

1 Introduction

1.1 Problem Context

In machine learning, particularly for high dimensional applications such as image analysis, it is often desirable to build generative models, so that we can represent the data in lower dimensions via representation learning, and generate new data similar to the examples in our dataset. Assume our dataset $X = \{x^{(i)}\}_{i=1}^N$ is N i.i.d. samples of random variables x . Also assume x can be generated by a stochastic process from a latent continuous random variable z . These models involve a posterior distribution parametrized by θ $p_\theta(z|x)$ that maps the dataset x to lower dimensional latent prior z (e.g. $z \sim N(\mu, \Sigma)$) then simulating from the prior $p(z)$ to generate new data through a decoder $p(x|z)$. In this particular field, there are three main problems to solve:

1. Estimation of θ , so that we can actually generate new data x
2. Evaluation of the posterior density $p_\theta(z|x) = \frac{p(z)p(x|z)}{p(x)} = \frac{p(x|z)p(z)}{\int_z p(x,z)dz}$, so we can encode our data x in an efficient representation z
3. Marginal inference of x ie. evaluating $p(x)$, so it can be used as a prior for other tasks

1.2 Objective Derivation

1.2.1 Implicit Prior

The posterior distribution cannot be computed efficiently if $\int_z p(x, z)dz$ is intractable, which is often the case with large datasets or high dimensional data. Furthermore, the prior $p(z)$ or the likelihood $p(x|z)$ may be implicit. In the framework of variational inference and machine learning, we approximate the posterior distribution $p_\theta(z|x)$ with a variational distribution parametrized by ϕ : $q_\phi(z)$. We aim to minimize the KL divergence between these distributions, defined as

$$KL(Q||P) = \int_{-\infty}^{\infty} Q(x) \log \frac{Q(x)}{P(x)} dx = \mathbb{E}_{Q(x)} \left[\log \frac{Q(x)}{P(x)} \right].$$

The parameters ϕ of our approximate distribution $q_\phi(z)$ are therefore chosen such that the KL divergence is minimised. Since the true posterior distribution $p_\theta(z|x)$ is dependent on x , it differs with each observation x_n that we take. This would mean we would have to define a variational distribution for each observation. Thus, we amortize our distribution, conditioning it on x and letting its parameters ϕ be constant for each observation:(or better to say condition on x and take expectation w.r.t. $q^*(x)$?)

$$q_\phi(z|x).$$

We therefore want to optimize ϕ across the observations from our dataset, so our objective now is to choose parameters ϕ such that the expected KL divergence with respect to $q^*(x)$ is minimized:

$$\begin{aligned} \phi &= \arg \min_{\phi} \mathbb{E}_{q^*(x)} KL(q_\phi(z|x)||p_\theta(z|x)) \\ &= \arg \min_{\phi} \mathbb{E}_{q^*(x)q_\phi(z|x)} [\log q_\phi(z|x) - \log p_\theta(z|x)] \\ &= \arg \min_{\phi} \mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log q_\phi(z|x) - \log \frac{p(x|z)p(z)}{p(x)} \right] \\ &= \arg \min_{\phi} (\mathbb{E}_{q^*(x)q_\phi(z|x)} [\log q_\phi(z|x) - \log p(x|z) - \log p(z)] + \log p(x)) \end{aligned}$$

Again, we cannot evaluate this expression as $\log p(x)$ is intractable, so we rearrange the terms to form what is called the ELBO (Evidence Lower Bound), which we want to maximise to minimise the KL divergence.

$$\begin{aligned} ELBO &= \mathbb{E}_{q^*(x)} [\log p(x) - KL(q_\phi(z|x) || p_\theta(z|x))] \\ &= -\mathbb{E}_{q^*(x)q_\phi(z|x)} [\log q_\phi(z|x) - \log p(x|z) - \log p(z)] \\ &= \mathbb{E}_{q^*(x)q_\phi(z|x)} [\log p(x|z) + \log p(z) - \log q_\phi(z|x)] \end{aligned}$$

If the prior distribution is implicit (but the likelihood $\log p(x|z)$ isn't), then we can still evaluate and optimize this expression by considering the density ratio between $p(z)$ and $q_\phi(z|x)$:

$$ELBO = \mathbb{E}_{q^*(x)q_\phi(z|x)} [\log p(x|z)] + \mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log \frac{p(z)}{q_\phi(z|x)} \right].$$

The $\mathbb{E}_{q^*(x)q_\phi(z|x)} [\log p(x|z)]$ term can be evaluated and optimised by applying the reparameterization trick (Kingma 2014; Rezende 2014; Tiao 2018), which involves taking random input noise $\epsilon \sim \pi(\epsilon)$ and observed variable $x \sim q^*(x)$, and passing it through the neural network representing the variational distribution $q_\phi(z|x)$, defined as the function $\mathcal{G}_\phi(\epsilon; x)$. We therefore have the following expression for the expected log-likelihood:

$$\mathbb{E}_{q^*(x)\pi(\epsilon)} [\log p_\theta(x|\mathcal{G}_\phi(\epsilon; x))].$$

The ELBO therefore becomes

$$ELBO = \mathbb{E}_{q^*(x)\pi(\epsilon)} [\log p_\theta(x|\mathcal{G}_\phi(\epsilon; x))] + \mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log \frac{p_\theta(z)}{q_\phi(z|x)} \right].$$

Our objective is to maximize the ELBO, equivalent to minimizing its negative:

$$\min_{\phi} NELBO = -\mathbb{E}_{q^*(x)q_\phi(z|x)} [\log p_\theta(x|z)] + \mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{p_\theta(z)} \right].$$

As previously stated, since the prior distribution is implicit, we use a ratio estimator $\hat{r}_\alpha(x)$ for the true density ratio $r^*(x) = \frac{q_\phi(z|x)}{p_\theta(z)}$.

1.2.2 Implicit Likelihood (& Prior)

Clearly, if our likelihood distribution is implicit, then this expression cannot be evaluated. Instead, we must rephrase the problem, considering a density ratio between two joint distributions (Tran, 2017). We begin by restating the original objective of choosing parameters ϕ to minimize the KL divergence between two distributions:

$$\begin{aligned} \phi &= \arg \min_{\phi} \mathbb{E}_{q^*(x)} KL(q_\phi(z|x) || p_\theta(z|x)) \\ &= \arg \min_{\phi} \mathbb{E}_{q^*(x)q_\phi(z|x)} \log \frac{q(z|x)p(x)}{p(x|z)p(z)} + \mathbb{E}_{q^*(x)} \log q^*(x) - \mathbb{E}_{q^*(x)} \log q^*(x) \\ &= \arg \min_{\phi} \mathbb{E}_{q^*(x)q_\phi(z|x)} \log \frac{q(z|x)q^*(x)}{p(x|z)p(z)} + \mathbb{E}_{q^*(x)} \log \frac{p(x)}{q^*(x)} \\ &= \arg \min_{\phi} \mathbb{E}_{q^*(x)q_\phi(z|x)} \log \frac{q(z, x)}{p(z, x)} - KL(q^*(x) || p(x)) \end{aligned}$$

We therefore have the expression for the 'evidence lower bound' (not actually an evidence lower bound but we keep the name for consistency):

$$\begin{aligned} ELBO &= KL(q^*(x)||p(x)) + \mathbb{E}_{q^*(x)} KL(q_\phi(z|x)||p_\theta(z|x)) \\ &= \mathbb{E}_{q^*(x)q_\phi(z|x)} \log \frac{q(z, x)}{p(x, z)}. \end{aligned}$$

Our objective now is to minimize this ELBO with respect to our variational parameters ϕ

$$\min_{\phi} \mathbb{E}_{q^*(x)q_\phi(z|x)} \log \frac{q(z, x)}{p(z, x)}$$

There are two main methods of estimating and optimizing the density ratio and subsequently, the ELBO: Class Probability Estimation and Divergence Minimisation. The goal of this thesis is to implement and compare these two methods (and hopefully ratio matching) over a variety of models (Hierarchical Lotka-Volterra, Image MNIST etc.) and conditions (implicit prior or likelihood), evaluating their rates of convergence, bias and mean squared error to determine which method is best in which conditions.

