Density Ratio Estimation in Variational Bayesian Machine Learning

Alexander Lam

Supervised by Prof. Scott Sisson and Doctor Edwin Bonilla

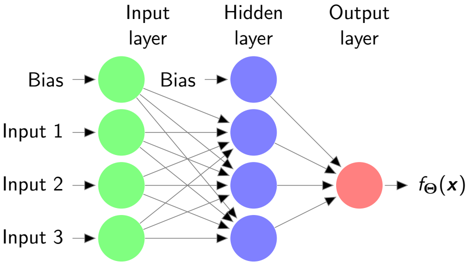
Statistics Honours, 2018

Outline

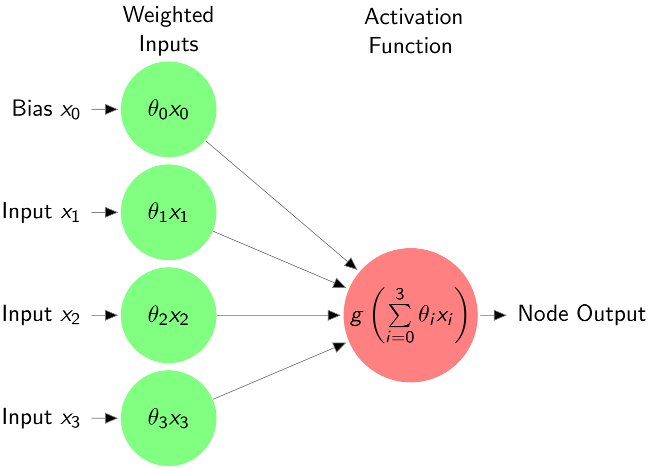
1. Background Information
   - Neural Networks
   - (Amortized) Variational Inference
   - Density Ratio Estimation
2. Undertrained Estimator Experiment
3. Autoencoder Experiment

Neural Networks
Overall Structure

- Objective is to approximate a function f^* using mapping with parameters $\boldsymbol{\Theta}$: $\mathbf{f}_{\boldsymbol{\Theta}}(\boldsymbol{x})$.
- Universal Approximation Theorem states a neural network can **approximate (almost) any function** if it is complex enough.
- Each node output is a transformed, weighted sum of previous node outputs.



Neural Networks
Individual Node



Neural Networks

Training

- Weights trained such that (ideally convex) loss function is minimized e.g. Mean Squared Error: $\min_\Theta \frac{1}{2}||\boldsymbol{y} - \boldsymbol{f}_\Theta(\boldsymbol{x})||_2^2$.
- Back-propagation finds partial derivatives of loss function with respect to weights.
- **Gradient descent** uses these partial derivatives to optimize network.

(Amortized) Variational Inference

Bayesian Inference

- Fundamental problem in Bayesian computation is to **estimate posterior densities** $p(z|x)$:

$$p(z|x) \propto \underbrace{p(z)}_{\text{Prior}} \underbrace{p(x|z)}_{\text{Likelihood}} .$$

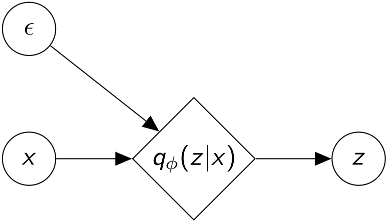
- Typical MCMC methods are slow with high dimensional data or large datasets.

- (Amortized) Variational Inference is a solution.

(Amortized) Variational Inference

Introduction

- Amortized variational inference approximates $p(z|x)$ with a different density $q_\phi(z|x)$.
- $q_\phi(z|x)$ is a **neural network** with parameters ϕ.
- Random noise ϵ makes the network probabilistic.



- Train network by minimizing the reverse KL divergence:

$$KL(q_\phi(z|x)\|p(z|x)) := \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(z|x)}\right].$$

- This is the same as solving:

$$\min_\phi \mathbb{E}_{q^*(x)}[\underbrace{-\mathbb{E}_{q_\phi(z|x)}[\log p(x|z)]}_{\text{Likelihood}} + \underbrace{KL(q_\phi(z|x)||p(z))}_{\text{Log Density Ratio}}].$$

- We call this $NELBO(q)$ as it is the **n**egative of **e**vidence **l**ower **bo**und.

$$NELBO(q) \geq -\log p(x)$$

- $q^*(x)$ is the density of the dataset.

Consider our log density ratio term

$$KL(q_\phi(z|x)||p(z)) = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(z)}\right].$$

- $q_\phi(z|x)$ is **implicit**.
- Use density ratio estimation to evaluate $\frac{q_\phi(z|x)}{p(z)}$ in $KL(q_\phi(z|x)||p(z))$.
- Density ratio estimation only requires **samples**.

