ADVANCES IN VARIATIONAL INFERENCE

A PREPRINT

Spandan Senapati
Department of Computer Science and Engineering
IIT Kanpur
spandans@iitk.ac.in

Avinandan Bose
Department of Computer Science and Engineering
IIT Kanpur
avibose@iitk.ac.in

Naman Biyani
Department of Computer Science and Engineering
IIT Kanpur
namanb@iitk.ac.in

July 9, 2019

ABSTRACT

The focus of the project will be on an important aspect of approximate inference in probabilistic machine learning namely Variational Inference and the advances in the feild. The algorithm begins with a trivial assumption similar to what is done in statistical physics by the Mean Feild Theory assumption.Advances in Variational inference include Black Box VI,Stochastic VI,Armotixed VI,Autoencoding VI etc.

1 Introduction

1.1 Variational Inference

Variational Inference and Markov Chain Monte Carlo Stimulation(MCMC)form the two most commonly used means of approximate inference.In MCMC we construct a Markov Chain over the hidden variables whose stationary distribution is the posterior of interest.We run the chain until it has (hopefully) reached equilibrium and collect samples to approximate the poxterior.In Variational Inference, we define a flexible family of distributions over the hidden variables, indexed by free parameters.We then find the setting if the parameters that is closest to the posterior,by minimising the Kullback-Leibler Divergence.

1.2 Mean Field Theory

The simplest variational family of distributions to work with is the *Mean Field Variational Family*,wherein each hidden variable is independent and governed by its own parameter.In mathematical terms

$$\prod_{i=1}^{N} q(z_i) = q(z) \tag{1}$$

. Assuming variables are governed by parameters 1 simplifies to

$$q(z) = \prod_{i=1}^{N} \prod_{j=1}^{J} q(z_{nj}|\phi_{nj}) \tag{2}$$

2 Optimisation

2.1 Evidence Lower Bound(ELBO)

Variational Inference minimises the KL divergence from the variational distribution to the posterior distribution.It thereby maximised the *evidence lower bound*(ELBO),a lower bound on the marginal probability.This is easy to derive using Jensen's inequality $\log \mathbb{E}[f(y)] \geq \mathbb{E}[\log f(y)]$.This gives the bound on the log marginal as

$$\begin{aligned}
\log p(x) &= \log \int p(z,x)dz \\
&= \log \int p(z,x)\frac{q(z)}{q(z)}dz \\
&= \log \left(\mathop{\mathbb{E}}_{q(z)} \left[\frac{p(x.z)}{q(z)} \right] \right) \\
&= \mathop{\mathbb{E}}_{q(z)} \Big[\log p(z,x) \Big] - \mathop{\mathbb{E}}_{q(z)} \Big[\log q(z) \Big] \\
&\equiv \mathcal{L}(q)
\end{aligned}$$

Further it is easy to prove by some trivial manipulation and by introducing the KL divergence the relation between the ELBO and KL divergence as,

$$\log p(x) = \mathcal{L}(q) + \mathrm{KL}(q(z)||p(z|x)) \tag{3}$$

In cases where the posterior distribution becomes intracable we introduce the variational family and optimise the ELBO with by using a step wise gradien ascent by updating each member of the family.The update rules can be obtained by considering the form of the ELBO for a particular member j as follows

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_{i=1}^{N} \left\{ \ln p(\mathrm{X,Z}) - \sum_{i=1}^{N} \ln q_i \right\} d\mathrm{Z} \\
&= \int q_j \left\{ \int \ln p(\mathrm{X,Z}) \prod_{i \neq j} q_i d\mathrm{Z}_i \right\} d\mathrm{Z}_j - \int q_j \ln q_j d\mathrm{Z}_j + \mathrm{const} \\
&= \int q_j \ln \hat{p}(\mathrm{X,Z}_j) d\mathrm{Z}_j - \int q_j \ln q_j d\mathrm{Z}_j + \mathrm{const}
\end{aligned}$$

where we have defined a new distribution

$$\hat{p}(\mathrm{X,Z}_j) = \mathop{\mathbb{E}}_{i \neq j} \Big[\ln p(\mathrm{X,Z}) \Big]$$

The optimal solution is obtained in the usual way by minimising the KL Divergence.The new distribution of the family hence becomes,

$$q_j^* = \mathop{\mathbb{E}}_{i \neq j} \Big[\ln p(\mathrm{X,Z}) \Big] \tag{4}$$

This is repeated until the ELBO converges to obatain a distribution closest to the posterior.

