On study of Variational Inference

Marshal Arijona Sinaga (2006560983) [1]

Abstract

The main problem of Bayesian inference is computing the posterior which is often intractable. In this paper, we review variational inference (VI) methods that aim to find a variational distribution to approximate the true posterior. The review starts with the original form of variational inference. Then we also discuss the extension of VI: mean-field, stochastic and black-box variational, and amortized inference. Finally, we review the recent improvement of variational inference.

1. Background

Uncertainty plays an important role in almost every aspect of our life. Oftentimes, we involve uncertainty in our decision-making. Therefore, we want to make the right estimation of our confidence in decision-making. Examples of decision-making include predicting the weather, predicting the next season's sale, and medical treatment for the patient.

Artificial intelligence and machine learning have a huge potential to help us estimate the uncertainty by involving the Bayes probability theorem. It turns out that Bayes probability theorem provides the right tool to estimate the probability. Specifically, machine learning utilizes the Bayes probability to perform inference. The objective of inference is that given the observed data, we want to infer the unobserved/latent variables (Li, 2020). Involving the Bayes probability, the goal of inference is to estimate the uncertainty of the latent variable. This uncertainty is represented by the posterior distribution of the latent variable given the observed data (Bishop, 2007).

However, computing the posterior distribution is often intractable. The source of intractability lies in the computation of the evidence, which involves integration (Li, 2020). Therefore, we need the approximate inference to estimate the true posterior distribution. The approximate inference involves a variational distribution that replaces the role of the true posterior distribution. This paper aims to review several techniques that have been proposed to perform the approximate inference.

This paper is arranged as follows. We first derive the general form of variational inference in Chapter (2). Chapter (3) tells us how to perform mean-field variational inference. We then introduce the stochastic variational inference (SVI) in Chapter (4) to solve the computation cost issue. Chapter (5) provides a more general form of SVI called black-box variational inference (BBVI). In Chapter (6) we derive the amortized inference and highlighting its drawback. The last two chapters discuss some modification proposed to improve the approximation result. Chapter (7) disccuss the design of the approximate inference while Chapter (8) discuss the design of the objective function.

2. Variational inference

The core of Bayesian statistics is computing the posterior distribution of parameters given the observed data (Bishop, 2007). Computing posterior distribution requires prior distribution $p(\theta)$, likelihood $p(D|\theta)$ and the evidence $p(D)$. The prior distribution encodes our belief about the parameter θ before involving the evidence. On the other hand, the likelihood describes the probability of the observed data given the parameter. Lastly, the evidence provide the probability of the observed data used to update the posterior.

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)} \tag{1}$$

We can write the evidence as $p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta$. However, computing the marginal distribution $p(D)$ is intractable, in the sense that the integration is high dimensional (Zhang et al., 2019). Instead, a variational distribution $q_\phi(\theta)$ with parameters ϕ is introduced to approximate the posterior. Mathematically, we can write $q_\phi(\theta) \approx p(\theta \mid D)$. The goal of variational inference is to find the parameters ϕ such that q gives the best matching (Bishop, 2007). We assume that the set of possible qs are lying in space Q (Blei et al., 2016). Commonly, this space describes a certain family of distribution. For example, suppose that Q is the space that consists all possible q parameterized by ϕ such that $0 \leq \phi \leq 1, \phi \in \mathbb{R}$. Then, Q represents the family of the Bernoulli distribution. Furthermore, it is possible that the posterior $p(\theta \mid D)$ is not lying in the space Q. Figure 1 shows the illustration of variational inference.

We start by introducing KL-divergence between two distributions. Given probability distribution $p(\theta)$ and $q(\theta)$, KL-divergence $KL[p(\theta) \,\|\, q(\theta)]$ quantifies the additional

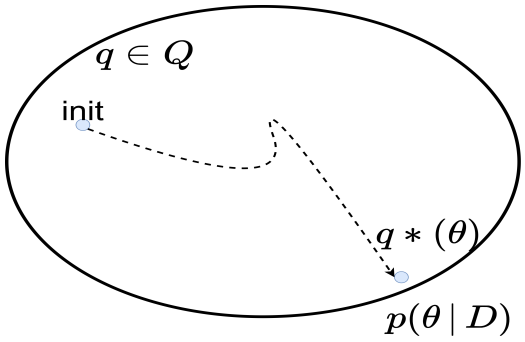

Figure 1. The Illustration of variational inference (Blei et al., 2016). The approximation starts with the initial q. Through the optimization process, we obtain q^* which gives the best matching.

amount of information to represent θ using q instead of p (Bishop, 2007). Mathematically, we can write KL-divergence as the expectation of log density ratio $\frac{q(\theta)}{p(\theta)}$.

$$KL[p(\theta) \,\|\, q(\theta)] = -\int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta \tag{2}$$

$$= \mathbb{E}_{q(\theta)} \left[\log \frac{q(\theta)}{p(\theta)} \right] \tag{3}$$

When $p = q$ then $KL[p \,\|\, q] = 0$, otherwise $KL[p \,\|\, q] \geq 0$. It is also important to emphasize that KL-divergence is not symmetric, means that $KL[p \,\|\, q] \neq KL[q \,\|\, p]$. The derivation of variational objective will be based on KL-divergence.

We start with the KL-divergence between $q(\theta)$ and $p(\theta)$, written as $KL[q(\theta) \,|\, p(\theta)]$. For simplicity, we omit the parameters ϕ from the notation. Then, by using the definition of conditional probability and the sum property of logarithm, we obtain:

$$\begin{aligned} KL[q(\theta) \,\|\, p(\theta \,|\, D)] &= -\mathbb{E}_{q(\theta)} \left[\log \frac{p(\theta \,|\, D)}{q(\theta)} \right] \\ &= -\mathbb{E}_{q(\theta)} \left[\log \frac{p(\theta,\, D)}{q(\theta)\, p(D)} \right] \\ &= -\mathbb{E}_{q(\theta)} \left[\log \frac{p(\theta,\, D)}{q(\theta)} - \log p(D) \right] \\ &= \log p(D) - \mathbb{E}_{q(\theta)} \left[\frac{p(\theta, D)}{q(\theta)} \right] \end{aligned} \tag{4}$$

Note that $\log p(D)$ comes from the fact that $\mathbb{E}_{q(\theta)}[\log p(D)] = \log p(D)$ (since the term does not depend on $q(\theta)$). By rearangging the Equation (4) and discarding the KL-divergence term we obtain the lower bound of $p(D)$

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\theta)} \left[\frac{p(\theta, D)}{q(\theta)} \right] \tag{5}$$

This lower bound is known as evidence lower bound (ELBO). Observe that maximizing ELBO is equivalent to minimizing the KL-divergence. Thus if $KL[q(\theta) \,\|\, p(\theta)] = 0$ then we have $\mathcal{L}_{\text{ELBO}} = \log p(D)$. Figure 2 illustrates the bound between ELBO and $\log\, p(D)$.