1.3 Class Probability Estimation

1.3.1 Implicit Prior

In class probability estimation, we use a discriminator function $D_\phi(z)$ denoted by a neural network that calculates the probability that a given sample z is from the variational posterior distribution $q_\phi(z|x)$ as opposed to the true prior distribution $p^*(x)$. The ratio estimator can be expressed in terms of the probability that a sample from the posterior distribution correctly classified by the discriminator. Optimization involves optimizing the discriminator by minimizing the discriminative loss (a Bernoulli loss dependent on the discriminators classification accuracy), then minimizing the evidence lower bound, and cycling between these two steps until convergence.

Firstly, consider the problem of estimating the density ratio between two arbitrary distributions $q(x)$ and $p(x)$: $\frac{q(x)}{p(x)}$. We take m samples from $p(x)$: $X_p = \{x_1^{(p)}, \dots, x_m^{(p)}\}$ and label them with $y = 0$, then we take n samples from $q(x)$: $X_q = \{x_1^{(q)}, \dots, x_n^{(q)}\}$. Therefore, $p(x) = P(x|y = 0)$ and $q(x) = P(x|y = 1)$. By applying Bayes' theorem, we derive an expression for the density ratio:

$$\begin{aligned} \frac{q(x)}{p(x)} &= \frac{P(x|y = 1)}{P(x|y = 0)} \\ &= \frac{P(y = 1|x)P(x)}{P(y = 1)} \bigg/ \frac{P(y = 0|x)P(x)}{P(y = 0)} \\ &= \frac{P(y = 1|x)}{P(y = 0|x)} \times \frac{P(y = 0)}{P(y = 1)} \\ &= \frac{P(y = 1|x)}{P(y = 0|x)} \times \frac{n}{m}. \end{aligned}$$

Often in practice, $m = n$, so the density ratio simplifies to:

$$\frac{q(x)}{p(x)} = \frac{P(y = 1|x)}{P(y = 0|x)}$$

which is the ratio of the probability that an arbitrary sample x was taken from the distribution $q(x)$ to the probability that it was taken from $p(x)$. If we define a discriminator function that finds these probabilities

$$D(x) = P(y = 1|x),$$

then our density ratio can be expressed in terms of this discriminator function:

$$\frac{q(x)}{p(x)} = \frac{D(x)}{1 - D(x)}.$$

We now turn to our problem of minimizing the negative ELBO:

$$\min_{\phi} -\mathbb{E}_{q^*(x)\pi(\epsilon)}[\log p_{\theta}(x|\mathcal{G}_{\phi}(\epsilon; x))] + \mathbb{E}_{q^*(x)q_{\phi}(z|x)} \left[\log \frac{q_{\phi}(z|x)}{p_{\theta}(z)} \right],$$

which requires us to find the density ratio between $q_\phi(z|x)$ and $p_\theta(z)$.

$$\begin{aligned}\mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{p_\theta(z)} \right] &= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z)}{p_\theta(z)} \right] \\ &= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{P(y=1|z)}{P(y=0|z)} \right] \\ &= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{D_\alpha(z)}{1 - D_\alpha(z)} \right] \\ &= \end{aligned}$$

We now define a discriminator function in the form of a neural network with parameters α that calculates the probability that an arbitrary sample z belongs to the variational posterior $q_\phi(z|x)$. We also amortize the discriminator (????) by taking an additional input from the dataset x , as opposed to training multiple discriminators for each observation x_n :

$$D_\alpha(z, x) = P(y=1|z).$$

As a binary classifier, this function can be trained by inputting an equal number of samples from both distributions and minimizing its Bernoulli loss:

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

We can express the expected log ratio estimator $\mathbb{E}_{q^*(x)q_\phi(z|x)} [\log \hat{r}(x)] \simeq \mathbb{E}_{q^*(x)q_\phi(z|x)} [\log \frac{q_\phi(z|x)}{p_\theta(z)}]$ in terms of the probability that a posterior sample is correctly classified by the discriminator:

$$\begin{aligned}\mathbb{E}_{q^*(x)q_\phi(z|x)} [\log \hat{r}(x)] &= \mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log \frac{P(y=1|z)}{P(y=0|z)} \right] \\ &= \mathbb{E}_{q^*(x)q_\phi(z|x)} \left[\log \frac{D_\alpha(z, x)}{1 - D_\alpha(z, x)} \right] \\ &= \mathbb{E}_{q^*(x)\pi(\epsilon)} \left[\log \frac{D_\alpha(\mathcal{G}_\phi(\epsilon; x), x)}{1 - D_\alpha(\mathcal{G}_\phi(\epsilon; x), x)} \right]\end{aligned}$$

Our optimization objective (negative ELBO) can now be written as:

$$\min_{\phi} -\mathbb{E}_{q^*(x)\pi(\epsilon)} [\log p_\theta(x|\mathcal{G}_\phi(\epsilon; x))] + \mathbb{E}_{q^*(x)\pi(\epsilon)} \left[\log \frac{D_\alpha(\mathcal{G}_\phi(\epsilon; x))}{1 - D_\alpha(\mathcal{G}_\phi(\epsilon; x))} \right].$$

In practice, the algorithm cycles between optimizing the discriminator until convergence (with the generator parameters fixed) and taking optimization steps of the negative evidence lower bound (with the discriminator parameters fixed).

1.3.2 Implicit Likelihood (& Prior)

As previously stated, if the likelihood distribution is implicit, we have the optimization objective of:

$$\min_{\phi} \mathbb{E}_{q^*(x)q_{\phi}(z|x)} \log \frac{q(z, x)}{p(z, x)}.$$

or

$$\min_{\phi} \mathbb{E}_{q^*(x)q_{\phi}(z|x)} \log \frac{q_{\phi}(z|x)q^*(x)}{p(x|z)p(z)}.$$

Like in the implicit prior case, we can label samples from $q(z, x)$ with $y = 1$ and samples from $p(z, x)$ with $y = 0$, leading to the density ratio expression:

$$\frac{q(z, x)}{p(z, x)} = \frac{P(y = 1|z, x)}{P(y = 0|z, x)}.$$

Again, we use a discriminator neural network parametrized by α to determine the probability that samples (z, x) came from the joint variational distribution $q(z, x)$:

$$D_{\alpha}(z, x) = P(y = 1|z, x).$$

Using this discriminator function, our density ratio expression becomes:

$$\frac{q(z, x)}{p(z, x)} = \frac{D_{\alpha}(z, x)}{1 - D_{\alpha}(z, x)}.$$

Overall, the class probability algorithm is (again) a cycle of optimizing the discriminator until convergence by minimizing its Bernoulli loss:

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

and taking an optimization step of the variational posterior:

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

Since our posterior is defined as a generative neural network model $\mathcal{G}_{\phi}(x, \epsilon)$, these optimization objectives become:

$$\min_{\alpha} \mathbb{E}_{q^*(x)\pi(\epsilon)} [-\log D_{\alpha}(\mathcal{G}_{\phi}(x, \epsilon), x)] + \mathbb{E}_{p(z)p(x|z)} [-\log(1 - D_{\alpha}(z, x))]$$

$$\min_{\phi} \mathbb{E}_{q^*(x)\pi(\epsilon)} \frac{D_{\alpha}(\mathcal{G}_{\phi}(x, \epsilon), x)}{1 - D_{\alpha}(\mathcal{G}_{\phi}(x, \epsilon), x)}$$

Note that the prior and likelihood terms don't appear in the ELBO training objective, so they can both be implicit. Another difference with the implicit prior algorithm is that instead of sampling x from the dataset, it is taken from the likelihood, conditional on the z sample.

