simplest way — approximate posterior with delta function in θ_{MP}:

$$\theta_{MP} = \arg\max p\left(\theta \mid X_{tr}, Y_{tr}\right) = \arg\max p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)$$

On the testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta \approx p\left(y | x, \theta_{MP}\right)$$

More advanced techniques are needed!

Key takeaways

- Basic probabilistic rules: product rule, sum rule, Bayes theorem
- Bayesian framework as an alternative approach to building probabilistic models
- Bayesian ML models: training and predictions
- Full Bayesian inference and conjugate distributions

Now let’s practice =)

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Practice
- Approximate Bayesian inference

The problem set is available here:

tiny.cc/WMLAP_bayes

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

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

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

Problem 2: frequentist framework

Setting

- $X = \{x_1, \ldots, x_N\}$ — independent dice rolls
- $N_k = \sum_{n=1}^{N} \mathbb{I}(x_n = k)$ — counts
- $p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$ — multinomial likelihood, $\theta \in S_K$

Question

Maximum likelihood estimate for $\theta_{ML} = \arg\max_{\theta \in S_K} \log p(X \mid \theta)$

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:

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

Problem 2: frequentist framework

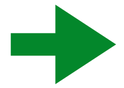
θ is restricted to simplex. To omit the inequality restrictions change parameterization to $\mu_k = \log \theta_k, \quad \mu_k \in \mathbb{R}$

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:

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



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

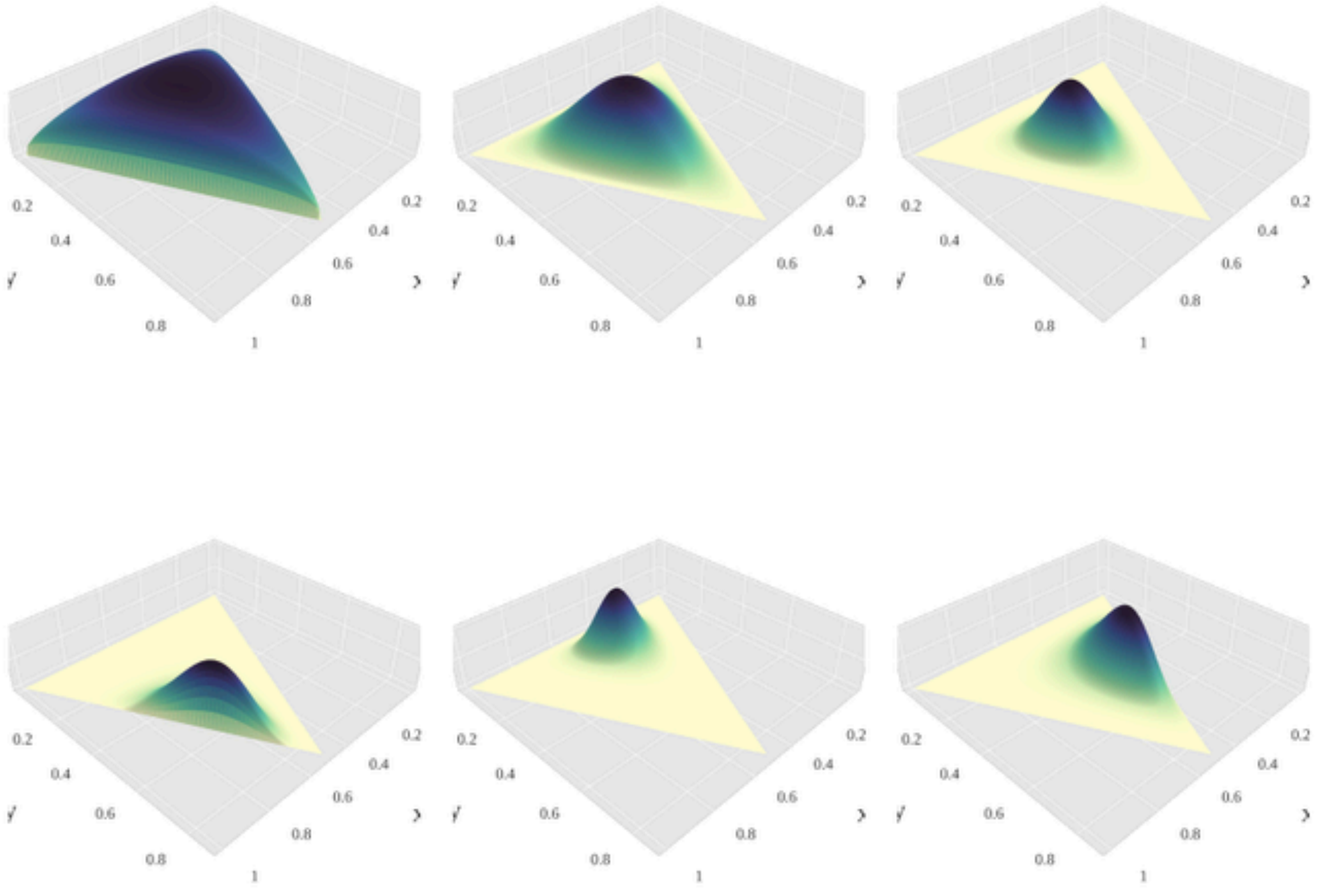
Problem 3: Bayesian framework

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:

