Introduction to Bayesian methods

Ekaterina Lobacheva

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

NATIONAL RESEARCH UNIVERSITY

Slides are partially based on lectures of Dmitry Vetrov, Dmitry Kropotov and Kirill Struminsky, deepbayes.ru/2018
Slides on BNNs are from the lecture of Nadezhda Chirkova.

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Approximate Bayesian inference
- Bayesian neural networks

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Approximate Bayesian inference
- Bayesian neural networks

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

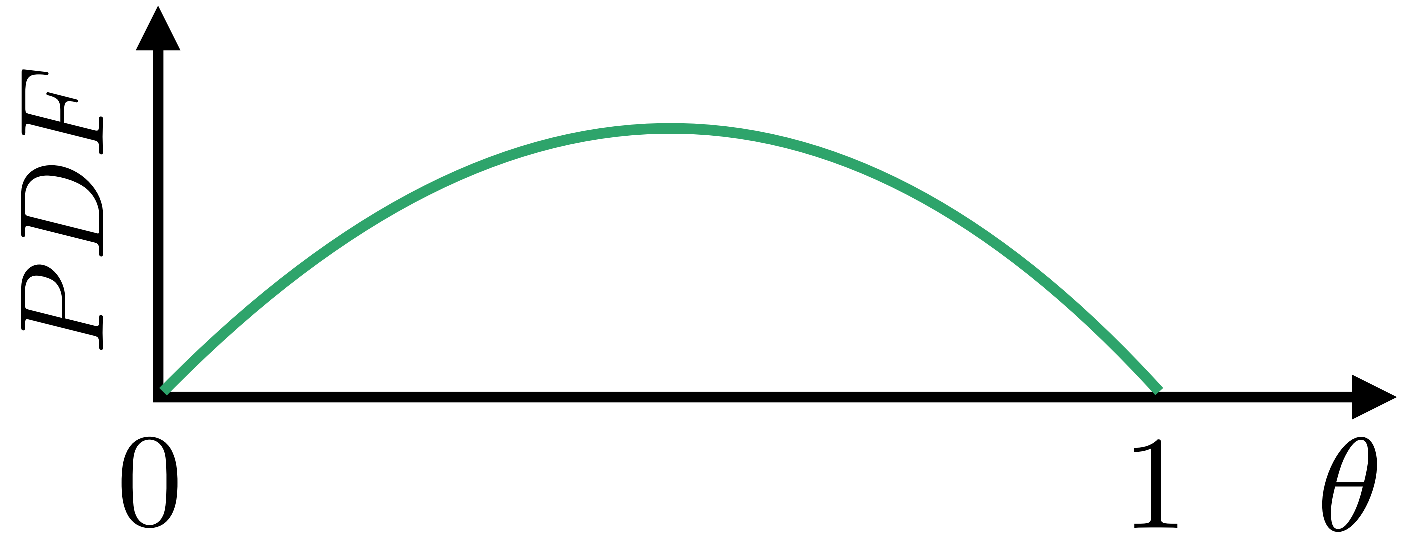
- We have a coin which may be fair or not
- The task is to estimate a probability θ of landing heads up
- Data: 2 tosses with a result (H,H)

Head (H)

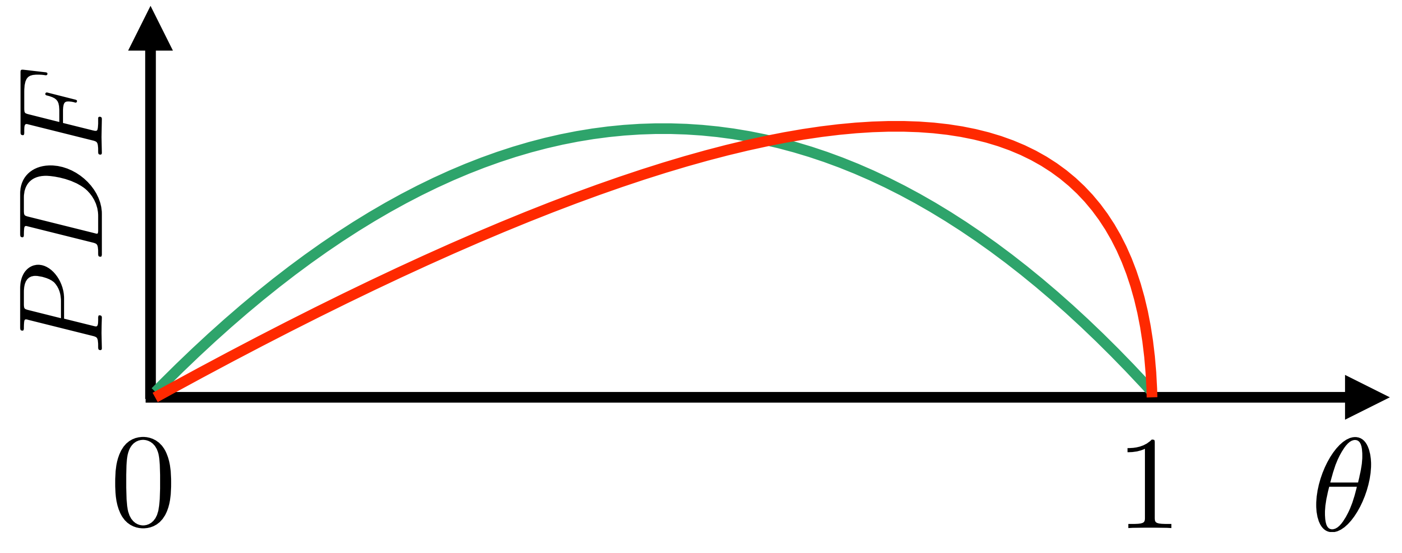
Tail (T)

Bayesian framework:

Prior



Posterior



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
- Approximate Bayesian inference
- Bayesian neural networks

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 $p(y, \theta \mid x)$ $\rightarrow$ 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:

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

Testing stage:

May be intractable

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

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
- Approximate Bayesian inference
- Bayesian neural networks

Conjugate distributions

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

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

Conjugate distributions

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

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

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 $p(\theta)$:

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

Intuition:

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

- 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 $p(\theta)$:

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

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:

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

Testing stage:

