Notes on Variational Lower Bound for Some Models

Anand K Subramanian[*]

Contents

1 Variational Lower Bound

Variational Inference (VI) is an approximate Bayesian technique that transforms Bayesian inference as an optimization problem. The motive behind VI is to approximate the often intractable posterior distribution using a parameterized variational distribution by minimizing the Kullback-Liebler divergence. The variational lower bound is given by

$$\mathcal{L} = \mathbb{E}_{q(z|\lambda)}\left[\log \frac{p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] \tag{1}$$

Where $p(\mathbf{y}, \mathbf{z})$ is the joint distribution between the output $\mathbf{y}$ and the latent variables $\mathbf{z}$ with some parameter $\boldsymbol{\theta}$; and $q(\mathbf{z}|\boldsymbol{\lambda})$ is the variational distribution with variational parameters $\boldsymbol{\lambda}$.

In classical VI, it is required that the above expectation must be tractable; implying that the likelihood must be conjugate to the prior distribution and the variational distribution must be able to factorized across the latent variables (mean-field assumption). The first problem has been addressed by the BlackBox VI (BBVI) algorithm [6] where the gradient of the expectation is computed via monte-carlo methods. In this note, we consider the variational lower bounds for some general classes of models and some methods to solve them.

1.1 Variational Lower Bound for Latent Conjugate Models

Consider the joint distribution of a latent conjugate model as

$$p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) = \left[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\right] p(\mathbf{z}|\boldsymbol{\theta}) \tag{2}$$

Where $\mathbf{z}$ is the latent function modelling the data as $z_i = f(\mathbf{x}_i)$ and the likelihood distributions $p(\mathbf{y}_n|\mathbf{z}_n)$ are conjugate to the prior $p(\mathbf{z}|\boldsymbol{\theta})$. A common example of such models is the Gaussian Process Regression model; where the likelihood and the prior are both Gaussians. Note that the prior distribution in case of GPs is written as $p(\mathbf{z}|\mathbf{0}, \mathbf{K})$, where the zero mean-function and the covariance function $\mathbf{K}$ are chosen beforehand. Now, the goal is to use a variational

[*] Work done at RIKEN AIP, Tokyo

distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ to approximate the posterior distribution by constructing the variational lower bound. Here $\boldsymbol{\lambda}$ are the natural parameters of the variational distribution. For this, we make the assumption that the variational distribution is of the same family as that of the prior distribution in the above equation.

In this case, the best variational distribution is essentially the true posterior distribution as the above model has a closed-form posterior, given as

$$q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta})) = \frac{1}{Z(\boldsymbol{\theta})}\left[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta}) \tag{3}$$

Where the normalizing term $Z(\boldsymbol{\theta})$ is given by

$$Z(\boldsymbol{\theta}) = \int\left[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta})d\mathbf{z} = p(\mathbf{y}|\boldsymbol{\theta}) \tag{4}$$

Which is the marginal likelihood of the dataset. With this variational distribution, the variational lower bound is given by

$$\begin{align}
\mathcal{L} &= \mathbb{E}_{q(z|\lambda)}\left[\log\frac{p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta}))}\right] \tag{5}\\
&= \int q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta}))\log\left[\frac{p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta}))}\right]d\mathbf{z} \tag{6}\\
&= \int q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta}))\log\left[\frac{\left[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta})}{\left[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta})}Z(\boldsymbol{\theta})\right]d\mathbf{z} \tag{7}\\
&= \int q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta}))\log Z(\boldsymbol{\theta})d\mathbf{z} \tag{8}\\
&= \log Z(\boldsymbol{\theta})\int q(\mathbf{z}|\boldsymbol{\lambda}(\boldsymbol{\theta}))d\mathbf{z} \tag{9}\\
&= \log p(\mathbf{y}|\boldsymbol{\theta}) \tag{10}
\end{align}$$

Thus, the variational inference essentially becomes maximizing the log marginal likelihood of the model, which is precisely what Gaussian process regression does.