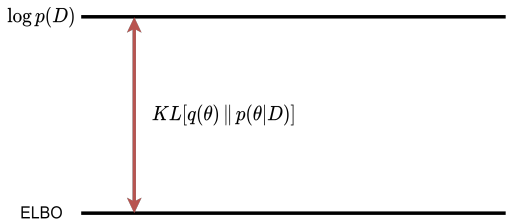

Figure 2. The illustration that shows the bound between $\log\, p(D)$ and ELBO. The gap between those two is represented by the KL-divergence (Li, 2020). When $KL[q(\theta) \,\|\, p(\theta|D)] = 0$ then $\mathcal{L}_{\text{ELBO}} = \log p(D)$.

We can also derive the variational objective by defining $p(D)$ as the integration of joint probability $p(\theta, D)$. Furthermore, we involve Jensen's inequality to derive the objective. Let f be a convex function and X be a random variable which is the input for f. Then the function f satisfies $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$, known as Jensen's inequality (Bishop, 2007). Subsequently, if f is a concave function then $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$. By combining all the information, we obtain $\mathcal{L}_{\text{ELBO}}$ as follows:

$$\begin{aligned} \log p(D) &= \log \int p(\theta, D)\, d\theta \\ &= \log \int \frac{p(\theta, D)\, q(\theta)}{q(\theta)}\, d\theta \\ &= \log \mathbb{E}_{q(\theta)} \left[\frac{p(\theta,\, D)}{q(\theta)} \right] \\ &\geq \mathbb{E}_{q(\theta)} \left[\log \frac{p(\theta,\, D)}{q(\theta)} \right] := \mathcal{L}_{\text{ELBO}} \end{aligned} \tag{6}$$

Observe that we augment the variational distribution $q(\theta)$ in the second row. Since logarithm is a concave function, then it enjoys the Jensen's inequality. Performing this inequality will obtain $\mathcal{L}_{\text{ELBO}}$.

3. Mean-Field Variational Inference (MFVI)

In the real-world application, we deal with high dimensional distribution. Such distribution is represented by multivariate parameters $\boldsymbol{\theta}$. The mean-field variational inference (MFVI) is introduced to approximate that kind of distribution. The

idea of MFVI originally comes from the mean-field theory of physics (Hogan, 2002). Let $q(\boldsymbol{\theta})$ be a variational distribution where $\boldsymbol{\theta}$ is a K-variate vector. Mean-field variational inference assumes that each component θ_i is independent one each other. Therefore we can express $q(\boldsymbol{\theta})$ as a product of $q(\theta_j)$. Mathematically, we can write $q(\boldsymbol{\theta})$ as:

$$q(\theta) = \prod_{j=1}^{K} q_{\phi_i}(\theta_i) \tag{7}$$

To give an example, suppose that $p(\mathbf{z})$ is a 2-variate Gaussian distribution parameterized by mean $\boldsymbol{\mu}$ and covariance Λ. Mathematically, we can write $p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \Lambda^{-1})$. Following Equation (7), we have $q(\mathbf{z})$ that approximates $p(\mathbf{z})$ such that $q(\mathbf{z}) = q(z_1)\, q(z_2)$, where $q(z_1) = \mathcal{N}(\mu, \sigma_1^{-1})$ and $q(z_2) = \mathcal{N}(\mu_2, \sigma_2^{-1})$ are a univariate Gaussian distribution. Figure 3 illustrates the principle of MFVI.

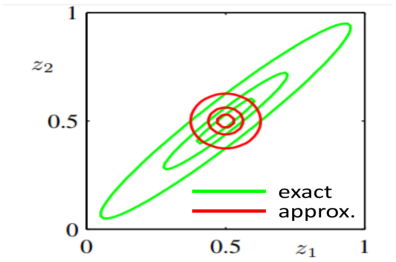

Figure 3. The illustration of MFVI (Bishop, 2007). The green curvature denotes the multivariate Gaussian distribution $p(\mathbf{z})$. The red curvature denotes the mean-field variational distribution $q(\mathbf{z})$ which comprises two independent univariate Gaussian distributions.

In order to derive the variational objective, we combine Equation (5) and Equation (7). Furthermore, we express $\mathcal{L}_{\text{ELBO}}$ by separating a particular θ_j from the other components $\theta_{\neg j}$ (Jordan et al., 1999). Mathematically, we can write ELBO as:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = &\int q(\theta_j)\mathbb{E}_{q(\theta_{\neg j})}\left[\log p(\theta_j, D|\theta_{\neg j})\right]\, d\theta_j \\ &- \int q(\theta_j)\log p(\theta_j) d\theta_j + c_j\end{aligned} \tag{8}$$

, with c_j denotes the constant. The motivation of separating the component is for ease of optimization. Recall that we rely on the optimization process to obtain an ideal variational distribution q. In the case of MFVI, we utilize coordinate-ascent to perform the optimization (Jordan et al., 1999). This optimization method works in a greedy manner by optimizing each component one by one. For each component, we fixed the other components and find the best solution for the interest component. For a particular component θ_j, we can write the optimal solution as:

$$q^*(\theta_j) \propto \exp(\mathbb{E}_{q(\theta_{\neg j})}\left[\log p(\theta_j|D, \theta_{\neg j})\right]) \tag{9}$$

with exp denotes the exponentiation w.r.t. natural number e.

4. Stochastic Variational Inference

One of the biggest challenges of variational inference is to scale the method on big dataset. Suppose that we have M observed data x_i, written as $\boldsymbol{x} = \{x_i\}_{i=1}^M$ and the corresponding local parameter ξ_i, written as $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^M$. We assume that M is very large (e.g. millions). Furthermore, we also have a global parameter θ which affects the probability of the local parameter and the observed data. Figure 4 shows the graphical model of the described problem.

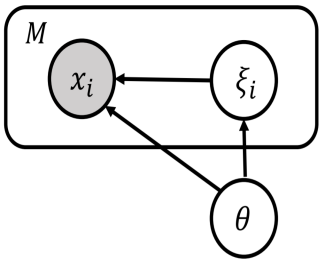

Figure 4. The graphical model of M observed data x_i (gray shaded), each depends on the local parameter ξ_i and the global parameter θ (Li, 2020). For simplification, we use plate to represent all M observed data and local parameters. The plate is denoted by the rectangle that covers x_i and ξ_i.

Recall that ELBO requires the joint distribution between the observed data and the parameter (Equation (5)). From Figure 4, we can write $p(\theta, \boldsymbol{\xi}, \boldsymbol{x})$ as:

$$p(\theta, \boldsymbol{\xi}, \boldsymbol{x}) = p(\theta) \prod_{i=1}^{M} p(\xi_i \mid \theta)\, p(x_i \mid \xi_i,\, \theta) \tag{10}$$

Subsequently, by substituting Equation (10) into Equation (5) we obtain:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} &= \mathbb{E}_q\left[\log \frac{p(\theta, \boldsymbol{\xi}, \boldsymbol{x})}{q(\theta, \xi)}\right] \\ &= \mathbb{E}_q\left[\log \frac{p(\theta)\prod_{i=1}^{M} p(\xi_i \mid \theta)\, p(x_i \mid \xi_i,\, \theta)}{q(\theta)\prod_{i=1}^{M} q(\xi_i)}\right] \\ &= \mathbb{E}_q\left[\log \frac{p(\theta)}{q(\theta)}\right] \\ &\quad + \sum_{i=1}^{M} \mathbb{E}_q\left[\log \frac{p(\xi_i \mid \theta)\, p(x_i \mid \xi_i, \theta)}{q(\xi_i)}\right] \end{aligned} \tag{11}$$

For simplicity, we exclude the parameters of q. Using the sum property of logarithm, we decompose the second row to obtain the third row. When M is very large, the computation might not be affordable. For K-variate $\boldsymbol{\xi}$, the time complexity of a single iteration coordinate-ascent which involves M observed data is $\mathcal{O}(MK)$. Fortunately, stochastic variational inference (SVI) can help estimating $\mathcal{L}_{\text{ELBO}}$.

