VARIATIONAL INFERENCE

Anand Khandekar
The Running Professor
Pune, 411038
khandekarsir@gmail.com

May 19, 2020

ABSTRACT

A central task in the application of probabilistic models is the evaluation of the posterior distribution p(Z|X) of the latent variables Z given the observed (visible) data variables X, and the evaluation of expectations computed with respect to this distribution. The model might also contain some deterministic parameters, which we will leave implicit for the moment, or it may be a fully Bayesian model in which any unknown parameters are given prior distributions and are absorbed into the set of latent variables denoted by the vector Z. For instance, in the EM algorithm we need to evaluate the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables. For many models of practical interest, it will be infeasible to evaluate the posterior distribution or indeed to compute expectations with respect to this distribution. This could be because the dimensionality of the latent space is too high to work with directly or because the posterior distribution has a highly complex form for which expectations are not analytically tractable.
In such cases wee need to rely on two approaches, according to whether they rely on Stochastic or Deterministic approximations.MCMC is a Stochastic technique, enabling Bayesian methods, which have a property that given infinite computational resources, they can generate EXACT results.
In this article it is intended to introduce Deterministic approximation schemes, some of which scale well to large applications. They are based on analytical approximations of the posterior distribution.As such, they can never generate exact results. Specifically, we review Variational Inference (VI), a method from machine learning that approximates probability densities through optimization. The idea behind VI is to first posit a family of densities and then to find the member of that family which is close to the target. Closeness is measured by Kullback-Leibler divergence.

1 Introduction

A central task in the application of probabilistic models is the evaluation of the posterior distribution p(Z|X) of the latent variables Z given the observed (visible) data variables X, and the evaluation of expectations computed with respect to this distribution.The model might also contain some deterministic parameters, which we will leave implicit for the moment, or it may be a fully Bayesian model in which any unknown parameters are given prior distributions and are absorbed into the set of latent variables denoted by the vector Z. For instance, in the EM algorithm we need to evaluate the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables. For many models of practical interest, it will be infeasible to evaluate the posterior distribution or indeed to compute expectations with respect to this distribution. This could be because the dimensionality of the latent space is too high to work with directly or because the posterior distribution has a highly complex form for which expectations are not analytically tractable.
One of the core problems of modern statistics is to approximate difficult-to-compute probability densities. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation about the posterior. Modern Bayesian statistics relies on models for which the posterior is not easy to compute and corresponding algorithms for approximating them.

2 General Set Up

Variational Inference is widely used to approximate posterior densities for Bayesian models, an alternative strategy to Markov chain Monte Carlo (MCMC) sampling.
First, we set up the general problem. Consider a joint density of latent variables z = z1:m and observations x = x1:n, p(z, x) = p(z)p(x|z).

Task modeling In Bayesian models, the latent variables help govern the distribution of the data. A Bayesian model draws the latent variables from a prior density p(z) and then relates them to the observations through the likelihood p(x|z). Inference in a Bayesian model amounts to conditioning on data and computing the posterior p(z|x). In complex Bayesian models, this computation often requires approximate inference.

KL divergence Equation Rather than use sampling, the main idea behind variational inference is to use optimization. First, we posit a family of approximate densities Q. This is a set of densities over the latent variables. Then, we try to find the member of that family that minimizes the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(z) argmin_{q(z)} KL(q(z)||p(z|x)) \tag{1}$$

q(z) = argmin Finally, we approximate the posterior with the optimized member of the family q().VI thus turns the Inference problem into that of Optimization.