1.2 Variational Lower Bound for Subset-of-Data Latent Conjugate Models

In this case, we consider a subset-of-data model where the training set for the model consists only a small subset $\mathcal{I} \in \mathcal{D}$ sampled from the original dataset $\mathcal{D}$. The Original dataset $\mathcal{D}$ contains N data points and the subset contains M data points where $M < N$. Making the same assumptions as before, the variational distribution is now given by

$$q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) = \frac{1}{Z_{\mathcal{I}}(\boldsymbol{\theta})}\left[\prod_{n\in\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta}) \tag{11}$$

Where the normalizing constant $Z_{\mathcal{I}}(\boldsymbol{\theta})$ can be derived as follows

$$\begin{align}
Z_{\mathcal{I}}(\boldsymbol{\theta}) &= \int\left[\prod_{n\in\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta})d\mathbf{z} \tag{12}\\
&= \int\int\left[\prod_{n\in\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\theta})p(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\theta})d\mathbf{z}_{-\mathcal{I}}d\mathbf{z}_{\mathcal{I}} \tag{13}\\
&= \int\left[\prod_{n\in\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\theta})d\mathbf{z}_{\mathcal{I}}\int p(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\theta})d\mathbf{z}_{-\mathcal{I}} \tag{14}\\
&= \int\left[\prod_{n\in\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\theta})d\mathbf{z}_{\mathcal{I}} \tag{15}\\
&= p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) \tag{16}
\end{align}$$

Which is the marginal likelihood of the subset of data. Here, we note that the notation $\mathbf{z}_{\mathcal{I}}$ refers to the latent components derived from the subset $\mathcal{I}$ and $\mathbf{z}_{-\mathcal{I}}$ represents the rest of the latent components. The variational lower bound can, similarly, be obtained as follows.

$$\mathcal{L} = \mathbb{E}_{q(z|\lambda_{\mathcal{I}})}\left[\log \frac{p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta}))}\right] \tag{17}$$

$$= \int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\frac{p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta}))}\right] d\mathbf{z} \tag{18}$$

$$= \int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\frac{[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)]p(\mathbf{z}|\boldsymbol{\theta})}{[\prod_{n\in\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)]p(\mathbf{z}|\boldsymbol{\theta})} Z_{\mathcal{I}}(\boldsymbol{\theta})\right] d\mathbf{z} \tag{19}$$

$$= \int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[[\prod_{n\notin\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)] Z_{\mathcal{I}}(\boldsymbol{\theta})\right] d\mathbf{z} \tag{20}$$

$$= \underbrace{\int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) d\mathbf{z}}_{\text{Term 1}} + \underbrace{\int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\prod_{n\notin\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z}}_{\text{Term 2}} \tag{21}$$

Term 1 can can be simplified as follows -

$$\int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) d\mathbf{z} = \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) \int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) d\mathbf{z} \tag{22}$$

$$= \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) \tag{23}$$

Term 2 can be derived as follows -

$$\int q(\mathbf{z}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\prod_{n\notin\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z} = \int\int q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\prod_{n\notin\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z}_{\mathcal{I}} d\mathbf{z}_{-\mathcal{I}} \tag{24}$$

$$= \int q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) d\mathbf{z}_{\mathcal{I}} \int q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\prod_{n\notin\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z}_{-\mathcal{I}} \tag{25}$$

$$= \int q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta})) \log\left[\prod_{n\notin\mathcal{I}} p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z}_{-\mathcal{I}} \tag{26}$$

$$= \sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta}))} \log\left[p(\mathbf{y}_n|\mathbf{z}_n)\right] \tag{27}$$

Therefore, the variational lower bound can be be written as

$$\mathcal{L} = \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) + \sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta}))} \log\left[p(\mathbf{y}_n|\mathbf{z}_n)\right] \tag{28}$$

The above equation indicates that the lower bound is simply the sum of the marginal likelihood of the model plus the expected log likelihood with respect to the subset-of-data posterior.