3 Black Box Variational Inference

3.1 Principle

Black Box VI can work with small minibatches of data rather than the entire dataset thereby increasing the efficiency of the VI algorithm.It uses a Monte Carlo Stimulation in evaluating the gradients of the ELBO as ,

$$\nabla_\phi \mathcal{L}(q) \approx \frac{1}{S}\sum_{i=1}^{S} \nabla_\phi \log q(\text{Z}_\phi)(\log p(\text{X,Z}_s) - \log q(\text{Z}_s|\phi)) \tag{5}$$

3.2 BBVI Identity

$$\nabla_\phi \mathcal{L}(q) = \mathbb{E}\left[\nabla_\phi \log q(\text{Z}|\phi)(\log p(\text{X,Z}) - \log q(\text{Z}|\phi))\right] \tag{6}$$

Since the required gradients are only a function of $q((Z|\phi))$ and not on the model it is called as Black Box VI.

3.3 Proof of the Identity

Using the Dominated Convergence Theorem,

$$\begin{aligned}\nabla_\phi \mathcal{L}(q) &= \nabla_\phi \int \left[\log q(\text{Z}|\phi)(\log p(\text{X,Z}) - \log q(\text{Z}|\phi))\right] \\ &= \mathop{\mathbb{E}}_q\left[-\nabla_\phi \log q(\text{Z}|\phi)\right] + \int \nabla_\phi q(\text{Z}|\phi)\left[\log p(\text{X,Z}) - \log q(\text{Z}|\phi)\right] \\ &= \mathbb{E}\left[\nabla_\phi \log q(\text{Z}|\phi)(\log p(\text{X,Z}) - \log q(\text{Z}|\phi))\right] && \left(\because \mathbb{E}_q\left[\nabla_\phi \log q(\text{Z}|\phi)\right] = 0\right) \\ &\approx \frac{1}{S}\sum_{i=1}^{S} \nabla_\phi \log q(\text{Z}_\phi)(\log p(\text{X,Z}_s) - \log q(\text{Z}_s|\phi))\end{aligned}$$

3.4 Inferences

1. BBVI allows VB inference for a wide variety of probabilistic models
2. Few Requirements
   - Should be able to sample from $q(\text{Z}|\phi)$
   - Should be able to compute the gradient $\nabla_\phi \log q(\text{Z}|\phi)$
   - Should be able to compute $p(\text{X,Z})$ and $\log q(\text{Z}|\phi)$
3. Reparametrization Trick to control Variance in the Monte Carlo estimate of gradient

4 Reparametrization Trick

1. LOTUS(Law of Unconscious Statistician) $\mathbb{E}(g(x)) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$

4.1 Example

Consider a Normal distribution given by $q \equiv \mathcal{N}(\theta, 1)$ and we wish to evaluate $\zeta = \nabla_\theta\, \mathbb{E}[x^2]$

$$\begin{aligned}\zeta &= \int x^2 \nabla_\theta q_\theta(x) dx = \int x^2 q_\theta(x) \nabla_\theta \log q_\theta(x) dx \\ &= \mathbb{E}\Big[x^2 \nabla_\theta \log q_\theta(x)\Big] \\ &= \mathop{\mathbb{E}}_{q_\theta(x)}\Big[x^2(x-\theta)\Big] && (\text{Assume } x \approx \theta + \epsilon \text{ where } \epsilon \equiv \mathcal{N}(0,1)) \\ &= \mathop{\mathbb{E}}_p\Big(2\theta + 2\epsilon\Big)\end{aligned}$$

Now the two distributions differ in Variance.

4.2 Use in BBVI

Numerical estimation of ELBO's gradient.Under a transformation $\mathrm{Z} = g(\epsilon, \phi)$ with $\epsilon = p(\epsilon)$ ELBO's gradient under Monte Carlo Stimulation can be written as

$$\nabla_\phi \underset{q_\phi(z)}{\mathbb{E}}(\log p(\mathrm{X},\mathrm{Z}_s) - \log q(\mathrm{Z}_s|\phi)) \approx \frac{1}{S}\sum_{s=1}^{S}\Big[\nabla_\phi \log p(\mathrm{X}, g(\epsilon_s, \phi)) - \nabla_\phi \log q_\phi(g(\epsilon_s, \phi))\Big]$$

Such a gradient is called *Pathwise Gradient*

4.3 Inferences

Limitations

- Isn't often applicable e.g when Z is discrete or categorical
- Transformation function g may be difficult to find for general distributions
- Transformation function g needs to be invertible
- Assumption of direct sampling from $p(\epsilon)$