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

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

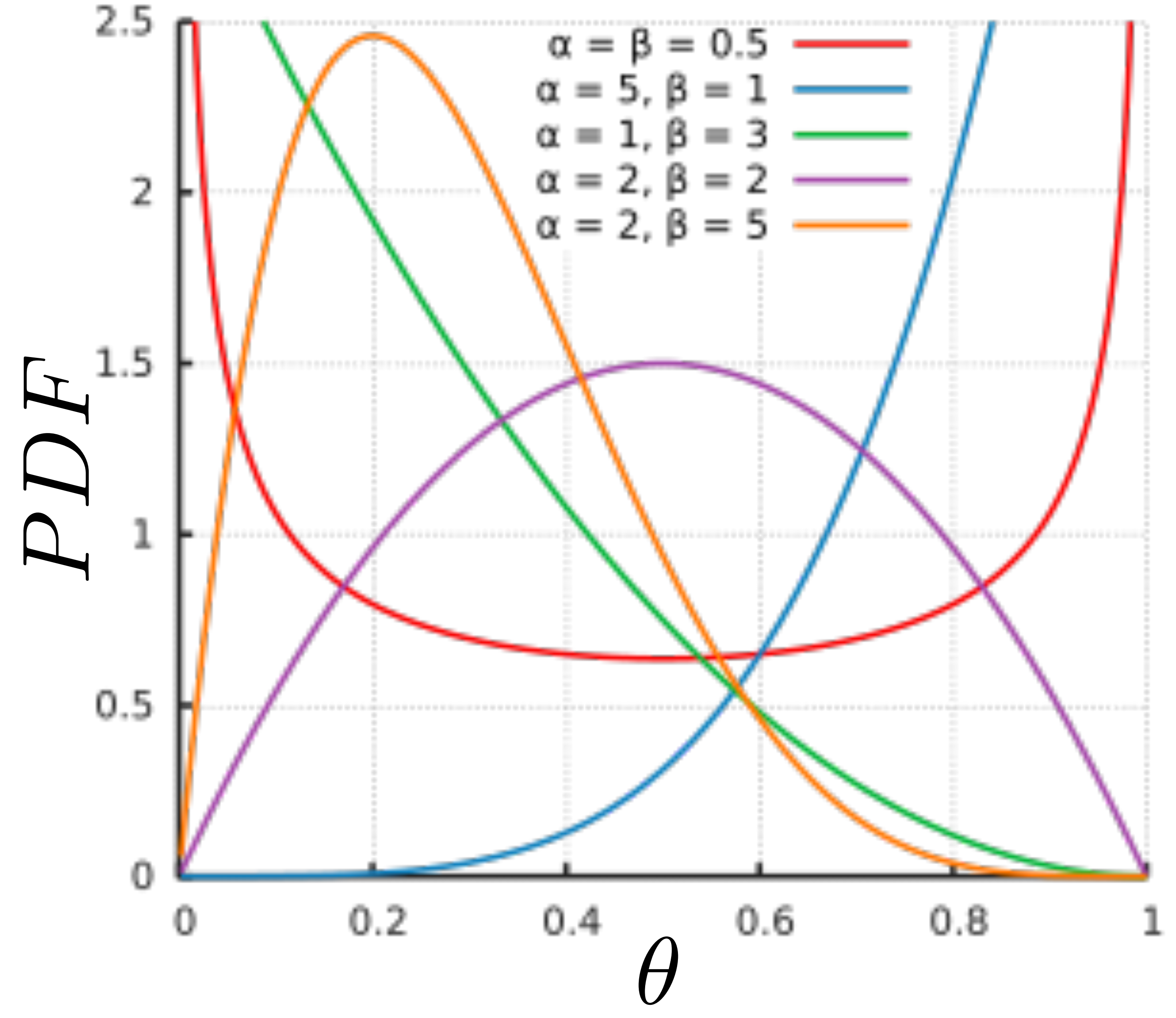
How to choose a prior?

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

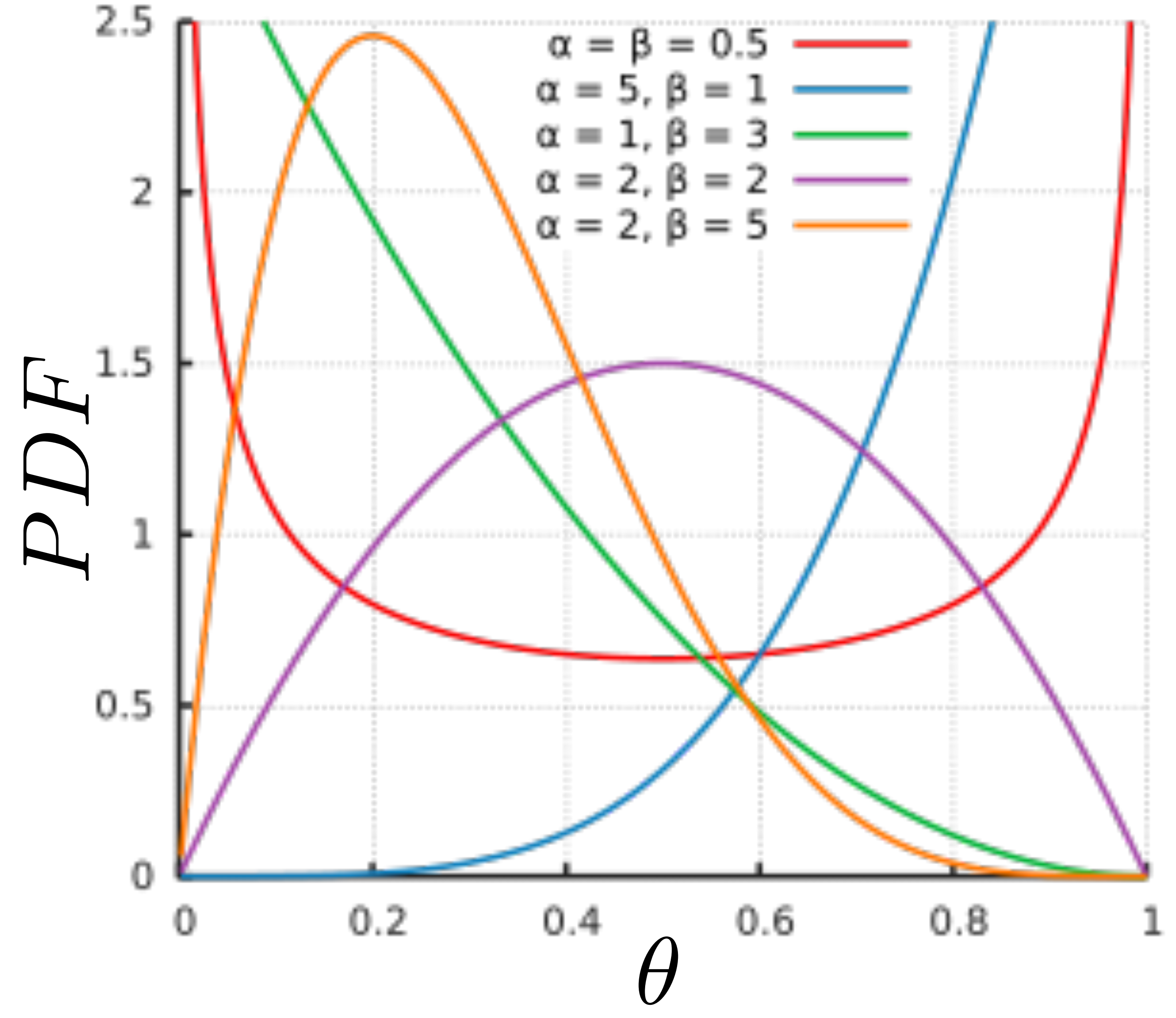


Example: coin tossing

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:

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

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:

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

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:

$$p(\theta) = \boxed{C\theta^C (1-\theta)^C}$$ 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 $X = (x_1, \ldots, x_n)$:

$$
\begin{aligned}
p(\theta \mid X) &= \frac{1}{Z} p(X \mid \theta) p(\theta) = \frac{1}{Z} \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right) \right] p(\theta) = \\
&= \frac{1}{Z} \left[\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right] \frac{1}{\mathrm{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} = \\
&= \frac{1}{Z'} \theta^{a + \sum_{i=1}^{n} x_i - 1} (1-\theta)^{b + n - \sum_{i=1}^{n} x_i - 1} = Beta\left(\theta \mid a', b'\right)
\end{aligned}
$$