- There exist different methods of estimating a density ratio $\frac{q(u)}{p(u)}$.
- Many of them use a neural network $r_\alpha(u) \simeq \frac{q(u)}{p(u)}$.
- Various different loss functions used to train network.

Example

The following loss function

$$\min_\alpha -\mathbb{E}_{q(u)}[\log D_\alpha(u)] - \mathbb{E}_{p(u)}[\log(1 - D_\alpha(u))]$$

trains a network $D_\alpha(u)$ estimating $\frac{q(u)}{q(u)+p(u)}$.

$\simeq$ is equality when the estimator is optimal, e.g. $D^*_\alpha(u) = \frac{q(u)}{q(u)+p(u)}$.

Theorem (Nguyen, 2010, Adapted)

If f is a convex function with derivative f' and convex conjugate f^*, and $r_\alpha(u)$ is a neural network, then we have the lower bound for the f-divergence between densities $p(u)$ and $q(u)$:

$$D_f[p(u)||q(u)] \geq \sup_\alpha \{\mathbb{E}_{q(u)}[f'(r_\alpha(u))] - \mathbb{E}_{p(u)}[f^*(f'(r_\alpha(u)))]\},$$

with equality when $r_\alpha(u) = q(u)/p(u)$.

Example

For the reverse KL divergence, we have:

$$KL[q(u)||p(u)] \geq \sup_\alpha \{\mathbb{E}_{q(u)}[1 + \log r_\alpha(u)] - \mathbb{E}_{p(u)}[r_\alpha(u)]\}.$$

- Apply theorem to generalise density ratio estimator loss functions.
- Choose ***f*-divergence bound**:
  - Reverse KL Divergence
  - GAN Divergence

  and **estimator parametrisation**:
  - Direct Ratio Estimator: $r_\alpha(u) \simeq \frac{q(u)}{p(u)}$
  - Class Probability Estimator: $D_\alpha(u) = \frac{r_\alpha(u)}{r_\alpha(u)+1} \iff \frac{q(u)}{p(u)} \simeq \frac{D_\alpha(u)}{1-D_\alpha(u)}$
  - Direct Log Ratio Estimator: $T_\alpha(u) = \log r_\alpha(u) \iff \frac{q(u)}{p(u)} \simeq \mathrm{e}^{T_\alpha(u)}$

Recap

To train a variational posterior network:

1. Train estimator network until convergence.
2. Use estimator network to calculate intractable term in $NELBO$.
3. Take one optimisation step of posterior network.
4. Repeat until posterior convergence.

$$(z_1, z_2) \sim \mathcal{N}(0, \sigma^2 I_{2\times 2})$$

$$x|\boldsymbol{z} \sim Exp(3 + \max(0, z_1)^3 + \max(0, z_2)^3)$$

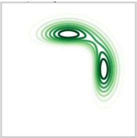
- Posterior is unusually-shaped and bimodal.
- Use Gaussian KDE to find 'true' KL divergence for $q_\phi(z|x = 0, 5, 8, 12, 50)$.

- In a previous experiment we found that all three estimator parametrisations lead to similar results when optimised effectively.
- What if they are poorly optimised?
- Training parameters:
  - High posterior training rate.
  - Low estimator training rate.
  - Low estimator to posterior iteration ratio (11:1).

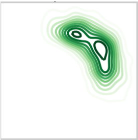
Undertrained Estimator Experiment

Results

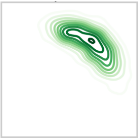
Algorithm		Mean KL Divergence	Standard Deviation
Reverse KL	$D_\alpha(z,x)$	**1.3786**	**0.0286**
	$r_\alpha(z,x)$	1.3934	0.0410
	$T_\alpha(z,x)$	1.4133	0.0597
GAN	$D_\alpha(z,x)$	1.4017	**0.0286**
	$r_\alpha(z,x)$	1.4086	0.0555
	$T_\alpha(z,x)$	1.4214	0.0518