In stochastic method, we assume that M observed variabels are i.i.d. Based on this assumption, SVI can split M observed data into S subsets, with $S \ll M$ (Hoffman et al., 2013). We call the subset a mini-batch. Each mini-batch consists of $\frac{M}{S}$ observed data. By incorporating the mini-batch into Equation (10), we can write SVI as:

$$\begin{aligned}\hat{\mathcal{L}}_{\text{ELBO}} &= \mathbb{E}_q\left[\log \frac{p(\theta)}{q(\theta)}\right] \\ &\quad + \frac{M}{S}\sum_{i=1}^{S} \mathbb{E}_q\left[\log \frac{p(\xi_i \mid \theta)\, p(x_i \mid \xi_i, \theta)}{q(\xi_i)}\right] \end{aligned} \tag{12}$$

When $S = M$, we have an ordinary/full-batch variational inference. Choosing the right S is a trade-off in SVI. Larger S produces a more accurate estimation of $\mathcal{L}_{\text{ELBO}}$ but requires more computation cost. Reciprocally, having a small S requires less computation cost but produces more noisy estimation in return.

The optimization of $\mathcal{L}_{\text{ELBO}}$ can be done through coordinate ascent or gradient ascent. Gradient ascent requires the gradient $\nabla_\phi \mathcal{L}_{\text{ELBO}}$ to update the parameters ϕ.

$$\begin{aligned}\nabla_\phi \mathcal{L}_{\text{ELBO}} &= \nabla_\phi \mathbb{E}_q\left[\log \frac{p(\theta)}{q(\theta)}\right] \\ &\quad + \sum_{i=1}^{M} \mathbb{E}_q\left[\nabla_\phi \log \frac{p(\xi_i \mid \theta)\, p(x_i \mid \xi_i, \theta)}{q(\xi_i)}\right] \end{aligned} \tag{13}$$

Since SVI splits the observed data into mini-batches, the gradient turns out into stochastic gradient $\nabla_\phi \hat{\mathcal{L}}_{\text{ELBO}}$. Instead of taking the full-batch gradient of all M observed data, stochastic gradient takes the mini-batch gradient of $\frac{M}{S}$ variables (Zhang et al., 2017). If the variance of each mini-batch gradient for any samples is low enough, then the average of the mini-batch gradient will converge to the full-batch gradient.

$$\begin{aligned}\nabla_\phi \hat{\mathcal{L}}_{\text{ELBO}} &= \nabla_\phi \mathbb{E}_q\left[\log \frac{p(\theta)}{q(\theta)}\right] \\ &\quad + \frac{M}{S}\sum_{i=1}^{S} \mathbb{E}_{q(\theta)}\left[\nabla_\phi \log \frac{p(\xi_i \mid \theta)\, p(x_i \mid \xi_i, \theta)}{q(\xi_i)}\right] \end{aligned} \tag{14}$$

Finally, we can write stochastic gradient ascent as:

$$\phi \leftarrow \phi + \alpha\, \nabla_\phi \hat{\mathcal{L}}_{\text{ELBO}} \tag{15}$$

with $\alpha \in \mathbb{R} > 0$ denotes the learning rate which scales the gradient. Another method is to update the parameters using natural gradient. Natural gradient considers the information geometry of the variational parameters (Amari, 1998). They are obtained by multiplying $\nabla_\phi \mathcal{L}_{\text{ELBO}}$ with the inverse Fisher information matrix. However, SVI is limited to conjugate exponential family models. Therefore, SVI can not handle non-conjugate model.

5. Black-Box Variational Inference

In this section, we introduce black-box variational inference (BBVI) which is more general than SVI. Black-box variational inference can deal with non-conjugate model by incorporating Monte Carlo estimation and the gradient estimator method.

Let f be a function which takes input x. Furthermore, suppose that x comes from probability distribution $p(x)$. Monte Carlo estimation method allows us to compute the expectation $\mathbb{E}_{p(x)}[f(x)]$ by taking K samples x_i from $p(x)$ (Bishop, 2007). We estimate the expectation by taking the average of $f(x_i)$. Algorithm 1 provides the complete steps of Monte Carlo estimation method.

Algorithm 1 Monte Carlo Estimation

- **To approximate** : $\mathbb{E}_{p(x)}[f(x)]$
- Sample $x_1, \cdots, x_K \sim p(x)$
- Evaluate $f(x_i)$ for each x_i
- Compute $\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{K}\sum_{i=1}^{K} f(x_i)$

The second recipe to build BBVI is gradient estimator method. This method allows us to compute both conjugate and non-conjugate model. In this paper, we explain two gradient estimators method: REINFORCE gradient and reparameterization trick.

5.1. REINFORCE Gradient

First, we introduce the log-derivative trick which is the key of deriving REINFORCE gradient. Given $\log p_\theta(x)$, we can derive $\nabla_\theta\, p_\theta(x)$ as:

$$\begin{aligned}\nabla_\theta \log p_\theta(x) &= \frac{1}{p_\theta(x)} \nabla_\theta\, p_\theta(x) \\ \nabla_\theta\, p_\theta(x) &= p_\theta(x)\, \nabla_\theta \log\, p_\theta(x)\end{aligned} \tag{16}$$

The first row is obtained by applying chain rule. We end up with Equation (16) by multiplying both sides with $p_\theta(x)$. By combining Equation (16) and the gradient of Equation (5), we derive REINFORCE gradient (Williams, 1992; Li, 2020) as:

$$\begin{aligned}\nabla\phi\, \mathcal{L}_{\text{ELBO}} &= \nabla_\phi\, \mathbb{E}_{q_\phi(\theta)} \left[\log \frac{p(\theta,\, D)}{q_\phi(\theta)}\right] \\ &= \int \nabla_\phi \left\{ q_\phi(\theta)\, \log \frac{p(\theta, D)}{q_\phi(\theta)} \right\} d\theta \\ &= \int \nabla_\phi\, q_\phi(\theta) \log \frac{p(\theta, D)}{q_\phi(\theta)} \\ &\quad + \int q_\phi(\theta)\, \nabla_\phi\, \log \frac{p(\theta, D)}{q_\phi(\theta)}\, d\theta \\ &= \int q_\phi(\theta)\, \nabla_\phi\, \log q_\phi(\theta)\, \log \frac{p(\theta,\, D)}{q_\phi(\theta)}\, d\theta \\ &\quad - \int \nabla_\phi\, q_\phi(\theta)\, d\theta \\ &= \mathbb{E}_{q_\phi(\theta)} \left[\nabla_\phi\, \log q_\phi(\theta)\, \log \frac{p(\theta,\, D)}{q_\phi(\theta)}\right]\end{aligned} \tag{17}$$

with

$$\nabla_\phi\, \log \frac{p(\theta,\, D)}{q_\phi(\theta)} = \frac{q_\phi(\theta)}{p(\theta, D)} \frac{-p(\theta,\, D)}{q_\phi(\theta)^2} \nabla_\phi\, q_\phi(\theta) \tag{18}$$

and

$$\int \nabla_\phi\, q_\phi(\theta)\, d\theta = \nabla_\phi \int q_\phi(\theta)\, d\theta = \nabla_\phi\, 1 = 0 \tag{19}$$

