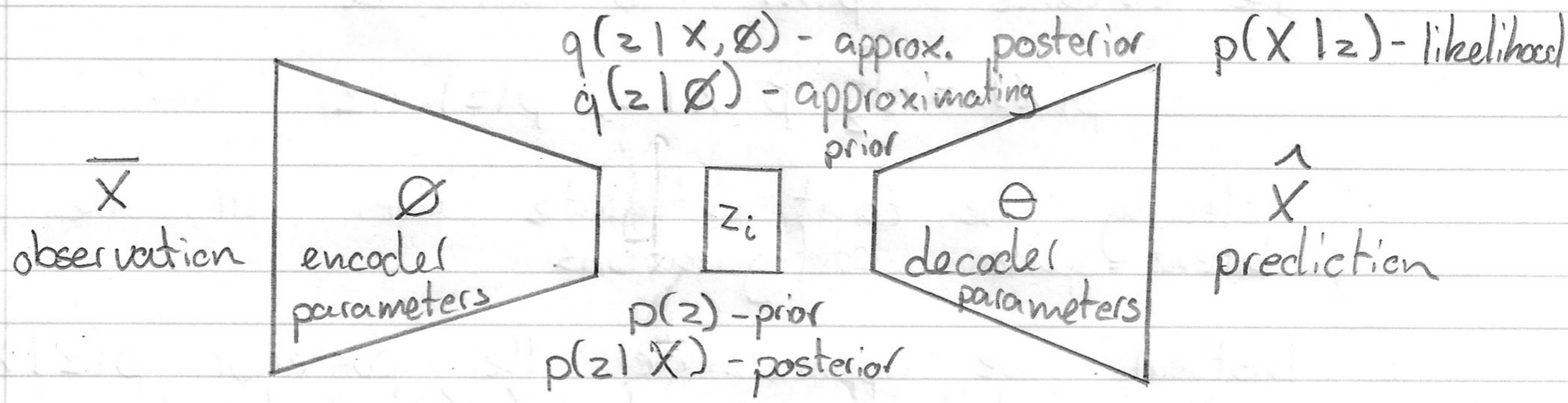
VAE 1

Training a VAE



$X = \{\bar{X}, \hat{X}\}$ the set of paired observations & predictions (NOTE the predictions aren't labels, they are generated!)

we "sample" a distribution $z_i \sim p(z)$
please look this up if you don't understand it!

It is not possible to take derivatives using SGD to train this, we will see why not mathematically

We can determine the posterior $p(z|X)$ using Bayes' equation

$$p(z|X) = \frac{p(X|z)\ p(z)}{p(X)}$$

where $p(\cdot)$ is a probability distribution

For example the likelihood can be a gaussian

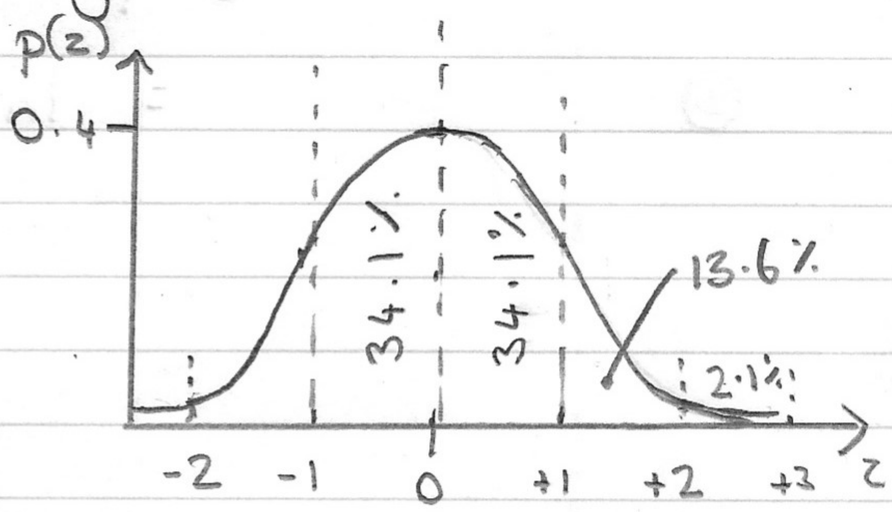
$$p(X|z) = \mathcal{N}(0, \Sigma)$$

$$= \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(\frac{1}{2}\frac{X^2}{\Sigma^2}\right) \quad \text{where } X = \bar{X} - \hat{X}$$

or the prior can be a gaussian

$$p(z) = \mathcal{N}(0, 1)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$$



where 95% values between [-2, 2]
99.6% values between [-3, 3]

We cant solve Bayes' equation directly because the evidence $p(X)$ is intractable

$$p(X) = \int_z p(X|z)\, p(z)\, dz$$

clearly we cant integrate over all the possible latent-space configurations

Instead we approximate the posterior $p(z|X)$ with another distribution $q(z|X, \phi)$ which depends on the output of the encoder.

