MA4270 Project – Variational Inference

Wang Zhenlin

October 22, 2020

1 Introduction

1.1 Background of Bayesian methods

In the field of machine learning, most would agree that frequentist approaches played a critical role in the development of early classical models. Nevertheless, we are witnessing the increasing significance of Bayesian methods in modern study of machine learning and data modelling. The simple-looking Bayes' rule $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$ has inspired a lot wonderful models in areas like topic modelling, representation learning and hyperparameter optimization. In these models, the latent variables $z = (z_1, z_2, ...z_n)$ are the focus of the study. By analysing several data on the observed variables $x = (x_1, x_2, ...x_m)$, we hope to get some meaningful information (for example, a point estimate or an entire distribution) about these latent variables.

1.2 Problem with Bayesian methods: intractable integral

While the rule looks easily understandable, the numerical computation is hard in reality. One major issue is the intractable integral $\int_z p(x|z)p(z)dz$ we need to compute in order to get the $p(x)$, which is often called the "model evidence"[**?**]. This is often because the search space for Z is combinatorially too large, making the computation extremely expensive. A common approach to deal with this problem is to approximate the posterior probability $p(z|x)$ directly. Some popular choices include Monte Carlo Sampling methods and variational inference. In this report, we will introduce the variational methods, which are perhaps the most widely used inference technique in machine learning[2]. We will analyse a particularly famous technique in variational methods, mean-field variational inference.[**?**]

1.3 Main idea of variational inference

In variational inference, we can avoid computing the intractable integral by magically modelling the posterior p directly. The main trick here is to approximate the unknown distribution p with some similar distribution q. Since we can choose the q to belong to a certain family of distribution (hence tractable), the problem is now transformed into an optimization problem about the parameters of q.

2 Understanding Variational Bayesian methods

In this section, we demonstrate the theory behind variational Bayesian methods.

2.1 Kullback-Leibler Divergence

As mentioned above, variational inference needs a distribution q to approximate the posterior distribution p. Therefore we need to gauge how well a candidate q approximates the posterior. A common measure is Kullback-Leibler Divergence (often called KL divergence).

KL divergence is defined as

$$KL(q||p) = \int_z q(z) \frac{q(z)}{p(z|x)} = E_q[\log \frac{q(z)}{p(z|x)}] \tag{1}$$

Where E_q means the expected value with respect to distribution q. The formula can be interpreted as follows:

- if q is low, the divergence is generally low.
- if q is high and p is high, the divergence is low.
- if q is high and p is low, the divergence is high, hence the approximation is not ideal.

Take note of the following about use of KL divergence in Variational Bayes:

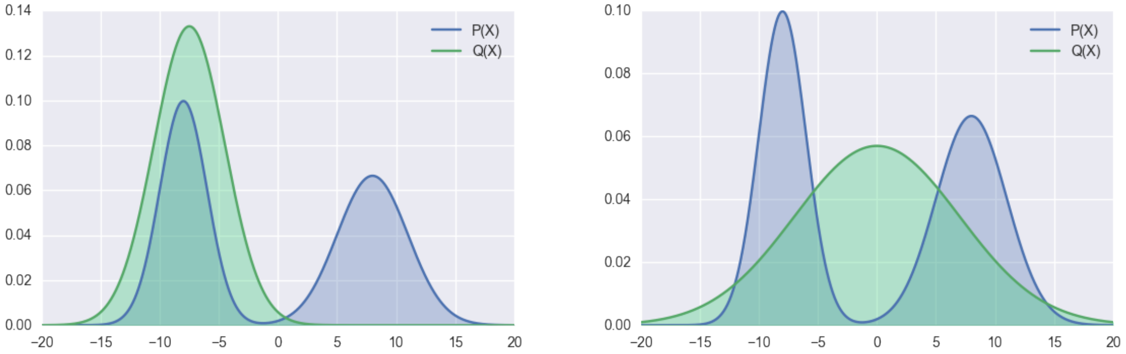
1. KL divergence is not symmetric, it's easy to see from the formula that $KL(p||q) \neq KL(q||p)$ as the approximation distribution q is usually different from the target distribution p.
2. In general, we focus on approximating some regions of p as good as possible (Figure 1(a)). It is not necessary for the q to nicely approximate every part of p. As a result $KL(p||q)$ (usually called forward KL divergence) is not ideal. Because for some regions $p > 0$ which we don't want to care, if $q \to 0$, the KL divergence will be very large, forcing q to take a different form even if it fits well with other regions of p (refer to Figure 1(b)). On the other hand, $KL(q||p)$ (usually called reverse KL divergence) has the nice property that only regions where $q > 0$ requires p and q to be similar. Consequently, reverse KL divergence is more commonly used in Variational Inference[1].

