# Introduction to Approximate Inference

**Advanced Statistical Inference**

Simone Rossi

## Bayesian inference

**Bayesian inference** allows to "transform" a prior distribution over the parameters into a posterior **after** observing the data

**Prior distribution** $p(\boldsymbol{w})$

**Posterior distribution** $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$

---

**Bayes' rule**:

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{X})}$$

- **Prior**: $p(\boldsymbol{w})$
  - Encodes our beliefs about the parameters **before** observing the data
- **Likelihod**: $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X})$
  - Encodes our model of the data
- **Posterior**: $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$
  - Encodes our beliefs about the parameters **after** observing the data (e.g. conditioned on the data)
- **Evidence**: $p(\boldsymbol{y} \mid \boldsymbol{X})$
  - Normalizing constant, ensures that $\int p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{w} = 1$

### Bayesian linear regression (review)

Modeling observation as noisy realization of a linear combination of the features As before, we assume a Gaussian likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I})$$

For the prior, we use a Gaussian distribution over the model parameters

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \boldsymbol{S})$$

In practice, we often use a diagonal covariance matrix $\boldsymbol{S} = \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}$

### When can we compute the posterior?

> Definition
>
> A prior is **conjugate** to a likelihood if the posterior is in the same family as the prior.

Only a few conjugate priors exist, but they are very useful.

Examples:

- Gaussian likelihood and Gaussian prior $\Rightarrow$ Gaussian posterior
- Binomial likelihood and Beta prior $\Rightarrow$ Beta posterior

Full table available on wikipedia

### Why is this useful?

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{X})}$$

- **Generally** the posterior is **intractable** to compute
  - We don't the form of the posterior $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$
  - The evidence $p(\boldsymbol{y} \mid \boldsymbol{X})$ is an integral
    - without closed form solution
    - high-dimensional and computationally intractable to compute numerically

. . .

- **Analytical solution** thanks to conjugacy:

  - We know the form of the posterior
  - We know the form of the normalization constant
  - We don't need to compute the evidence, just some algebra to get the posterior

---

From the likelihood and prior, we can write the posterior as

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w}) &\propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 - \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{S}^{-1}\boldsymbol{w}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{w}^\top\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{S}^{-1}\right)\boldsymbol{w} - \frac{2}{\sigma^2}\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y}\right)\right) \end{aligned}$$

. . .

From conjugacy, we know that the posterior is Gaussian

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)$$

We can identify the posterior mean and covariance

**Posterior covariance**

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{S}^{-1}\right)^{-1}$$

**Posterior mean**

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{y}$$

## Exact inference is rare

- **Exact inference** is possible when the posterior distribution can be computed analytically

**Example**: Linear regression with Gaussian likelihood and Gaussian prior

- This is the case for simple models with conjugate priors, …
- … but most of the time, the posterior distribution is intractable

**Examples**: Logistic regression (binary classification), neural networks, …

## Introduction to Approximate Inference

- In this lecture, we will introduce the concept of **approximate inference** in the context of Bayesian models.
- Approximate inference methods provide a way to approximate the posterior distribution when it is intractable
- For the next 2 weeks, we will be **model-agostic** and focus on the methods used to perform inference in complex and intractable models.

. . .

**Why model-agnostic?**

Solving a machine learning problem involves multiple steps:

1. **Modeling**: Define a model that captures the underlying structure of the data
2. **Inference**: Estimate the parameters of the model
3. **Prediction**: Use the model to make predictions on new data

In these two weeks, we will focus on the **inference** step

## Problem definition

In probabilistic models, all unknown quantities are treated as **random variables**

- **Observed quantities**: $\boldsymbol{y} \in \mathbb{R}^N$ (vector of $N$ observations);
- **Unknown variables**: $\boldsymbol{\theta} \in \mathbb{R}^D$ (vector of $D$ parameters)

Given a likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and a prior $p(\boldsymbol{\theta})$, the goal is to compute the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

**Note**: We drop the conditioning on the data $\boldsymbol{X}$ for simplicity, but it is present in the likelihood as input of the model

## Approximate inference

**Approximate inference** methods provide a way to approximate distributions when the exact computation is intractable

$$p(\theta \mid y) \approx q(\theta)$$

We will study two main classes of approximate inference methods

Sampling-based methods

- **Monte Carlo methods**
- **Markov Chain Monte Carlo (MCMC)**
- **Hamiltonian Monte Carlo (HMC)**

Parametric methods

- **Variational inference**
- **Laplace approximation**

## Grid approximation

Divide the parameter space into $K$ regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$ of equal volume $\Delta$. For each region, approximate the posterior mass by a Riemann approximation:

$$p(\theta \in \mathcal{R}_k \mid y) = \int_{\mathcal{R}_k} p(\theta \mid y), \mathrm{d}\theta \approx p(\theta_k \mid y)\,\Delta.$$

Use Bayes' rule at each grid point:

$$p(\theta_k \mid y) = \frac{p(y \mid \theta_k)p(\theta_k)}{\int p(y \mid \theta)p(\theta)\,\mathrm{d}\theta}.$$

Define unnormalized terms $\widetilde{p}_k = p(y \mid \theta_k)p(\theta_k)$. Normalization gives

$$p(\theta_k \mid y) = \frac{\widetilde{p}_k}{\sum_{j=1}^{K} \widetilde{p}_j}, \qquad p(\theta \in \mathcal{R}_k \mid y) \approx \frac{\widetilde{p}_k}{\sum_{j=1}^{K} \widetilde{p}_j}\Delta.$$

The marginal likelihood follows from the same Riemann sum $p(y) \approx \sum_{k=1}^{K} \widetilde{p}_k \Delta$