

# Density Ratio Estimation in Variational Methods in Bayesian Neural Networks

Alexander Lam

Department of Mathematics and Statistics  
UNSW

Statistics Honours, 2018

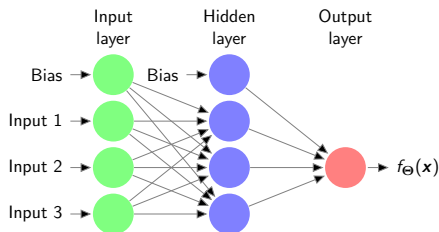
# Outline

- 1 Background Info
  - Neural Networks
  - (Amortized) Variational Inference
  - Density Ratio Estimation
- 2 Activation Function Experiment
- 3 Optimal Estimator Experiment
- 4 Undertrained Estimator Experiment
- 5 Autoencoder Experiment
- 6 Theory

# Neural Networks

## Overall Structure

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



# Neural Networks

## Training

- Weights trained such that (ideally convex) loss function is minimized  
e.g. Mean Squared Error:  $\min_{\Theta} \frac{1}{2} \|\mathbf{y} - \mathbf{f}_{\Theta}(\mathbf{x})\|_2^2$ .
- 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  $p(z|x)$ :

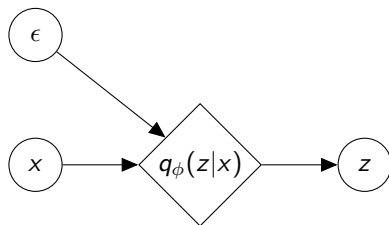
$$p(z|x) \propto \underbrace{p(z)}_{\text{Prior}} \underbrace{p(x|z)}_{\text{Likelihood}} .$$

- Typical MCMC methods are slow with large datasets or high dimensional data.
- Variational Inference is a solution.

# (Amortized) Variational Inference

## Introduction

- Amortized variational inference approximates  $p(z|x)$  with a different distribution  $q_\phi(z|x)$ .
- $q_\phi(z|x)$  is a neural network with parameters  $\phi$  that takes in data  $x$  and random noise  $\epsilon \sim \pi(\epsilon)$  and outputs samples  $z \sim q_\phi(z|x)$ .
- Typically  $\pi(\epsilon) = \mathcal{N}(0, I_{n \times n})$ .



# (Amortized) Variational Inference

## Network Training

- Minimize the **negative** of the **evidence lower bound**  $NELBO(q)$ :

$$\min_{\phi} NELBO(q) = -\mathbb{E}_{q_{\phi}(z|x)q^{*}(x)}[\log p(x|z)] + \mathbb{E}_{q^{*}(x)}[KL(q_{\phi}(z|x)||p(z))].$$

- This is the same as minimizing the reverse KL divergence between the two distributions:

$$\mathbb{E}_{q^{*}(x)}[KL(q(z|x)||p(z|x))] = \mathbb{E}_{q^{*}(x)q(z|x)} \left[ \log \left( \frac{q(z|x)}{p(z|x)} \right) \right]$$

- Taking expectation with respect to dataset distribution  $q^{*}(x)$  allows model to work for different data points.

# (Amortized) Variational Inference

## Problems with Implicit Distributions

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

- $q_{\phi}(z|x)$  is a neural network so extremely difficult to evaluate density function but easy to draw samples, we therefore say that it is **implicit**.
- Use density ratio estimation to evaluate  $\frac{q_{\phi}(z|x)}{p(z)}$  in  $KL(q_{\phi}(z|x)||p(z))$ .
- The prior  $p(z)$  can also be implicit.
- We call this the “prior-contrastive” formulation.
- “Joint-contrastive” formulation used with implicit likelihood minimizes  $NELBO(q) = \underbrace{KL(q(z, x)||p(z, x))}_{\text{Log Density Ratio}}$ .



# Density Ratio Estimation

## Class Probability Estimation

We want to estimate  $\frac{q(u)}{p(u)}$ .

- 1 Define discriminator function that finds probability that a sample  $u$  came from  $q(u)$ :  $D_\alpha(u) \simeq P(u \sim q(u))$ , so that  $\frac{q(u)}{p(u)} \simeq \frac{D_\alpha(u)}{1-D_\alpha(u)}$ .
- 2  $D_\alpha(u)$  is neural network parametrised by  $\alpha$ , sigmoid activation function used for output layer
- 3 Train discriminator with Bernoulli loss:  
$$\min_{\alpha} -\mathbb{E}_{q(u)}[\log D_\alpha(u)] - \mathbb{E}_{p(u)}[\log(1 - D_\alpha(u))].$$
- 4 Optimal discriminator is  $D_\alpha^*(u) = \frac{q(u)}{q(u)+p(u)}$ .