2.2 Evidence lower bound

Usually we don't directly minimizing KL divergence to obtain a good approximated distribution. This is because computing KL divergence still depends on the posterior $p(z|x)$. The computation involves the "evidence" term $p(x)$ which is expensive to compute, as shown in the formula below:

$$\begin{aligned} KL(q||p) &= E_q[\log \frac{q(z)}{p(z|x)}] \\ &= E_q[\log q(z)] - E_q[\log p(z|x)] \\ &= E_q[\log q(z)] - E_q[\log p(z,x)] + \log p(x) \\ &= -(E_q[\log p(z,x)] - E_q[\log q(z)]) + \log p(x) \end{aligned} \tag{2}$$

[1] The approximation using reverse KL divergence usually gives good empirical results, even though some regions of p may be compromised



(a) Reverse KL simulation

(b) Forward KL simulation

Figure 1: Reverse KL vs Forward KL divergence: The left image has a better approximation Q(x) on part of P(x) (source: Agustinus Kristiadi's Blog - KL Divergence: Forward vs Reverse?).

We can directly conclude by the fact $KL(q||p) \geq 0$ that the term $E_q[\log p(z,x)] - E_q[\log q(z)]$ is less than the log of evidence. We can also proof this result using Jensen's inequality as follows:

- By the definition of marginal probability, we have $p(x) = \int_z p(x,z)$, take log on both side we have:

$$\begin{aligned} \log p(x) &= \log \int_z p(x,z) \\ &= \log \int_z p(x,z) \frac{q(z)}{q(z)} \\ &= \log(E_q[\frac{p(z,x)}{q(z)}]) \\ &\geq E_q[\log p(z,x)] - E_q[\log q(z)] \end{aligned} \tag{3}$$

- The last 2 lines follow from Jensen's Inequality which states that for a convex function f, we have $f(E[X]) \geq E[f(X)]$.

This term $E_q[\log p(z,x)] - E_q[\log q(z)]$ is known as the Evidence Lower Bound, or $ELBO$. Since $\log p(x)$ does not depend on q, we can treat it as a constant from the perspective of optimizing q. Hence, minimizing $KL(q||p)$ is now equivalent to maximizing $ELBO$.

2.3 General procedure

In general, a variational inference starts with a family of variational distribution (such as the mean-field family described below) as the candidate for q. We can then use the manually chosen q to compute the ELBO terms $E_q[\log p(z,x)]$ and $E_q[\log q(z)]$. Afterwards, we optimize the parameters in q to maximize the ELBO value using some optimization techniques (such as coordinate ascent and gradient methods).

3 Mean Field Variational Family

3.1 The "Mean Field" Assumptions

As shown above, the particular variational distribution family we use to approximate the posterior p is chosen by ourselves. A popular choice is called the mean-field variational family. This family of distribution assumes joint approximation distribution $q(Z)$ to be factorized over some partition of the latent variables. This implies mutual independence among the n fractions in the partition. In particular: we have

$$p(z|X) \approx q(z) = \prod_{i=1}^{n} q_i(z_i) \tag{4}$$

where z is factorized into $z_1, ..., z_n$. For simplicity, we assume that each fraction only contains 1 latent variable ($n = |z|$), it is often referred as "naive mean field". This family is nice to analyse because we can model each distribution with a tractable distribution based on the problem set-up. Do note that a limitation of this family is that we cannot easily capture the interdependence among the latent variables.

