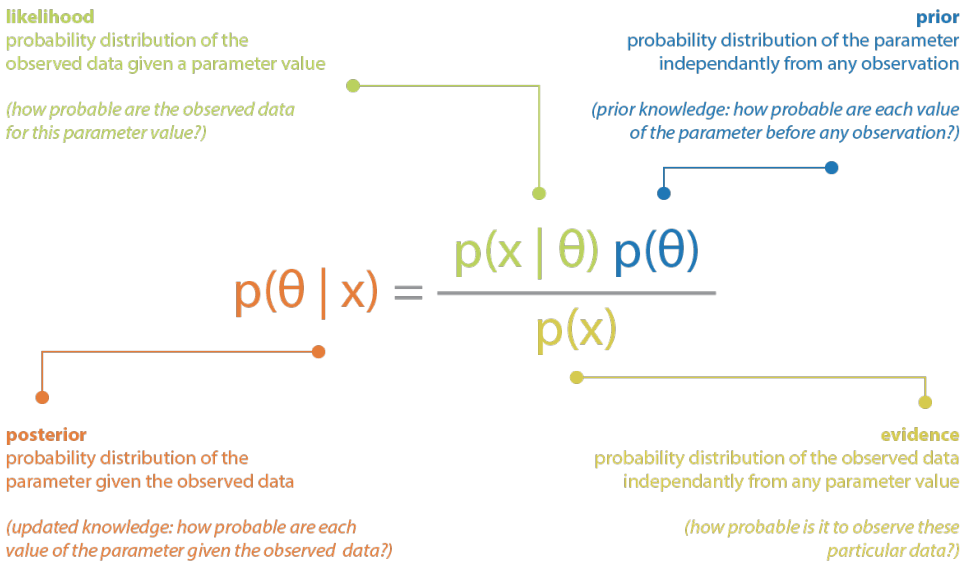
Variational Inference

Motivation

Bayesian inference is the process of producing statistical inference taking a Bayesian point of view.

A classical example is the **Bayesian inference of parameters**. Let's assume a model where data x are generated from a probability distribution depending on an unknown parameter θ. Let's also assume that we have a prior knowledge about the parameter θ that can be expressed as a probability distribution p(θ). Then, when data x are observed, we can update the prior knowledge about this parameter using the Bayes theorem as follows



The Bayes theorem tells us that the computation of the posterior requires three terms: a prior, a likelihood and an evidence. The first two can be expressed easily as they are part of the assumed model (in many situation, the prior and the likelihood are explicitly known). However, the third term, that is a normalisation factor, requires to be computed such that

$$p(x) = \int_\theta p(x|\theta)p(\theta)d\theta$$

Although in low dimension this integral can be computed without too much difficulties, **it can become intractable in higher dimensions**. In this last case, the exact computation of the posterior distribution is practically infeasible and some approximation techniques have to be used to get solutions to problems that require to know this posterior

Variational Bayes

Among the approaches that are the most used to overcome these difficulties we find **Markov Chain Monte Carlo** and **Variational Inference** methods.

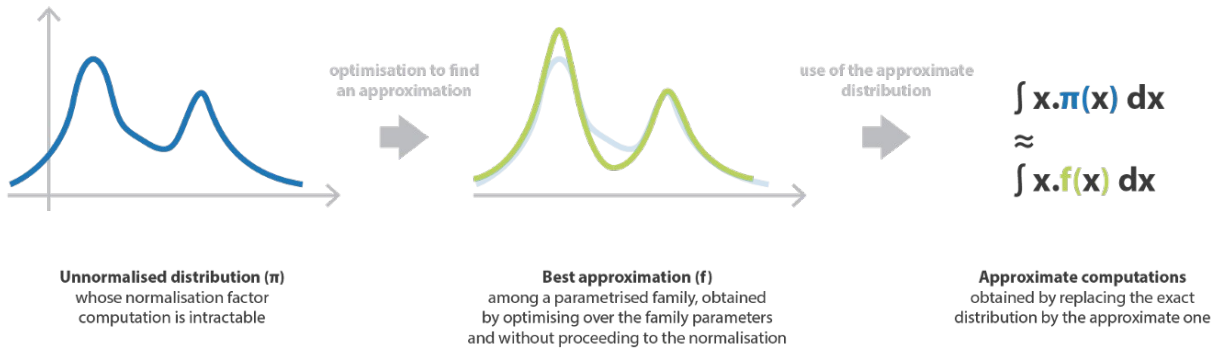
Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. They are typically used in complex statistical

models consisting of observed variables (usually termed "data") as well as unknown parameters and latent variables, with various sorts of relationships among the three types of random variables. As typical in Bayesian inference, the parameters and latent variables are grouped together as "unobserved variables". Variational Bayesian methods are primarily used for two purposes:

1. To provide an analytical approximation to the posterior probability of the unobserved variables, in order to do statistical inference over these variables.
2. To derive a lower bound for the marginal likelihood (sometimes called the *evidence*) of the observed data (i.e. the marginal probability of the data given the model, with marginalization performed over unobserved variables). This is typically used for performing model selection.