We acquire Equation (18) by performing the chain rule gradient. By using the fact that we can pull out the gradient outside the expectation and the integral of probability distribution equals to one, we obtain Equation (19). We obtain the third row of Equation (17) by performing partial derivative. Subsequently, we substitute Equation (16) into the first term and Equation (18) into the second term of third row to obtain the fourth row. Incorporating (19) into the fourth row, we are completely derive the REINFORCE gradient method. Finally, we combine REINFORCE gradient method with Monte Carlo Estimation to obtain a complete black-box variational inference (Ranganath et al., 2014). Specifically, this BBVI aims to estimate $\mathbb{E}_{q_\phi(\theta)} \left[\nabla_\phi\, \log q_\phi(\theta)\, \log \frac{p(\theta, D)}{q_\phi(\theta)}\right]$. Algorithm (2) provides a complete steps to perform BBVI.

Algorithm 2 Black-Box Variational Inference

- Sample $\theta_1, \cdots, \theta_K \sim q_\phi(\theta)$
- Evaluate $\nabla_\phi\, \log q_\phi(\theta_i)\, \log \frac{p(\theta_i, D)}{q_\phi(\theta_i)}$ for each θ_i
- Compute $\nabla \hat{\mathcal{L}}_{\text{ELBO}} = \frac{1}{K} \sum_{i=1}^{K} \nabla_\phi \log\, q_\phi(\theta_i)\, \log\, \frac{p(\theta_i, D)}{q_\phi(\theta_i)}$

5.2. Reparameterization Trick

Besides REINFORCE gradient, another common gradient estimator method is the reparameterization trick. Suppose that the variational distribution $q_\phi(\theta)$ is a Gaussian distribution, parameterized by μ and σ. Reparameterization trick suggests a function g that transforms a noise ϵ to express θ (Kingma & Welling, 2014). This function multiplies ϵ with σ and adds the result into μ. Commonly, we choose ϵ comes from the standard Gaussian distribution. Since the variational parameters ϕ are static, we can tune ϕ via gradient-based optimization. Specifically, we can perform back-propagation to obtain the variational parameters ϕ. Figure (5) illustrates the reparameterization trick.

$$\begin{aligned}\theta &\sim \mathcal{N}(\mu, \sigma^2) \\ \epsilon &\sim r(\epsilon) = \mathcal{N}(0, 1) \\ \theta &= g(\epsilon, \phi) = \mu + \sigma\epsilon\end{aligned} \tag{20}$$

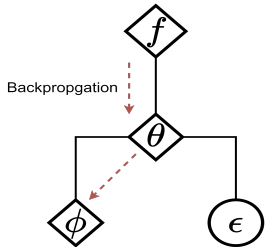

Figure 5. Illustration of reparameterization trick (Kingma & Welling, 2014). Function f requires θ. We generate θ through reparameterization trick. This trick enables us to compute the gradient of function f w.r.t. the variational parameters ϕ.

Incorporating Equation (20), we can express the ELBO as follows:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{r(\epsilon)} \left[\log \frac{p(g(\epsilon, \phi), D)}{q_\phi(g(\epsilon, \phi))} \right] \tag{21}$$

Observe that we replace the expectation over q_ϕ with the expectation over ϵ. Subsequently, we can write $\nabla_\phi \mathcal{L}_{\text{ELBO}}$ as:

$$\begin{aligned} \nabla_\phi \mathcal{L}_{\text{ELBO}} &= \nabla_\phi \mathbb{E}_{r(\epsilon)} \left[\log \frac{p(g(\epsilon, \phi), D)}{q_\phi(g(\epsilon, \phi))} \right] \\ &= \mathbb{E}_{r(\epsilon)} \left[\nabla_\phi \log \frac{p(g(\epsilon, \phi), D)}{q_\phi(g(\epsilon, \phi))} \right] \end{aligned} \tag{22}$$

With the help of Monte Carlo estimation, we obtain the black-box gradient estimator as:

$$\nabla_\phi \hat{\mathcal{L}} = \frac{1}{K} \sum_{i=1}^{K} \nabla_\phi \log \frac{p(g(\epsilon_k, \phi), D)}{q_\phi(g(\epsilon_k, \phi))}, \epsilon_k \sim r(\epsilon) \tag{23}$$

Observe that for every k, we sample ϵ_k to generate θ_k.

5.3. Variance Reduction Method for Black-Box Variational Inference

The main challenge of REINFORCE gradient method is that this method has a high variance (Li, 2020). This condition leads to an inaccurate ELBO estimation. There are two approaches to solve the issue. The first approach is to use low variance unbiased estimators with control variance. Another approach is using a biased estimator to enable a reparameterization trick. In this paper, we only discuss the first method.

We start with the control variance method (Paisley et al., 2012). Suppose that we want to estimate a function $F(\theta)$ with Monte Carlo method. Mathematically, we can write $\mathbb{E}_{q(\theta)}[F(\theta)] \approx \frac{1}{K} \sum_{i=1}^{K} F(\theta_k)$, $\theta_k \sim q(\theta)$ (for simplicity, we exclude the parameters of q). Now, we define a control variate function $G(\theta)$ which possesses two properties. The first property is that $V_{q(\theta)}[G(\theta)] < \infty$, with $V[.]$ denotes the variance. Another property is $\mathbb{E}_{q(\theta)}[G(\theta)]$ known or fast computable. Using $F(\theta)$ and $G(\theta)$, we construct a new function $\hat{F}(\theta)$ to approximate $\mathbb{E}_{q(\theta)}[F(\theta)] \approx \frac{1}{K} \sum_{i=1}^{K} \hat{F}(\theta_k)$, $\theta_k \sim q(\theta)$. Mathematically, we can write $\hat{F}(\theta)$ as:

$$\hat{F}(\theta) = F(\theta) - G(\theta) + \mathbb{E}_{q(\theta)} [G(\theta)] \tag{24}$$

Since computing $\mathbb{E}_{q(\theta)}[G(\theta)]$ is relatively fast, then computing $\hat{F}(\theta)$ does not affect the computation time too much. Also, it can be shown that $\mathbb{E}_{q(\theta)}[\hat{F}(\theta)] = \mathbb{E}_{q(\theta)}[F(\theta)]$, which means that $\hat{F}(\theta)$ is an unbiased estimator of $F(\theta)$. By the properties of variance, we can write the variance $V_{q(\theta)}(\hat{F}(\theta))$ as:

$$\begin{aligned} V_{q(\theta)}[\hat{F}(\theta)] = V_{q(\theta)}[F(\theta)] + V_{q(\theta)}[G(\theta)] \\ - 2\, Cov_{q(\theta)}[F(\theta),\, G(\theta)] \end{aligned} \tag{25}$$

From Equation (25), observe that if $F(\theta)$ and $G(\theta)$ is strongly correlated (the covariance of $F(\theta)$ and $G(\theta)$ is high), then $V_{q(\theta)}[G(\theta)] - 2\, Cov_{q(\theta)}[F(\theta),\, G(\theta)]$ will be negative. Therefore, we have $Var[F(\hat{\theta})] < Var[F(\theta)]$. In this fashion, we are successfully estimate $F(\theta)$ while reducing the variance at the same time.