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

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}:

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

We do not need to calculate the normalisation constant

On the testing stage:

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

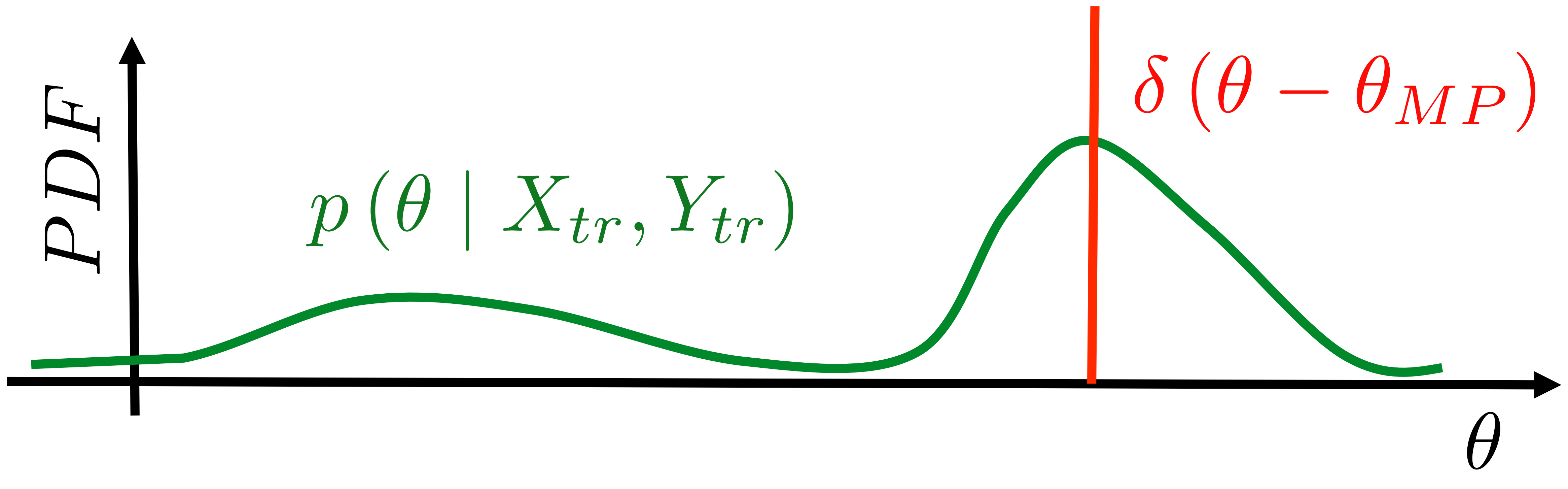
What to do if there is no conjugacy?

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}:

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

On the testing stage:

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

More advanced techniques are needed!

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- $\boxed{\text{Approximate Bayesian inference}}$
- Bayesian neural networks

Approximate inference

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

Variational Inference

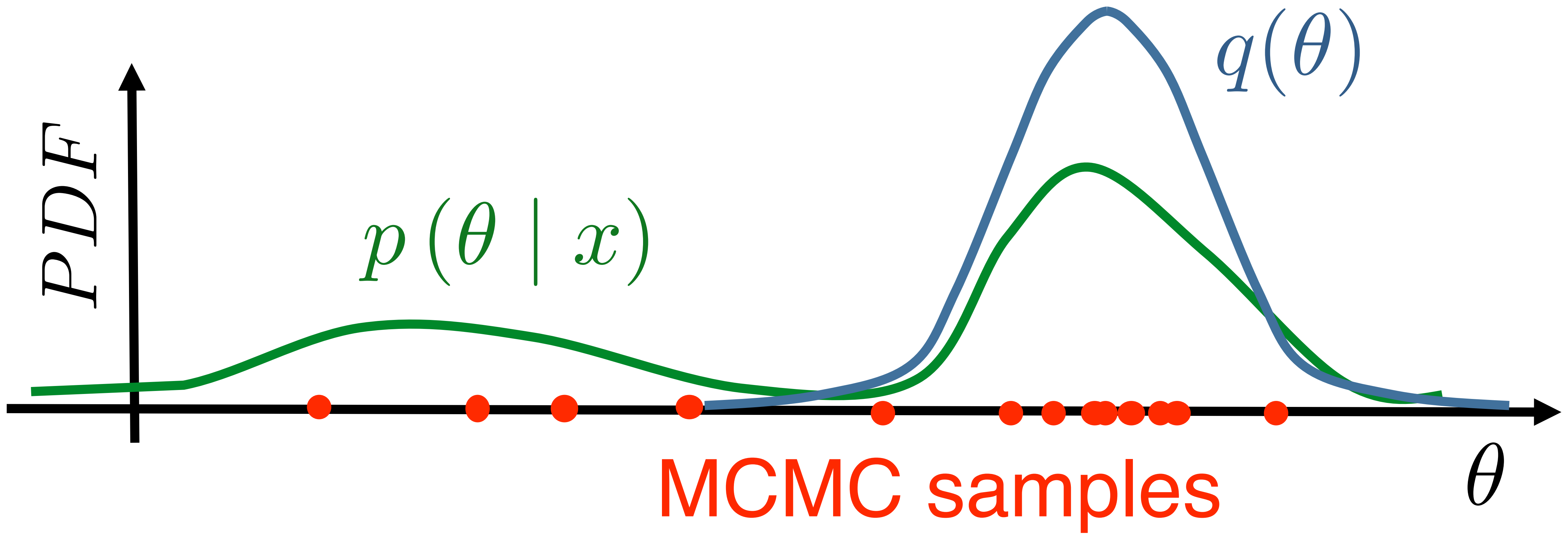
Approximate $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$