# Density Ratio Estimation

## Class Probability Estimation

Prior-Contrastive Application:

$$\min_{\alpha} -\mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log D_{\alpha}(z, x)] - \mathbb{E}_{q^*(x)p_{\theta}(z)}[\log(1 - D_{\alpha}(z, x))]$$
$$\min_{\phi} \underbrace{-\mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log p(x|z)]}_{\text{Likelihood}} + \underbrace{\mathbb{E}_{q^*(x)q_{\phi}(z|x)} \left[ \log \frac{D_{\alpha}(z, x)}{1 - D_{\alpha}(z, x)} \right]}_{\text{Log Density Ratio}}$$

Joint-Contrastive Application:

$$\min_{\alpha} -\mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log D_{\alpha}(z, x)] - \mathbb{E}_{p(z)p(x|z)}[\log(1 - D_{\alpha}(z, x))]$$
$$\min_{\phi} \mathbb{E}_{q^*(x)q_{\phi}(z|x)} \log \frac{D_{\alpha}(z, x)}{1 - D_{\alpha}(z, x)}$$

Program alternates between several optimisation steps of discriminator and one optimisation step of posterior.

# Density Ratio Estimation

## Divergence Minimisation

### Theorem

*If  $f$  is a convex function with derivative  $f'$  and convex conjugate  $f^*$ , and  $\mathcal{R}$  is a class of functions with codomains equal to the domain of  $f'$ , then we have the lower bound for the  $f$ -divergence between distributions  $p(u)$  and  $q(u)$ :*

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

*with equality when  $r(u) = q(u)/p(u)$ .*

For the reverse KL divergence,  $f(u) = u \log u$  so we have

$$KL[q(u)||p(u)] \geq \sup_{r \in \mathcal{R}} \{ \mathbb{E}_{q(u)}[1 + \log r(u)] - \mathbb{E}_{p(u)}[r(u)] \}$$

# Density Ratio Estimation

## Divergence Minimisation

- Let our ratio estimator be a neural network parametrised by  $\alpha$ :  
 $r_\alpha(u) \simeq \frac{q(u)}{p(u)}$ .
- Maximise the lower bound w.r.t.  $\alpha$  until equality, which is when  $r_\alpha(u) = \frac{q(u)}{p(u)}$ . The optimisation problem for this is

$$\min_{\alpha} -\mathbb{E}_{q(u)}[\log r_\alpha(u)] + \mathbb{E}_{p(u)}[r_\alpha(u)].$$

- Obviously our optimal ratio estimator is  $r_\alpha^*(u) = \frac{q(u)}{p(u)}$ .

# Density Ratio Estimation

## Divergence Minimisation

Prior-Contrastive Application:

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

Joint-Contrastive Application:

$$\min_{\alpha} -\mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log r_{\alpha}(z, x)] + \mathbb{E}_{p(z)p(x|z)}[r_{\alpha}(z, x)]$$
$$\min_{\phi} \mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log r_{\alpha}(z, x)]$$

# Density Ratio Estimation

## Algorithm Generalisation

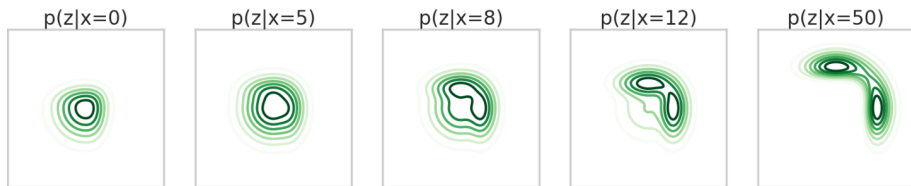
- Actually,  $f(u) = u \log u - (u + 1) \log(u + 1)$  and  $D(u) = \frac{r(u)}{r(u)+1}$  leads to class probability estimation equations.
- The upper bound f-divergence is  $2JS(p(u)||q(u)) - \log 4$ , we call this the GAN divergence.
- To formulate optimisation problems for density ratio estimation, choose either reverse KL or GAN f-divergence upper bound and estimator parametrisation:
  - Class Probability Estimator  $D_\alpha(u) \simeq \frac{q(u)}{q(u)+p(u)}$
  - Direct Ratio Estimator  $r_\alpha(u) \simeq \frac{q(u)}{p(u)}$
  - Direct Log Ratio Estimator  $T_\alpha(u) \simeq \log \frac{q(u)}{p(u)}$ .