Now, let us use trick for REINFORCE gradient. Let $F(\theta) = f(\theta)\, \nabla_\phi \log q_\phi(\theta)$, with $f(\theta) = \log \frac{p(D,\theta)}{q_\phi(\theta)}$. Subsequently, we define control variate function $G(\theta) = g(\theta) \log q_\phi(\theta)$. For convenience, we define $g(\theta)$ later. Following Equation (25), we can write the new estimator function $\hat{F}(\theta)$ as:

$$\hat{F}(\theta) = (f(\theta) - g(\theta))\, \nabla_\phi \log q_\phi(\theta) + \mathbb{E}_{q_\phi(\theta)}[G(\theta)] \tag{26}$$

Now, let us set $g(\theta) = b$. Using the log derivative trick, we obtain the expectation $\mathbb{E}_{q(\theta)}$ as follows:

$$\begin{aligned} \mathbb{E}_{q(\theta)}\, [G(\theta)] &= b\, \mathbb{E}_{q(\theta)}[\nabla_\phi \log q(\theta)] \\ &= b\, \nabla_\phi \int q_\phi(\theta)\, d\theta \\ &= b\, \nabla_\phi 1 = 0 \end{aligned}$$

Since the expectation of $G(\theta)$ is zero, then Equation (26) turns into:

$$\hat{F}(\theta) = \Delta(\theta)\, \nabla_\phi \log q_\phi(\theta) \tag{27}$$

Intuitively, this method works as follows. Suppose that we have $\theta_1, \theta_2 \sim q_\phi(\theta)$. Furthermore, suppose that $f(\theta_1) - g(\theta_1) > 0$ and $f(\theta_2) - g(\theta_2) < 0$. This method tries to find q_ϕ that increases the probability of θ_1 while decreasing the probability of θ_2. Another common choice of $g(\theta)$ is based on Taylor expansion (Gu et al., 2015), that is $g(\theta) = f(\theta_0) + \nabla_{\theta_0} f(\theta_0)(\theta - \theta_0)$.

6. Amortized Inference

Despite being able to handle a large amount of data, SVI still suffers from the memory issue. Note that we need to optimize the variational parameters ϕ_i of the latent variable z_i

independently. This can be memory-consuming, especially when we have a huge amount of observed data. Instead, amortized inference views the variational parameters ϕ as a function which takes x_i as the input. Therefore the latent variables z_i depends on x_i. Figure (6) shows the illustration of amortized inference.

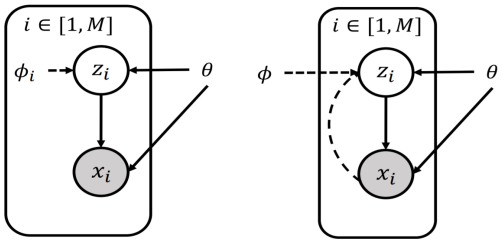

Figure 6. left Probabilistic graphical model of ordinary SVI. Observe that each latent variable z_i depends on variational parameters ϕ_i (Li, 2020). *right* Probabilistic graphical model of amortized inference. The variational parameters ϕ is now outside the plate. Therefore, the variational parameters ϕ becomes a global parameter. Furthermore, the latent variables z_i now depends on x_i.

Based on Figure (6), we aim to derive $\mathcal{L}_{\text{ELBO}}$ for amortized inference case. We start with the definition of ELBO in term of KL-divergence (Equation (4)), $\mathcal{L}_{\text{ELBO}} = \log\, p(x) - KL[q(z) \,\|\, p(z|x)]$. Since z_i depends on x_i, we replace $q(z)$ with $q(z|x)$ (Kingma & Welling, 2014; Rezende et al., 2014). Therefore, we can write $\mathcal{L}_{\text{amortized}}$ as:

$$\begin{aligned}\mathcal{L}_{\text{amortized}} &= \log\, p(x) - KL[q(z|x) \,\|\, p(z|x)] \\ &= \mathbb{E}_{q_\phi(z\,|\,x)}\left[\log p_\theta(x\,|\,z)\right] \\ &\quad - KL\left[q(z\,|\,x) \,\|\, p(z\,|\,x)\right] \end{aligned} \tag{28}$$

One of the renowned applications of amortized inference is the variational autoencoder (VAE) (Kingma & Welling, 2014). VAE feeds image $\mathbf{x}$ through the encoder E and outputs $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. Subsequently, VAE generates the latent variable $\mathbf{z}$ by performing a reparameterization trick on the distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The decoder takes $\mathbf{z}$ and outputs $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. Commonly, VAE assumes that $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is following factorized multivariate Gaussian distribution with the parameters mean $\boldsymbol{\mu}$ and . Furthermore, the prior distribution $p(\mathbf{z})$ is assumed to be a standard Gaussian distribution $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. Figure (7) shows the illustration of VAE. Another application of amortized inference including amortized sequential Monte Carlo (SMC) (Naesseth et al., 2018; Le et al., 2017), amortized Markov chain Monte Carlo (MCMC) (Li et al., 2017), and amortized Monte Carlo integration (Golinski et al., 2019).

Although reducing the computation cost, the amortized inference also leaves a drawback. Naturally, amortized approximate posterior is sub-optimal (Cremer et al., 2018).

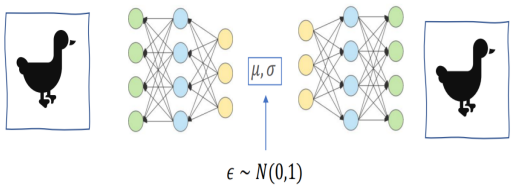

Figure 7. The illustration of variational autoencoder (Li, 2020). Commonly, VAE assumes that the prior distribution $p(\mathbf{z})$ is following standard Gaussian distribution $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$.

Figure (8) shows the inference result with and without amortization. From that figure, it is obvious that in some cases, the amortization is far from the true posterior distribution, especially when the region of the posterior is relatively small and when the posterior is a mixture distribution.