2.0.1 Comparing MCMC and VI

- MCMC methods tend to be more computationally intensive.
- MCMC also provide guarantees of producing (asymptotically) exact samples from the target density.
- VI does not guarantee this exactness- it can only find a close density but tends to be faster than MCMC.
- Because VI is based on Optimisation, therefore it take the advantages offered by methods like Stochastic Optimization.
- VI is suited for large data sets and cases where exploration needs to be done quickly.
- MCMC is suited to smaller data sets and scenarios where we happily pay a heavier computational cost for more precise samples.
- We use VI when fitting a probabilistic model of text to one billion text documents and where inferences will be served to serve search results to a large population of users.
- We do know that variational inference generally underestimates the variance of the posterior density.This is the consequence of its objective function.
- MCMC is a tool for simulating from densities while VI is a tool for approximating densities.
- Apart from the data size, geometry of the posterior distribution is also a factor

3 Variational Inference

The goal of VI is to approximate the conditional density of the latent variables given observed variables. This is achieved by considering ti as a problem of Optimization.We use a family of densities over the latent variables,parameterized by free " variational parameters". The optimization finds the member of this family,i.e. the setting of the parameters, that is closest in KL divergence to the conditional density.

The problem of approximate Inference Let x = x1:n be a set of observed variables and z = z1:m be a set of latent variables, with joint density p(z, x).The inference problem is to compute the conditional density of the latent variables given the observations, p(z|x). This conditional (posterior) can be used to create points or interval estimates of the latent variables, form predictive densities of new data and much more. We can write the conditional density as

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{2}$$

The denominator contains the marginal density of the observations, also called the evidence. We calculate it by marginalizing out the latent variables from the joint density

$$p(x) \;=\; \int p(z,x)dz \tag{3}$$

For many models the evidence integral is unavailable in the closed form or requires exponential time to compute.The evidence is what we need to compute the conditional form the joint; this is why inference in such models is HARD.

3.1 The Evidence Lower Bound

ELBO : In VI we specify the family Q of densities over the latent variables. Each $q(z) \in Q$ is a candidate approximation to the exact conditional. Our goal is to find the best candidate, the one closest in KL Divergence to the exact conditional. Inference now amounts to solving the following Optimization problem,

$$q^*(z) \ = \ argmin_{q(z)\in Q} KL(q(z)||p(z|x)) \tag{4}$$

Once found, q(.) is the best approximation of the conditional, within the family Q. The complexity of the family determines the complexity of this optimization.However, this objective is not computable because it requires computing the evidence log p(x).(That the evidence is hard to compute is why we appeal to approximate inference in the first place.) To see why, recall that KL divergence is

$$KL(q(z)||p(z|x)) \ = \ \mathbb{E}[log\ q(z)] - \mathbb{E}[log\ p(z|x)] \tag{5}$$

where all the Expectations are taken with respect to $q(z)$. Now, expand the conditional

$$KL(q(z)||p(z|x)) \ = \ \mathbb{E}[log\ q(z)] - \mathbb{E}[log\ p(z,x)] + log\ p(x) \tag{6}$$

This clearly reveals its dependence on $p(x)$. Because we cannot compute the KL , we optimize and ALTERNATIVE objective function that is equivalent to the KL upto the added constant.

$$ELBO(q) \ = \ \mathbb{E}[log\ p(z,x)] - \mathbb{E}[log\ q(z)] \tag{7}$$

This function is called the evidence lower bound (ELBO). The ELBO is the negative KL divergence of Equation (6) plus log p(x), which is a constant with respect to q(z). Maximizing the ELBO is equivalent to minimizing the KL divergence.
Examining the ELBO gives intuitions about the optimal variational density. We re write the ELBO as the sum of the expected log likelihood o the data and the KL divergence between the prior $p(z)$ and $q(z)$. Examining the ELBO gives intuitions about the optimal variational density. We rewrite the ELBO as a sum of the expected log likelihood of the data and the KL divergence between the prior $p(z)$ and $q(z)$.,

$$ELBO(q) \ = \ \mathbb{E}[log\ p(z)] + \mathbb{E}[log\ p(x|z)] - \mathbb{E}[log\ q(z)] \tag{8}$$

$$ELBO(q) \ = \ \mathbb{E}[log\ p(x|z)] \ - \ KL(q(z)||p(z)) \tag{9}$$

W hich values of z will this objective encourage q(z) to place its mass on? The first term is an expected likelihood; it encourages densities that place their mass on configurations of the latent variables that explain the observed data. The second term is the negative divergence between the variational density and the prior; it encourages densities close to the prior.
Thus the variational objective mirrors the usual balance between likelihood and prior. Another property of the ELBO is that it lower-bounds the (log) evidence, log p(x) $\geq$ ELBO(q) for any q(z). This explains the name.To see this , notice tat the equations (6) and(7) give the following expression

