Auto-Encoding Variational Bayes

Yeseul Jeon

July 23, 2021

Contents

Paper

Auto-Encoding Variational Bayes

Diederik P. Kingma
Machine Learning Group
Universiteit van Amsterdam
dpkingma@gmail.com

Max Welling
Machine Learning Group
Universiteit van Amsterdam
welling.max@gmail.com

Abstract

How can we perform efficient inference and learning in directed probabilistic models, in the presence of continuous latent variables with intractable posterior distributions, and large datasets? We introduce a stochastic variational inference and learning algorithm that scales to large datasets and, under some mild differentiability conditions, even works in the intractable case. Our contributions is two-fold. First, we show that a reparameterization of the variational lower bound yields a lower bound estimator that can be straightforwardly optimized using standard stochastic gradient methods. Second, we show that for i.i.d. datasets with continuous latent variables per datapoint, posterior inference can be made especially efficient by fitting an approximate inference model (also called a recognition model) to the intractable posterior using the proposed lower bound estimator. Theoretical advantages are reflected in experimental results.

Goal; Problem Settings

- **Density Inference**
  - Efficient approximate inference and learning with directed probabilistic model
  - But have intractable posterior distributions?
  - **Intractable**; Requires lot of integration or normalizing constant contains parameters

Ways to estimate posterior distributions?

- **Markov chain Monte Carlo**
  - Sample and update based on likelihood ratio
- **Variational Inference**
  - Estimate $q_\phi(x)$; Indirect way

Given conditions

- **Dataset**

  - $X = (x_i)_{i=1}^N$ consisting of N i.i.d samples of some continuous or discrete variable x

  - Assumption: Data are generated by some random process, involving an unobserved continuous random variable **z**

  - Prior distribution $p_\theta(z)$ and $p_\theta(x|z)$

- **Unknown**

  - True parameter θ

  - Latent variables z_i

Limitations:

- **Intractability**
  - Ture posterior density $p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$ is intractable
  - complicated likelihood functions $p_\theta(x|z)$; e.g a neural network with a nonlinear hidden layer
- **A large dataset**
  - expensive sampling loop cost

Scenario

- Efficient approximate ML or MAP estimation for the parameters θ
- Efficient approximate posterior inference of the latent variable z given an observed value x for a choice of parameters θ.
- Efficient approximate marginal inference of the variable x.

Solving problems scenario

- Random variable z
- Encoder: recognition model $q_\phi(z|x)$ (Posterior)
- Decoder: $p_\theta(z|x)$, z produces a distribution over the possible corresponding values of x. (Likelihood)
- Update using gradient descent

Variational bound

- Marginal likelihood
  - $\log p_\theta(x_1, x_2, ..., x_N) = \sum \log p_\theta(x_i)$

$$\log p_\theta(x_i) = D_{KL}(q_\phi(z|x_1)||p_\theta(z|x_i)) + L(\theta, \phi; x_i) \tag{1}$$

- KL-divergence is non-negative
- (variational) *lower bound* on the marginal likelihood of datapoint i

$$L(\theta, \phi; x_i) = E_{q_\phi(z|x)}[-\log q_\phi(z|x) + \log p_\theta(x, z)] \tag{2}$$

- Optimize the lower bound (ELBO) w.r.t ϕ, θ

SGVB estimator

$$L(\theta, \phi; x_i) = -D_{KL}(q_\phi(z|x_i)||p_\theta(z)) + E_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] \tag{3}$$

- **Impractical Issue**
  - Problem about p_θ and $q_\phi(z|x_i)$
  - Gradient of the lower bound w.r.t. ϕ is problematic
  - Sampling is not differential

Reparameterization trick

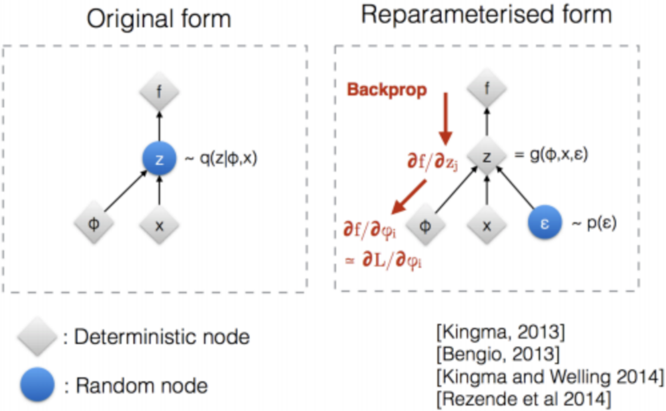
- Random variable z as a deterministic variable $z = q_\phi(\epsilon, x)$
- ϵ is an auxiliary variable with independent marginal $p(\epsilon)$
- Given the deterministic mapping $z = q_\phi(\epsilon, x)$, the MC estimate of the expectation is differentiable w.r.t. ϕ
- some function, $f(x)$