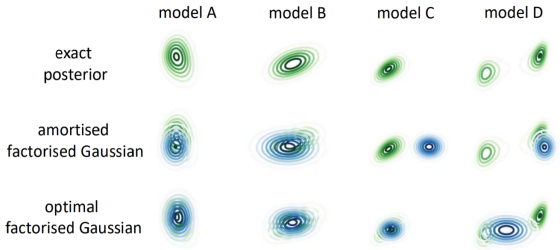

Figure 8. The limitation of amortized inference (Cremer et al., 2018). It turns out that amortized inference produces a far result from the true posterior when the region of the posterior distribution is small and when the posterior is a mixture distribution.

Furthermore, we can also show the gap of amortized inference by visualizing the bound of $\log\, p(x)$, $\mathcal{L}_{\text{ELBO}}$ and $\mathcal{L}_{\text{amortized}}$, shown by Figure (9). Unless using a very deep neural network as the universal approximator, otherwise there is a gap between the optimal amortized inference result and the optimal individual inference result.

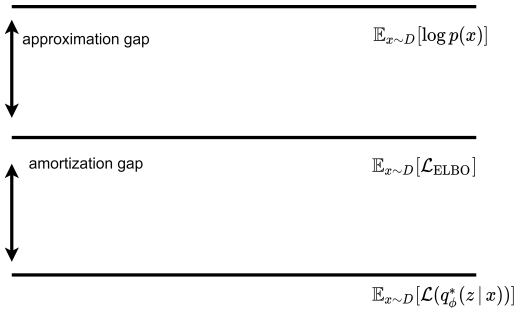

Figure 9. The gap between $\mathcal{L}_{\text{ELBO}}$ and $\mathcal{L}_{\text{amortized}}$ (Li, 2020). Commonly, we have $\mathcal{L}_{\text{ELBO}} \geq \mathcal{L}_{\text{amortized}}$ since there is an amortization gap.

One of the solution to close the amortization gap is by intro-

ducing the refinement method (Marino et al., 2018). This method starts by initializing $q_\phi(z|x)$ through the amortized inference. Then, we run several individual gradient steps to update the variational parameters ϕ.

7. Approximate distribution design

In this section, we will discuss four techniques that are commonly used to design the approximate distribution. The goal of the distribution design is to close the gap with the true posterior distribution. Those techniques include structured approximation, normalizing flows, auxiliary & mixture distribution, and Implicit approximate posterior.

7.1. Structured Approximations

Structured approximation distribution suggests grouping the latent variables that generate the same sequences while ignoring the dependency between the groups that generate different sequences. Furthermore, structured approximation also defines the dependency within latent variables in the same group (Li, 2020). Figure (10) illustrates the structured distribution. Given the approximation distribution q with S groups, we define the structured approximation distribution as:

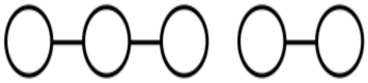

Figure 10. The structured distribution consists of two groups with three and two members, respectively. There is no dependency between groups.

$$q(z) = \prod_s q(z_s)$$
$$q(z_s) = q(\{z_i\}_{i \in s})$$

with $q(z_s)$ defines the dependency within the latent variables that lies in group s.

7.2. Normalizing Flows

Researches use the normalizing flows to obtain a flexible approximation posterior. This method is based on the sequence of invertible function on random variables. Given an invertible functions f that transforms random variable X into random variable Y, we want to compute the corresponding density p_Y given p_X. If we map the region in $\mathcal{X}$ space to $\mathcal{Y}$ space using f, then we want the corresponding twisted region in $\mathcal{Y}$ to have the same probability mass. On the other hand, we can think $p_X(x)dx$ as the probability mass of the corresponding region around x. Similarly, $p_Y(y)dy$ is the probability mass of the corresponding region around y. Therefore, we want to have:

$$p_X(y)dy = p_Y(x)dx \tag{29}$$

Based on Equation (29), we can write the density p_X and p_Y as :

$$p_Y(y) = p_X(x) \left|\det(\frac{dx}{dy})\right| \tag{30}$$
$$p_X(x) = p_Y(y) \left|\det(\frac{dy}{dx})\right| \tag{31}$$

The determinant term is responsible to represent the volume changes. Equation (30) and Equation (31) are known as the change of variable formula.

Now, we aim to build a variational inference with normalizing flows (Rezende & Mohamed, 2015). We start with simple distribution $q_0 = \mathcal{N}(z_0; 0, I)$ and an invertible function parameterized by ϕ which maps $z = f_\phi(z_0)$. By the change of variable we can write $q(z)$ as:

$$q(z) = q_0(z_0) \left|\det(\frac{dz}{dz_0})\right|^{-1} \tag{32}$$

with $z_0 = f_\phi^{-1}(z)$. Recall that we can write $\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(z)}[\log\, p(x|z) + \log\, p(z) - \log\, q(z)]$. Substituting Equation (32) and the definition of z into $\mathcal{L}_{\text{ELBO}}$, we obtain:

$$\begin{aligned}
\mathcal{L}(q(z)) &= \mathbb{E}_{q(z)}\left[\log\, p(x\,|\,z) + \log\, p(z) - \log\, q(z)\right] \\
&= \mathbb{E}_{q(z)}\left[\log\, p(x,\, z) - \log\, q_0(z_0)\, \left|\det(\frac{dz}{dz_0})\right|^{-1}\right] \\
&= \mathbb{E}_{q(z)}\left[\log\, p(x,\, f_\phi(z_0)) - \log\, q_0(z_0)\right] \\
&\quad + \mathbb{E}_{q(z)}\left[\log \left|\det(\frac{dz}{dz_0})\right|^{-1}\right]
\end{aligned} \tag{33}$$

The challenge is to define f_ϕ such that $\log \left|\det(\frac{df_\phi}{dz_0})\right|$ is easy to compute. Common implementation takes chain simple invertible mappings to allow flexibility.

$$f_\phi = f_K \circ \cdots \circ f_1, f_k(.) = f_{\phi_k}(.), \phi = \phi_k{}_{i=1}^{K}$$

For each simple mapping, we hope that the Jacobian log determinant is fast to compute, such that:

$$\log \left|\det(\frac{df_\phi}{dz_0})\right| = \sum_{k=1}^{K} \log \left|\det(\frac{df_{\phi_k}}{dz_{k-1}})\right|$$

7.3. Auxiliary Variables and Mixture Distributions

We can construct a mixture distribution $q(\theta)$ by introducing an auxiliary distribution $q(a)$ such that:

$$q(\theta) = \int q(\theta|a)q(a)da \tag{34}$$

One of the classic example is the Gaussian mixture distribution (Li, 2020). We start by defining the auxiliary distribution q_a as a categorical distribution: $q \sim q(a) = \text{Cat}(\pi_1, \cdots, \pi_k)$. Conditioned on a, θ is Gaussian distributed: $\theta \sim q(\theta|a) = \mathcal{N}(\theta; m_a, \Sigma_a)$. Here, m and Σ denote the mean and the covariance, respectively. If we have many components of a, we can build a very flexible and accurate approximation distribution. However, performing variational inference with the mixture-distribution q can result into an intractable problem due to the fact that $q(\theta) = \int q(\theta|a)q(a)da$.