- 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**

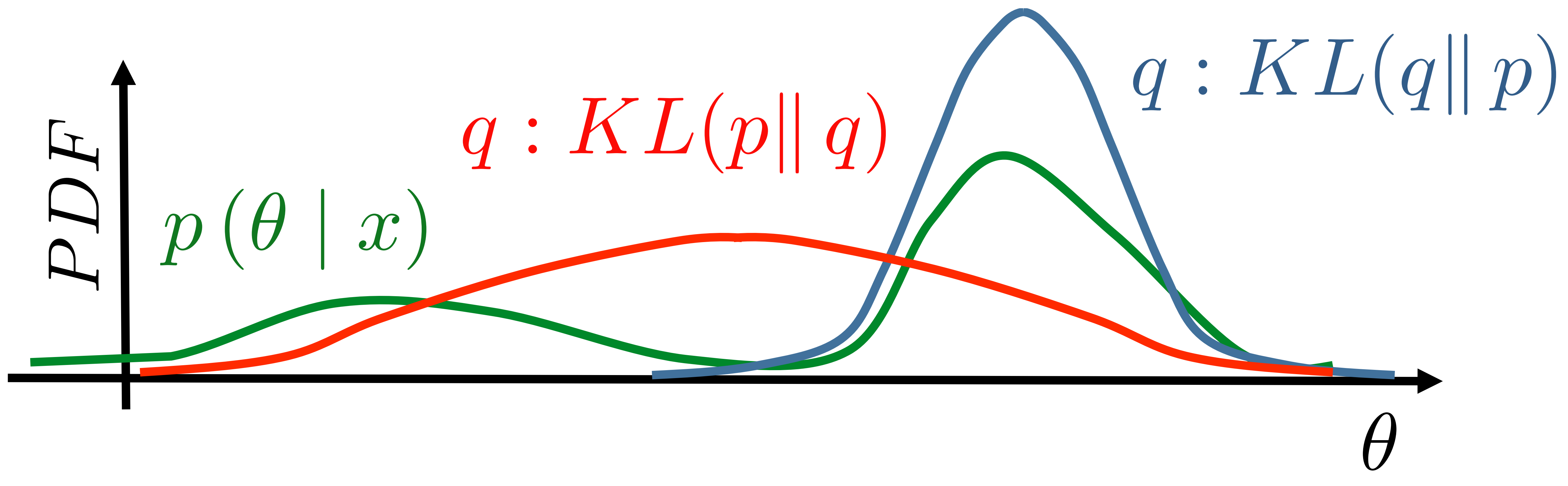
Kullback-Leibler divergence

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:

- $KL(q \| p) \geq 0$
- $KL(q \| p) = 0 \iff q = p$
- $KL(q \| p) \neq KL(p \| q)$



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 $\longrightarrow$ $\log p(x) \geq \mathcal{L}(q(\theta))$

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:

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

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

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

Variational inference: ELBO interpretation

Final optimisation problem:

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

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

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

this is not the same KL-divergence!

Variational inference: ELBO interpretation

Final optimisation problem:

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

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

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

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: $p(\theta \mid x)$

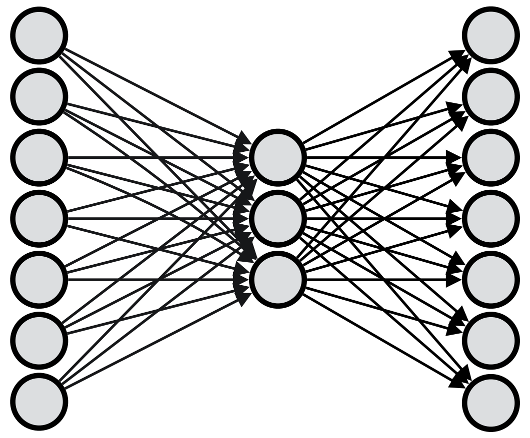
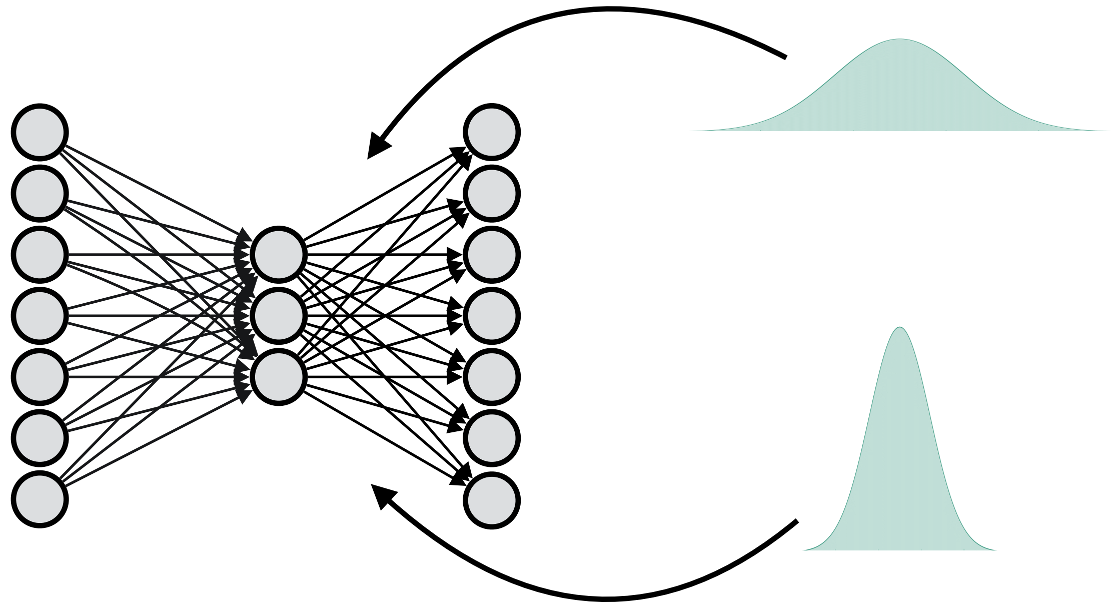
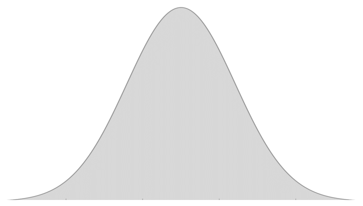
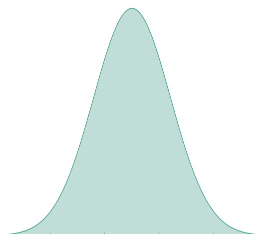
Approximation		Inference
Exact	$p(\theta \mid x)$	Full Bayesian inference
Parametric	$p(\theta \mid x) \approx q(\theta) = q(\theta \mid \lambda)$	Parametric VI
Delta function	$p(\theta \mid x) \approx \delta(\theta - \theta_{MP})$	MP inference
No prior	θ_{ML}	MLE

Outline

- Bayesian framework
- Bayesian ML models
- Full Bayesian inference and conjugate distributions
- Approximate Bayesian inference
- Bayesian neural networks

