Bayesian framework

Dmitry Vetrov

Research professor at HSE

Head of ML lab in SAIC-Moscow

Special thanks to Ekaterina Lobacheva for assistance with slides

Outline: Intro to Bayesian methods

- Bayesian framework
- Bayesian ML models and full Bayesian inference
- Conjugate distributions

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 tries 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 tries with a result (H,H)

Head (H) Tail (T)

Frequentist framework:

In all experiments the coin landed heads up

$$\theta_{ML} = 1$$

→

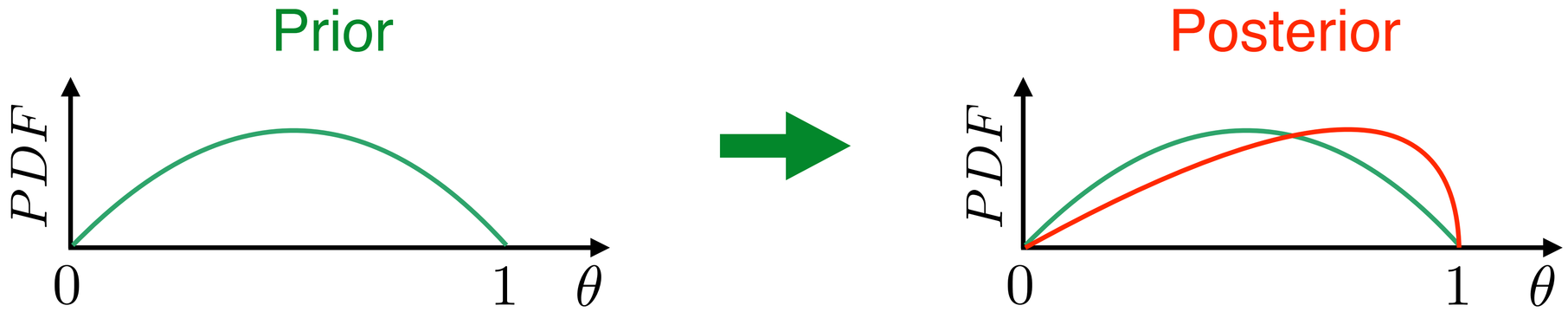
The coin is not fair and always lands heads up

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 tries 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 tries 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 tries 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

Probabilistic ML model

For each object in the data:

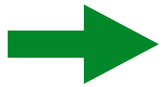
- x — set of observed variables (features)
- y — set of hidden / latent variables (class label / hidden representation, etc.)

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:

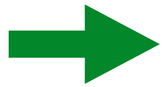
$$p(y, \theta \mid x) = p(y \mid x, \theta)p(\theta)$$

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:

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:

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

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:

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

Full Bayesian inference

Training stage:

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

May be intractable

Testing stage:

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

Conjugate distributions

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

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

Conjugate distributions

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

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

Intuition:

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

Conjugate distributions

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

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

Intuition:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int p(x \mid y)p(y)dy} \leftarrow \text{conjugate}$$

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

Conjugate distributions

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

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

Intuition:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int p(x \mid y)p(y)dy} \propto p(x \mid y)p(y)$$

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

Conjugate distributions

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

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

Intuition:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int p(x \mid y)p(y)dy} \propto p(x \mid y)p(y)$$

In this case Bayesian inference can be done in closed form

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

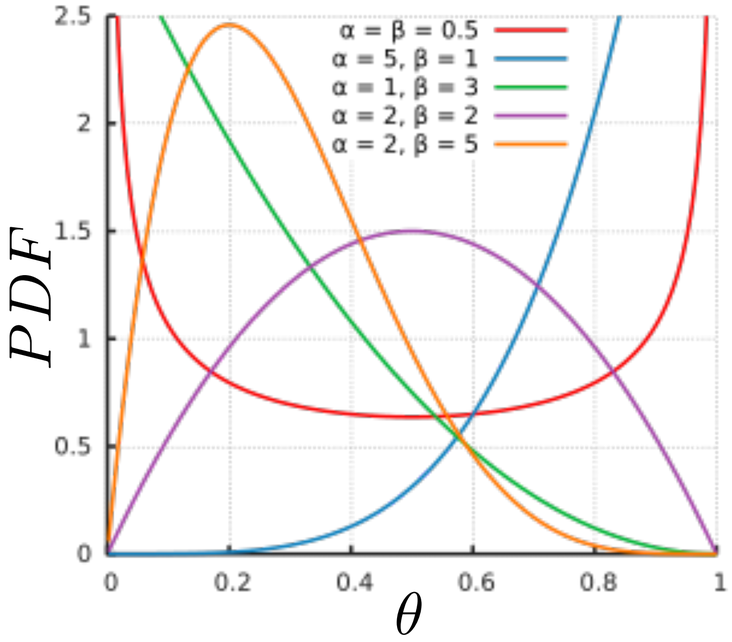
How to choose a prior?