1.4 Divergence Minimisation

1.4.1 Implicit Prior

In divergence minimisation, we convert the expected density ratio expression into a KL divergence, and use a theorem (name of theorem?) to find a lower bound of the divergence in terms of a ratio estimator denoted by a neural network. We then perform bi-level optimization of maximizing the KL divergence lower bound and minimizing the negative evidence lower bound.

Firstly, note that maximising the ELBO is equivalent to minimising its negative, so our overall objective is (Ask scott about amortized inference: optimizing across $x \sim q^*(x)$)

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

or

$$\min_{\phi} -\mathbb{E}_{q^*(x)\pi(\epsilon)} [\log p_{\theta}(x|\mathcal{G}_{\phi}(\epsilon; x))] + \mathbb{E}_{q^*(x)} KL[q_{\phi}(z|x)||p_{\theta}(z)]$$

Denoting the true density ratio $\frac{q_{\phi}(z|x)}{p_{\theta}(z)}$ as $r^*(z)$, we aim to find an estimator $\hat{r}(z)$. To do so, we apply a theorem (Nguyen et al. 2010) to find a lower bound to the f-divergence between the two distributions in terms of the ratio estimator.

Theorem 1. If f is a convex function with derivative f' and convex conjugate f^* , then we have the lower bound for the f-divergence between distributions $p(u)$ and $q(u)$:

$$\mathcal{D}_f[q(u)||p(u)] \geq \sup_{\hat{r}} \{ \mathbb{E}_{q(u)}[f'(\hat{r}(u))] - \mathbb{E}_{p(u)}[f^*(f'(\hat{r}(u)))] \}$$

with equality when $\hat{r}(u) = q(u)/p(u)$.

We let our ratio estimator be parametrized by α and apply this theorem to $p_{\theta}(z)$ and $q_{\phi}(z|x)$ to obtain a lower bound for the divergence. We also condition the ratio estimator on \mathbf{x} , optimizing it across all $\mathbf{x} \sim q^*(\mathbf{x})$:

$$\mathcal{D}_f[q_{\phi}(z|x)||p_{\theta}(z)] \geq \sup_{\alpha} \{ \mathbb{E}_{q^*(x)q_{\phi}(z|x)}[f'(r_{\alpha}(z; x))] - \mathbb{E}_{q^*(x)p_{\theta}(z)}[f^*(f'(r_{\alpha}(z; x)))] \}$$

We then specify this divergence as a KL divergence by letting $f = u \log u$, forming the following lower bound for the KL divergence:

$$KL[q_{\phi}(z|x)||p_{\theta}(z)] \geq \sup_{\alpha} \{ \mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log r_{\alpha}(z; x)] - \mathbb{E}_{q^*(x)p_{\theta}(z)}[r_{\alpha}(z; x) - 1] \}$$

To optimize our ratio estimator, we fix ϕ and optimize α such that the lower bound is maximised, reducing the gap between the lower bound and the true KL divergence.

Simultaneously, we want to minimise our overall objective, which is the negative evidence lower bound. Noting that $KL[q_{\phi}(z|x)||p_{\theta}(z)] \simeq E_{q_{\phi}(z|x)}[\log r_{\alpha}(z; x)]$, accomplish this by fixing α and optimizing ϕ to minimize the lower bound. We therefore have the bi-level optimization problem:

$$\begin{aligned} & \max_{\alpha} \mathbb{E}_{q^*(x)q_{\phi}(z|x)}[\log r_{\alpha}(z; x)] - \mathbb{E}_{q^*(x)p_{\theta}(z)}[r_{\alpha}(z; x) - 1] \\ & \min_{\phi} -\mathbb{E}_{q^*(x)\pi(\epsilon)} [\log p_{\theta}(x|\mathcal{G}_{\phi}(\epsilon; x))] + E_{q^*(x)q_{\phi}(z|x)}[\log r_{\alpha}(z; x)] \end{aligned}$$

Note that this estimator is biased as it follows a lower bound.

(Also note that there is an f-divergence that leads to the class probability optimization expressions)

1.4.2 Implicit Likelihood (& Prior)

2 Learning

2.1 Variational Inference

2.1.1 Context

In Bayesian statistics, a common problem is to estimate posterior densities, so that we may perform inference to determine an unknown parameter. Consider a set of unknown, latent variables $\mathbf{Z} = \{z_i\}_{i=1}^M$ and a dataset of known variables $\mathbf{X} = \{x_i\}_{i=1}^N$. These sets have a joint density of $P(\mathbf{Z}, \mathbf{X})$. In the Bayesian framework, inference is often performed on the posterior density (the distribution of the parameters \mathbf{Z} after the data \mathbf{X} is observed) $P(\mathbf{Z}|\mathbf{X})$, which, after applying Bayes' theorem, can be written as:

$$P(\mathbf{Z}|\mathbf{X}) = \frac{P(\mathbf{Z})P(\mathbf{X}|\mathbf{Z})}{P(\mathbf{X})} = \frac{P(\mathbf{Z})P(\mathbf{X}|\mathbf{Z})}{\int_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X})d\mathbf{Z}}$$

where

- $P(\mathbf{Z})$ is the prior distribution: the initial distribution of \mathbf{Z} before the data \mathbf{X} is observed. This can be initialised to represent our initial beliefs, or it can be parametrised randomly,
- $P(\mathbf{X}|\mathbf{Z})$ is the likelihood: the distribution of data \mathbf{X} conditioned on the parameters \mathbf{Z} ,
- $P(\mathbf{X}) = \int_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X})d\mathbf{Z}$ is the marginal likelihood, or the evidence: the density of the data averaged across all possible parameter values.

If the evidence integral $P(\mathbf{X}) = \int_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X})d\mathbf{Z}$ is impossible or difficult to compute (possibly because it is unavailable in closed form or the dimensionality is too high), then we are unable to evaluate the posterior density. Traditionally, MCMC(Markov Chain Monte Carlo) methods overcome this obstacle by constructing a Markov chain that converges to the stationary distribution $P(\mathbf{Z}|\mathbf{X})$, then sampling from the chain to create an empirical estimate for the posterior distribution. However, these methods rely on the speed of convergence, which can be slow for large datasets or complex models. When faced with these issues or when desiring a faster computation, one may instead apply variational inference, an alternative approach to density estimation.