In the former purpose (that of approximating a posterior probability), **Variational Bayes** is an alternative to Monte Carlo sampling methods — particularly, Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling — for taking a fully Bayesian approach to statistical inference over complex distributions that are difficult to evaluate directly or sample. In particular, whereas Monte Carlo techniques provide a numerical approximation to the exact posterior using a set of samples, variational Bayes provides a locally-optimal, exact analytical solution to an approximation of the posterior.

VI Approach



General procedure

Problem

In variational inference, the posterior distribution over a set of unobserved variables $\mathbf{Z} = \{Z_1 \dots Z_n\}$ given some data $\mathbf{X}$ is approximated by a so-called **variational distribution**, $Q(\mathbf{Z})$:

$$P(\mathbf{Z} \mid \mathbf{X}) \sim Q(\mathbf{Z})$$

The distribution $Q(\mathbf{Z})$ is restricted to belong to a family of distributions of simpler form than $P(\mathbf{Z} \mid \mathbf{X})$ (e.g. a family of Gaussian distributions), selected with the intention of making $Q(\mathbf{Z})$ similar to the true posterior, $P(\mathbf{Z} \mid \mathbf{X})$.

The similarity (or dissimilarity) is measured in terms of a dissimilarity function $d(Q; P)$ and hence inference is performed by selecting the distribution $Q(\mathbf{Z})$ that minimizes $d(Q; P)$.

Distribution Family selection

KL divergence

The most common type of variational Bayes uses the Kullback–Leibler divergence (KL-divergence) of *Q* from *P* as the choice of dissimilarity function. This choice makes this minimization tractable. The KL-divergence is defined as

$$D_{\mathrm{KL}}(Q \parallel P) \triangleq \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z} \mid \mathbf{X})}.$$

Note that *Q* and *P* are reversed from what one might expect. This use of reversed KL-divergence is conceptually similar to the expectation-maximization algorithm. (Using the KL-divergence in the other way produces the expectation propagation algorithm.)

As a side fact, we can conclude this subsection by noticing that the KL divergence is the cross-entropy minus the entropy.

Intractability

Variational techniques are typically used to form an approximation for:

$$P(\mathbf{Z} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{Z})}{P(\mathbf{X})} = \frac{P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{Z})}{\int_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z}') \, d\mathbf{Z}'}$$

The marginalization over **Z** to calculate *P*(**X**) in the denominator is typically intractable, because, for example, the search space of **Z** is combinatorially large. Therefore, we seek an approximation, using *Q*(**Z**) ~ *P*(**Z** | **X**).

Evidence Lower Bound (ELBO)

We can rewrite the KL-divergence formula as:

$$D_{\mathrm{KL}}(Q \parallel P) = \mathbb{E}_{\mathbf{Q}} \left[\log Q(\mathbf{Z}) - \log P(\mathbf{Z}, \mathbf{X}) \right] + \log P(\mathbf{X})$$

Which can be rearranged as:

$$\log P(\mathbf{X}) = D_{\mathrm{KL}}(Q \parallel P) - \mathbb{E}_{\mathbf{Q}} \left[\log Q(\mathbf{Z}) - \log P(\mathbf{Z}, \mathbf{X}) \right] = D_{\mathrm{KL}}(Q \parallel P) + \mathcal{L}(Q)$$

As the *log-evidence* log $P(\mathbf{X})$ is fixed with respect to Q, maximizing the final term $L(Q)$ minimizes the KL divergence of Q from P. By appropriate choice of Q, $L(Q)$ becomes tractable to compute and to maximize. Hence we have both an analytical approximation Q for the posterior $P(\mathbf{Z} \mid \mathbf{X})$, and a lower bound $L(Q)$ for the log-evidence log $P(\mathbf{X})$ (since the KL-divergence is non-negative).

The term $L(Q)$ is also known as **Evidence Lower BOund**, abbreviated as **ELBO**, to emphasize that it is a lower bound on the log-evidence of the data.

References

1. https://towardsdatascience.com/bayesian-inference-problem-mcmc-and-variational-inference-25a8aa9bce29
2. https://en.wikipedia.org/wiki/Variational_Bayesian_methods