3.2 Derivation of optimal $q_j(z_j)$

Now in order to derive the the optimal form of distribution for $q_j(z_j)$ and thus the overall q, we need to go back to the ELBO optimization with this mean-field family assumption. Recall the formula for ELBO (we use $\mathcal{L}$ here as it is the convention): $\mathcal{L} = E_q[\log p(x, z)] - E_q[\log q(z)]$. We express this formula in terms of $q_j(z_j)$ as using functional integral (see appendix A):

$$\mathcal{L} = \int_{z_j} q_j(z_j) E_{q_{-j}}[\log p(x, z)]\, dz_j + E_{q_j}[\log q_j(z_j)] + G(q_1, ..., q_{j-1}, q_{j+1}, ..., q_n)$$

With this new expression, we can consider maximizing $\mathcal{L}$ with respect to each of the $q_j(z_j)$. The optimal form of $q_j(z_j)$ is the one which maximizes $\mathcal{L}$, that is:

$$\begin{aligned}\arg\max_{q_j} \mathcal{L} &= \arg\max_{q_j} \int_{z_j} q_j(z_j) E_{q_{-j}}[\log p(x, z)]\, dz_j + E_{q_j}[\log q_j(z_j)] + G(q_1, ..., q_{j-1}, q_{j+1}, ..., q_n) \\ &= \arg\max_{q_j} \int_{z_j} q_j(z_j)(E_{q_{-j}}[\log p(x, z)] + \log q_j(z_j))\, dz_j \end{aligned} \tag{5}$$

We take the derivative with respect to $q_j(z_j)$ using Lagrange multipliers λ_j [2] and set to 0 yields:

$$\begin{aligned}\log q_j(z_j) &= E_{q_{-j}}[\log p(x, z)] - 1 - \lambda_j \\ q_j(z_j) &= \frac{exp(E_{q_{-j}}[\log p(x, z)])}{C_j}\end{aligned} \tag{6}$$

where C_j is a normalization constant that plays minimal role in the variable update.

[2] The funtional derivative of this expression actually requires some knowledge about calculus of variations, specifically Euler–Lagrange equation.`https://en.wikipedia.org/wiki/Calculus_of_variations`

3.3 Variable update with Coordinate Ascent

From equation (9) we found that $q_j(z_j) \propto exp(E_{q_{-j}}[\log p(x,z)])$. Therefore iterative optimization algorithms like Coordinate Ascent[3] can be applied to update the latent variables to reach their optimal form. Note that all the $q_j(z_j)$'s are interdependent during the update, hence in each iteration, we need to update all the $q_j(z_j)$'s. As short description for the coordinate ascent in this setup will be:

1. Compute values (if any) that can be directly obtained from data and constants
2. Initialize a particular z_j to an arbitrary value [4]
3. Update each variable with the step function ($\propto exp(E_{q_{-j}}[\log p(x,z)])$)
4. Repeat step 3 until the convergence of ELBO

A more detailed example of coordinate ascent will be shown in section 4 with the univariate gaussian distribution example. A point to take note that in general, we cannot guarantee the convexity of ELBO function. Hence, the convergence is usually to a local maximum.

4 Example with Univariate Gaussian

We demonstrate the mean-field variational inference with a simple case of observations from univariate Gaussian model. We first assume there are N observations $X = (x_1, ...x_N)$ from a Gaussian distribution satisfying:

$$\begin{aligned} &x_i \sim \mathcal{N}(\mu, \tau^{-1}), i = 1, ..., N \\ &where\ \mu \sim \mathcal{N}(\mu_0, (\kappa_0\tau)^{-1})\ \&\ \tau \sim Gamma(a_0, b_0) \end{aligned}$$