1.2.1 Gradients of the Variational Lower Bound

In this section, we shall derive the gradients of the above lower bound with respect to the variational parameters.

$$\nabla_{\boldsymbol{\lambda}_{\mathcal{I}}} \mathcal{L} = \nabla_{\boldsymbol{\lambda}_{\mathcal{I}}} \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) + \nabla_{\boldsymbol{\lambda}_{\mathcal{I}}} \sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}_{\mathcal{I}}(\boldsymbol{\theta}))} \log\left[p(\mathbf{y}_n|\mathbf{z}_n)\right] \tag{29}$$

Here $\boldsymbol{\lambda}_{\mathcal{I}}$ represents the set of all parameters in the variational distribution, i.e. prior and the likelihood parameters. For instance. in case of Gaussian Process regression with Gaussian likelihoods, $\boldsymbol{\lambda}_{\mathcal{I}} = \{\boldsymbol{\theta}, \sigma_n^2\}$ (assuming zero mean), where $\boldsymbol{\theta}$ is the set of parameters of the kernel $\mathbf{K}$ and σ_n^2 is the likelihood variance.

1.2.2 Equivalence to Hensman's Lower Bound

The variational lower bound given in Hensman et.al [2] for Gaussian Processes is given by

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{p(\mathbf{z}|\mathbf{z}_{\mathcal{I}})q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] - \mathbb{D}_{KL}[q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})\|p(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\theta})] \tag{30}$$

The above lower bound can be split as follows:

$$\mathcal{L} = \underbrace{\sum_{n\notin\mathcal{I}} \mathbb{E}_{p(\mathbf{z}|\mathbf{z}_{\mathcal{I}})q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right]}_{\text{Term 1}} + \underbrace{\sum_{n\in\mathcal{I}} \mathbb{E}_{p(\mathbf{z}|\mathbf{z}_{\mathcal{I}})q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right]}_{\text{Term 2}} - \mathbb{D}_{KL}[q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})\|p(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\theta})] \tag{31}$$

Note that we can combine the expectation conditionals as $p(\mathbf{z}|\mathbf{z}_{\mathcal{I}})q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda}) = q(\mathbf{z}|\boldsymbol{\lambda})$; i.e. the prediction over all latent variables given the learnt parameters over the subset of data. This is essentially equal to the variational distribution given in equation (11). From this, we can simply the above expectations in the following manner. Term 1 in the above equation can be simplified as

$$\begin{aligned}
\sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] &= \sum_{n\notin\mathcal{I}} \int q(\mathbf{z}|\boldsymbol{\lambda}) \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z} && (32)\\
&= \sum_{n\notin\mathcal{I}} \int\int q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda})q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda}) \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z}_{\mathcal{I}} d\mathbf{z}_{-\mathcal{I}} && (33)\\
&= \sum_{n\notin\mathcal{I}} \int q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda}) \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] d\mathbf{z}_{-\mathcal{I}} \int q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda}) d\mathbf{z}_{\mathcal{I}} && (34)\\
&= \sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] && (35)
\end{aligned}$$

Similarly, The term 2 in the above equation can be simplified as

$$\sum_{n\in\mathcal{I}} \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] = \sum_{n\in\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] \tag{36}$$

Putting it all together, the Hensman's bound in equation (30) can be rewritten as

$$\mathcal{L} = \sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] + \underbrace{\sum_{n\in\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] - \mathbb{D}_{KL}[q(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\lambda})\|p(\mathbf{z}_{\mathcal{I}}|\boldsymbol{\theta})]}_{=\log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta})} \tag{37}$$

Observe that the last two terms in the above equation is essentially the variational lower bound for a model over only the subset of data. This is equivalent to the model discussed in the previous section where the latent conjugate model where the joint is defined on the subset $\mathcal{I}$ and the variational lower bound is essentially the log marginal likelihood (Refer equation (10)). Therefore, the Hensnman's lower bound can be written as

$$\mathcal{L} = \sum_{n\notin\mathcal{I}} \mathbb{E}_{q(\mathbf{z}_{-\mathcal{I}}|\boldsymbol{\lambda})} \log \left[p(\mathbf{y}_n|\mathbf{z}_n)\right] + \log p(\mathbf{y}_{\mathcal{I}}|\boldsymbol{\theta}) \tag{38}$$

Comparing the above equation with the subset of data lower bound in equation (29), we show that they are equivalent.