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

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

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

Problem 3: 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 3: 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 3: 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: $\mathrm{Dir}(\theta \mid \alpha) = \dfrac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$

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

Problem 3: Bayesian framework

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

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

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

Problem 3: 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 3: 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: $p(\theta \mid X) \propto p(X \mid \theta)p(\theta) = \prod_{k=1}^{K} \theta_k^{N_k} \cdot \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} =$

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

conjugate

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

Problem 3: 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 3: Bayesian framework

Likelihood and prior are conjugate → 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 3: 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 3: 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 3: 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 → Bayesian estimate is mostly based on prior

Large K → Bayesian estimate is very similar to ML estimate

Problem 3: 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 3: Bayesian framework

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

Useful formulas:

$$B(\alpha_1, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \qquad \Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$$

Problem 3: Bayesian framework

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

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

Problem 3: Bayesian framework

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

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

$$B(\alpha_1, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

Problem 3: Bayesian framework

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

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

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

Problem 3: Bayesian framework

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

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

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

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$$

Problem 3: Bayesian framework

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

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

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

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

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Practice
- Approximate Bayesian inference

Approximate inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Variational Inference

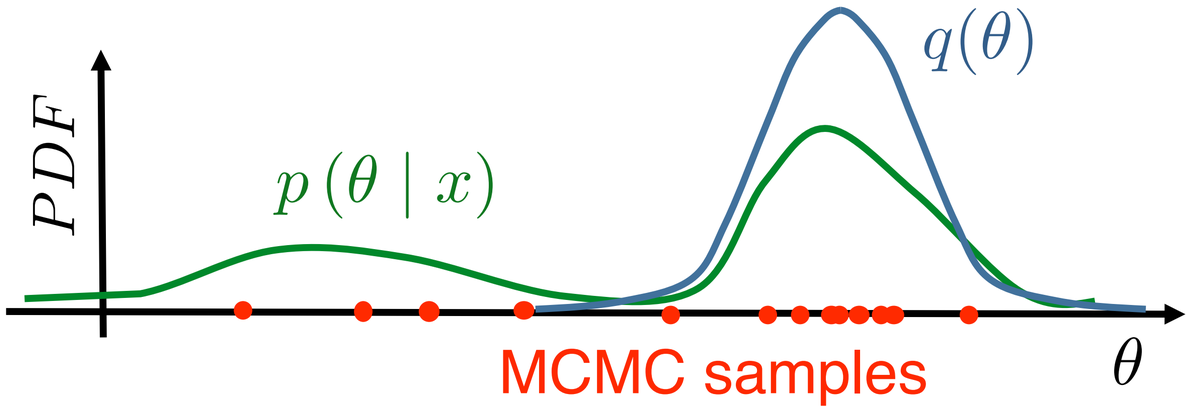
Approximate $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$

- Biased
- Faster and more scalable

MCMC

Samples from unnormalized $p(\theta \mid x)$

- Unbiased
- Need a lot of samples



Variational inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Main idea: find posterior approximation $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$, using the following criterion function:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Kullback-Leibler divergence
a good mismatch measure between
two distributions over the **same domain**