Here τ is inverse of variance σ^2 (hence one-to-one correspondence). From the derivation of $q_j(z_j)$ we know we need to compute the log joint probability $\log p(X, \mu, \tau)$. We will first derive an explicit formula for it by expanding the join probability into conditional probability:

$$\begin{aligned} \log p(X,\mu,\tau) =\ & \log p(X|\mu,\tau) + \log p(\mu|\tau) + \log p(\tau) \\ =\ & \log(\sqrt{\frac{\tau^N}{2\pi}} e^{-\frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2}) + \log(\sqrt{\frac{\kappa_0\tau}{2\pi}} e^{-\frac{\kappa_0\tau}{2}(\mu-\mu_0)^2}) + \log(\frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}e^{-b_0\tau}) \\ =\ & \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\log(\kappa_0\tau) - \frac{\kappa_0\tau}{2}(\mu-\mu_0)^2 \\ & + (a_0-1)\log\tau - b_0\tau + C \end{aligned} \tag{7}$$

where C is a constant

[3] https://en.wikipedia.org/wiki/Coordinate_descent

[4] Note that sometimes some latent variable has higher priority that others. The choice of this variable depends on the exact question in hand.

4.1 Compute independent $q_\mu(\mu)$ and $q_\tau(\tau)$

Next, we apply approximation via $p(\mu,\tau|X) \approx q(\mu,\tau)$. By the mean-field assumption, we have $q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$. We proceed to find the optimal form of $q_\mu(\mu)$ and $q_\tau(\tau)$:

- Compute the expression for $q_\mu(\mu)$:

$$\begin{aligned}
\log q_\mu(\mu) &= E_\tau[\log p(X,\mu,\tau)] + C_1 \\
&= E_\tau[\frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\log(\kappa_0\tau) - \frac{\kappa_0\tau}{2}(\mu-\mu_0)^2 + (a_0-1)\log\tau - b_0\tau] + C_2 \\
&= -\frac{E_\tau[\tau]}{2}[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2] + C_3 \\
&\quad \textit{(completing the square for the term inside the square bracket)} \\
&= -\frac{(\kappa_0+N)E_\tau[\tau]}{2}(\mu - \frac{\kappa_0\mu_0 + \sum_{i=1}^{N}x_i}{\kappa_0+N})^2 + C_4 \qquad (8)
\end{aligned}$$

Note that here E_τ is a shortcut representation for $E_{q_\tau(\tau)}$, and all $C_1, ...C_k$ are constant terms not involved in the optimization update. From the expression above, it's easy to observe that $q_\mu(\mu)$ follows a Gaussian distribution with $q_\mu(\mu) \sim \mathcal{N}(\hat{\mu}, \hat{\tau}^{-1})$, where:

$$\begin{aligned}
\hat{\mu} &= \frac{\kappa_0\mu_0 + \sum_{i=1}^{N}x_i}{\kappa_0+N} \\
\hat{\tau} &= (\kappa_0+N)E_\tau[\tau]
\end{aligned}$$

- Compute the expression for $q_\tau(\tau)$:

$$\begin{aligned}
\log q_\tau(\tau) &= E_\mu[\frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\log(\kappa_0\tau) - \frac{\kappa_0\tau}{2}(\mu-\mu_0)^2 + (a_0-1)\log\tau - b_0\tau] + C_5 \\
&= (a_0-1)\log(\tau) - b_0\tau + \frac{1+N}{2}\log(\tau) - \frac{\tau}{2}E_\mu[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2] + C_6 \\
&= (a_0 - 1 + \frac{1+N}{2})\log(\tau) - (b_0 + \frac{1}{2}E_\mu[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2])\tau + C_6 \qquad (9)
\end{aligned}$$

A closer look at the result (9) suggest that $q_\mu(\mu)$ follows a Gaussian distribution with $q_\tau(\tau) \sim Gamma(\hat{a}, \hat{b})$, where:

$$\begin{aligned}
\hat{a} &= a_0 + \frac{1+N}{2} \\
\hat{b} &= b_0 + \frac{1}{2}E_\mu[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2]
\end{aligned}$$