2.1.2 Introduction to Variational Inference

Variational inference chooses another distribution $Q(\mathbf{Z})$ from a select family of variational distributions (approximate densities) \mathcal{Q} to serve as an approximation to $P(\mathbf{Z}|\mathbf{X})$, and then minimizes the divergence between the two distributions in an optimization problem:

$$Q^*(\mathbf{Z}) = \arg \min_{Q(\mathbf{Z}) \in \mathcal{Q}} D(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X})) \quad (1)$$

where D denotes an f-divergence (a measure of how divergent two probability distributions are, it is minimized if $Q = P$). This results in an analytical approximation to the posterior density. Additionally, a lower bound for the marginal likelihood of the dataset is derived, which can be used as a model selection criterion. Due to the stochastic nature of the optimization, variational inference methods can be much faster than MCMC, but the solution is only locally optimal as there is no guarantee of global convergence.

2.1.3 Derivation of the ELBO

The most common f-divergence used in variational inference is the KL(Kullback-Leibler) divergence, defined as the expected logarithmic difference between two distributions Q and P with respect to Q :

$$KL(Q||P) = \int_{-\infty}^{\infty} Q(x) \log \frac{Q(x)}{P(x)} dx = \mathbb{E}_{Q(x)} \left[\log \frac{Q(x)}{P(x)} \right].$$

Note that the KL divergence is not symmetric. We use the reverse KL divergence instead of the forward KL divergence $KL(P||Q)$ because it leads to an expectation maximization algorithm as opposed to an expectation propagation algorithm.

Using this expression, we can rewrite equation (1) as:

$$\begin{aligned} Q^*(\mathbf{Z}) &= \arg \min_{Q(\mathbf{Z}) \in \mathcal{Q}} KL(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X})) \\ &= \arg \min_{Q(\mathbf{Z}) \in \mathcal{Q}} \mathbb{E}_{Q(\mathbf{Z})} [\log Q(\mathbf{Z}) - \log P(\mathbf{Z}|\mathbf{X})] \\ &= \arg \min_{Q(\mathbf{Z}) \in \mathcal{Q}} \mathbb{E}_{Q(\mathbf{Z})} \left[\log Q(\mathbf{Z}) - \log \frac{P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z})}{P(\mathbf{X})} \right] \\ &= \arg \min_{Q(\mathbf{Z}) \in \mathcal{Q}} (\mathbb{E}_{Q(\mathbf{Z})} [\log Q(\mathbf{Z}) - \log P(\mathbf{X}|\mathbf{Z}) - \log P(\mathbf{Z})] + \log P(\mathbf{X})). \end{aligned}$$

Note in the last line $\mathbb{E}_{Q(\mathbf{Z})} [P(\mathbf{X})] = P(\mathbf{X})$ as it is not dependent on $Q(\mathbf{Z})$. Also note that the KL divergence is dependent on $P(\mathbf{X})$, which we have determined to be intractable, so this optimization problem cannot be solved in this form. This issue is resolved by rearranging the KL divergence expression as follows:

$$\begin{aligned} KL(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X})) &= \mathbb{E}_{Q(\mathbf{Z})} [\log Q(\mathbf{Z}) - \log P(\mathbf{X}|\mathbf{Z}) - \log P(\mathbf{Z})] + \log P(\mathbf{X}) \\ \log P(\mathbf{X}) - KL(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X})) &= -\mathbb{E}_{Q(\mathbf{Z})} [\log Q(\mathbf{Z}) - \log P(\mathbf{X}|\mathbf{Z}) - \log P(\mathbf{Z})] \\ &= \mathbb{E}_{Q(\mathbf{Z})} [\log P(\mathbf{X}|\mathbf{Z})] - \mathbb{E}_{Q(\mathbf{Z})} [\log Q(\mathbf{Z}) - \log P(\mathbf{Z})] \\ &= \mathbb{E}_{Q(\mathbf{Z})} [\log P(\mathbf{X}|\mathbf{Z})] - KL(Q(\mathbf{Z})||P(\mathbf{Z})). \end{aligned} \tag{2}$$

We refer to $\log P(\mathbf{X}) - KL(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X}))$ as $ELBO(Q)$ (evidence lower bound), as it is equal to the marginal probability of the data subtracted by a constant error term. This error term $KL(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X}))$ becomes 0 when $Q(\mathbf{Z}) = P(\mathbf{Z}|\mathbf{X})$, maximizing the ELBO. Note that since $P(\mathbf{X})$ is constant, maximizing the $ELBO$ is equal to minimizing the KL divergence between $Q(\mathbf{Z})$ and $P(\mathbf{Z}|\mathbf{X})$, and that the expression on line (2) is entirely computable. We can therefore rewrite our optimization problem from equation (1) as:

$$\begin{aligned} Q^*(\mathbf{Z}) &= \arg \min_{Q(\mathbf{Z}) \in \mathcal{Q}} D(Q(\mathbf{Z})||P(\mathbf{Z}|\mathbf{X})) \\ &= \arg \max_{Q(\mathbf{Z}) \in \mathcal{Q}} ELBO(Q) \\ &= \arg \max_{Q(\mathbf{Z}) \in \mathcal{Q}} (\mathbb{E}_{Q(\mathbf{Z})} [\log P(\mathbf{X}|\mathbf{Z})] - KL(Q(\mathbf{Z})||P(\mathbf{Z}))). \end{aligned}$$

2.1.4 Mean-Field Variational Family

The family of variational distributions \mathcal{Q} is typically a 'mean-field variational family', in which the distribution $Q(\mathbf{Z})$ factorizes over the latent variables $\{z_i\}_{i=1}^M$:

$$Q(\mathbf{Z}) = \prod_{i=1}^M q_i(z_i). \tag{3}$$

The individual factors $q_i(z_i)$ can take any form. We want to choose these factors so that $ELBO(Q)$ is maximized. To derive an expression for the optimal factor $q_i^*(z_i)$, we substitute equation (3) into the $ELBO$, factor out a specific $q_j(z_j)$ and equate the functional derivative of the resulting Lagrangian equation with 0. Firstly, we express $ELBO(Q)$ in an integral form as follows:

$$\begin{aligned} ELBO(Q) &= \mathbb{E}_{Q(\mathbf{Z})}[\log P(\mathbf{X}|\mathbf{Z})] - KL(Q(\mathbf{Z})||P(\mathbf{Z})) \\ &= \mathbb{E}_{Q(\mathbf{Z})}[\log P(\mathbf{X}|\mathbf{Z}) + \log P(\mathbf{Z}) - \log Q(\mathbf{Z})] \\ &= \mathbb{E}_{Q(\mathbf{Z})}[\log P(\mathbf{X}, \mathbf{Z}) - \log Q(\mathbf{Z})] \\ &= \int_{\mathbf{Z}} Q(\mathbf{Z})(\log P(\mathbf{X}, \mathbf{Z}) - \log Q(\mathbf{Z}))d\mathbf{Z}. \end{aligned}$$