We find the best approximating distribution when the KL divergence is minimised

$$\text{(A)} \qquad q(X|z, \phi^*) = \min_{\phi} KL\left[q(z|X, \phi) \,\|\, p(z|X)\right]$$

Clearly this expression still has issues because it contains the posterior

ELBO

The Evidence Lower Bound is a way to rearrange Eq A so we can solve it

$$KL\left[q_\phi \,|\, p(z|X)\right] = \int_z q_\phi \ln\left[\frac{q_\phi}{p(z|X)}\right] dz$$

$$= \int_z q_\phi \ln\left[\frac{q_\phi\ p(X)}{p(z, X)}\right] dz$$

$$= \int_z q_\phi \ln\left[p(X)\right] + q_\phi \ln\left[\frac{q_\phi}{p(z, X)}\right] dz$$

Note $\int q(z|X, \phi)\, dz = 1$

$$\text{(B)} \qquad = \ln\left[p(X)\right] + \int_z q_\phi \ln\left[\frac{q_\phi}{p(z, X)}\right] dz$$

We also know that $KL[q_\phi \| p(z|X)] \geq 0$ by definition, so we can rewrite Eq B

$$0 \leq \ln[p(X)] + \int_z q_\phi \ln\left[\frac{q_\phi}{p(z,X)}\right] dz$$

(C) $$\ln[p(X)] \geq -KL[q_\phi \| p(z,X)]$$

So the evidence is bounded below by the negative of the KL divergence of the approximating & joint distributions

REMEMBER probability distributions (ie probabilities) are always less than 1 so

$$p(X) \leq 1 \quad \text{therefore} \quad \ln[p(X)] \leq 0$$

(Eq A) It should be possible to see that minimising $KL[q_\phi | p(z|X)]$ involves minimising the $\log_x$ evidence, which can be found by minimising the Right Hand Side of Eq C.

We want to use a gradient based optimiser, but how do we take a gradient of Eq C?

Gradient of the KL divergence

We can rewrite Eq A using Eq C.

$$q(z|X,\phi^*) = \min_\phi -KL[q_\phi \| p(z,X)]$$

(D) $$= \min_{\phi,\theta} \int_z \underbrace{q_\phi \ln[p(X|z,\theta,\phi)]}_{\text{Term 1}} - \underbrace{q_\phi \ln\left[\frac{q_\phi}{p(z)}\right]}_{\text{Term 2}} dz$$

When we perform SGD we want the gradient estimate at each iteration to be correct "on average"

The mathematical way to do this is say that "the gradient of the expectation of $p(X|z,\theta,\phi)$" should be the same as "the expectation of the gradient of $p(X|z,\theta,\phi)$"

Unfortunately we can't do this transform directly as it leads to a "Biased estimator"

Note the definition of the Expectation of $f(y,x)$ with respect to a probability $p(x)$

$$\mathbb{E}_{p(x)}[f(y,x)] = \int_x p(x)\, f(y,x)\, dx$$

So for Term 1 in Eq D

$$\nabla_\phi \mathbb{E}_{q_\phi}[p(X|z,\theta,\phi)] = \nabla_\phi \int_z q(z|X,\phi) \ln[p(X|z,\theta,\phi)]\, dz$$
$$\neq \int_z q(z|X,\phi) \nabla_\phi \ln[p(X|z,\theta,\phi)]\, dz$$
$$\neq \mathbb{E}_{q_\phi}[\nabla_\phi \ln[p(X|z,\theta,\phi]]$$

This means when we have a specific observation $X_i \sim p(X)$ we cant use Monte Carlo methods to take the derivative wrt. ϕ

wrt "with respect to"

How do we get to an "unbiased estimator of the gradient"?