# Activation Function Experiment

## Experiment Outline

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

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



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

# Activation Function Experiment

## Experiment Outline

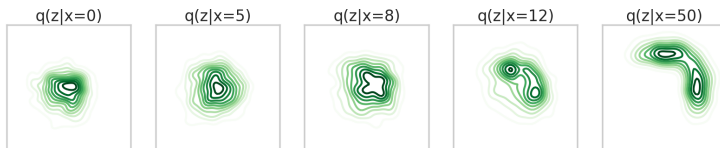
- Common to use ReLU  $g(x) = \max\{0, x\}$  as activation function for output layer of direct ratio estimator  $r_\alpha(u) \simeq \frac{q(u)}{p(u)}$ .
- Experiences 'dying ReLU problem'.
- Linearity of ReLU activation causes imbalance between ratios in  $(0, 1)$  and  $(1, \infty)$ .
- We propose exponential activation function  $g(x) = e^x$ .
- Compare them with reverse KL divergence upper bound.
- Low training rate, high iterations.



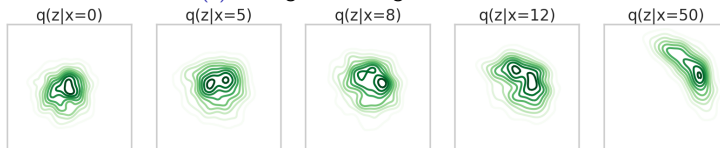
# Activation Function Experiment

## Results

Algorithm	Mean KL Divergence	Standard Deviation
Prior Contrastive - ReLU	1.3807	0.0391
Prior Contrastive - Exp	1.3265	0.0045



(a) Average KL Divergence of 1.3288



(b) Average KL Divergence of 1.3963

# Optimal Estimator Experiment

## Experiment Outline

- Same inference problem as before.
- Aim of this experiment is to verify that choice of estimator does not matter as long as it reaches equality.
- Low training rate with high estimator to posterior optimisation ratio (100:1).
- High posterior iterations.

# Optimal Estimator Experiment

## Results

Algorithm	Mean KL Divergence	Standard Deviation
JC Reverse KL - $D_\alpha(z, x)$	1.3416	0.0068
JC Reverse KL - $r_\alpha(z, x)$	1.3397	0.0066
JC Reverse KL - $T_\alpha(z, x)$	1.3446	0.0108
JC GAN - $D_\alpha(z, x)$	1.3648	0.0242
JC GAN - $r_\alpha(z, x)$	1.3657	0.0302
JC GAN - $T_\alpha(z, x)$	1.3670	0.0387

- Prior-contrastive posteriors fully converged at  $\approx 1.325$ .
- No significant difference in convergence between estimators in each f-divergence.
- Reverse KL converged faster in joint-contrastive context.

# Undertrained Estimator Experiment

## Experiment Outline

- Estimators are similar when they are optimal but what if they are not optimal?
- Same inference experiment again.
- Significantly reduce amount of estimator training between posterior iterations.
- Increased posterior training rate.

# Undertrained Estimator Experiment

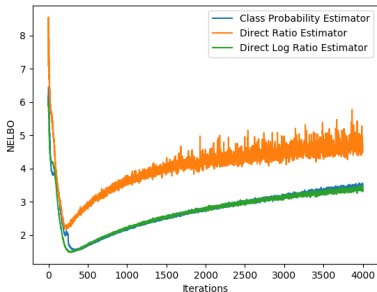
## Results

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

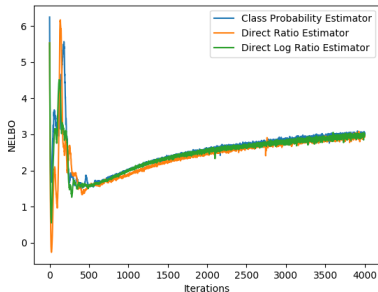
- Reverse KL divergence significantly better than GAN divergence.
- $D_{\alpha}(z, x) < r_{\alpha}(z, x) < T_{\alpha}(z, x)$

# Undertrained Estimator Experiment

## Joint-Contrastive NELBO Plots



(a) GAN Divergence



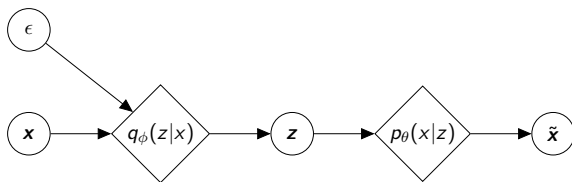
(b) Reverse KL Divergence

- Unclear why direct ratio estimator has unusual NELBO plot: posterior convergence was not affected.
- Reverse KL Divergence has initial instability.