$$log\ p(x) \ = \ KL(q(z)||p(z|x)) + ELBO(q) \tag{10}$$

The bound follows that KL()$\geq$ 0. In the original paper submitted by Kullback and Leibler on VARIATIONAL INFERENCE this was proved using JENSON's INEQUALITY.
The relationship between the ELBO and log p(x) has led to using the variational bound as a model selection criterion. Finally, notice that the first term of the ELBO in Equation (9) is the expected complete log-likelihood, which is optimized by the EM algorithm.The EM algorithm was designed for finding maximum likelihood estimates in models with latent variables. It uses the fact that the ELBO is equal to the log likelihood log p(x) (i.e., the log evidence) when q(z) = p(z |x). EM alternates between computing the expected complete log likelihood according to p(z |x) (the E step) and optimizing it with respect to the model parameters (the M step).
Unlike variational inference, EM assumes the expectation under p(z|x) is computable and uses it in otherwise difficult parameter estimation problems. Unlike EM, variational inference does not estimate fixed model parameters—it is often used in a Bayesian setting where classical parameters are treated as latent variables. Variational inference applies to models where we cannot compute the exact conditional of the latent variables.

3.2 The Mean-Field Variational Family

Overcoming the ELBO difficulty We described the ELBO, the variational objective function in the optimization of Equation (7). We now describe a variational family Q, to complete the specification of the optimization problem. The complexity of the family determines the complexity of the optimization; it is more difficult to optimize over a complex

family than a simple family.
In this review we focus on the mean-field variational family, where the latent variables are mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational family is

$$q(z) \ = \ \Pi_{j=1}^{m} q_j(z_j) \tag{11}$$

Each latent variable z_j is governed by its own variational factor, the density $q_j(z_j)$. In optimization, these variational factors are chosen to maximize the ELBO of Equation (7).
We emphasize that the variational family is not a model of the observed data—indeed, the data x does not appear in Equation (11). Instead, it is the ELBO, and the corresponding KL minimization problem, that connects the fitted variational density to the data and model.
Notice we have not specified the parametric form of the individual variational factors. In principle, each can take on any parametric form appropriate to the corresponding random variable.

Visualising the Mean-Field Approximation The mean-field family is expressive because it can capture any marginal density of the latent variables. However, it cannot capture correlation between them. Seeing this in action reveals some of the intuitions and limitations of mean-field variational inference.

3.3 Co-ordinte Ascent mean-field variational inference

The ALGORITHM Using the ELBO and the mean-field family, we have cast approximate conditional inference as an optimization problem. In this section, we describe one of the most commonly used algorithms for solving this optimization problem, coordinate ascent variational inference.CAVI iteratively optimizes each factor of the mean-field variational density, while holding the others fixed. It climbs the ELBO to a local optimum.
We understand this algorithm the other way round. Let us understand the RESULT first. Consider the j^thlatent variable z_j. The $complete\ conditional$ of $z_j$is its conditional density given of all the other latent variables in the model and the observations, $p(z_j|z_{-j}, x)$. Fix the other variational factors $q_l(z_l), l \neq j$. The optimal $q_j(z_j)$ is then proportional to the expontiated expected log of the complete conditional,

$$q_j^*(z_j) \ \propto \ exp\{\mathbb{E}_{-j} log\ p(z_j|z_{-j}, x)]\} \tag{12}$$

The expectation in equation (12) is with respect to (currently fixed)variational density over z_{-j} that is, $\prod_{l\neq j} q_l(z_l)$.
Equivalently, equation (12) is proportional to the exponentiated log of the joint likelihood

$$q_j^*(z_j) \ \propto \ exp\{\mathbb{E}_{-j} log\ p(z_j, z_{-j}, x)]\} \tag{13}$$