Slides for this part are from the lecture of Nadezhda Chirkova

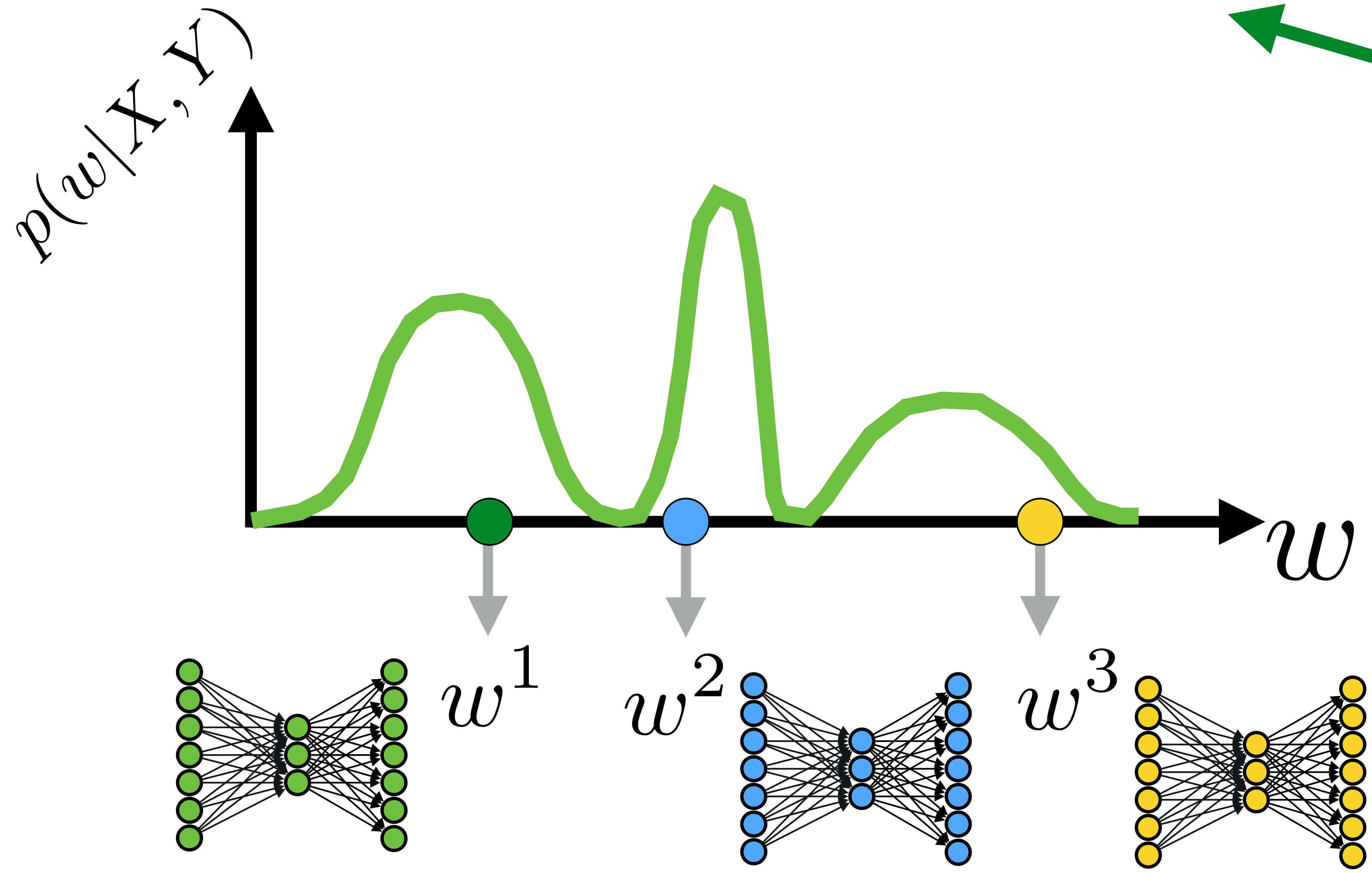
Bayesian neural networks

	Deterministic neural network	Bayesian neural network
Weights:	deterministic weights 	stochastic weights 
Training:	Data X, Y → point estimate w	 Prior $p(w)$ + Data X, Y →  Posterior $p(w \mid X, Y)$
Prediction on new object x_* :	$p(y_* \mid x_*, w)$	$\mathbb{E}_{p(w \mid X,Y)} p(y_* \mid x_*, w)$

BNN as an ensemble of neural networks

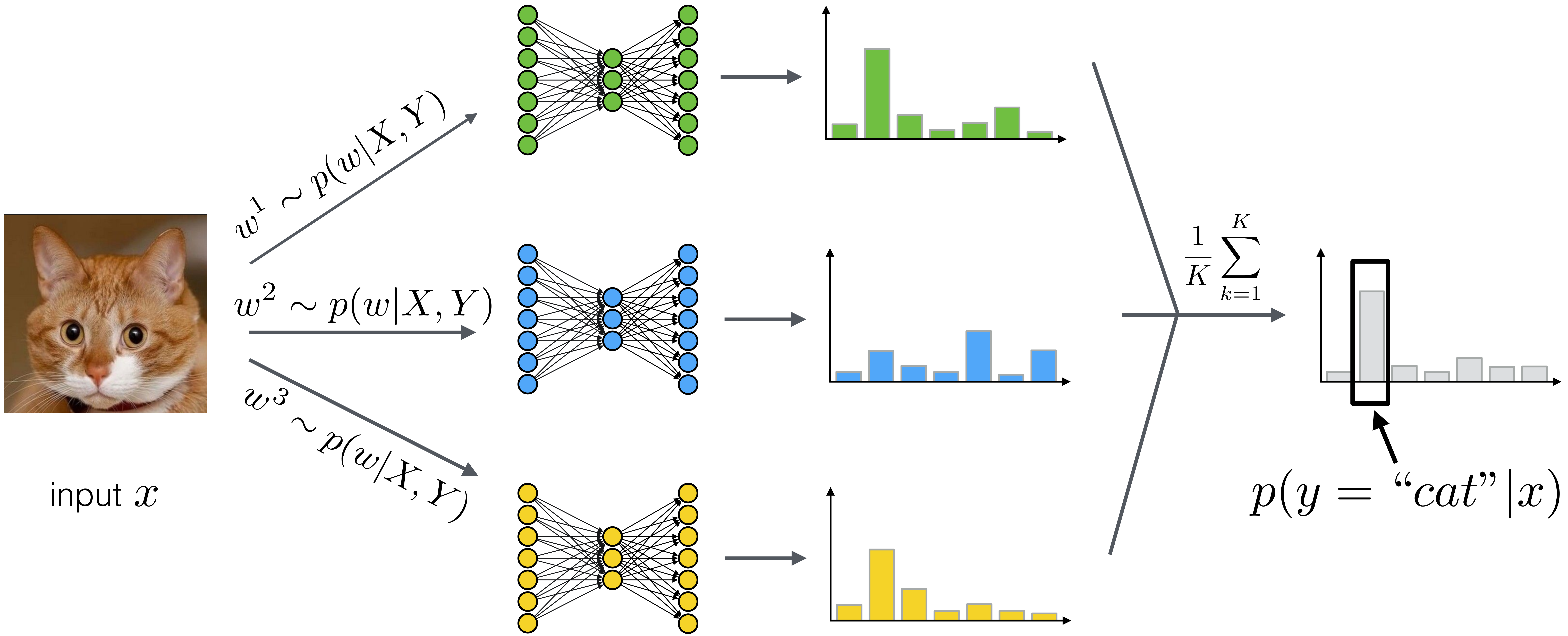
Prediction on a new object x_*:

$$\mathbb{E}_{p(w|X,Y)}p(y_*|x_*,w) \approx \frac{1}{K}\sum_{k=1}^{K} p(y_*|x_*,w^k), \quad w^k \sim p(w|X,Y)$$



average net's output across several weight samples

BNN as an ensemble of neural networks



BNN as an ensemble of neural networks

Prediction on a new object x_*:

$$\mathbb{E}_{p(w|X,Y)} p(y_*|x_*, w) \approx \frac{1}{K} \sum_{k=1}^{K} p(y_*|x_*, w^k), \quad w^k \sim p(w|X,Y)$$

- Higher quality (models compensate each other's errors)
- Better uncertainty estimation

Slide From Dmitry Molchanov's presentation, deepbayes.ru/2018

Why go Bayesian?

