Variational Methods For Statistics

Jeremy Austin

2024-01-27

Contents

Introduction

TODO: finish introduction

Some Motivations For Variational Methods

A Problem in Bayesian Statistics

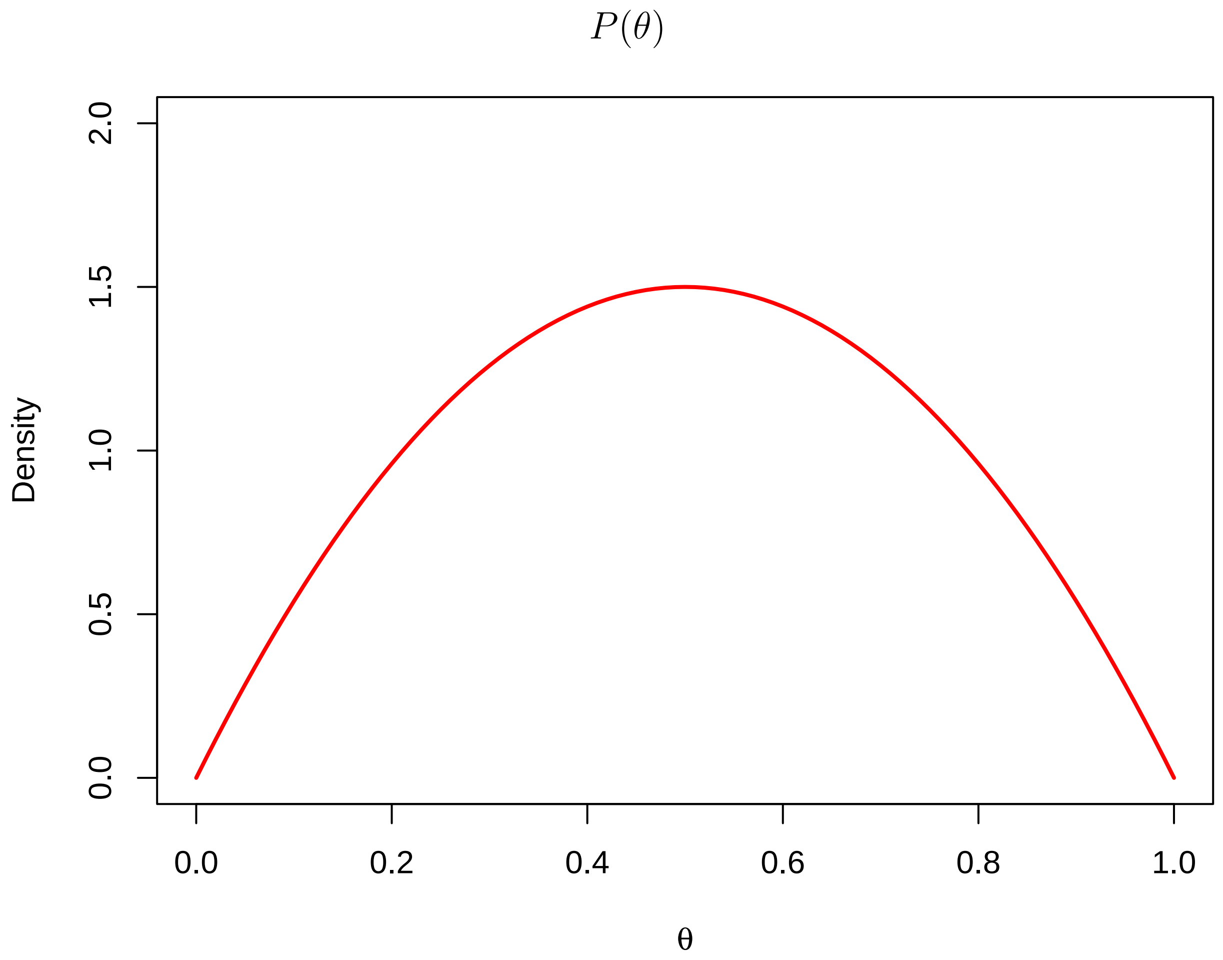
Typically, in Bayesian statistics, the objective is to find a posterior distribution of parameters and then perform inference. This distribution can be difficult to solve and sample from. The intuition behind it is based on Bayes' formula, which can be thought of as a rule for updating beliefs about a system, provided some additional information or data.