The solution is introducing an auxiliary variational lower bound $\mathcal{L}(\phi, r)$ with an auxiliary distribution $r(a|\theta)$ (Agakov & Barber, 2004). Mathematically, we can write:

$$\mathcal{L}(\phi, r) = \mathbb{E}_{q(\theta,a)}[\log p(D|\theta)] - KL[q(\theta, a)\|p(\theta)r(a|\theta)] \tag{35}$$

Figure 11 shows the gap between ELBO and auxiliary variational lower bound. This method aims to optimize $r(a|\theta)$ to close the gap shown in the Figure (11).

7.4. Implicit Approximate Posteriors

In modern method, we rely on Monte Carlo estimation to perform Bayesian inference. This estimation requires fast sampling process from the distribution q. We can replace q with a neural network such that we do not need the analytical form of q anymore (Li, 2020). We derive the variational objective by expressing ELBO as:

$$\mathcal{L}_\phi = \mathbb{E}_{q(\theta)}[\log\, p(D|\theta)] - \mathbb{E}_{q(\theta)}\left[\log \frac{p(\theta)}{q(\theta)}\right] \tag{36}$$

The first term can be solve easily using Monte Carlo estimation. However, since q is now a neural network, the second term becomes intractable. Instead, the research introduces a discriminator that aims to diffrentiate whether θ comes from p or q (Li & Turner, 2017). Figure (12) shows the illustration of implicit approximate posterior.

8. Objective Function Design

In this section, we will discuss the objective function design for fitting the approximate posterior distribution. Variational inference mainly requires three components: the posterior distribution p, the variational distribution q, and the KL divergence $KL[p\,\|\,q]$. We rely on KL-divergence as the criterion to optimize q towards p. However, KL-divergence often suffers from mode seeking property which leads q to underestimate the variance of the true distribution p (Li, 2020). This condition can be dangerous in real world application which requires a calibrated uncertainty. Some researches has been conducted to see the impact of using different divergence.

8.1. Rényi α-Divergence

The first approach is to use α-divergence. Given distribution $p(\theta)$, $q(\theta)$, and the parameter α, the α-divergence can be written as:

$$D_\alpha[p(\theta)\|q(\theta)] = \frac{1}{\alpha - 1} \log \int p(\theta)^\alpha\, q(\theta)^{1-\alpha} d\theta \tag{37}$$

with $\alpha > 0, \alpha \neq 1$. Observe that when $\lim_{\alpha \to 1}$, we have $D_\alpha[p(\theta)\,\|\,q(\theta)] = KL[p(\theta)\,\|\,q(\theta)]$. Expressing $\mathcal{L}_{\text{ELBO}}$ by replacing KL-divergence with α-divergence we obtain variational Rényi bound (Li & Turner, 2016):

$$\begin{aligned}\mathcal{L}_\alpha &= \frac{1}{\alpha - 1}\, \mathbb{E}_{q(\theta)}\left[\left(\log \frac{p(D, \theta)}{q(\theta)}\right)^{1-\alpha}\right] \\ &= \log\, p(D) - D_\alpha\left[q(\theta)\,\|\,p(\theta\,|\,D)\right]\end{aligned} \tag{38}$$

when $lim_{\alpha \to 1}$, we have $\mathcal{L}_\alpha = \mathcal{L}_{\text{ELBO}}$. Therefore, $\mathcal{L}_\alpha$ is a generalization of variational inference. Subsequently, Figure (13) shows the inference result of variational Rényi bound with different α. The black ellipsoid denotes the true posterior, which is a correlated Gaussian distribution. The circles with different colors denote the approximation with different α. We can conclude that different α can affect the approximation results. For example choosing $\alpha \geq 1$ leads to an approximation that underestimates the variance of the true posterior. In general, choosing α too big or too small leads to an exteremely big variance.

8.2. Perturbation in Black-Box Variational Inference (PBBVI)

We start by defining the $\log\, p(x)$ as the expectation over $q_\phi(z)$ of the log density ratio $\frac{p(x,z)}{q_\phi(z)}$. Subsequently, PBBVI (Bamler et al., 2017) introduces a function $V(x, z)$:

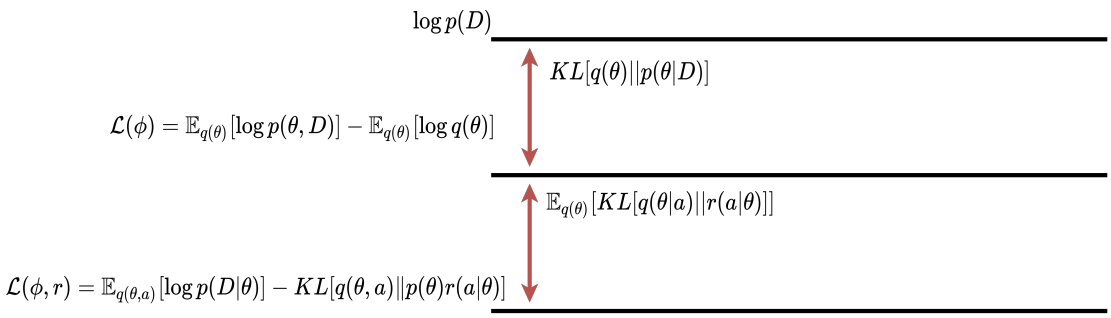

Figure 11. The inference gap as the result of involving auxiliary distribution $r(a \mid \theta)$ (Li, 2020)

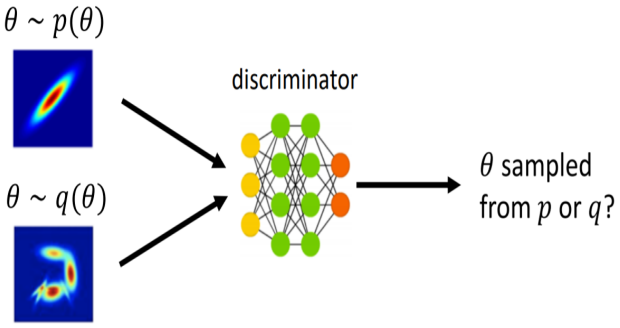

Figure 12. Implicit approximate posterior which relies on a discriminator (Li & Turner, 2017). The discriminator is responsible to differentiate whether θ comes from p or q.