$$E_{q_\phi(z|x_i)}[f(z)] = E_{p(\epsilon)}[f(g_\phi(\epsilon, x_i)] = \frac{1}{L}\sum f(g_\phi(\epsilon_i, x_i)) \tag{4}$$

$$\tilde{L}(\theta, \phi; x_i) = -D_{KL}(q_\phi(z|x_i)||p_\theta(z)) + \frac{1}{L}\sum(\log p_\theta(x_i|z_{i,l}))) \tag{5}$$

where $z_{i,l} = g_\phi(\epsilon_{i,l}, x_i)$

Reparameterization trick



[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

Reparameterization trick

- Tractable Inverse CDF: $g_\phi(\epsilon_{i,l}, x_i)$ be the inverse CDF of $q_\phi(z|x)$
- Gaussian example for any "location-scale" family: Laplace, Elliptical, Student's t, Logistic
- Composition: Log-Normal, Gamma, ...

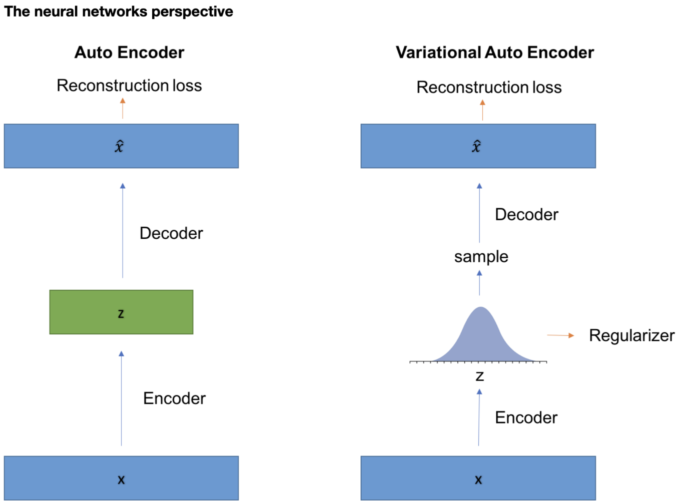
Variational Auto-Encoder

- Let the variational approximate posterior $q_\theta(z|x_i)$ be a multivariate Gaussian with a diagonal covariance structure
- Prior z, $p_\theta(z)$ and $q_\theta(z|x)$ are Gaussian

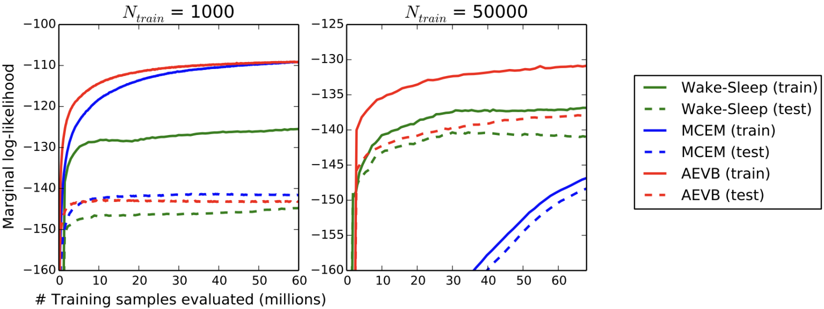
$$\tilde{L}(\theta, \phi; x_i) \approx \frac{1}{2}\sum_{j=1}^{J}(1 + \log((\sigma_i^j)^2) - (\mu_i^j)^2 - (\sigma_i^j)^2) + \frac{1}{L}\sum(\log p_\theta(x_i|z_{i,l}))) \tag{6}$$

where $z_{i,l} = \mu_i + \sigma_i \odot \epsilon_l$

Variational Auto-Encoder



Experiments Result



Conclusions

- Auto Encoder using Variational Inference (reparametrization trick)
- Using the Latent variable
  - Explainable AI
  - Detection using latent distribution of z
- How about combining spatial information?
- How about ignoring i.i.d assumption?