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{\Sigma_i P(B|A_i)P(A_i)}$$

Some definitions:

- θ:
  - The vector of model parameters
- D:
  - The data. This is usually a combination of independent and dependent variables.

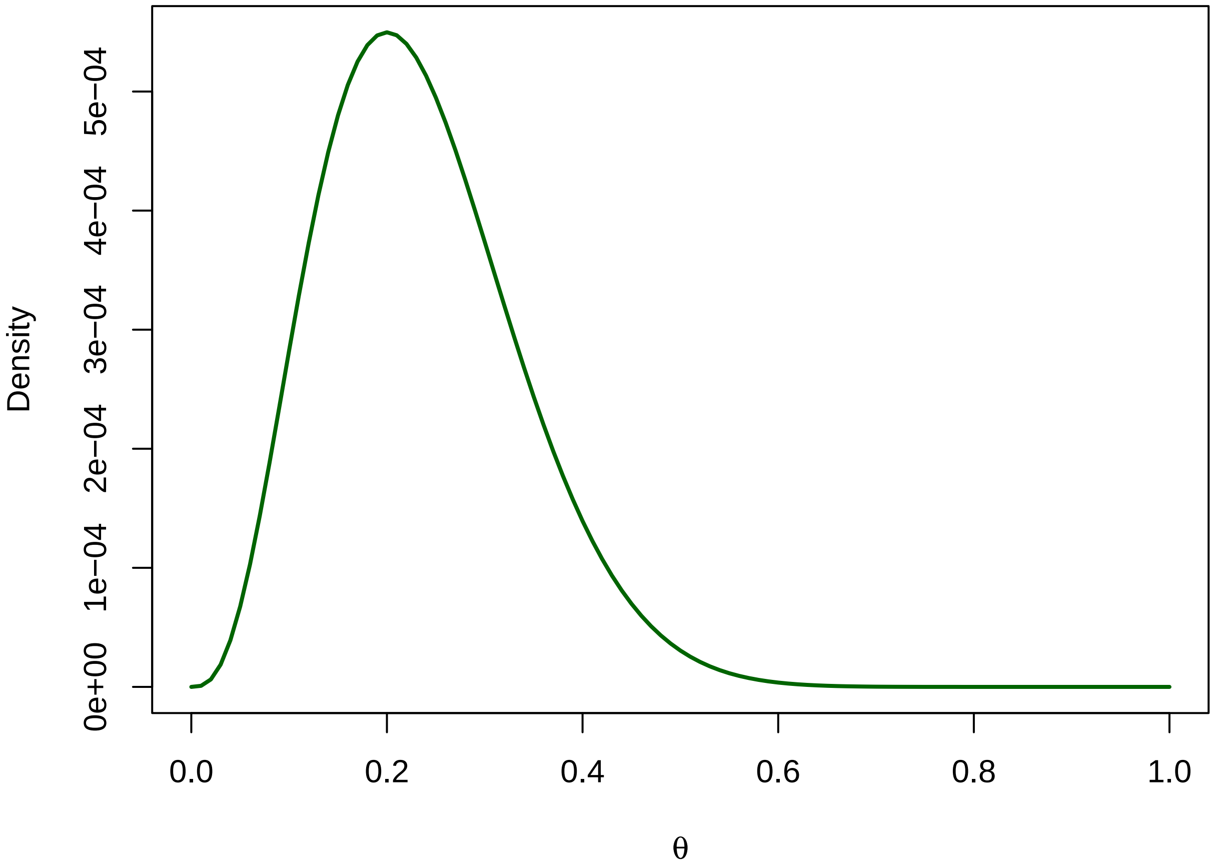
Prior distribution of parameters:



The distribution of the parameters prior to obtaining information through collecting data. Typically, the prior distribution is not known. Choices for the prior may have a variety of motivating factors. If there is some knowledge about the parameters before collecting data, it would be useful to encode it as the prior. This knowledge could come from expert opinion or past experiments. However, it is common that there is no knowledge to begin with. An honest choice in this situation could be to assign the prior distribution to be one with high entropy. In other words, a distribution that makes very little assumptions about likely parameter settings; one that is flat or uniform across the parameter space. Another popular choice for priors are ones that encourage shrinkage. These would be distributions with centers near 0 in the parameter space. It turns out that for many complex problems, biasing towards 0 yields better estimators.

