Paper Summary: Auto-Encoding Variational Bayes

Sergio Charles
Department of Mathematics
Stanford University
sergioc1@stanford.edu

June 2023

1 Introduction

We present a summary of *Auto-Encoding Variatioanl Bayes* due to Kingma and Welling [1]. This presents a framework for efficiently approximating inference and learning with directed probabilistic models whose continuous latent variables have intractable posterior distributions. The variational Bayesian (VB) arpproach uses optimization of an approximation to the posterior. However, the standard mean-field technique requires analytical solutions to an expectation with respect to an approximate posterior.

This paper shows how a reparameterization of the variational lower bound gives a differentiable unbiased estimator of the lower bound, which is termed *Stochastic Gradient Variational Bayes* (SGVB). The SGVB estimator can be leveraged for approximate posterior inference in probabilistic models with continuous latent parameters.

2 Method

2.1 Problem Setup

Let $X = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ be a dataset of i.i.d. samples for discrete variable x. The data is generated by a random process which involves an unobserved continuous random variable z. This process is as follows:

1. A quantity $\mathbf{z}^{(i)}$ is sampled from a prior distribution $p_{\theta^*}(\mathbf{z})$.
2. A quantity $\mathbf{x}^{(i)}$ is sampled from a conditional distribution $p_{\theta^*}(\mathbf{x}|\mathbf{z})$.

The prior $p_{\theta^*}(\mathbf{z})$ and likelihood $p_{\theta^*}(\mathbf{x}|\mathbf{z})$ come from parametric families of distributions $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{x}|\mathbf{z})$, respectively. Moreover, their PDFs are differentiable almost everywhere. The true parameters θ^* annd latent variables $\mathbf{z}^{(i)}$ are unknown.

We are interested in efficient approximation and inference of maximum likelihood (ML) or maximum a posteriori (MAP) estimation of the global parameters in the setting where the following conditions hold:

1. **Intractability**: the integral of the marginal

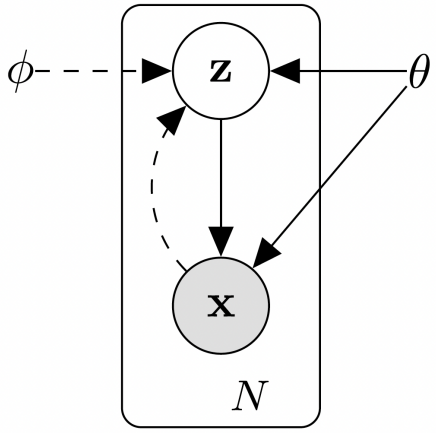
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z} \tag{1}$$

   is not tractable, i.e. the posterior $p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) / p_\theta(\mathbf{x})$ is intractable. This is common for complex likelihood functions, e.g. $p_\theta(\mathbf{x}|\mathbf{z})$ is a neural network.

2. **Large dataset**: Batch optimization is too costly due to large amount of data. Instead, we want to make parameter updates via small minibatches.

As such, we define a **recognition model** $q_\phi(\mathbf{z}|\mathbf{x})$, which approximates true posterior $p_\theta(\mathbf{x}|\mathbf{z})$. The unobserved variables $\mathbf{z}$ are latent representations or *codes*. The recognition model is a probabilistic encoder, since it produces a probability distribution over possible $\mathbf{z}$ from which $\mathbf{x}$ was generated. On the other hand, the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ is a probabilistic decoder, i.e. produces distribution over possible $\mathbf{x}$ given an unobserved $\mathbf{z}$. The probabilistic graphical model representing this problem is shown in Figure 1.

Figure 1: Solid lines represent the generative model $p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})$. Dashed lines represent variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ to true posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The variational parameters ϕ are jointly learned with generative model parameters θ.