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

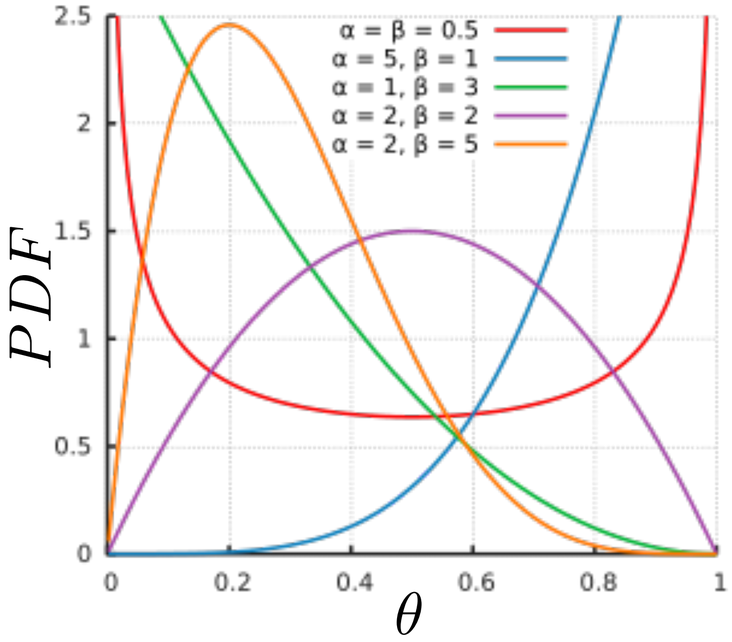


Example: coin tossing

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



Beta distribution matches all requirements:

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

* May be also used for the case of most likely unfair coin

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 lie in the same parametric family:

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 lie in the same parametric family:

$$p(\theta) = C\theta^{\alpha}(1-\theta)^{\beta}$$

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 lie in the same parametric family:

$$p(\theta) = C\theta^{\alpha}(1-\theta)^{\beta}$$

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

$$= C''\theta^{\alpha'}(1-\theta)^{\beta'}$$

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 lie in the same parametric family:

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

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

$$= \boxed{C''\theta^{\alpha'}(1-\theta)^{\beta'}} \quad \text{conjugacy}$$

Example: coin tossing

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

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

Example: coin tossing

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

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

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

Example: coin tossing

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

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

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

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

Example: coin tossing

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

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

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

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

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

Conjugate distributions

Likelihood $p(x\|y)$	y	Conjugate prior $p(y)$
Gaussian	μ	Gaussian
Gaussian	σ^{-2}	Gamma
Gaussian	(μ, σ^{-2})	Gaussian-Gamma
Multivariate Gaussian	Σ^{-1}	Wishart
Bernoulli	p	Beta
Multinomial	$(p_1, \ldots, p_m)$	Dirichlet
Poisson	λ	Gamma
Uniform	θ	Pareto

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}:

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

We do not need to calculate normalization constant

On the testing stage:

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

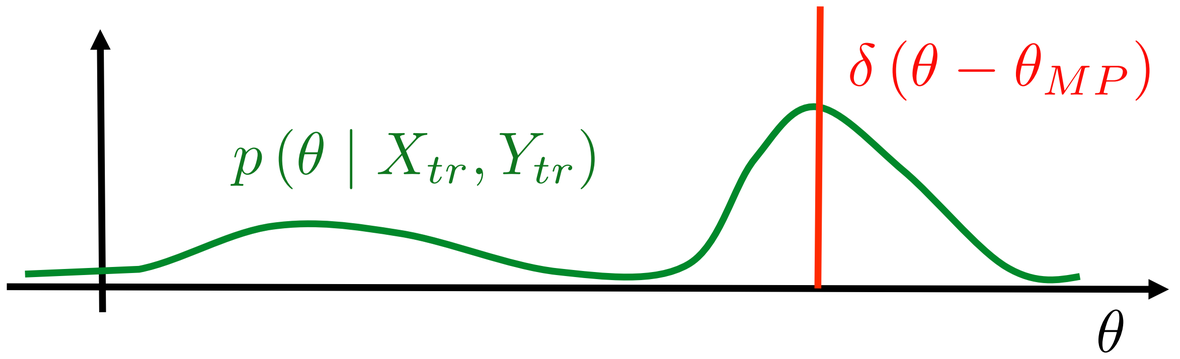
What to do if there is no conjugacy?

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:

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

We cannot compute the true posterior