A principled framework with many useful applications

- Regularization
- Ensembling
- Uncertainty estimation
- On-line / continual learning
- Automatic hyperparameter choice
- Different prior $\Rightarrow$ different properties of the network

Slide From Dmitry Molchanov's presentation, deepbayes.ru/2018

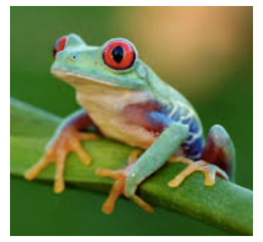
On-line / continual learning

Prior 1 → Task 1: cats vs dogs

→ Posterior 1

=

Prior 2 → Task 2: pigs vs frogs



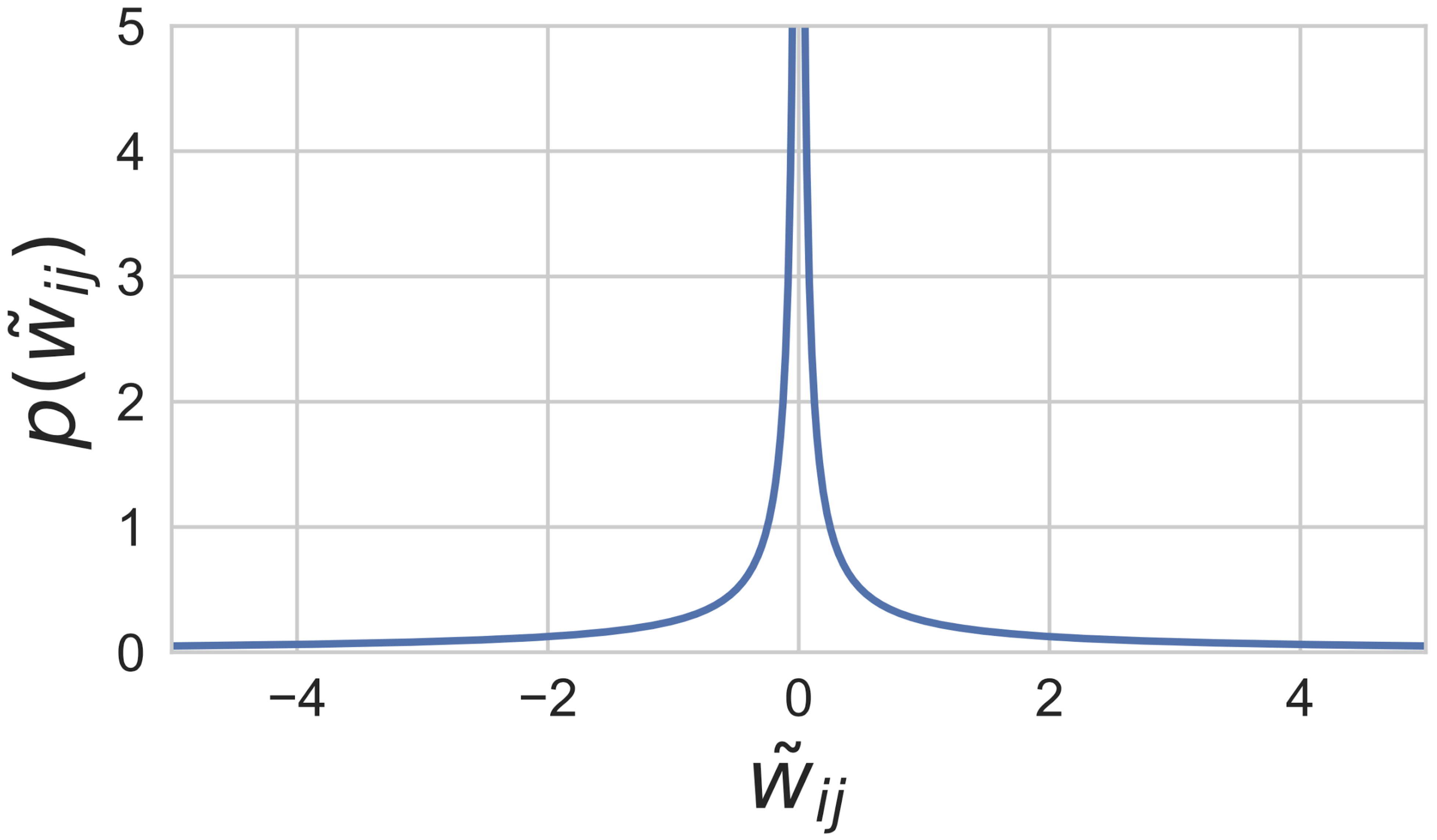
→ Posterior 2

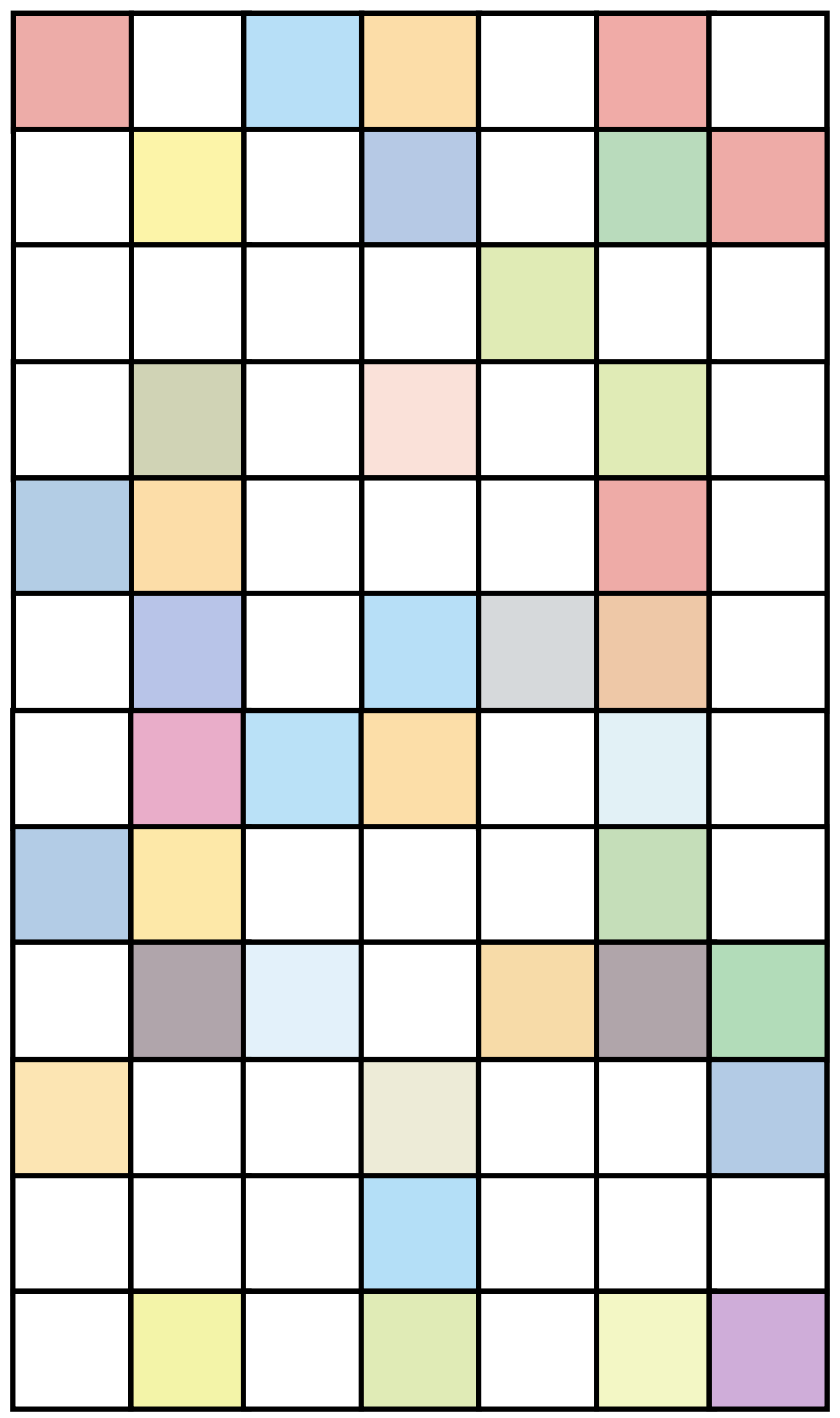
=

Prior 3 →

Prior can encode our desirable model properties

Prior concentrated at zero → A lot of zero weights





Weight matrix W

Outline

- Practice on full Bayesian inference

The problem set is available here:

http://tiny.cc/RB_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: Linear regression - frequentist

Setting

- Training data: $\mathcal{D} = \{X, y\}, \quad X \in \mathbb{R}^{K \times d}$, $y \in \mathbb{R}^K$
- Normal likelihood:

$$p(y|X, w) = \prod_{i=1}^{K} \mathcal{N}(y_i | w^T x_i, 1) = \mathcal{N}(y | Xw, I)$$

Question

Maximum likelihood estimate for $w_{ML} = \arg\max_{w \in \mathbb{R}^d} \log p(y|X, w)$

Problem 2: Linear regression - frequentist

ML estimation:

$$w_{ML} = \arg\max_{w \in \mathbb{R}^d} \log p(y|X, w)$$

Normal distribution:

$$\mathcal{N}(x|\mu, \Lambda^{-1}) = \sqrt{\frac{|\Lambda|}{(2\pi)^d}} exp\left(-\frac{1}{2}(x-\mu)^T \Lambda (x-\mu)\right)$$

Optimization task:

$$\mathcal{L}(w) = \log \mathcal{N}(y|Xw, I) = -(y - Xw)^T (y - Xw)$$

Problem 2: Linear regression - frequentist

Optimization task:

$$\mathcal{L}(w) = \log \mathcal{N}(y|Xw, I) = -(y - Xw)^T (y - Xw)$$

Differentiation:

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \frac{\partial}{\partial w} \left(-y^T y + 2w^T X^T y - w^T X^T X w \right) = 2X^T y - 2X^T X w$$

$$\frac{\partial \mathcal{L}(w)}{\partial w} = 0 \qquad \Rightarrow \qquad w_{ML} = (X^T X)^{-1} X^T y$$

Problem 2: Linear regression - Bayesian

Setting

- Normal likelihood: $p(y|X, w) = \prod_{i=1}^{K} \mathcal{N}(y_i|w^T x_i, 1) = \mathcal{N}(y|Xw, I)$
- Normal prior: $p(w) = \mathcal{N}(w|0, A^{-1}), \quad A = \alpha I, \alpha \in \mathbb{R}^+$

Questions

- $\boxed{\text{Check that likelihood and prior are conjugate}}$
- Compute the posterior $p(w|X, y)$
- Compare w_{MP} and w_{ML}
- Compute the predictive posterior $p(y_{new}|x_{new}, X, y)$

Problem 2: Linear regression - Bayesian

Probabilistic model: $p(y|X,w) = p(y|X,w)p(w) = \mathcal{N}(y|Xw, I)\mathcal{N}(w|0, A^{-1})$

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

Problem 2: Linear regression - Bayesian

Probabilistic model: $p(y|X,w) = p(y|X,w)p(w) = \mathcal{N}(y|Xw, I)\mathcal{N}(w|0, A^{-1})$

Prior: $p(w) = \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right) = Cexp\left(w^T C w\right)$