2.2 Variational Bound

The marginal likelihood is a sum over marginal likelihood of the i.i.d. samples:

$$\log p_\theta(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}) = \sum_{i=1}^{N} \log p_\theta(\mathbf{x}^{(i)}). \tag{2}$$

Hence consider the follow KL divergence:

$$
\begin{aligned}
&D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z}|\mathbf{x}^{(i)})) \\
&= \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}{p_\theta(z|\mathbf{x}^{(i)})} d\mathbf{z} \\
&= -\int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \log \frac{p_\theta(\mathbf{z}|\mathbf{x}^{(i)})}{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})} d\mathbf{z} \\
&- \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \log \frac{p_\theta(\mathbf{z}, \mathbf{x}^{(i)})}{p_\theta(\mathbf{x}^{(i)}) q_\phi(\mathbf{z}|\mathbf{x}^{(i)})} d\mathbf{z} \\
&= -\left(\int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \log \frac{p_\theta(\mathbf{z}, \mathbf{x}^{(i)})}{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})} d\mathbf{z} - \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \log p_\theta(\mathbf{x}^{(i)}) dz \right) \\
&= \log p_\theta(\mathbf{x}^{(i)}) \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) d\mathbf{z} - \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \log \frac{p_\theta(\mathbf{z}, \mathbf{x}^{(i)})}{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})} d\mathbf{z} \\
&= \log p_\theta(\mathbf{x}^{(i)}) + \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})
\end{aligned} \tag{3}
$$

That is,

$$
\log p_\theta(\mathbf{x}^{(i)}) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}), \tag{4}
$$

where the first term on the right-hand side is the KL divergence of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$ from the true posterior $p_\theta(\mathbf{z}|\mathbf{x}^{(i)})$. The second term $\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ is called the **variational lower bound** on the marginal likelihood of datapoint i. Therefore,

$$
\log p_\theta(\mathbf{x}^{(i)}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\theta(\mathbf{x}, \mathbf{z})], \tag{5}
$$

where

$$
\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(z)) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}, \mathbf{z})\right]. \tag{6}
$$

Here, $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}, \mathbf{z})\right]$ on the right-hand side is referred to as the **reconstruction error**. We need to optimize the lower bound $\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ with respect to variational parameters ϕ and generative parameters θ. However, computing the gradient with respect to ϕ is non-trivial. In particular, we use a Monte Carlo gradient estimate:

$$
\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})] &= \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z}) \nabla_{q_\phi(\mathbf{z})} \log q_\phi(\mathbf{z})] \\
&\approx \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}) \nabla_{q_\phi(\mathbf{z}^{(l)})} \log q_\phi(\mathbf{z}^{(l)})
\end{aligned} \tag{7}
$$

where $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$. However, this is a high variance and unstable method.

2.3 SGVB estimator

We can reparameterize the random variable $\tilde{\mathbf{z}} \sim q_\phi(\mathbf{z}|\mathbf{x})$ using a differentiable transformation $g_\phi(\epsilon, \mathbf{x})$ of an auxiliary noise variable ϵ:

$$
\tilde{\mathbf{z}} = g_\phi(\epsilon, \mathbf{x}) \text{ where } \epsilon \sim p(\epsilon). \tag{8}
$$

We estimate the expectation of $f(\mathbf{z})$ with respect to recognition model $q_\phi(\mathbf{z}|\mathbf{x})$ using Monte Carlo estimates:

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[f(\mathbf{z})] &= \mathbb{E}_{p(\epsilon)}[f(g_\phi(\epsilon, \mathbf{x}^{(i)}))] \\ &\approx \frac{1}{L}\sum_{i=1}^{L} f(g_\phi(\epsilon^{(l)}, \mathbf{x}^{(i)}))\end{aligned} \tag{9}$$

where $\epsilon^{(l)} \sim p(\epsilon)$. Invoking the variational lower bound in Equation 5, we obtain the **Stochastic Gradient Variational Bayes estimator** $\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) \approx \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$:

$$\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}^{(i)}, z^{(i,l)}) - \log q_\phi(z^{(i,l)}|\mathbf{x}^{(i)}) \tag{10}$$

such that $z^{(i,l)} = g_\phi(\epsilon^{(i,l)}, \mathbf{x}^{(i)})$ and $\epsilon^{(l)} \sim p(\epsilon)$.