The log derivative trick

The "log derivative trick" or "score function method" is the first of two approaches to finding the unbiased gradient estimator

$$\nabla_\phi \mathbb{E}_{q_\phi}[\ln[p(X|z,\theta,\phi)]] = \int_z \nabla_\phi [q(z|X,\phi) \ln[p(X|z,\theta,\phi)]]\, dz$$
$$= \int_z q(z|X,\phi)\, \nabla_\phi \ln[p(X|z,\theta,\phi)] + \ln[p(X|z,\theta,\phi)] \nabla_\phi q(z|X,\phi)\, dz$$
$$= \int q_\phi \nabla_\phi \ln[p(X|z,\theta,\phi)] + \ln[p(X|z,\theta,\phi)]\, q_\phi\, \nabla_\phi \ln[q_\phi]\, dz$$
$$F = \mathbb{E}_{q_\phi}[\nabla_\phi \ln[p(X|z,\theta,\phi)] + p(X|z,\theta,\phi)\, \nabla_\phi \ln[q_\phi]]$$

where we use the rule $\nabla_x f(x) = f(x)\, \nabla_x \ln[f(x)]$

Great!

The problem is that this estimator has a <u>high variance</u> because $0 \leq q_\phi \leq 1$ which means $\ln[q_\phi] << 0$ (sometimes)

This is a problem once we sample to take the gradients

We have as Eq F

$$z_i \sim q(z \mid X, \emptyset) \qquad \hat{X}_i \sim p(X \mid z_i, \theta, \emptyset)$$

$$\nabla_\emptyset \mathbb{E}_{q_\emptyset}[p(X \mid z, \theta, \emptyset)] = n\left[\nabla_\emptyset (\bar{X}_i - \hat{X}_i)^2 + (\bar{X}_i - \hat{X}_i)^2 \nabla_\emptyset z_i^2\right]$$

which is the MC gradient estimate, $\bar{X}_i$ for example is a single observed image

The reparameterisation trick

This is an approach which reduces the variance of the gradient estimates

Sometimes the distribution chosen for the latent variable can be rewritten as a "deterministic function" of another distribution with constant parameters

Mathematically $q(z \mid X, \emptyset) = g(p(\epsilon), \{\lambda\})$.

An example is necessary to understand this. Consider a normally distributed latent space

$$q(z \mid X, \emptyset) = \mathcal{N}(\mu, \Sigma)$$

In general normal distributions

$$g(p(\epsilon), \{\lambda\}) = \mu + \Sigma \mathcal{N}(0, 1)$$

where $p(\epsilon) = \mathcal{N}(0, 1)$

$\{\epsilon\} = \{0, 1\}$ NOTE these are constants!

$\{\lambda\} = \{\mu, \Sigma\}$ (the parameters of the gaussian)

This means we now have differentiable parameters in the latent space, where the parameters are each functions of the encoder $\{\lambda\} = \{\mu(\emptyset), \Sigma(\emptyset)\}$

This means Term 1 Eq D becomes

$$q(z \mid X, \emptyset^*) = \min_{\emptyset, \theta} \int_z q(z \mid X, \emptyset) \ln[p(X \mid z, \theta, \emptyset)] \, dz$$

$$= \min_{\emptyset, \theta} \int_z p(\epsilon) g(\mathcal{N}(0,1), \mu, \Sigma) \ln[p(X \mid z, \theta, \emptyset)] \, dz$$

Differentiating this expression is really straight-forward

Note
$q(\epsilon) = N(0,1)$

$$\nabla_{\phi} E_{q(\epsilon)}[g(N(0,1), \mu, \Sigma) \ln[p(X|z,\theta,\phi)]]$$

$$= \int_{\epsilon} q(\epsilon) \nabla_{\phi}[g(q(\epsilon), \mu, \Sigma) \ln[p(X|z,\theta,\phi)]] \, d\epsilon$$

$$= \int_{\epsilon} q(\epsilon) [\ln[p(X|z,\theta,\phi)] \underline{\underline{\nabla_{\mu}}} g(q(\epsilon), \mu, \Sigma) \nabla_{\phi} \mu(\phi) + \ldots$$

$$\ldots g(q(\epsilon), \mu, z) \nabla_{\phi} \ln[p(X|z,\theta,\phi)]] \, d\epsilon$$

but does look a bit complicated

Note that an additional expression is required for each additional parameter of $g(\cdot)$. In this case we will have one additional expression where $\underline{\underline{\nabla_{\mu}}}$ will be replaced by $\underline{\underline{\nabla_{\Sigma}}}$