4.2 Variable update until ELBO convergence

Now that we have $q_\mu(\mu) \sim \mathcal{N}(\hat{\mu}, \hat{\tau}^{-1})$ and $q_\tau(\tau) \sim Gamma(\hat{a}, \hat{b})$, we only need to update their parameters:

$$\hat{\mu} = \frac{\kappa_0\mu_0 + \sum_{i=1}^N x_i}{\kappa_0 + N} \tag{10}$$

$$\hat{\tau} = (\kappa_0 + N)E_\tau[\tau] = (\kappa_0 + N)\frac{\hat{a}}{\hat{b}} \tag{11}$$

$$\hat{a} = a_0 + \frac{1+N}{2} \tag{12}$$

$$\hat{b} = b_0 + \frac{1}{2}[\kappa_0(\frac{1}{\hat{\tau}} + \hat{\mu}^2 + \mu_0^2 - 2\hat{\mu}\mu_0) + \sum_{i=1}^N (x_i^2 - 2\hat{\mu}x_i + \frac{1}{\hat{\tau}} + \hat{\mu}^2)] \tag{13}$$

Using the updated $q_\mu(\mu)$ and $q_\tau(\tau)$, we can then compute $ELBO = E_q[\log p(X, \mu, \tau)] - E_q[\log q_\mu(\mu) + \log q_\tau(\tau)]$ with $q = q_\mu q_\tau$. Hence the coordinate ascent algorithm can be applied here:

1. Compute $\hat{\mu}$ and $\hat{a}$ as they can be derived directly from the data and constants based on their formula

2. Initialize $\hat{\tau}$ to some random value

3. Update $\hat{b}$ with current values of $\hat{\mu}$, $\hat{a}$ and $\hat{\tau}$

4. Update $\hat{\tau}$ with current values of $\hat{\mu}$, $\hat{a}$ and $\hat{b}$

5. Compute ELBO value with the variables μ & τ updated with the parameters in step 1 - 4

6. Repeat the last 3 steps until ELBO value doesn't vary by much

As a result of the algorithm, we obtain an approximation $q = q_\mu q_\tau$ for the posterior distribution of μ and τ given observations X.

5 Extension and Further result

In this section, we briefly outline some more theory and reflection about general variational Bayesian methods. Due to space limitations, we only provide a short discussion on each of these.

5.1 Exponential family distributions in Varational Inference

A nice property of the exponential family distribution is the presence of conjugate priors in closed forms. This allows for less computationally intensive approaches when approximating posterior distributions (due to reasons like simpler optimization algorithm applicable and better analytical forms). Further more, Gharamani & Beal even suggested in 2000 that if all the $q_j(z_j)$ belong to the same exponential family, the update of latent variables in the optimization procedure can be exact.

A great achievement in the field of variational inference is the generalized update formula for Exponential-family-conditional models. These models has conditional densities that are in exponential family. The nice property of exponential family leads to an amazing result that the optimal approximation form for posteriors are in the same exponential family as the conditional. This has benefits a lot of well-known models like Markov random field and Factorial Hidden Markov Model.

5.2 Comparison to other Inference methods

The ultimate results of variational inference are the approximation for the entire posterior distribution about the parameters and variables in the target problem with some observations instead of just a single point estimate. This serves the purpose of further statistical study of these latent variables, even if their true distributions are analytically intractable. Another group of inference methods commonly used to achieve the similar aim is Markov chain Monte Carlo (MCMC) methods like Gibbs sampling, which seeks to produce reliable resampling of given observations that help to approximate latent variables well. Another common Bayesian method that has a similar iterative variable update procedure is Expectation Maximization (EM). For EM, however, only point estimates of posterior distribution are obtained. The estimates are "Expectation maximizing" points, which means any information about the distribution around these points (or the parameters they estimate) are not preserved. On the other hand, despite the advantage of "entire distribution" Variational inference has, its point estimates are often derived just by the mean value of the approximated distributions. Such point estimates are often less significant compared to those derived using EM, as the optimum is not directly achieved from the Bayesian network itself, but the optimal distributions inferred from the network.

5.3 Popular algorithms applying variational inference

The popularity of variational inference has grown to even surpass the classical MCMC methods in recent years. It is particularly successful in generative modeling as a replacement for Gibbs sampling. The methods often show better empirical result than Gibbs sampling, and are thus more well-adopted. We here showcase some popular machine learning models and even deep learning models that heavily rely on variational inference methods and achieved great success:

- Latent Dirichlet Allocation: With the underlying Dirichlet distribution, the model applies both variational method (for latent variable distribution) and EM algorithm to obtain an optimal topic separation and categorization.