2.4 Integration of $-D_{KL}(q_\phi(z)||p_\theta(z))$, Gaussian

We give an example where $-D_{KL}(q_\phi(\mathbf{z})||p_\theta(\mathbf{z}))$ in Equation 6 can be integrated analytically. Suppose $p_\theta(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ and the recognition model $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$ are Gaussian. Let $\mathbf{z} \in \mathbb{R}^d$. Likewise, let μ and σ denote the variational mean and standard deviation, respectively. Then

$$\begin{aligned}\int q_\phi(z) \log p(z) dz &= \int \mathcal{N}(\mathbf{z}; \mu, \sigma^2) \log \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) d\mathbf{z} \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}(\mu_i^2 + \sigma_i^2)\end{aligned} \tag{11}$$

and

$$\begin{aligned}\int q_\phi(\mathbf{z}) \log q_\phi(\mathbf{z}) d\mathbf{z} &= \int \mathcal{N}(\mathbf{z}; \mu, \sigma^2) \log \mathcal{N}(\mathbf{z}; \mu, \sigma^2) d\mathbf{z} \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}(1 + \log \sigma_i^2).\end{aligned} \tag{12}$$

Hence,

$$\begin{aligned}-D_{KL}(q_\phi(z)||p_\theta(z)) &= \int q_\phi(z)(\log p_\theta(z) - \log q_\phi(z)) dz \\ &= \frac{1}{2}\sum_{i=1}^{N}(1 - \mu_i^2 - \sigma_i^2 \log(\sigma_i^2))\end{aligned} \tag{13}$$

2.5 Remark on SGVB estimator

Since we can sometimes find an analytic solution for the KL-divergence $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(z))$ in Equation 6, whereby only the reconstruction error $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}, \mathbf{z})]$ requires estimation via sampling. As such, the KL-divergence term is a regularizer of ϕ, forcing

the recognition model to be closer to the prior. Thus, assuming analytic solution, we get another SGVB estimator $\tilde{\mathcal{L}}^B(\theta, \phi; \mathbf{x}^{(i)}) \approx \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$, which has less variance than the $\mathcal{L}^A$ estimator:

$$\mathcal{L}^B(\theta, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}^{(i)}|z^{(i,l)}) \tag{14}$$

where $z^{(i,l)} = g_\phi(\epsilon^{(i,l)}, x^{(i)})$ and $\epsilon^{(l)} \sim p(\epsilon)$.

Analogous to auto-encoders, Equation 14 has a first term that behaves as a regularizer and second term that is the expected negative reconstruction error. In particular, $g_\phi(\cdot)$ maps a datapoint $\mathbf{x}^{(i)}$ and random noise $\epsilon^{(l)}$ to a sample from the recognition model $z^{(i,l)} = g_\phi(\epsilon^{(l)}, \mathbf{x}^{(i)})$ where $z^{(i,l)} \sim q_\phi(z|\mathbf{x}^{(i)})$. Then we form $\log p_\theta(\mathbf{x}^{(i)}|z^{(i,l)})$ which is the probability density of $\mathbf{x}^{(i)}$ under generative model, given $z^{(i,l)}$, i.e. the negative reconstruction error.

2.6 The Reparameterization Trick

Let $\mathbf{z}$ be a continuous random variable and $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ be a conditional distribution. One can express the random variable $\mathbf{z}$ as a determinsitic variable $\mathbf{z} = g_\phi(\epsilon, \mathbf{x})$ where $g_\phi(\cdot)$ is a vector-valued function and ϵ is an auxiliary variable with independent marginal $p(\epsilon)$. Hence, one case rewrite the expectation with respect to $q_\phi(\mathbf{z}|\mathbf{x})$ such that the Monte Carlo estimate is differentiable with respect to ϕ. Denote by $d\mathbf{z} = \prod_i dz_i$ the infinitesimal of an n-dimensional vector. For a deterministic function $\mathbf{z} = g_\phi(\epsilon, \mathbf{x})$, it follows that $q_\phi(\mathbf{z}|\mathbf{x}) \prod_i dz_i = p(\epsilon) \prod_i d\epsilon_i$. Hence,

$$\begin{aligned} \int q_\phi(\mathbf{z}|\mathbf{x}) f(\mathbf{z}) d\mathbf{z} &= \int p(\epsilon) f(\mathbf{z}) d\epsilon \\ &= \int p(\epsilon) f(g_\phi(\epsilon, \mathbf{x})) d\epsilon. \end{aligned} \tag{15}$$