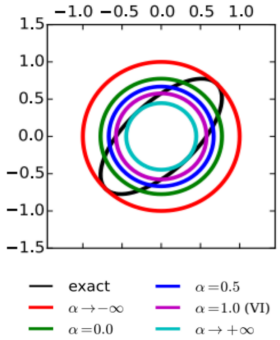

Figure 13. The approximation results of variational Rényi bound with different α (Li & Turner, 2016)

$$V(x, z) \equiv \log\, q_\phi(z) - \log\, p(x, z) \tag{39}$$

Combining the definition of $\log\, p(x)$ with Equation (39), we can write $\log\, p(x)$ as follows:

$$\begin{aligned} \log\, p(x) &= \log \left(\mathbb{E}_{q_\phi(z)} \left[\frac{p(x, z)}{q_\phi(z)} \right] \right) \\ &= \log \mathbb{E}_{q_\phi(z)} \left[\exp(-\beta V(x, z)) \right] \end{aligned} \tag{40}$$

Note that we have an auxiliary parameter β in Equation (41). If we set $\beta = 1$, then the equation holds. Now, we aim to take the Taylor expansion of $\log\, p(x)$ around $\beta = 1$:

$$\begin{aligned} \log\, p(x) \approx \mathbb{E}_{q_\phi}[-V] + \frac{1}{2} \left[(V - \mathbb{E}_{q_\phi}[-V])^2 \right] + \\ - \frac{1}{3!} \left[(V - \mathbb{E}_{q_\phi}[-V])^3 \right] + \cdots \end{aligned} \tag{41}$$

In particular, if we truncate the expansion at the first term, we are obtaining the familiar ELBO:

$$\mathbb{E}_{q_\phi(z)}[-V(x, z)] = \mathbb{E}_{q_\phi(z)}[\log\, p(x, z) - \log\, q_\phi(z)] \tag{42}$$

Generally, truncating the Taylor expansion in the odd term provides a lower bound approximation of the model evidence $p(x)$. Figure (14) shows the comparison of PBBVI approximation, with α-divergence. It turns out if we truncate in a higher order e.g. 3, PBBVI results a better approximation comparing to standard variational inference, which is equivalent to truncating the expansion in one term. Results also show that PBBVI has a better bias-variance trade-off comparing to α-divergence. The remain challenge is to find the ideal position to truncate the expansion.

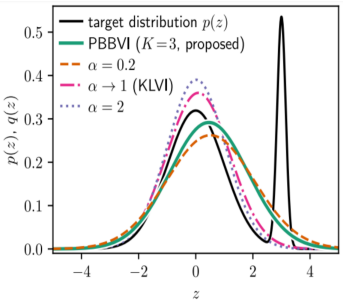

Figure 14. The approximation results of PBBVI and α-divergence.

8.3. f-Divergence

There is a more flexible divergence than α-divergence called f-divergence (Wan et al., 2020). Mathematically, we can write f-divergence $D_f[p(\theta) \,\|\, q_\phi(\theta)]$ as:

$$D_f[p(\theta) \,\|\, q_\phi(\theta)] = \mathbb{E}_{q_\phi(\theta)}\left[f(\frac{p(\theta)}{q_\phi(\theta)}) - f(1)\right] \tag{43}$$

Here, we can choose function f to be any convex function. Different f yields different divergence.

$$f(t) = -\log t \to KL[q \,\|\, p]$$
$$f(t) = t \log t \to KL[p \,\|\, q]$$
$$f(t) = \frac{t^\alpha}{\alpha(\alpha-1)} \to D_\alpha[p \,\|\, q]$$

8.4. Integral Probability Metric (IPM)

Integral probability metric requires a test function f to describe the difference between the variational distribution $q(z)$ and the true posterior $p(z|x)$ (Li, 2020). Specifically, IPM aims to find the best test function $f^* \in \mathcal{F}$ that maximize the difference between $q(z)$ and $p(z|x)$. Mathematically, we can write:

$$D[q(z),\, p(z \,|\, x)] = \sup_{f \in \mathcal{F}} \left|\mathbb{E}_{q(z)}[f(z)] - \mathbb{E}_{p(z \,|\, x)}[f(z)]\right| \tag{44}$$

different test function f defines a different integral probability metric. One of the special case of integral probability metric is Stein discrepancy (Liu & Wang, 2016). Stein discrepancy only requires sample $z \sim p(z)$ and the score function of the posterior, written as:

$$\nabla_z \log p(z \,|\, x) = \nabla_z \log p(z, x) \tag{45}$$

Finally, we can write Stein discrepancy as:

$$S[q(z),\, p(z \,|\, x)]$$
$$= \sup_{f \in F} \left|\mathbb{E}_{q(z)}[\nabla_z \log p(z, x)^T f(z) + \nabla_z^T f(z)]\right| \tag{46}$$

Stein discrepancy has been applied not only for approximate inference, but also for goodness of fit and fitting the entropy model.

9. Conclusion

In this paper, we review the definition of various form of variational inference method. Specifically, we review variational inference, mean-field variational inference, stochastic variational inference, and black-box variational inference. Later, we also discuss recent techniques to improve the approximation result. In general, we can improve the approximation result by either designing a new variational distribution or designing a new objective function.

References

Agakov, F. V. and Barber, D. An auxiliary variational method. In *International Conference on Neural Information Processing*, pp. 561–566. Springer, 2004.

Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Bamler, R., Zhang, C., Opper, M., and Mandt, S. Perturbative black box variational inference. *arXiv preprint arXiv:1709.07433*, 2017.

Bishop, C. M. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007. ISBN 9780387310732. URL https://www.worldcat.org/oclc/71008143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *CoRR*, abs/1601.00670, 2016. URL http://arxiv.org/abs/1601.00670.

Cremer, C., Li, X., and Duvenaud, D. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, pp. 1078–1086. PMLR, 2018.

Golinski, A., Wood, F., and Rainforth, T. Amortized monte carlo integration. In *International Conference on Machine Learning*, pp. 2309–2318. PMLR, 2019.

Gu, S., Levine, S., Sutskever, I., and Mnih, A. Muprop: Unbiased backpropagation for stochastic neural networks. *arXiv preprint arXiv:1511.05176*, 2015.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Hogan, J. M. Advanced mean field methods: theory and practice: M. opper and d. saad (eds.); MIT press, cambridge, ma, 2001, 300pp. ISBN 0-262-15054-9. *Neurocomputing*, 48(1-4):1057–1060, 2002. doi: 10.1016/S0925-2312(02)00612-4. URL https://doi.org/10.1016/S0925-2312(02)00612-4.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Le, T. A., Igl, M., Rainforth, T., Jin, T., and Wood, F. Auto-encoding sequential monte carlo. *arXiv preprint arXiv:1705.10306*, 2017.

Li, Y. *Topics in Approximate Inference*. 2020. URL http://yingzhenli.net/home/pdf/topics_approx_infer.pdf.

Li, Y. and Turner, R. E. Rényi divergence variational inference. arxiv e-prints, page. *arXiv preprint arXiv:1602.02311*, 2016.

Li, Y. and Turner, R. E. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

Li, Y., Turner, R. E., and Liu, Q. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

Marino, J., Yue, Y., and Mandt, S. Iterative amortized inference. In *International Conference on Machine Learning*, pp. 3403–3412. PMLR, 2018.

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. Variational sequential monte carlo. In *International conference on artificial intelligence and statistics*, pp. 968–977. PMLR, 2018.

Paisley, J., Blei, D., and Jordan, M. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822. PMLR, 2014.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Wan, N., Li, D., and Hovakimyan, N. f-divergence variational inference. *Advances in Neural Information Processing Systems*, 33, 2020.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Zhang, C., Kjellstrom, H., and Mandt, S. Determinantal point processes for mini-batch diversification. *arXiv preprint arXiv:1705.00607*, 2017.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):2008–2026, 2019. doi: 10.1109/TPAMI.2018.2889774. URL https://doi.org/10.1109/TPAMI.2018.2889774.