Because of the mean field family assumption - that all the latent variables are independent - the expectations on the right hand side do not involve the j^{th} variational factor. Thus, this is a valid corodinate update.
Note that these equations underlie the CAVI (coordinate ascent Variational Inference)Algorithm presented below.We iterate through them, updating, $q_j(z_j)$ using the equation (13)

Algorithm 1: Corodinate Ascent Varaitional Inference (COVI)

Input: A model $p(x, z)$, a data set x
Output: A variational density $q(z) \ = \ \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while *the ELBO has not converged* **do**
  for $j \in \{1, 2, ...m\}$ **do**
    Set $q_j^*(z_j) \ \propto \ exp\{\mathbb{E}_{-j} log\ p(z_j|z_{-j}, x)]\}$;
  end
  Compute ELBO(q) = $\mathbb{E}[log\ p(z, x)] - \mathbb{E}[log\ q(z)]$
end
return q(z)

CAVI goes uphill on equation(7), eventually finding a local optimum.

The DERIVATION (Bishop's point of view) Factorized Distributions : Suppose wee partition the elements of Z into disjoint groups that we denote by Z_i where $i = 1, 2, ...M$. We assume that the q distributions factorizes with respect to these groups, so that

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i) \tag{14}$$

Note that we place no restrictions of the functional forms of the individual factors $q_i(Z_i)$. This factorised form of VI is rooted in a similar framework developed in physics call the MEAN-FIELD THEORY.
Amongst all the distributions q(Z) having the form of equation (14) , we now seek that particular distribution for which the Lower Bound L(W) is the largest. We intend to make a free form (variational) optimization of L(q) wrt all the distributions $q_i(Z_i)$. Let us denote $q_i(Z_i)$ as a simple q_i.

$$L(q) \; = \; \int \prod_i q_i \; \{ \; ln \; p(X, Z) \; - \; \Sigma_i \; ln(q_i) \; \} dZ \tag{15}$$

$$L(q) \; = \; \int \; q_j \{ \int \; ln \; p(X, Z) \; \prod_{i \neq j} \; q_i \; dZ_i \} \; dZ_j \; - \; \int \; q_j \; ln(q_j) \; dZ_j \; + \; constant \tag{16}$$

$$L(q) \; = \; \int \; q_j \; ln \; \overline{\mathrm{p}}(X, Z_j) \; dZ_j \; - \; \int \; q_j \; ln(q_j) \; dZ_j \; + \; constant \tag{17}$$

where we have defined the new distribution $\overline{\mathrm{p}}(X, Z_j)$ by the relation

$$ln \; \overline{\mathrm{p}}(X, Z_j) \; = \; \mathbb{E}_{i \neq j}[ln \; p(X, Z)] \; + \; constant \tag{18}$$

Here the notation $\mathbb{E}_{i \neq j}[...]$ denotes an expectation wrt the q distributions over all variables z_i for $i \neq j$, so that

$$\mathbb{E}_{i \neq j}[ln \; p(X, Z)] \; = \; \int \; ln \; p(X, Z) \; \prod_{i \neq j} q_i \; dZ_i \tag{19}$$

Now suppose we keep the $q_{i \neq j}$ fixed and maximize $L(q)$ in the equation(17) with respect to all the possible forms for the distribution $q_j(Z_j)$. This is very easy since we recognise that equation(17) is the negative Kullback-Leibler divergence between $q_j(Z_j)$ and $\overline{\mathrm{p}}(X, Z_j)$. Thus maximizing equation(17) is equivalent to minimizing the Kullback-Leibler divergence, and the minimum occurs when $q_j(Z_j) \; = \; \overline{\mathrm{p}}(X, Z_j)$. Thus we obtain a general expression for the optimal solution $q_j^*(Z_j)$ given by the equation

$$ln \; q_j^*(Z_j) \; = \; \mathbb{E}_{i \neq j}[ln \; p(X, Z)] \; + \; constant \tag{20}$$