Thus, we obtain a differentiable estimator:

$$p(\epsilon) f(g_\phi(\epsilon, \mathbf{x})) d\epsilon = \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(\mathbf{x}, \epsilon^{(l)})) \tag{16}$$

such that $\epsilon^{(l)} \sim p(\epsilon)$.

For instance, if $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu, \sigma^2)$, then we can define $\mathbf{z} = \mu + \sigma\epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$. Thus,

$$\begin{aligned} \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mu,\sigma^2)}[f(\mathbf{z})] &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}[f(\mu + \sigma\epsilon)] \\ &= \frac{1}{L}\sum_{l=1}^{L} f(\mu + \sigma\epsilon^{(l)}) \end{aligned} \tag{17}$$

where $\epsilon^{(l)} \sim \mathcal{N}(0, 1)$.

3 Variational Auto-Encoder

Suppose we use a neural network for the probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$. Let the prior over latent variables be a standard multivariate Gaussian $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. Let $p_\theta(\mathbf{x}|\mathbf{z})$ be a multivariate Gaussian whose distribution parameters, given $\mathbf{z}$ are the output of a full-connected neural network. Suppose the variational approximate posterior is a multivariate Gaussian with diagonal covariance:

$$\log q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) = \log \mathcal{N}(\mathbf{z}; \mu^{(i)}, \sigma^{\mathbf{2}^{(i)}}\mathbf{I}) \tag{18}$$

with mean $\mu^{(i)}$ and standard deviation $\sigma^{\mathbf{2}^{(i)}}\mathbf{I}$ given by the output of a neural network as a function of datapoint $\mathbf{x}^{(i)}$ and variational parameters ϕ.

We sample from the posterior $\mathbf{z}^{(i,l)} \sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$ with $\mathbf{z}^{(i,l)} = g_\phi(\mathbf{x}^{(i)}, \epsilon^{(l)}) = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)}$ and $\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\odot$ denotes element-wise multiplication. Since $p_\theta(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$ are assumed to be Gaussian, we can use the estimator in Equation 14. Hence, it follows that for datapoint $\mathbf{x}^{(i)}$:

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) \approx \frac{1}{2}\sum_{j=1}^{N}(1 - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2 + \log((\sigma_j^{(i)})^2) + \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}) \tag{19}$$

whereby $\mathbf{z}^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)}$ and $\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Remark 1 *As mentioned, we can use a Gaussian encoder or decoder parameterized by a neural network. That is:*

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{z}) &= \log \mathcal{N}(\mathbf{x}; \mu, \sigma^2\mathbf{I}) \\ \mu &= \mathbf{W}_4\mathbf{h} + \mathbf{b}_4 \\ \log \sigma^2 &= \mathbf{W}_5\mathbf{h} + \mathbf{b}_5 \\ \mathbf{h} &= \mathit{tanh}(\mathbf{W}_3\mathbf{z} + \mathbf{b}_3) \end{aligned} \tag{20}$$

where $\{\mathbf{W}_3, \mathbf{W}_4, \mathbf{W}_5, \mathbf{b}_3, \mathbf{b}_4, \mathbf{b}_5\}$ *are weights and biases of the neural network.*