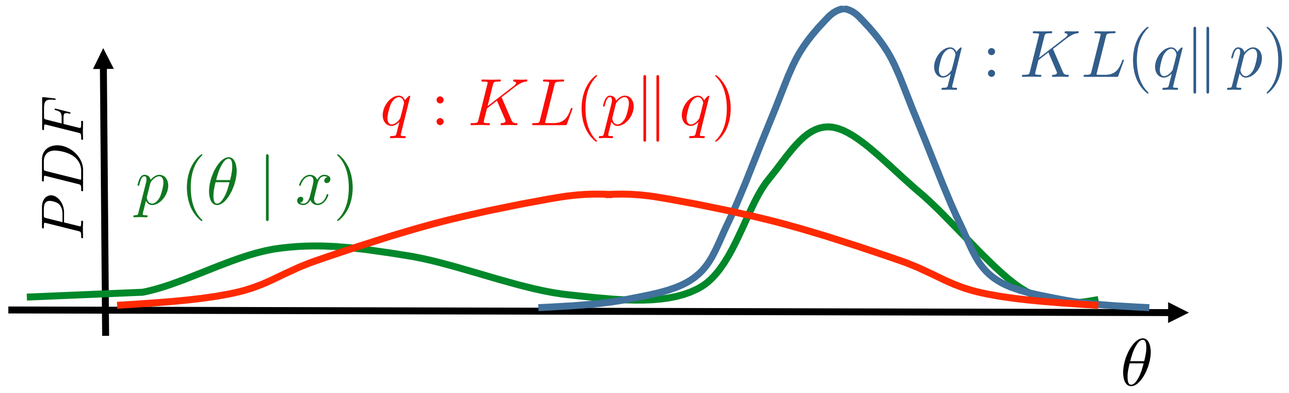
Kullback-Leibler divergence

A good mismatch measure between two distributions over the **same domain**

$$KL(q(\theta) \| \, p(\theta \mid x)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta$$

Properties:

- $KL(q \| \, p) \geq 0$
- $KL(q \| \, p) = 0 \;\Leftrightarrow\; q = p$
- $KL(q \| \, p) \neq KL(p \| \, q)$



Variational inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Main idea: find posterior approximation $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$, using the following criterion function:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Variational inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Main idea: find posterior approximation $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$, using the following criterion function:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

We could not compute the posterior in the first place

How to perform an optimization w.r.t. a distribution?

Mathematical magic

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x, \theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x, \theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =$$

Mathematical magic

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x,\theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =$$

$$= \boxed{\mathcal{L}(q(\theta))} + \boxed{KL(q(\theta) \| p(\theta \mid x))}$$

Evidence lower bound (ELBO)

KL-divergence we need for VI

ELBO = Evidence Lower Bound

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

Evidence:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Evidence of the probabilistic model shows the total probability of observing the data.

Lower Bound: KL is non-negative $\rightarrow$ $\log p(x) \geq \mathcal{L}(q(\theta))$

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| \, p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Let's use our magic:

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| \, p(\theta \mid x))$$

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| \, p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Let's use our magic:

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| \, p(\theta \mid x))$$

does not depend on q

depend on q

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Let's use our magic:

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

does not depend on q — depend on q

$$KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}} \quad \Leftrightarrow \quad \mathcal{L}(q(\theta)) \to \max_{q(\theta) \in \mathcal{Q}}$$

Variational inference

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta \to \max_{q(\theta) \in \mathcal{Q}}$$

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \boxed{\mathbb{E}_{q(\theta)} \log p(x \mid \theta)} - \boxed{KL(q(\theta) \| p(\theta))}$$

data term regularizer

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \boxed{\mathbb{E}_{q(\theta)} \log p(x \mid \theta)} - \boxed{KL(q(\theta) \| p(\theta))}$$

data term

regularizer

this is not the same KL-divergence!

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \mathbb{E}_{q(\theta)} \log p(x \mid \theta) - KL(q(\theta) \| p(\theta))$$

data term

regularizer

this is not the same KL-divergence!

$$\log p(x) = \mathbb{E}_{q(\theta)} \log p(x \mid \theta) - KL(q(\theta) \| p(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

Variational inference

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta \to \max_{q(\theta) \in \mathcal{Q}}$$

How to perform an optimization w.r.t. a distribution?

Variational inference

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta \to \max_{q(\theta) \in \mathcal{Q}}$$

How to perform an optimization w.r.t. a distribution?

Parametric approximation

Parametric family

$$q(\theta) = q(\theta \mid \lambda)$$

Parametric approximation

Parametric family of variational distributions:

$$q(\theta) = q(\theta \mid \lambda), \qquad \lambda \text{ — some parameters}$$

Why is it a restriction? We choose a family of some fixed form:

- It may be too simple and insufficient to model the data
- If it is complex enough then there is no guaranty we can train it well to fit the data

Parametric approximation

Parametric family of variational distributions:

$$q(\theta) = q(\theta \mid \lambda), \qquad \lambda \text{ — some parameters}$$

Variational inference transforms to parametric optimization problem:

$$\mathcal{L}(q(\theta \mid \lambda)) = \int q(\theta \mid \lambda) \log \frac{p(x, \theta)}{q(\theta \mid \lambda)} d\theta \to \max_{\lambda}$$

If we're able to calculate derivatives of ELBO w.r.t. λ then we can solve this problem using some numerical optimization solver.

Inference methods: summary

Probabilistic model: $p(x, \theta)$ We want to compute: $p(\theta \mid x)$

Approximation		Inference
Exact	$p(\theta \mid x)$	Full Bayesian inference
Parametric	$p(\theta \mid x) \approx q(\theta) = q(\theta \mid \lambda)$	Parametric VI
Delta function	$p(\theta \mid x) \approx \delta(\theta - \theta_{MP})$	MP inference
No prior	θ_{ML}	MLE