1.3 Variational Lower Bound for Latent Non-Conjugate Models

For non-conjugate latent models, the joint distribution is not a conjugate to the prior distribution and therefore, no closed form exists for the posterior. An example of such a model is the Gaussian process classification, where the prior is a Gaussian and but the likelihood is usually a Bernoulli or Multinoulli (categorical) distribution. For the case of Gaussian processes classification, there are various methods often used in practice such as Laplace approximation of the posterior or the Expectation Propagation (EP) algorithm. But a more generic method would be to use the Stochastic Variational Inference (SVI) method where the non-conjugate likelihood distribution is approximated by a Gaussian distribution $q(\mathbf{z}|\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With this, the variational lower bound can be written as follows.

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] && (39)\\
&= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{\left[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\right]p(\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] && (40)\\
&= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\log\Big[\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n)\Big] + \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p(\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] && (41)\\
&= \sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p(\mathbf{y}_n|\mathbf{z}_n)\big] - \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{q(\mathbf{z}|\boldsymbol{\lambda})}{p(\mathbf{z}|\boldsymbol{\theta})}\right] && (42)\\
&= \sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p(\mathbf{y}_n|\mathbf{z}_n)\big] - \mathbb{D}_{KL}\big[\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\|p(\mathbf{z}|\boldsymbol{\theta})\big] && (43)
\end{aligned}$$

Where $\mathbb{D}_{KL}$ is the Kullback-Liebler divergence. The above expression is generally intractable for non-Gaussian likelihoods and even for non-Gaussian priors. However, optimization methods like SGD can be used for optimizing such lower bounds where the stochastic gradient can be computed using monte-carlo methods. These are referred under the umbrella term stochastic variational inference.

1.4 Variational Lower Bound for Latent Non-Conjugate Models with Conjugate Components

A special case of the latent non-conjugate models is the one with *some* conjugate components. In the likelihood distribution, some terms in the product belong to the conjugate family with respect to the prior distribution. The naive usage of the SVI for these kinds of models might prove inefficient as they don't leverage the advantage of such conjugate terms. Thus, along these lines, we construct a lower bound that can provide a way to leverage these conjugate terms.

Consider the decomposition of joint distribution as $p(\mathbf{y},\mathbf{z}) \propto p_{NC}(\mathbf{y},\mathbf{z})p_C(\mathbf{y},\mathbf{z})$ where p_C, p_{NC} are the conjugate and non-conjugate components of the joint distribution respectively. The proportional symbol indicates that these can be unnormalized in terms of $\mathbf{z}$. Here, we emphasize that by conjugate terms, we mean that $p_C(\mathbf{y},\mathbf{z})$ belongs to the exponential family. This follows from the fact that the conjugate distribution of any exponential-family distribution, belongs to the exponential family. Now, we derive the lower bound as follows.

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] && (44)\\
&= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p_{NC}(\mathbf{y},\mathbf{z})p_C(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] + c && (45)\\
&= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] + \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{p_C(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] + c && (46)\\
&= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] - \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log \frac{q(\mathbf{z}|\boldsymbol{\lambda})}{p_C(\mathbf{y},\mathbf{z})}\right] + c && (47)\\
&= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] - \mathbb{D}_{KL}\big[q(\mathbf{z}|\boldsymbol{\lambda})\|p_C(\mathbf{y},\mathbf{z})\big] + c && (48)
\end{aligned}$$

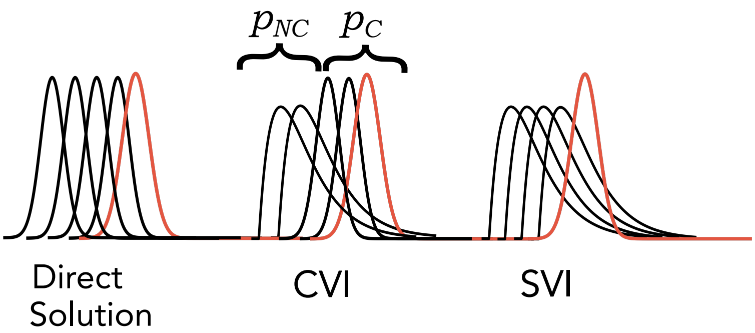

Figure 1: Illustration of methods used for various latent models. The likelihood distributions are given in black and the prior is given in red. *(from left to right)* If the likelihood is conjugate to the prior, then a direct closed-form solution is possible. In case of models with some conjugate components, the joint can be split into conjugate and non-conjugate components as shown. the CVI algorithm solves these niche problems. In case of complete non-conjugate models, the only solution is to use the SVI algorithm using monte-carlo estimators for computing the gradients.