# Autoencoder Experiment

## Autoencoders

- Likelihood  $p_{\theta}(x|z)$  is now a neural network.
- Posterior  $q_{\phi}(z|x)$  represents data  $x$  as lower dimensional latent  $z$ .
- Likelihood  $p_{\theta}(x|z)$  reconstructs data  $\tilde{x}$  from  $z$ .
- Generate new data  $\tilde{x}$  using  $z$  from  $p(z)$ .



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

# Autoencoder Experiment

## Experiment Outline

- MNIST dataset -  $28 \times 28$  grey-scale images of handwritten digits
- Joint-contrastive context not tested here.
- Again use undertrained estimator.
- Use reconstruction error  $\|x - \tilde{x}\|^2$  as metric.
- Perform experiment with low dimensional latent space (2 dimensions) and high dimensional latent space (20 dimensions).
- Low dimensional case had similar results to previous experiment.



# Generation Experiment

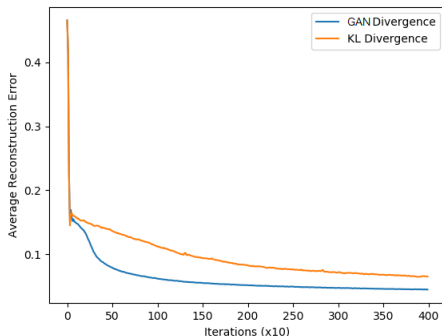
Results - high dimensional latent space

Algorithm	Mean Reconstruction Error	Standard Deviation
PC GAN - $D_\alpha(z, x)$	0.0444	0.0017
PC Reverse KL - $D_\alpha(z, x)$	0.0647	0.0019

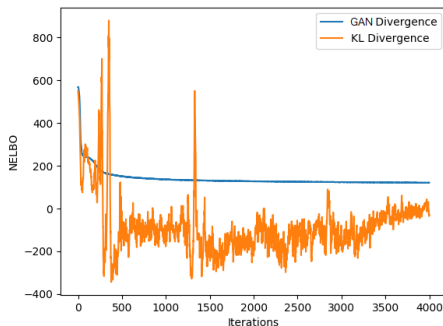
- Direct ratio and direct log ratio estimators attempted to store numbers exceeding float64(max).
- Exponential of  $T_\alpha(z, x)$  taken in loss function.
- $D_\alpha(z, x)$  ranges in  $(0, 1)$ .
- Value before sigmoid activation function for  $D_\alpha(z, x)$  is log density ratio.
- This time GAN divergence leads to better convergence than reverse KL.

# Autoencoder Experiment

Results - high dimensional latent space



(a) Reconstruction Error



(b) NELBO

- As before, GAN divergence is more stable.
- Recall reverse KL divergence is initially unstable but stabilizes later.
- In this case it fails to stabilise by the end of the program runtime.

- Nowozin's f-GAN paper shows empirically that the reverse KL divergence is superior when it is additionally used to optimize the posterior.
- Intuitive that the f-divergence used to optimize posterior is the best upper bound for estimator.

# Further Estimator Loss Function Analysis

## Estimator Parametrisation

- $D_\alpha(u)$  has smallest bound of  $(0, 1)$ , followed by  $r_\alpha(u) \in \mathbb{R}^+$  and  $T_\alpha(u) \in \mathbb{R}$ .
- The density ratio changes every time the posterior is optimised, and the estimator must catch up.
- $D_\alpha(u)$  has a strictly lower displacement than  $r_\alpha(u)$ , that is,  
 $|D_\alpha^{(n+1)}(u) - D_\alpha^{(n)}(u)| < |r_\alpha^{(n+1)}(u) - r_\alpha^{(n)}(u)|$ .

- 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 demonstrates **initial instability** (especially when estimator is undertrained) but leads to **faster convergence** when it stabilizes.
- Outlook
  - 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.