3.1 Full Variational Bayes

We can perform variational inference on both parameters θ and latent variables z. Let $p_\alpha(\theta)$ be a hyperprior for parameters θ, parameterized by α. Thus, the marginal likelihood is:

$$\log p_\alpha(\mathbf{X}) = D_{KL}(q_\phi(\theta)||p_\alpha(\theta|X)) + \mathcal{L}(\phi; \mathbf{X}) \tag{21}$$

where the first term on the right-hand side is the KL divergence of the approximate from the true posterior, and second term $\mathcal{L}(\phi; \mathbf{X})$ is the variational lower bound of the marginal likelihood given by:

$$\mathcal{L}(\phi; \mathbf{X}) = \int q_\phi(\theta)(\log p_\theta(\mathbf{X}) + \log p_\alpha(\theta) - \log q_\phi(\theta))d\theta. \tag{22}$$

Here, the $\log p_\theta(\mathbf{X})$ is a sum over maginal likelihoods of individual datapoints $\log p_\theta(\mathbf{X}) = \sum_{i=1}^{N} \log p_\theta(\mathbf{x}^{(i)})$ such that

$$\log p_\theta(\mathbf{x}^{(i)}) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}||p_\theta(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) \tag{23}$$

where the first term is the KL divergence of the approximate to true posterior and $\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ is the variational lower bound of the marginal likelihood for $\mathbf{x}^{(i)}$:

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = \int q_\phi(\mathbf{z}|\mathbf{x}) \Big(\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) \Big) d\mathbf{z}. \tag{24}$$

Furthermore, we can reparameterize conditional samples $\tilde{\mathbf{z}} \sim q_\phi(\mathbf{z}|\mathbf{x})$ as:

$$\tilde{\mathbf{z}} = g_\phi(\epsilon, \mathbf{x}), \text{ where } \epsilon \sim p(\epsilon). \tag{25}$$

The prior $p(\epsilon)$ and function $g_\phi(\epsilon, \mathbf{x})$ are chosen such that

$$\begin{aligned} \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) &= \int q_\phi(\mathbf{z}|\mathbf{x})(\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}))d\mathbf{z} \\ &= \int p(\epsilon)(\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}))|_{\mathbf{z}=g_\phi(\epsilon, \mathbf{x}^{(i)})} d\epsilon \end{aligned} \tag{26}$$

Likewise, for the approximate posterior $q_\phi(\theta)$:

$$\tilde{\theta} = h_\phi(\zeta) \text{ where } \zeta \sim p(\zeta). \tag{27}$$

As before, we chose prior $p(\zeta)$ and $h_\phi(\zeta)$ such that:

$$\begin{aligned} \mathcal{L}(\phi; \mathbf{X}) &= \int q_\phi(\theta)(\log p_\theta(\mathbf{X}) + \log p_\alpha(\theta) - \log q_\phi(\theta))d\theta \\ &= \int p(\zeta)(\log p_\theta(\mathbf{X}) + \log p_\alpha(\theta) - \log q_\phi(\theta))|_{\theta=h_\phi(\zeta)} d\zeta. \end{aligned} \tag{28}$$

Furthermore, let

$$f_\phi(\mathbf{x}, \mathbf{z}, \theta) = N(\log p_\theta(\mathbf{x}|\mathbf{z}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})) + \log p_\alpha(\theta) - \log q_\phi(\theta). \tag{29}$$

Invoking Equation 26 and 28, the Monte Carlo estimate of the variational lower bound for $\mathbf{x}^{(i)}$ is:

$$\mathcal{L}(\phi; \mathbf{X}) \approx \frac{1}{L} \sum_{l=1}^{L} f_\phi(\mathbf{x}^{(l)}, g_\phi(\epsilon^{(l)}, \mathbf{x}^{(l)}), h_\phi(\zeta^{(l)})) \tag{30}$$

where $\epsilon^{(l)} \sim p(\epsilon)$ and $\zeta^{(l)} \sim p(\zeta)$. This estimator is only a function of samples from $p(\epsilon)$ and $p(\zeta)$, independent of the variational parameters ϕ. Hence, we can differentiate this Monte Carlo estimator with respect to ϕ and use stochastic optimization methods to compute gradients in tandem.

References

[1] D. Kingma and M. Welling. Auto-encodingvariationalbayes. 2022. Available at `https://arxiv.org/abs/1312.6114`.