Where c is a constant term independent of the variational parameters. Now, the above lower bound can be made simpler in two ways - 1) Making the variational distribution of the same family as that of the conjugate part, thereby forming a closed-form solution for the second term; 2) Borrowing the idea from the latent conjugate models - making the variational distribution proportional to the closed-form posterior of the conjugate part $p_C(\mathbf{y}, \mathbf{z})$ (Similar to equation (3)). Other terms can be added to approximate the non-conjugate part. This ensures certain optimality to the final solution, based on the results for the latent conjugate models. In either case, we can obtain a closed-form expression for the gradient of the second term and we only have to compute the gradient of the first term using other techniques.

1.4.1 Conjugate Computation Variational Inference (CVI)

In this section, we discuss reducing the number of free parameters while performing gradient descent on the above lower bound. Note that the choice and parameteriztion of the variational distribution can have a huge impact on the efficiency and convergence of the optimization method. As discussed in the previous paragraph, the CVI method [3] chooses the variational distribution to be proportional to the conjugate part as

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto e^{\langle \phi(\mathbf{z}), \tilde{\boldsymbol{\lambda}}_t \rangle} p_C(\mathbf{y}, \mathbf{z}) \tag{49}$$

Simply put, the variational distribution is simply a product of the conjugate terms and an exponential family which approximates the non-conjugate terms. The natural parameters of the above variational distribution are the natural parameters of the exponential family $\tilde{\boldsymbol{\lambda}}_t$ at time t and the natural parameters of the conjugate terms $\boldsymbol{\eta}$.

With the above variational distribution, we can try to derive the gradient of the lower bound given in equation (46). To improve upon the traditional SGD, CVI uses natural gradients which require lower memory and computations than the former. They also exploit the information geometry of the model for better convergence. To provide a simple contrast between SGD and NGD (natural gradient descent), consider the usual SGD step written as

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \alpha_t \hat{\nabla}_\lambda \mathcal{L} \tag{50}$$

Where $\mathcal{L}$ can be any loss function or the lower bound, discussed earlier. The $\hat{\nabla}$ signifies that the gradient is stochastic. Note that the above update is in natural parameter space of the variational distribution. Now, the NDG update is given by

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \beta_t \mathbf{F}(\boldsymbol{\lambda})^{-1} \hat{\nabla}_\lambda \mathcal{L} \tag{51}$$

Where $\mathbf{F}(\boldsymbol{\lambda})$ is the Fisher information matrix with respect to the natural parameters $\boldsymbol{\lambda}$. The inverse Fisher information matrix actually captures the local Riemannian curvature of the optimization landscape and scales the local gradient accordingly. This is similar to the Newton's method where the gradients are the scaled by the inverse Hessian. This however is an approximation because, the inverse Hessian is a quadratic approximation of the local curvature. Fisher information, on the other hand, captures the exact geometry and therefore converges faster. This gradient scaled by the inverse Fisher information is called as *natural gradient.*

The Fisher information is difficult to compute and invert, making the NGD highly undesirable. However, for exponential family, there exists beautiful relation between the natural gradients with respect to the natural parameters and the gradients with respect to the expectation $\mathbf{m}$, as follows.

$$\mathbf{F}(\boldsymbol{\lambda})^{-1}\hat{\nabla}_{\lambda}\mathcal{L} = \hat{\nabla}_{\mathbf{m}}\mathcal{L} \tag{52}$$

$$\mathbf{F}(\mathbf{m})^{-1}\hat{\nabla}_{\mathbf{m}}\mathcal{L} = \hat{\nabla}_{\lambda}\mathcal{L} \tag{53}$$

The natural gradients with respect to the natural parameters of an exponential family is equal to the gradients with respect to its expectation (mean) parameters and vice versa [7], [4]. Or simply put, $\mathbf{F}^{-1}(\boldsymbol{\lambda}) = \mathbf{F}(\mathbf{m})$. Using equation (52), the NGD update in equation (51), for the exponential family, can be written as