- variational autoencoder: The latent Gaussian space (a representation for the input with all the latent variables and parameters) is derived from observations, and fine-tuned to generate some convincing counterparts (a copy for instance) of the input.

These models often rely on a mixture of statistical learning theories, but variational inference is definitely one of the key function within them.

Appendix

A Derivation of ELBO expression in terms of $q_j(z_j)$

- By the independence assumption, we can swap the order of integral in the expectation formula:

$$\begin{aligned} E_q[\log p(x,z)] &= \int_{z_1,\ldots,z_n} q(z) \log p(x,z)\, dz_1...dz_n \\ &= \int_{z_1,\ldots,z_n} \prod_{i=1}^{n} q_i(z_i) \log p(x,z)\, dz_1...dz_n \\ &= \int_{z_j} q_j(z_j) \left[\int_{z_{i|i\neq j}} \prod_{i\neq j} q_i(z_i) \log p(x,z)\, dz_1, ..., dz_{j-1}, dz_{j+1}, ..., dz_n \right] dz_j \end{aligned} \tag{14}$$

note that for ease of evaluation, we express the part inside square bracket as an expectation across all variables except j as $E_{q_{-j}}[\log p(x,z)]$. Then:

$$E_q[\log p(x,z)] = \int_{z_j} q_j(z_j) E_{q_{-j}}[\log p(x,z)]\, dz_j$$

- Next, we can decompose $E_q[\log q(z)]$ into:

$$\begin{aligned} E_q[\log p(x,z)] &= \int_{z_1,\ldots,z_n} q(z) \log q(z)\, dz_1...dz_n \\ &= \int_{z_1,\ldots,z_n} \prod_{i=1}^{n} q_i(z_i) \log \left(\prod_{k=1}^{n} q_k(z_k) \right) dz_1...dz_n \\ &= \int_{z_1,\ldots,z_n} \prod_{i=1}^{n} q_i(z_i) \sum_{k=1}^{n} \log q_k(z_k)\, dz_1...dz_n \end{aligned} \tag{15}$$

now, using the same trick as above, and the fact that integral of a probability density function is 1, we obtain

$$\begin{aligned} E_q[\log p(x,z)] &= \int_{z_j} q_j(z_j) \log q_j(z_j) \left[\int_{z_{i|i\neq j}} \prod_{i\neq j} q_i(z_i) \sum_{k=1}^{n} \log q_k(z_k)\, dz_1, ..., dz_{j-1}, dz_{j+1}, ..., dz_n \right] dz_j \\ &= \int_{z_j} q_j(z_j) \log q_j(z_j)\, dz_j \int_{z_{i|i\neq j}} \prod_{i\neq j} q_i(z_i)\, dz_1, ..., dz_{j-1}, dz_{j+1}, ..., dz \\ &\quad + \int_{z_j} q_j(z_j)\, dz_j \int_{z_{i|i\neq j}} \prod_{i\neq j} q_i(z_i) \sum_{k\neq j} \log q_k(z_k)\, dz_1, ..., dz_{j-1}, dz_{j+1}, ..., dz \\ &= \int_{z_j} q_j(z_j) \log q_j(z_j)\, dz_j + \int_{z_{i|i\neq j}} \prod_{i\neq j} q_i(z_i) \sum_{k\neq j} \log q_k(z_k)\, dz_1, ..., dz_{j-1}, dz_{j+1}, ..., dz \\ &= E_{q_j}[\log q_j(z_j)] + G(q_1, ..., q_{j-1}, q_{j+1}, ..., q_n) \end{aligned} \tag{16}$$

where G is a functional of latent variables except j.

- Therefore, for any particular q_j, the ELBO can be expressed as

$$\mathcal{L} = \int_{z_j} q_j(z_j) E_{q_{-j}}[\log p(x,z)]\, dz_j + E_{q_j}[\log q_j(z_j)] + G(q_1, ..., q_{j-1}, q_{j+1}, ..., q_n)$$