IMPORTANT: equation(20) provides the necessary basis for applications of variational methods.
It says that the log of the optimal solution for the factor q_j is obtained simply by considering the log of the joint distribution over all hidden and visible variables and then taking the expectation with respect to all the other factors $\{q_i\}$ for $i \neq j$. The additive constant in equation(20) is set by normalizing the distribution $q_j^*(Z_j)$. Thus if we take the exponential on both the sides and normalize, wee have

$$q_j^*(Z_j) \; = \; \frac{exp \; (\mathbb{E}_{i \neq j}[ln \; p(X, Z)])}{\int \; exp \; (\mathbb{E}_{i \neq j}[ln \; p(X, Z)]) dZ_j} \tag{21}$$

The set of equations given by (19) for j = 1, . . . , M represent a set of consistency conditions for the maximum of the lower bound subject to the factorization constraint. However, they do not represent an explicit solution because the expression on the right-hand side of (19) for the optimum $q_j^*(Zj)$ depends on expectations computed with respect to the other factors $q_i(Z_i)$ for $i \neq j$. We will therefore seek a consistent solution by first initializing all of the factors $q_i(Z_i)$ appropriately and then cycling through the factors and replacing each in turn with a revised estimate given by the right-hand side of (19) evaluated using the current estimates for all of the other factors. Convergence is guaranteed because bound is convex with respect to each of the factors $q_i(Z_i)$.

The DERIVATION (David Blei- Columbia point of view) Recall the equation(7) above as

$$ELBO(q) \; = \; \mathbb{E}[log \; p(z, x)] - \mathbb{E}[log \; q(z)] \tag{22}$$

We rewrtite this equation as as the function of the J^th variational factor $q_j(Z_j)$ absorbing into a constant the rems that do no depend on it,

$$ELBO(q_j) \; = \; \mathbb{E}_j[\; \mathbb{E}_{-j}[\; ln \; p(z_j, z_{-j},)]] \; - \; E_j[log \; q_j(z_j)] \; + \; constant \tag{23}$$

We have rewritten the first term of the ELBO using iterated expectation. The second term we have decomposed, using the independence of the variables (i.e., the mean-field assumption) and retaining only the term that depends on $q_j(z_j)$. Up to an added constant, the objective function in Equation (23) is equal to the negative KL divergence between $q_j(z_j)$ and $q_j^*(z_j)$ from equation (13). Thus we maximize the ELBO with respect to q_j when we set $q_j(z_j) \; = \; q_j^*(z_j)$.

Practicalities Here are a few points to highlight when implementing and using variational inference in practise.

- **Initialization** The ELBO is generally a non-convex objective function. CAVI only guarantees convergence to a local optimum, which can be sensitive to initialization. But note that this is not always a disadvantage. Somme models, such as a mixture of Gaussians, and mixed membership models, exhibit many posterior models die to label switching: swapping cluster assignment labels includes many symmetric posterior models. Representing one of these models is sufficient for exploring latent clusters or predicting new observations.

- **Assessing Convergence** Monitoring the ELBO in CAVI is simple; we typically assess convergence once the change in ELBO has fallen below some small threshold. However, computing the ELBO of the full dataset may be undesirable. Instead, we suggest computing the average log predictive of a small held-out dataset. Monitoring changes here is a proxy to monitoring the ELBO of the full data. (Unlike the full ELBO, held-out predictive probability is not guaranteed to monotonically increase across iterations of CAVI.)

- **Numerical stability** Probabilities are constrained to live within [0,1]. Precisely manipulating and performing arithmetic of small numbers requires additional care. When possible, we recommend working with logarithms of probabilities. One useful identity is the "log-sum-exp" trick,

$$log[\ \sigma_i\ exp(x_i)]\ =\ \alpha\ +\ log[\Sigma_i\ exp(x_i - \alpha] \tag{24}$$

  The constant α is typically set to $max_i x_i$. This provides numerical stability to common computations in variational inference procedures.

4 A Complete Example : Bayesian Mixture of Gaussian

4.1 The Variational Density of Mixture Assignments

4.2 The Variational Density of mixture-component means

4.3 CAVI for Mixture of Guassians