Substituting $Q(\mathbf{Z}) = \prod_{i=1}^M q_i(z_i)$ and factoring out $q_j(z_j)$ yields:

$$\begin{aligned} ELBO(Q) &= \int_{\mathbf{Z}} \left[\prod_{i=1}^M q_i(z_i) \right] \left(\log P(\mathbf{X}, \mathbf{Z}) - \sum_{i=1}^M \log q_i(z_i) \right) d\mathbf{Z} \\ &= \int_{\dagger_j} q_j(z_j) \left(\int_{\dagger_{-j}} \log P(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i(z_i) d\mathbf{z}_{-j} \right) dz_j \\ &\quad - \int_{\dagger_j} q_j(z_j) \left(\int_{\dagger_{-j}} \left[\prod_{i \neq j} q_i(z_i) \right] \sum_{i=1}^M q_i(z_i) d\mathbf{z}_{-j} \right) dz_j \\ &= \int_{\dagger_j} q_j(z_j) \mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] dz_j - \int_{\dagger_j} q_j(z_j) \log q_j(z_j) \left(\int_{\dagger_{-j}} \prod_{i \neq j} q_i(z_i) d\mathbf{z}_{-j} \right) dz_j \\ &\quad - \int_{\dagger_j} q_j(z_j) \left(\int_{\dagger_{-j}} \left[\prod_{i \neq j} q_i(z_i) \right] \sum_{i \neq j} q_i(z_i) d\mathbf{z}_{-j} \right) dz_j \\ &= \int_{\dagger_j} q_j(z_j) \mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] dz_j - \int_{\dagger_j} q_j(z_j) \log q_j(z_j) dz_j \\ &\quad - \int_{\dagger_{-j}} \left[\prod_{i \neq j} q_i(z_i) \right] \sum_{i \neq j} q_i(z_i) d\mathbf{z}_{-j} \end{aligned} \tag{4}$$

$$= \int_{\dagger_j} q_j(z_j) (\mathbb{E}_{\dagger_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] - \log q_j(z_j)) dz_j + \text{const.} \tag{5}$$

The term in line (4) becomes a constant as it does not depend on $q_j(z_j)$. Now our Lagrangian equation with the constraint that $q_i(z_i)$ are probability density functions is:

$$ELBO(Q) - \sum_{i=1}^M \lambda_i \int_{\dagger_i} q_i(z_i) dz_i = 0$$

or using our expression for $ELBO(Q)$ in line (5),

$$\int_{\dagger_j} q_j(z_j) (\mathbb{E}_{\dagger_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] - \log q_j(z_j)) dz_j - \sum_{i=1}^M \lambda_i \int_{\dagger_i} q_i(z_i) dz_i + \text{const} = 0. \tag{6}$$

Using the Euler-Lagrange equation (need to put this in), we then take the functional derivative of (6) with respect to $q_j(z_j)$ (in this case, the partial derivative with respect to $q_j(z_j)$ of the expression

inside the integral):

$$\begin{aligned}\frac{\partial ELBO(q)}{\partial q_j(z_j)} &= \frac{\partial}{\partial q_j(z_j)} [q_j(z_j) (\mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] - \log q_j(z_j)) - \lambda_j q_j(z_j)] \\ &= \mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] - \log q_j(z_j) - 1 - \lambda_j\end{aligned}\quad (7)$$

Equating expression (7) to 0 and letting $1 + \lambda_j$ be a constant (as it is independent of z_j), we have:

$$\begin{aligned}\log q_j^*(z_j) &= \mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})] - \text{const} \\ q_j^*(z_j) &= \frac{e^{\mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})]}}{\text{const}} \\ &= \frac{1}{\int e^{\mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})]} dz_j}.\end{aligned}$$

The normalization constant on the denominator can be easily derived by observing $q_j^*(z_j)$ as a density. Lastly, we derive a simpler expression of $q_j^*(z_j)$ by observing that terms independent of z_j can be treated as a constant:

$$\begin{aligned}q_j^*(z_j) &\propto \exp(\mathbb{E}_{\mathbf{z}_{-j}}[\log P(\mathbf{X}, \mathbf{Z})]) \\ &\propto \exp(\mathbb{E}_{\mathbf{z}_{-j}}[\log P(z_j | \mathbf{z}_{-j}, \mathbf{X})]).\end{aligned}\quad (8)$$

This expression can be used in an expectation-maximization algorithm, in which the $q_j^*(z_j)$ is evaluated and iterated from $j = 1 \dots M$. This particular algorithm is called coordinate ascent variational inference (CAVI) (Algorithm 1):

Data: Dataset \mathbf{X} and Model $P(\mathbf{X}, \mathbf{Z})$

Result: Approximate density $Q(\mathbf{Z}) = \prod_{i=1}^M q_i(z_i)$

begin

```

    Initialize random variational factors  $q_j(z_j)$ ;
    while  $ELBO(Q)$  has not converged do
        for  $j = 1$  to  $m$  do
            Set  $q_j(z_j) \propto \exp(\mathbb{E}[\log P(z_j | \mathbf{z}_{-j}, \mathbf{X})])$ ;
        end
        Calculate  $ELBO(Q) = \mathbb{E}[\log P(\mathbf{Z}, \mathbf{X})] - \mathbb{E}[\log Q(\mathbf{Z})]$ ;
    end
    Return  $Q(\mathbf{Z})$ ;
```

end

Algorithm 1: Coordinate Ascent Variational Inference (CAVI)

2.1.5 Example: Bayesian mixture of Gaussians

To illustrate the variational inference approach, we will use the Bayesian mixture of Gaussians example from (Blei, 2018/16 idk).

Consider the hierarchical model

$$\begin{aligned}\mu_k &\sim N(0, \sigma^2), & k &= 1, \dots, K, \\ c_i &\sim \text{Categorical}\left(\frac{1}{K}, \dots, \frac{1}{K}\right), & i &= 1, \dots, n, \\ x_i | c_i, \boldsymbol{\mu} &\sim N(c_i^\top \boldsymbol{\mu}, 1), & i &= 1, \dots, n.\end{aligned}$$

This is a Bayesian mixture of univariate Gaussian random variables with unit variance. In this model, we draw K μ_k variables from a prior Gaussian distribution $N(0, \sigma^2)$ (σ^2 is a hyperparameter), forming the vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^\top$. We then generate an indicator vector c_i of length K from a prior categorical distribution. This vector has zeros for every element except for one element, where it is a 1. Each element has equal probability $1/K$ of being the element that contains the 1. The transpose of this c_i is then multiplied by $\boldsymbol{\mu}$, essentially choosing one of the $\boldsymbol{\mu}$ elements at random. We then draw x_i from the resulting $N(c_i^\top \boldsymbol{\mu}, 1)$.

Here, our latent variables are $\mathbf{z} = \{\mathbf{c}, \boldsymbol{\mu}\}$. Assuming n samples, our joint density is

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^n p(c_i) p(x_i | c_i, \boldsymbol{\mu}). \quad (9)$$

From this, we derive the marginal likelihood

$$p(\mathbf{x}) = \int p(\boldsymbol{\mu}) \prod_{i=1}^n \sum_{c_i} p(c_i) p(x_i | c_i, \boldsymbol{\mu}) d\boldsymbol{\mu}.$$

This integral is intractable, as the time complexity of evaluating it is $\mathcal{O}(K^n)$, which is exponential in K . To evaluate the posterior distribution over the latent variables $p(\boldsymbol{\mu}, \mathbf{c} | \mathbf{x})$, we would have to apply variational inference, approximating it with a variational distribution $q(\boldsymbol{\mu}, \mathbf{c})$. We will assume this distribution follows the mean-field variational family:

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^K q(\mu_k; m_k, s_k^2) \prod_{i=1}^n q(c_i; \boldsymbol{\phi}_i).$$

In this distribution, we have K Gaussian factors with mean μ_k and variance s_k^2 , and n categorical factors with index probabilities defined by the vector $\boldsymbol{\phi}_i$, such that

$$\begin{aligned} \mu_k &\sim N(m_k, s_k^2), & k &= 1, \dots, K, \\ x_i &\sim \text{Categorical}(\boldsymbol{\phi}_i), & i &= 1, \dots, n. \end{aligned}$$