Joint distribution of the data given parameters (likelihood):

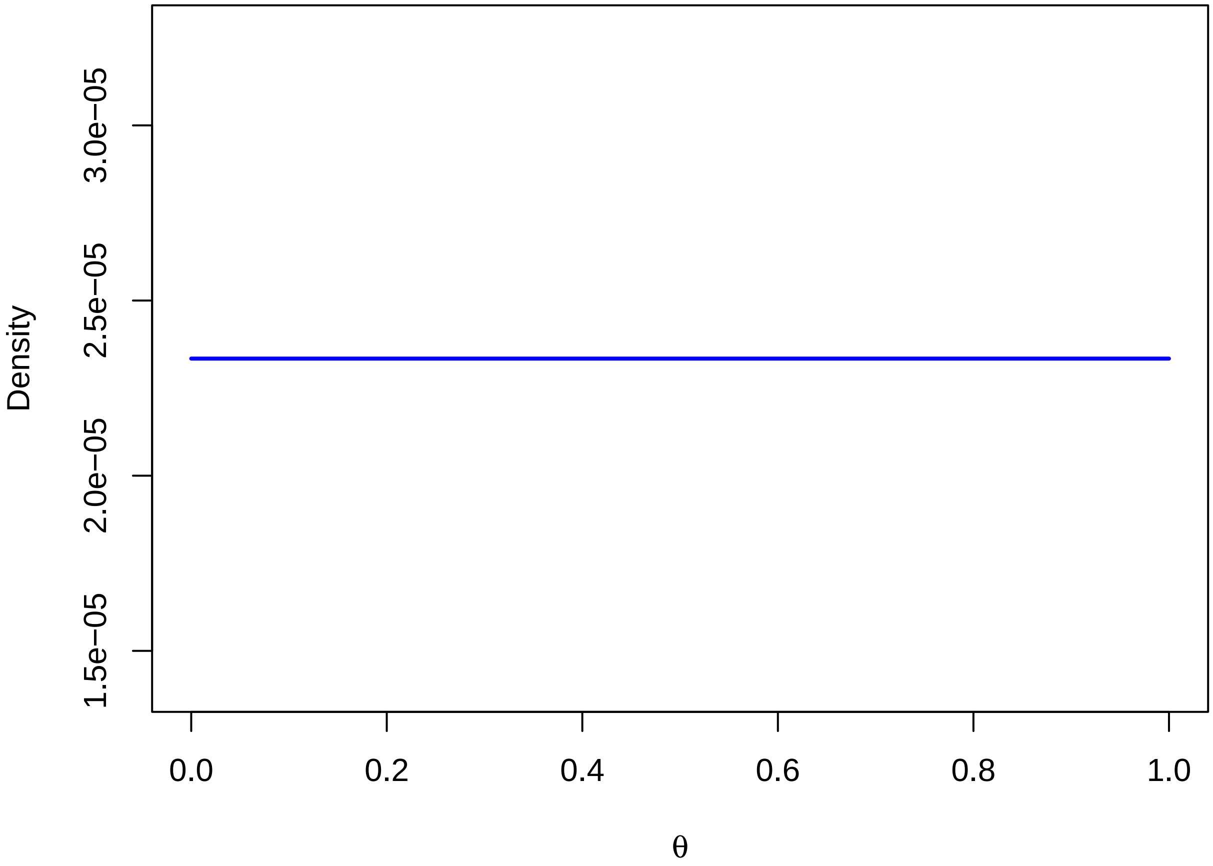
$$P(D|\theta)$$



The distribution of the data when the parameter setting is known. In classical statistics, finding the parameter setting which achieves the highest likelihood is a primary objective. In order to calculate the likelihood for a given parameter setting, it is typically assumed that all observations of the data are independent of each other, and all follow the same distribution. This distribution is usually determined by the type of the responses. For example, linear regression assumes $y_i \sim Normal(X_i\beta, \sigma^2)\ \forall i = 1, ..., n$ and logistic regression assumes $y_i \sim Bernoulli(plogis(X_i\beta))\ \forall i = 1, ..., n$. The independence property allows us to decompose the high dimensional joint distribution as the product of all the marginal distributions. When there is a lot of data, evaluating this distribution at a parameter setting usually yields exceptionally small numbers. To circumvent this for maximization, log is usually taken and the product becomes sum of logs.