True



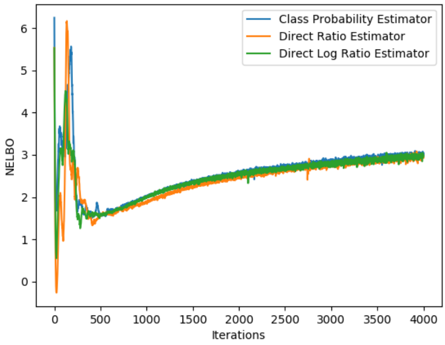
Reverse KL



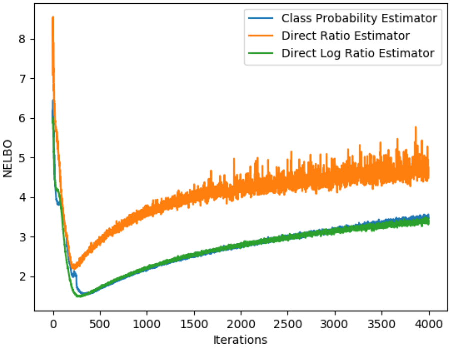
GAN

- Reverse KL divergence better than GAN divergence.
- $D_\alpha(z,x) < r_\alpha(z,x) < T_\alpha(z,x)$

Undertrained Estimator Experiment
NELBO Plots



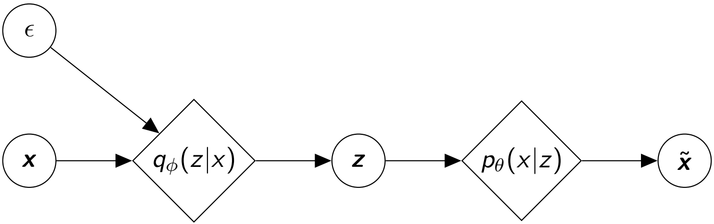
(a) Reverse KL Divergence



(b) GAN Divergence

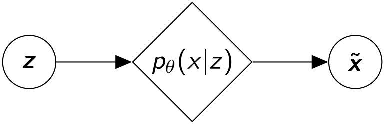
- Reverse KL Divergence has initial instability.

- Posterior $q_\phi(z|x)$ 'compresses' data x into z.
- Likelihood $p_\theta(x|z)$ 'reconstructs' data $\tilde{x}$ from z.



$$\min_{\theta,\phi} \mathbb{E}_{q^*(x)}[\underbrace{-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{Likelihood}} + \underbrace{KL(q_\phi(z|x)||p(z))}_{\text{Log Density Ratio}}]$$

- Generate z from $p(z)$.
- Typically $p(z)$ is $\mathcal{N}(0, I)$.

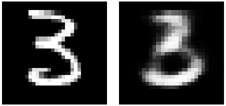


- MNIST dataset - 28 $\times$ 28 grey-scale images of handwritten digits
- Again use undertrained estimator.
- Use reconstruction error $\|x - \tilde{x}\|^2$ as metric.

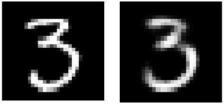
Generation Experiment

Results - 20-dimensional latent space

Algorithm	Mean Reconstruction Error	Standard Deviation
Reverse KL - $D_\alpha(z,x)$	0.0647	0.0191
GAN - $D_\alpha(z,x)$	**0.0444**	**0.0017**



Reverse KL

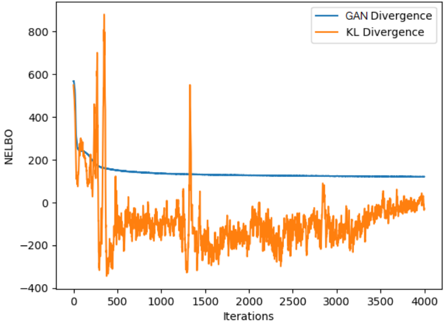


GAN

- Density ratios too big for direct ratio and log ratio estimators.
- Exponential of $T_\alpha(z,x)$ taken in loss function.
- $D_\alpha(z,x)$ ranges in $(0,1)$.

Autoencoder Experiment

NELBO plot



- Reverse KL divergence fails to stabilise by the end of runtime.

Summary

- The class probability estimator $D_\alpha(u) \simeq \frac{q(u)}{q(u)+p(u)}$ is the 'best' parametrisation as it can store the highest density ratios.
- Reverse KL divergence upper bound may be unstable but leads to faster convergence when stable.
- Future Research
  - Still unclear exactly why reverse KL divergence is more unstable but more accurate when stable.
  - Several more f-divergences exist which have unknown stability when undertrained.
  - Alternate density ratio estimation algorithms e.g. denoisers, k-nearest neighbours

For Further Reading I

M. Sugiyama, T. Suzuki, T. Kanamori
Density Ratio Estimation in Machine Learning.
Cambridge University Press, 2012.

D.M. Blei, A. Kucukelbir, D.J. McAuliffe
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518):859-877 2017.

F. Huszar
Variational Inference Using Implicit Distributions.
ArXiv e-prints, 2017.