Using this and equation (9), we can derive the evidence lower bound as a function of the variational parameters:

$$\begin{aligned} ELBO(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\phi}) &= \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})] \\ &= \mathbb{E}[\log p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x})] - \mathbb{E}[\log q(\boldsymbol{\mu}, \mathbf{c})] \\ &= \sum_{i=1}^K \mathbb{E}[\log p(\mu_k; m_k, s_k^2)] + \sum_{i=1}^n (\mathbb{E}[\log p(c_i); \boldsymbol{\phi}_i] + \mathbb{E}[\log p(x_i | c_i, \boldsymbol{\mu}); \boldsymbol{\phi}_i, \mathbf{m}, \mathbf{s}^2]) \\ &\quad - \sum_{k=1}^K \mathbb{E}[\log q(\mu_k; m_k, s_k^2)] - \sum_{i=1}^n \mathbb{E}[\log q(c_i; \boldsymbol{\phi}_i)] \end{aligned}$$

From equation (8), we derive the optimal categorical factor by only considering terms from the true distribution $p(\cdot)$ dependent on c_i :

$$q^*(c_i; \boldsymbol{\phi}_i) \propto \exp(\log p(c_i) + \mathbb{E}[\log p(x_i | c_i, \boldsymbol{\mu}); \mathbf{m}, \mathbf{s}^2]). \quad (10)$$

Now since c_i is an indicator vector,

$$p(x_i | c_i, \boldsymbol{\mu}) = \prod_{k=1}^K p(x_i | \mu_k)^{c_{ik}}.$$

We can now evaluate the second term of equation (10):

$$\begin{aligned}
\mathbb{E}([\log p(x_i|c_i, \boldsymbol{\mu}); \mathbf{m}, \mathbf{s}^2]) &= \sum_{k=1}^K c_{ik} \mathbb{E}[\log p(x_i|\mu_k); m_k, s_k^2] \\
&= \sum_{k=1}^K c_{ik} \mathbb{E}[-(x_i - \mu_k)^2/2; m_k, s_k^2] + \text{const} \\
&= \sum_{k=1}^K c_{ik} (\mathbb{E}[\mu_k; m_k, s_k^2] x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2) + \text{const}.
\end{aligned}$$

In each line, terms constant with respect to c_{ik} have been taken out of the expression. Our optimal categorical factor becomes

$$q^*(c_i; \boldsymbol{\phi}_i) \propto \exp \left(\log p(c_i) + \sum_{k=1}^K c_{ik} (\mathbb{E}[\mu_k; m_k, s_k^2] x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2) \right).$$

By proportionality, we then have the variational update

$$\phi_{ik} \propto \exp (\mathbb{E}[\mu_k; m_k, s_k^2] x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2).$$

Now we find the variational density of the k th mixture component, again using equation (8) with the ELBO and ignoring terms independent of $p(\cdot)$ and μ_k :

$$q(\mu_k; m_k, s_k^2) \propto \exp \left(\log p(\mu_k) + \sum_{i=1}^n \mathbb{E}[\log p(x_i|c_i, \boldsymbol{\mu}); \phi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2] \right).$$

The log of this density is

$$\begin{aligned}
\log q(\mu_k) &= \log p(\mu_k) + \sum_{i=1}^n \mathbb{E}[\log p(x_i|c_i, \boldsymbol{\mu}); \phi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2] + \text{const} \\
&= \log p(\mu_k) + \sum_{i=1}^n \mathbb{E}[c_{ik} \log p(x_i|\mu_k); \phi_i] + \text{const} \\
&= -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^n \mathbb{E}[c_{ik}; \phi_i] \log p(x_i|\mu_k) + \text{const} \\
&= -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^n \phi_{ik} \frac{-(x_i - \mu_k)^2}{2} + \text{const} \\
&= -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^n \phi_{ik} x_i \mu_k - \frac{\phi_{ik} \mu_k^2}{2} + \text{const} \\
&= \mu_k \left(\sum_{i=1}^n \phi_{ik} x_i \right) - \mu_k^2 \left(\frac{1}{2\sigma^2} + \frac{\sum_{i=1}^n \phi_{ik}}{2} \right) + \text{const} \\
&= -\frac{1}{2} \left(\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik} \right) \left(\mu_k^2 - \frac{2 \sum_{i=1}^n \phi_{ik} x_i}{1/\sigma^2 + \sum_{i=1}^n \phi_{ik}} \mu_k \right) + \text{const}
\end{aligned}$$

The density is therefore

$$q(\mu_k) \propto \sqrt{\frac{1/\sigma^2 + \sum_{i=1}^n \phi_{ik}}{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{1}{\sigma^2} + \sum_{i=1}^n \phi_{ik} \right) \left(\mu_k - \frac{\sum_{i=1}^n \phi_{ik} x_i}{1/\sigma^2 + \sum_{i=1}^n \phi_{ik}} \right)^2 \right)$$

It can be seen that $q(\mu_k)$ is a Gaussian distribution, so our variational updates for m_k and s_k^2 are its mean and variance:

$$m_k = \frac{\sum_{i=1}^n \phi_{ik} x_i}{1/\sigma^2 + \sum_{i=1}^n \phi_{ik}}, \quad s_k^2 = \frac{1}{1/\sigma^2 + \sum_{i=1}^n \phi_{ik}}.$$

We can now formulate the CAVI algorithm (Algorithm 2), which simply iterates the cluster assignment probabilities ϕ_{ik} and the variational density parameters m_k and s_k^2 until the ELBO converges.

Data: Data \mathbf{x} , Number of Gaussian components K , Hyperparameter value σ^2

Result: Optimal variational factors $q(\mu_k; m_k, s_k^2)$ and $q(c_i; \phi_i)$

begin

 Randomly initialize parameters \mathbf{m}, \mathbf{s}^2 and ϕ ;

while *ELBO has not converged* **do**

for $i = 1$ **to** n **do**

 Set $\phi_{ik} \propto \exp(\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2)$;

end

for $k = 1$ **to** K **do**

 Set $m_k = \frac{\sum_i \phi_{ik} x_i}{1/\sigma^2 + \sum_i \phi_{ik}}$;

 Set $s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \phi_{ik}}$;

end

 Compute $ELBO(\mathbf{m}, \mathbf{s}^2, \phi)$;

end

 Return $q(\mathbf{m}, \mathbf{s}^2, \phi)$;

end

Algorithm 2: CAVI Algorithm for Bayesian mixture of Gaussians

2.1.6 References

To be organised properly and moved to the end later:

https://en.wikipedia.org/wiki/Variational_Bayesian_methods

<http://bjlkeng.github.io/posts/variational-bayes-and-the-mean-field-approximation/>

<https://arxiv.org/pdf/1601.00670.pdf>

Pattern recognition and machine learning by Bishop (2006) pages 461-dunno

<https://www.cs.cmu.edu/~epxing/Class/10708-17/notes-17/10708-scribe-lecture13.pdf>

<http://dept.stat.lsa.umich.edu/~xuanlong/Papers/Nguyen-Wainwright-Jordan-10.pdf>

2.2 Neural Networks

2.2.1 Motivation

Originally, neural networks were an attempt creating an algorithm that mimics the human brain's method of solving problems. The first machines using a neural network structure were created in the 1950s, and they were used widely from the 1980s onwards, as computational power became sufficient for most applications at the time.

One key feature of the brain structure is the capability for the neurons to adapt to suit any purpose. Neuroscientists have conducted experiments on animals where they rewired the optic nerve from the eye to the auditory cortex. They found that the auditory cortex eventually adapted to process the visual signals, and the animals were able to perform tasks requiring sight. This experiment can be repeated for almost any input sensor and the neurons will adjust accordingly to process the signals in a useful manner.