Marginal distribution of the data (evidence):

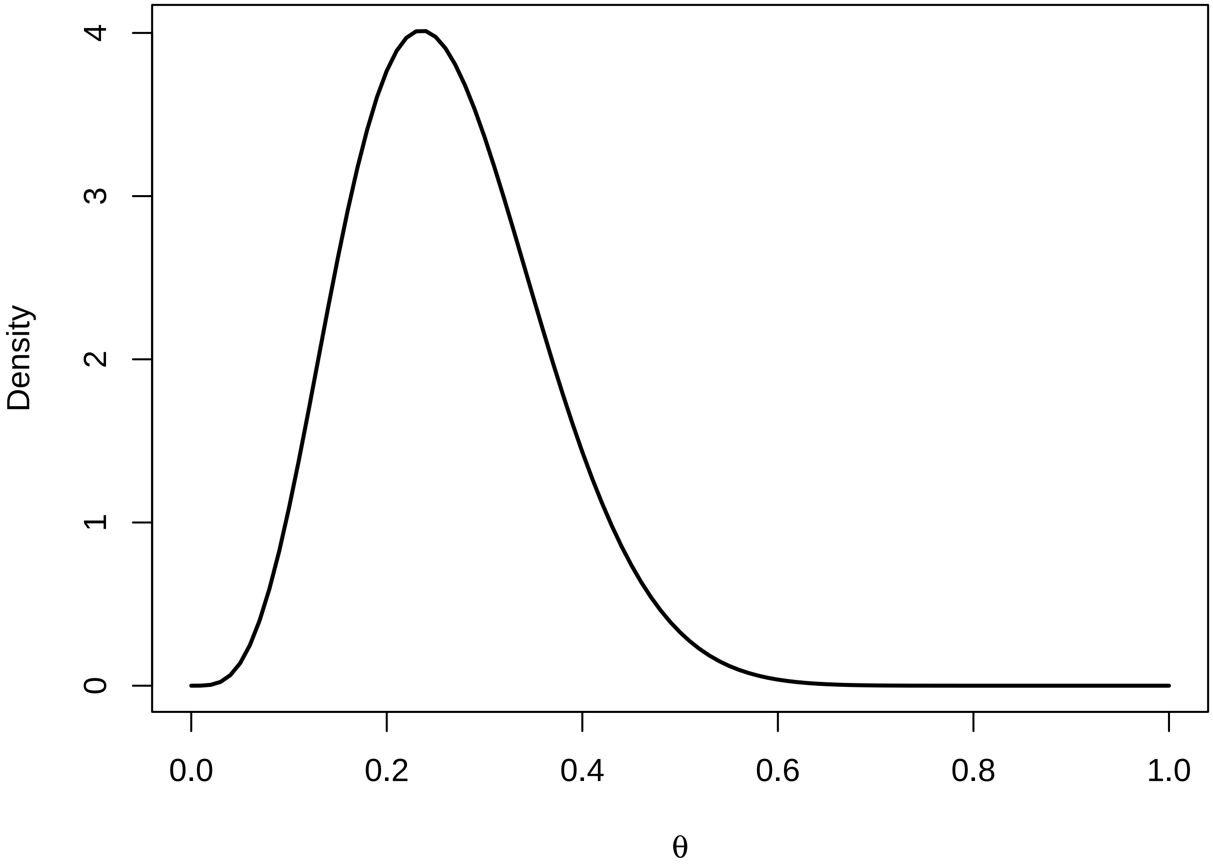
$$P(D) = \int_\theta P(D|\theta)P(\theta)d\theta$$



The distribution of the data under the model after marginalizing away the model's parameters. To find the marginal distribution of the data for a given model, it is useful to apply the law of total probability and the product rule to decompose it. $P(D) = \int_\theta P(D, \theta)d\theta = \int_\theta P(D|\theta)P(\theta)d\theta$. Because the data is known, this ends up being a constant with respect to the parameters. It is often times intractable, thus creating some problems for finding the exact posterior distribution. Another motivation for finding this quantity is in Bayesian model comparison as it is the value used in Bayes Factors. A Bayes factor is simply a ratio of the evidences of two competing models, without considering any particular parameter settings.