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

Problem 2: Linear regression - Bayesian

Probabilistic model: $p(y|X, w) = p(y|X, w)p(w) = \mathcal{N}(y|Xw, I)\mathcal{N}(w|0, A^{-1})$

Prior: $p(w) = \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T Aw\right) = Cexp\left(w^T Cw\right)$

Posterior: $p(w|X, y) \propto p(y|w, X)p(w) =$

$$= \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T(y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T Aw\right)$$

$$= C\exp(w^T Cw + w^T C + C)$$

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

Problem 2: Linear regression - Bayesian

Probabilistic model: $p(y|X,w) = p(y|X,w)p(w) = \mathcal{N}(y|Xw, I)\mathcal{N}(w|0, A^{-1})$

Prior: $p(w) = \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2} w^T A w\right) = \boxed{Cexp\left(w^T C w\right)}$

Posterior: $p(w|X,y) \propto p(y|w,X)p(w) =$

$$= \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T (y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2} w^T A w\right)$$

$$= \boxed{C \exp(w^T C w + w^T C + C)}$$

not conjugate???

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

Problem 2: Linear regression - Bayesian

Probabilistic model: $p(y|X,w) = p(y|X,w)p(w) = \mathcal{N}(y|Xw, I)\mathcal{N}(w\boxed{|0,}A^{-1})$

Prior: $p(w) = \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right) = \ Cexp\left(w^T C w + \boxed{w^T \cdot 0 + 0)}\right.$

Posterior: $p(w|X,y) \propto p(y|w,X)p(w) =$

$$= \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T(y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right)$$

$$= C\exp(w^T C w + w^T C + C)$$

zero is a parameter of the prior

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

Problem 2: Linear regression - Bayesian

Probabilistic model: $p(y|X,w) = p(y|X,w)p(w) = \mathcal{N}(y|Xw, I)\mathcal{N}(w|0, A^{-1})$

Prior: $p(w) = \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right) = \boxed{Cexp\left(w^T C w + w^T \cdot 0 + 0\right)}$

Posterior: $p(w|X,y) \propto p(y|w,X)p(w) =$

$$= \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T(y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right)$$

$$= \boxed{C \exp(w^T C w + w^T C + C)}$$

conjugate

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

Problem 2: Linear regression - Bayesian

Setting

- Normal likelihood: $p(y|X,w) = \prod_{i=1}^{K} \mathcal{N}(y_i|w^T x_i, 1) = \mathcal{N}(y|Xw, I)$
- Normal prior: $p(w) = \mathcal{N}(w|0, A^{-1}), \quad A = \alpha I, \alpha \in \mathbb{R}^+$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(w|X,y)$
- Compare w_{MP} and w_{ML}
- Compute the predictive posterior $p(y_{new}|x_{new}, X, y)$

Problem 2: Linear regression - Bayesian

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

Form of normal distr.: $\mathcal{N}(x|\mu, \Lambda^{-1}) \propto \exp\left(-\frac{1}{2}x^T\Lambda x + x^T\Lambda\mu - \frac{1}{2}\mu^T\mu\right)$