It can be deduced that each neuron has a similar structure, regardless of its location in the brain, in which inputs in the form of electrical signals are changed in some way and outputted to other neurons. Furthermore, a network of neurons is capable of processing almost any input electrical signal in almost any way. These are the core principles behind neural networks.

2.2.2 Neural Network Structure

The primary goal of a neural network is to approximate some function $f^*(\mathbf{x})$ using a mapping with parameters $\boldsymbol{\theta}$ from the input \mathbf{x} to the output \mathbf{y} : $\mathbf{y} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$. In fact, it is known that neural networks can approximate any function (universal approximation theorem). For example, a typical regression problem of estimating housing prices would have the network inputting the values of certain predictors such as size (continuous) and type of building (categorical), and outputting the price. Another example is the classification problem of recognizing handwritten digits (0-9) in a black and white image. There would be many inputs corresponding to the value of each pixel, and the network would have 10 outputs corresponding to the probability of each digit. Another function would be used to select the digit with the highest probability.

Before discussing the overall structure of the neural network, we describe the structure of an individual node. **(I need to insert figure and refer to it)** A typical node takes in inputs from either the external input, or the outputs from other nodes, in addition to a 'bias' node, which is the equivalent of the intercept term in a regression problem. We label these inputs as $\mathbf{x} = [x_0 \ x_1 \ x_2 \ x_3]^\top$, with x_0 corresponding to the bias node. These values are multiplied by weights $\boldsymbol{\theta} = [\theta_0 \ \theta_1 \ \theta_2 \ \theta_3]$, and then passed through an activation function that normalizes the result to a particular range. Denoting the overall node function with $h_{\boldsymbol{\theta}}(\mathbf{x})$ we illustrate the three most common activation functions:

- The rectified linear unit or ReLU activation function output is within the range $[0, \infty)$. It has the formula $h_{\boldsymbol{\theta}}(\mathbf{x}) = \max\{0, \boldsymbol{\theta}^\top \mathbf{x}\}$.
- The sigmoid or logistic activation function outputs are restricted to $[0, 1)$, with the formula $h_{\boldsymbol{\theta}}(\mathbf{x}) = (1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x}))^{-1}$.
- The hyperbolic tangent function output ranges between $(-1, 1)$, denoted as $h_{\boldsymbol{\theta}}(\mathbf{x}) = \tanh(\boldsymbol{\theta}^\top \mathbf{x})$.

A typical neural network is made up of layers of interconnected nodes. The first layer, called the input layer, does not have an activation function or weights, rather it simply acts as an input interface for the network. The outputs from the nodes can only be sent to other nodes in succeeding layers, with

the exception of the final output layer; its result is simply the output of the network. The layers of nodes between the input and output layer are called the hidden layers, as its weights and outputs are not useful to the user. Hidden layers can have any number of nodes, whilst the nodes in the input and output layers are restricted to the number of inputs and outputs the program has. The figure below illustrates a simple neural network with 3 inputs, 1 hidden layer with 3 nodes and 1 output node. **(Steal content from Andrew Ng's Machine Learning course, use same notation is ok?)**

In this example, we denote the activation function as g , the output of unit i in layer j as $a_i^{(j)}$, and the matrix of weights from layer j to $j + 1$ as $\Theta^{(j)}$. We also use the subscript $\Theta_{m,n}^{(j)}$ where m is the row of the matrix corresponding to the unit m in layer $j + 1$, and n is the column of the matrix relating to unit n in layer j .

(Time to learn how to make a dag!)

Individually, the outputs in the hidden nodes and the output node are:

$$\begin{aligned}a_1^{(2)} &= g(\Theta_{1,0}^{(1)}x_0 + \Theta_{1,1}^{(1)}x_1 + \Theta_{1,2}^{(1)}x_2 + \Theta_{1,3}^{(1)}x_3) \\a_2^{(2)} &= g(\Theta_{2,0}^{(1)}x_0 + \Theta_{2,1}^{(1)}x_1 + \Theta_{2,2}^{(1)}x_2 + \Theta_{2,3}^{(1)}x_3) \\a_3^{(2)} &= g(\Theta_{3,0}^{(1)}x_0 + \Theta_{3,1}^{(1)}x_1 + \Theta_{3,2}^{(1)}x_2 + \Theta_{3,3}^{(1)}x_3) \\f_{\Theta}(\mathbf{x}) = a_1^{(3)} &= g(\Theta_{1,0}^{(2)}a_0^{(2)} + \Theta_{1,1}^{(2)}a_1^{(2)} + \Theta_{1,2}^{(2)}a_2^{(2)} + \Theta_{1,3}^{(2)}a_3^{(2)})\end{aligned}$$

Denoting the weights outputting to unit i in layer $j + 1$ as $\boldsymbol{\theta}_i^{(j)} = [\Theta_{i,0}^{(j)} \quad \Theta_{i,1}^{(j)} \dots \Theta_{i,k}^{(j)}]^{\top}$ where $k + 1$ is the number of inputs, we have the vectorized notation:

$$\begin{aligned}a_1^{(2)} &= g(\boldsymbol{\theta}_1^{(1)\top} \mathbf{x}) \\a_2^{(2)} &= g(\boldsymbol{\theta}_2^{(1)\top} \mathbf{x}) \\a_3^{(2)} &= g(\boldsymbol{\theta}_3^{(1)\top} \mathbf{x}) \\f_{\boldsymbol{\theta}}(\mathbf{x}) = a_1^{(3)} &= g(\boldsymbol{\theta}_1^{(2)\top} \mathbf{a}^{(2)})\end{aligned}$$

where $\mathbf{a}^{(2)} = [a_1^{(2)} \quad a_2^{(2)} \quad a_3^{(2)}]^{\top}$.

An even simpler notation is:

$$\begin{aligned}\mathbf{a}^{(2)} &= g(\Theta^{(1)\top} \mathbf{x}) \\f_{\Theta}(\mathbf{x}) = \mathbf{a}^{(3)} &= g(\Theta^{(2)\top} \mathbf{a}^{(2)})\end{aligned}$$

2.2.3 Back-Propagation

The goal of back-propagation is to train the weights of the network such that the squared error cost function, which we will denote as L is minimized:

$$\min_{\Theta} L(\Theta) = \frac{1}{2}(\mathbf{y} - \mathbf{f}_{\Theta}(\mathbf{x}))^{\top}(\mathbf{y} - \mathbf{f}_{\Theta}(\mathbf{x})).$$

The $\frac{1}{2}$ factor is included to eliminate the factor of 2 in the derivative, simplifying the derivations. The derivative is multiplied by an arbitrary training rate during optimization so there is no significant impact of including that term.