Posterior distribution of the parameters, given the data:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$



The distribution of the parameters, after considering the information provided from the collected data. This is the primary objective in Bayesian model fitting and inference. Using Bayes' formula we have: posterior = prior * likelihood / marginal. This distribution can be maximized with respect to the parameters to find maximum a posteriori estimators. Depending on the choice of prior, it can also effectively describe the remaining uncertainty in the model parameters and be sampled from. However, calculating the marginalization factor and sampling from this distribution can be difficult, especially when the number of parameters is high. This is the problem that needs to be addressed in Bayesian statistics.

The most frequently discussed way for handling the issues of the posterior distribution are Markov chain Monte Carlo sampling (MCMC) methods. These include algorithms like Gibbs' Sampling, and Metropolis-Hastings variants. These methods allow one to sample from the posterior distribution, by leveraging the stationary distribution of a markov chain. In fact, the posterior distribution only needs to be known up to proportionality to use these methods. This means we can avoid the tricky marginalization factor. The samples can be used to find estimators, describe uncertainty of the parameters, and perform other kinds of inference. Unfortunately, MCMC methods are not very scalable. They are also sensitive to proposal distributions, and convergence can be slow and difficult to achieve in high dimensions. The main advantage of MCMC is its simplicity, and ability to retrieve samples from the true posterior distribution. An alternative to MCMC is variational inference. Instead of sampling or describing the posterior exactly, the idea is to fit a surrogate in its place. If the surrogate is chosen carefully and properly optimized, it can be used to perform all the necessary inference that would be performed on the true posterior. The main advantage is the scalability of these methods, at the cost of not modeling the true posterior, and often-times underestimating covariance.

TODO: finish Bayesian motivation

Extending The Functionality of Autoencoders

TODO: finish autoencoder motivation

Distances Between Distributions

In order to fit a surrogate distribution in place of another, one must define a quantity that measures how far apart two distributions are from each other. Minimizing this quantity will be the objective for the surrogate. For the following discussion, let $p(x)$ represent a probability distribution, and $q(x)$ to be a surrogate.

Sum of Absolute Differences

A fairly intuitive measure might be:

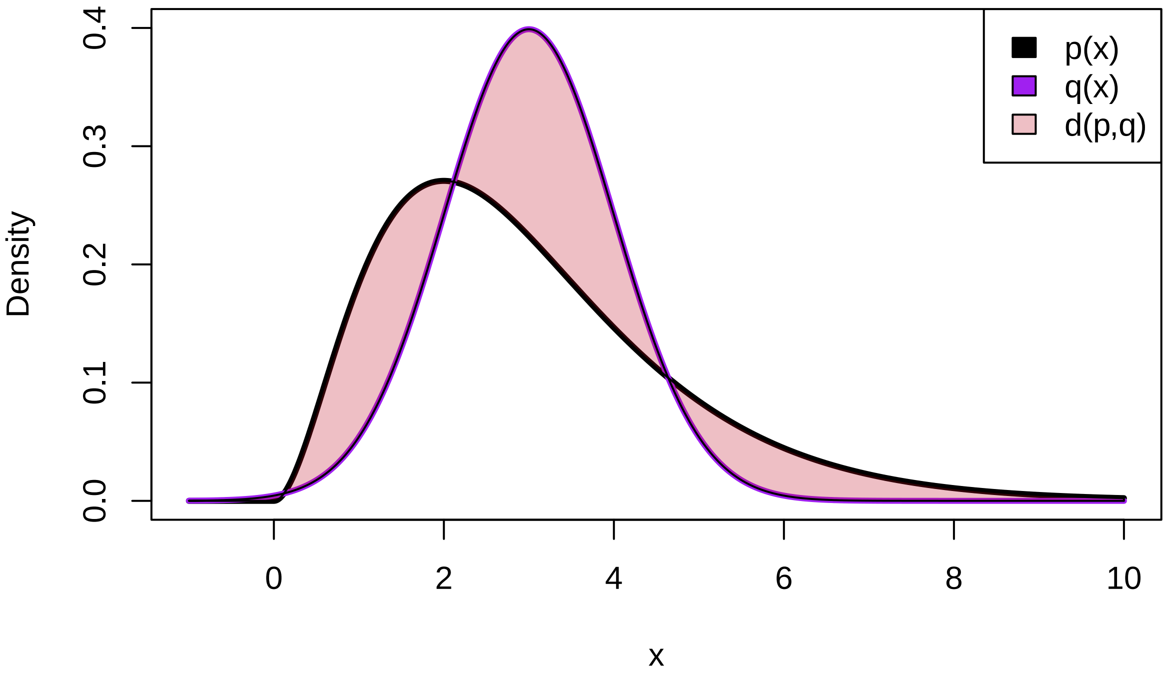
Discrete:

$$d_1(p,q) = \sum_x |p(x) - q(x)|$$

Continuous:

$$d_1(p,q) = \int_x |p(x) - q(x)| dx$$

This quantity measures the size of the region between the two probability distributions. If p and q are the same, the size of the region between them will be 0. If p and q are very different (ie probability regions do not overlap), the quantiy will approach 2. Due to the absolute values, this quantity will also be symmetric (d(p,q)=d(q,p)) and satisfy triangle inequality properties. Below is a nice visualization of this quantity:



The area of the shaded region is the same as the measure discussed above. One noticeable problem with this quantity is that the probabilities are stuck inside the absolute value in the sum. If p and q were high-dimensional joint distributions, p(x) and q(x) may be extremely small, possibly much smaller than a computer can reasonable represent. Here is a link to an interactive graph in desmos. The interactive graph represents the shaded region in a slightly different way, but is equivalent in terms of area.

Kullback-Liebler Divergence

This last quantity comes from information theory. It lends itself as one of the most useful quantities of divergence due to its convenient properties. Remarkably, this can be represented as a expected value with respect to the q distribution, so estimation is as easy as sampling from q. Another advantage of this formulation is that the rest of the density evaluations are wrapped inside logs. This means that the distributions can be broken into a sum of logs if each component variable of x are independent.

Discrete:

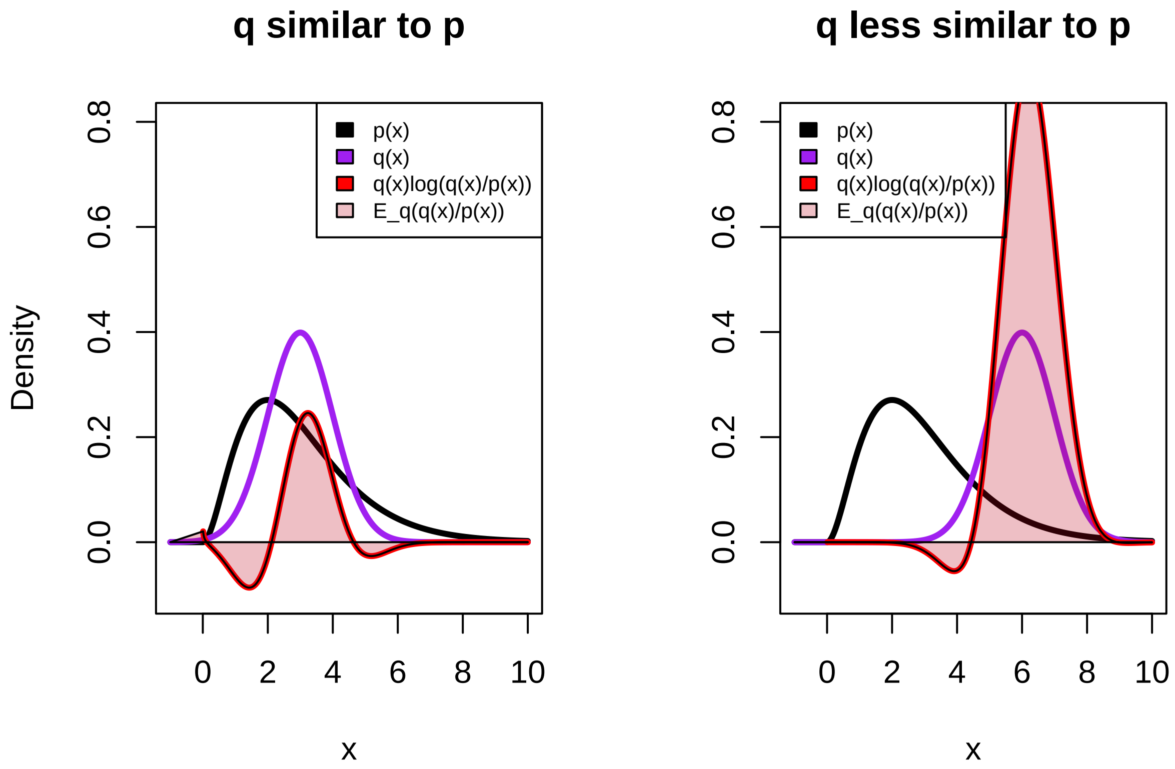
$$D_{KL}(q||p) = \sum_x q(x) log(\frac{q(x)}{p(x)})$$