Problem 2: Linear regression - Bayesian

Likelihood and prior are conjugate $\rightarrow$ posterior is normal

Form of normal distr.: $\mathcal{N}(x|\mu, \Lambda^{-1}) \propto \exp\left(-\frac{1}{2}x^T \Lambda x + x^T \Lambda \mu - \frac{1}{2}\mu^T \mu\right)$

$$p(w|X, y) \propto p(y|w, X)p(w) \propto$$

$$\propto \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T (y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right) \propto$$

$$\propto \exp\left(-\frac{1}{2}w^T (X^T X + A)w + w^T X^T y\right)$$

Problem 2: Linear regression - Bayesian

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

Form of normal distr.: $\mathcal{N}(x|\mu, \Lambda^{-1}) \propto \boxed{\exp\left(-\frac{1}{2}x^T\Lambda x + x^T\Lambda\mu - \frac{1}{2}\mu^T\mu\right)}$

$$p(w|X,y) \propto p(y|w,X)p(w) \propto$$

$$\propto \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T(y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T Aw\right) \propto$$

$$\propto \boxed{\exp\left(-\frac{1}{2}w^T(X^TX + A)w + w^TX^Ty\right)}$$

Problem 2: Linear regression - Bayesian

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

Form of normal distr.: $\mathcal{N}(x|\mu, \Lambda^{-1}) \propto \boxed{\exp\left(-\frac{1}{2}x^T \Lambda x + x^T \Lambda \mu - \frac{1}{2}\mu^T \mu\right)}$

$$p(w|X, y) \propto p(y|w, X)p(w) \propto$$

$$\propto \sqrt{\frac{1}{(2\pi)^K}} \exp\left(-\frac{1}{2}(y - Xw)^T (y - Xw)\right) \sqrt{\frac{|A|}{(2\pi)^d}} \exp\left(-\frac{1}{2}w^T A w\right) \propto$$

$$\propto \boxed{\exp\left(-\frac{1}{2}w^T (X^T X + A)w + w^T X^T y\right)}$$

$$p(w|X, y) = \mathcal{N}(w|(X^T X + A)^{-1} X^T y, (X^T X + A)^{-1})$$

Problem 2: Linear regression - Bayesian

Setting

- Normal likelihood: $p(y|X,w) = \prod_{i=1}^{K} \mathcal{N}(y_i|w^T x_i, 1) = \mathcal{N}(y|Xw, I)$
- Normal prior: $p(w) = \mathcal{N}(w|0, A^{-1}), \quad A = \alpha I, \alpha \in \mathbb{R}^+$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(w|X,y)$
- Compare w_{MP} and w_{ML}
- Compute the predictive posterior $p(y_{new}|x_{new}, X, y)$

Problem 2: Linear regression - Bayesian

Maximum likelihood estimate: $w_{ML} = (X^T X)^{-1} X^T y$

Expectation of the posterior: $w_{MP} = (X^T X + A)^{-1} X^T y$

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

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

+ the prior regularizes the model

Problem 2: Linear regression - Bayesian

Setting

- Normal likelihood: $p(y|X, w) = \prod_{i=1}^{K} \mathcal{N}(y_i|w^T x_i, 1) = \mathcal{N}(y|Xw, I)$
- Normal prior: $p(w) = \mathcal{N}(w|0, A^{-1}), \quad A = \alpha I, \alpha \in \mathbb{R}^+$

Questions

- Check that likelihood and prior are conjugate
- Compute the posterior $p(w|X, y)$
- Compare w_{MP} and w_{ML}
- Compute the predictive posterior $p(y_{new}|x_{new}, X, y)$

Problem 2: Linear regression - Bayesian

Method 1 — ensembling:

$$p(y_{new}|x_{new}, X, y) = \int p(y_{new}|x_{new}, w)p(w|X, y)\mathrm{d}w$$

Method 2 — from Bayes theorem:

$$p(y_{new}|x_{new}, X, y) = \frac{p(y_{new}|x_{new}, w)p(w|X, y)}{p(w|x_{new}, y_{new}, X, y)}$$

Problem 2: Linear regression - Bayesian

$$p(y_{new}|x_{new}, X, y) = \frac{p(y_{new}|x_{new}, w)p(w|X, y)}{p(w|x_{new}, y_{new}, X, y)}$$

normal

$p(y|x, w), p(w)$

normal

$p(w|X, y), p(w|x_{new}, y_{new}, X, y)$

$$p(w|X, y) = \mathcal{N}(w|w_{MP}, \hat{A}^{-1}) = \mathcal{N}(w|(X^TX + A)^{-1}X^Ty, (X^TX + A)^{-1})$$

$$p(w|x_{new}, y_{new}, X, y)\mathcal{N}(w|(\tilde{X}^T\tilde{X} + A)^{-1}\tilde{X}^T\tilde{y}, (\tilde{X}^T\tilde{X} + A)^{-1})$$

$$\tilde{X} = \{X, x_{new}\}, \quad \tilde{y} = \{y, y_{new}\}$$

Problem 2: Linear regression - Bayesian

$$p(y_{new}|x_{new}, X, y) = \frac{p(y_{new}|x_{new}, w)p(w|X, y)}{p(w|x_{new}, y_{new}, X, y)}$$

normal
$p(y|x, w), p(w)$

normal
$p(w|X, y), p(w|x_{new}, y_{new}, X, y)$

normal
$p(y_{new}|x_{new}, X, y)$

Problem 2: Linear regression - Bayesian

$p(y_{new}|x_{new}, X, y)$ is normal, so let's compute its expectation and variance:

$$y_{new} = x_{new}^T w + \epsilon \qquad w \sim \mathcal{N}(w|w_{MP}, \hat{A}^{-1}) \qquad \epsilon \sim \mathcal{N}(\epsilon|0, 1)$$

$$\mathbb{E}y_{new} = \mathbb{E}(x_{new}^T w + \epsilon) = x_{new}^T \mathbb{E}w = x_{new}^T w_{MP}$$

$$\begin{aligned}\mathbb{D}y_{new} &= \mathbb{E}(y_{new} - x_{new}^T w_{MP})^2 = \mathbb{E}(x_{new}^T w + \epsilon - x_{new}^T w_{MP})^2 = \\ &= x_{new}^T \mathbb{E}(w - w_{MP})^2 x_{new} + 2\mathbb{E}\epsilon\mathbb{E}(x_{new}^T w - x_{new}^T w_{MP}) + \mathbb{E}\epsilon^2 = \\ &= x_{new}^T \hat{A}^{-1} x_{new} + 1\end{aligned}$$

Problem 2: Linear regression - Bayesian

$p(y_{new}|x_{new}, X, y)$ is normal, so let's compute its expectation and variance:

$$y_{new} = x_{new}^T w + \epsilon \qquad w \sim \mathcal{N}(w|w_{MP}, \hat{A}^{-1}) \qquad \epsilon \sim \mathcal{N}(\epsilon|0, 1)$$

$$p(y_{new}|x_{new}, X, y) = \mathcal{N}(y_{new}|x_{new}^T w_{MP}, x_{new}^T \hat{A}^{-1} x_{new} + 1)$$

$$w_{MP} = (X^T X + A)^{-1} X^T y, \quad \hat{A} = (X^T X + A)$$

More informative predictive uncertainty