In the back-propagation algorithm, the weights are initialized randomly, and the goal is to find the partial derivative of the loss function with respect to the individual weights

$$\frac{\partial}{\partial \Theta_{m,n}^{(j)}} L(\Theta),$$

so that gradient descent can be performed to optimize the weights. For each training sample $(\mathbf{x}^{(I)}, \mathbf{y}^{(I)})$, $I = 1, \dots, N$, the input signal is propagated forward throughout the network to calculate $\mathbf{a}^{(j)}$ for $j = 2, \dots, J$, where J is the total number of layers. The difference between the network output and the ideal result is calculated with

$$\boldsymbol{\delta}^{(J)} = \mathbf{a}^{(J)} - \mathbf{y}^{(I)},$$

and this error is propagated backwards through the network to find $\boldsymbol{\delta}^{(J-1)}, \dots, \boldsymbol{\delta}^{(2)}$ by using the formula

$$\boldsymbol{\delta}^{(j)} = ((\boldsymbol{\Theta}^{(j)})^\top \boldsymbol{\delta}^{(j+1)}) .* g'(\boldsymbol{\Theta}^{(j)\top} \mathbf{a}^{(j)}),$$

where $.*$ denotes element-wise multiplication and g' is the derivative of the activation function. In this case, g' takes in the sum of its weighted inputs, and as an example, the sigmoid activation function has the derivative $g'(\boldsymbol{\Theta}^{(j)\top} \mathbf{a}^{(j)}) = \mathbf{a}^{(j)} .* (1 - \mathbf{a}^{(j)})$. Note that $\boldsymbol{\delta}^{(1)}$ does not need to be calculated as the input layer is not weighted.

The errors for each layer are multiplied by each of the preceeding layer's activation outputs to form the estimated partial derivative for the training sample. This result is added to an accumulator matrix, so that the average partial derivative from all the training samples can be computed:

$$\Delta_{m,n}^{(j)} := \Delta_{m,n}^{(j)} + a_n^{(j)} \delta_m^{(j+1)}$$

or in matrix-vector form.

$$\Delta^{(j)} := \Delta^{(j)} + \boldsymbol{\delta}^{(j+1)} (\mathbf{a}^{(j)})^\top.$$

Finally, we divide the accumulator matrix entries by the number of training samples to find the average partial derivative of the cost function with respect to the weights:

$$\frac{\partial}{\partial \Theta_{m,n}^{(j)}} L(\Theta) = \frac{1}{N} \Delta_{m,n}^{(j)}$$

When $n \neq 0$ (i.e. not considering the bias node), we can optionally add a regularizer term $\lambda > 0$ which decreases the magnitude of the weights, preventing overfitting.

$$\frac{\partial}{\partial \Theta_{m,n}^{(j)}} L(\Theta) = \frac{1}{N} (\Delta_{m,n}^{(j)} + \lambda \Theta_{m,n}^{(j)})$$

There is no significant change when the bias node is regularized.

Pseudocode for the back-propagation algorithm is shown below.

Having derived the partial derivatives of the loss function with respect to the individual weights, we can use gradient descent or some other optimization method to update the weights. The partial derivatives are re-calculated after each optimization update until convergence.

Data: Training Data $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$, Regularizer Term λ
Result: Cost Function Partial Derivatives $\frac{\partial}{\partial \Theta_{m,n}^{(j)}} L(\Theta)$

```

begin
  Randomly initialize weights  $\Theta$ ;
  Set  $\Delta_{m,n}^{(j)} = 0 \quad \forall j, m, n$ ;
  for  $I = 1$  to  $N$  do
    Set  $\mathbf{a}^{(1)} = \mathbf{x}^{(I)}$ ;
    for  $j = 2$  to  $J$  do
      Set  $\mathbf{a}^{(j)} = \Theta^{(j-1)\top} \mathbf{a}^{(j-1)}$ ;
    end
    Set  $\boldsymbol{\delta}^{(J)} = \mathbf{a}^{(J)} - \mathbf{y}^{(I)}$ ;
    for  $j = J - 1$  to  $2$  do
      Set  $\boldsymbol{\delta}^{(j)} = ((\Theta^{(j)})^\top \boldsymbol{\delta}^{(j+1)}) \cdot g'(\Theta^{(j)\top} \mathbf{a}^{(j)})$ ;
    end
    for  $j = 1$  to  $J - 1$  do
      Set  $\Delta^{(j)} = \Delta^{(j)} + \boldsymbol{\delta}^{(j+1)} (\mathbf{a}^{(j)})^\top$ ;
    end
  end
  for all  $j, m, n$  do
    if  $n = 0$  then
      Set  $\frac{\partial}{\partial \Theta_{m,n}^{(j)}} L(\Theta) = \frac{1}{N} \Delta_{m,n}^{(j)}$ ;
    else
      Set  $\frac{\partial}{\partial \Theta_{m,n}^{(j)}} L(\Theta) = \frac{1}{N} (\Delta_{m,n}^{(j)} + \lambda \Theta_{m,n}^{(j)})$ ;
    end
  end
end

```

Algorithm 3: Back-Propogation Algorithm

2.2.4 Weight Initialization

Proper initialization of the weights is ideal to improve convergence, as if the weights are too low, then the nodal outputs will continually decrease through the layers and become very small, requiring many iterations of back-propagation training to fix. Similarly, if the weights are too high, then the result output of forward propagation will be extremely large. In this section we discuss Xavier Initialization, which aims to keep the signal variance constant throughout the network. To derive the initialization algorithm, first consider a single node with n inputs, and let z denote the weighted sum of the inputs $\boldsymbol{\theta}^\top \mathbf{x}$ before it is passed through the activation function. This is written as

$$z = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n.$$

x_0 is a constant term, so $\text{Var}(\theta_0 x_0) = 0$. Now under the assumption that the inputs and weights have 0 mean, we find the variance of the other terms:

$$\text{Var}(\theta_i x_i) = \mathbb{E}[x_i]^2 \text{Var}(\theta_i) + \mathbb{E}[\theta_i]^2 \text{Var}(x_i) + \text{Var}(\theta_i) \text{Var}(x_i) = \text{Var}(\theta_i) \text{Var}(x_i).$$

Assuming that the weights and inputs are also independent and identically distributed, we have

$$\text{Var}(z) = n \text{Var}(\theta_i) \text{Var}(x_i).$$

Since we want constant variance of the signals throughout the network, we set $\text{Var}(z) = \text{Var}(x_i)$ and the result follows:

$$\text{Var}(\theta_i) = \frac{1}{n}.$$

However, this result only considers forward propagation of the signal. A variation of this result accounts for back propagation by averaging the number of input and output nodes:

$$\text{Var}(\theta_i) = \frac{2}{n_{in} + n_{out}}.$$

Thus, to enforce constant signal variance throughout the network, the ideal initialization of weights is to sample from a distribution, typically uniform or Gaussian, with 0 mean and $\frac{2}{n_{in} + n_{out}}$ variance:

$$\theta_i \sim U\left(-\sqrt{\frac{6}{n_{in} + n_{out}}}, \sqrt{\frac{6}{n_{in} + n_{out}}}\right)$$

or

$$\theta_i \sim N\left(0, \frac{2}{n_{in} + n_{out}}\right).$$

2.2.5 References

<http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf>
Machine Learning by Andrew Ng Stanford online course

3 Experiments

3.1 "Sprinkler" Example

3.2 Lotka-Volterra Predator-Prey model

3.3 Autoencoding Variational Bayes (MNIST image generation)

Kingma 2013

3.3.1 Problem Context

Introduce notations, similar to VB but with NN parametrization. Discuss problems with intractability and large dataset, introduce the 3 main things we want to solve.

3.3.2 Structure

x) Diagram of VAE

3.3.3 Variational Bound

Take ELBO from VB section but parametrize it with NN, discuss the typical useless Monte Carlo gradient estimator.

3.3.4 Algorithm

Propose better estimator of lower bound and its derivatives, derive AEVB algorithm.

3.3.5 Reparametrization Trick

Explain the trick and show changes in structure.

3.3.6 Example: Variational Autoencoder (for MNIST? or cats :) or human faces D:)

Show pseudocode and diagram for VAE (if different from before) and show output from own VAE code (gonna have to code my own VAE).

3.4 no idea lol