Continuous:

$$D_{KL}(q||p) = \int_x q(x) log(\frac{q(x)}{p(x)}) dx$$

Another representation using expectation:

$$D_{KL}(q||p) = E_q[log(\frac{q(x)}{p(x)})] \geq 0$$



Note that when $q(x)log(\frac{q(x)}{p(x)})$ is negative, the negative area between the curve and 0 actually shrinks the divergence. This happens when p puts probability in locations where q does not assign it. In general, when p fails to assign high probability in high probability locations of q, the divergence will be large. In turn, this quantity does not penalize when p places higher probability in locations where q does not have it. This also means that KL-divergence is not symmetric and does not qualify as a metric. The reason $D_KL(q||p)$ was discussed instead of $D_KL(p||q)$ is because usually p, the true distribution, is difficult to sample from for the estimation of the expected value. Here is a link to desmos.

Kullback-Liebler Divergence Continued: Evidence Lower Bound Frequently, when employing variational methods, a quantity known as evidence lower bound plays a key role. It is useful when working with latent variables or when performing Bayesian inference. Below is a manipulation of probability equations that derives it, using symbols from the Bayesian setting described in the first section. Only the continuous case is discussed.

$$D_{KL}(q(\theta)||p(\theta|D)) = E_q[log(\frac{q(\theta)}{p(\theta|D)})]$$

$$= E_q[log(\frac{q(\theta)}{\frac{p(D|\theta)p(\theta)}{p(D)}})]$$

$$= E_q[log(\frac{q(\theta)}{p(D|\theta)p(\theta)}) + log(p(D))]$$

$$= E_q[log(\frac{q(\theta)}{p(D|\theta)p(\theta)})] + log(p(D))$$

$$E_q[log(\frac{q(\theta)}{p(D|\theta)p(\theta)})] + log(p(D)) \geq 0$$

$$log(p(D)) \geq -E_q[log(\frac{q(\theta)}{p(D|\theta)p(\theta)})]$$

$$log(p(D)) \geq E_q[log(\frac{p(D|\theta)p(\theta)}{q(\theta)})]$$

Thus, $E_q[log(\frac{p(D|\theta)p(\theta)}{q(\theta)})]$ is a lower bound on the log of the model evidence. Maximizing the evidence lower bound has the same effect as minimizing the KL divergence between $q(\theta)$ and $p(\theta|D)$ because the log model evidence can be removed from the KL divergence as an additive constant. Maximizing the evidence lower bound also gives us a better idea for the actual value of the log evidence, although it will still be biased less than the true log evidence:

$$log(p(D)) = log(\int_x p(D|\theta)p(\theta)d\theta)$$

$$= log(\int_x q(\theta)\frac{p(D|\theta)p(\theta)}{q(\theta)}d\theta)$$

$$= log(E_q[\frac{p(D|\theta)p(\theta)}{q(\theta)}])$$

By concavity of logarithm and Jensen's inequality...

$$log(p(D)) = log(E_q[\frac{p(D|\theta)p(\theta)}{q(\theta)}]) \geq E_q[log(\frac{p(D|\theta)p(\theta)}{q(\theta)})]$$

TODO: finish KL-divergence section

Citations

TODO: finish citations