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \beta_t\hat{\nabla}_{\mathbf{m}}\mathcal{L} \tag{54}$$

Now, the task is to simply compute the gradients of the lower bound in equation (46) with respect to the expectation parameters, since our variational distribution belongs to the exponential family.

$$\begin{aligned}
\hat{\nabla}_{\mathbf{m}}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log\frac{p_C(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] &= \hat{\nabla}_{\mathbf{m}}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\langle\phi(\mathbf{z}),\boldsymbol{\eta}-\boldsymbol{\lambda}\rangle + A(\boldsymbol{\lambda})\big] && (55)\\
&= \hat{\nabla}_{\mathbf{m}}\big[\mathbf{m}^T(\boldsymbol{\eta}-\boldsymbol{\lambda}) + A(\boldsymbol{\lambda})\big] && (56)\\
&= (\boldsymbol{\eta}-\boldsymbol{\lambda}) - \mathbf{m}^T\hat{\nabla}_{\mathbf{m}}\boldsymbol{\lambda} + \hat{\nabla}_{\mathbf{m}}A(\boldsymbol{\lambda}) && (57)\\
&= (\boldsymbol{\eta}-\boldsymbol{\lambda}) - \mathbf{m}^T\hat{\nabla}_{\mathbf{m}}\boldsymbol{\lambda} + \hat{\nabla}_{\boldsymbol{\lambda}}A(\boldsymbol{\lambda})\hat{\nabla}_{\mathbf{m}}\boldsymbol{\lambda} && (58)\\
&= (\boldsymbol{\eta}-\boldsymbol{\lambda}) - \mathbf{m}^T\hat{\nabla}_{\mathbf{m}}\boldsymbol{\lambda} + \mathbf{m}^T\hat{\nabla}_{\mathbf{m}}\boldsymbol{\lambda} && (59)\\
&= \boldsymbol{\eta}-\boldsymbol{\lambda} && (60)
\end{aligned}$$

Where $\phi(\mathbf{z})$ is the sufficient statistic of the variational distribution, $\boldsymbol{\eta}$ are the natural parameters of the conjugate distribution and $A(\boldsymbol{\lambda})$ is the log-partition function; we have used the fact that $\mathbf{m} = \mathbb{E}[\phi(\mathbf{z})]$ in equation (56) and $\mathbf{m} = \nabla_{\boldsymbol{\lambda}}A(\boldsymbol{\lambda})$ in equation (59).

Now, putting all of it together, the NGD update for CVI can be written as

$$\begin{aligned}
\boldsymbol{\lambda}_{t+1} &= \boldsymbol{\lambda}_t + \beta_t\hat{\nabla}_{\mathbf{m}}\mathcal{L} && (61)\\
&= \boldsymbol{\lambda}_t + \beta_t\hat{\nabla}_{\mathbf{m}}\left[\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] + \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log\frac{p_C(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right] + c\right] && (62)\\
&= \boldsymbol{\lambda}_t + \beta_t\left[\hat{\nabla}_{\mathbf{m}}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] + \hat{\nabla}_{\mathbf{m}}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\left[\log\frac{p_C(\mathbf{y},\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\lambda})}\right]\right] && (63)\\
&= \boldsymbol{\lambda}_t + \beta_t\left[\hat{\nabla}_{\mathbf{m}}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] + \boldsymbol{\eta} - \boldsymbol{\lambda}_t\right] && (64)\\
&= (1-\beta_t)\boldsymbol{\lambda}_t + \beta_t\big[\hat{\nabla}_{\mathbf{m}}\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\lambda})}\big[\log p_{NC}(\mathbf{y},\mathbf{z})\big] + \beta_t\boldsymbol{\eta} && (65)
\end{aligned}$$

Thus CVI provides a neat formulation of the NGD update for latent non-conjugate models with some conjugate components.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[2] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

[3] Mohammad Emtiyaz Khan and Wu Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *arXiv preprint arXiv:1703.04265*, 2017.

[4] Mohammad Emtiyaz Khan and Didrik Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. *arXiv preprint arXiv:1807.04489*, 2018.

[5] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

[6] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[7] Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

[8] Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural gradients in practice: Non-conjugate variational inference in gaussian process models. *arXiv preprint arXiv:1803.09151*, 2018.