Variational Autoencoder (VAE) Mathematics: A Revision Guide

This document summarizes the key mathematical concepts behind Variational Autoencoders (VAEs), focusing on the latent space, normal distributions, KL-divergence, and the reparameterization trick. It is designed for revision, with intuitive explanations and examples.

1. What is a VAE?

A VAE is a neural network that:

- Encodes input data x (e.g., images) into a compressed latent representation z.
- Decodes z to reconstruct x or generate new data.
- Balances reconstruction accuracy with a structured latent space for generation.

2. Core Goal: Modeling $p(z|x)$

The goal is to find the posterior distribution $p(z|x)$, the probability of latent variables z given input x:

$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)}$$

- $p(x|z)$: Likelihood (decoder), how likely x is given z.
- $p(z)$: Prior, typically a standard normal $N(0, I)$.
- $p(x) = \int p(x|z) \cdot p(z)\, dz$: Evidence, often intractable, making $p(z|x)$ hard to compute.

Solution: Approximate $p(z|x)$ with a simpler distribution $q(z|x)$, learned by the encoder.

3. Key Distributions

Both $q(z|x)$ and $p(z)$ are normal distributions for computational simplicity.

3.1 $q(z|x)$: Approximate Posterior

- Form: Multivariate normal, $q(z|x) = N(\mu(x), \text{diag}(\sigma^2(x)))$.
- Parameters: Encoder outputs $\mu(x)$ (mean) and $\sigma(x)$ (standard deviation) for each x.
- Normalization: Integrates to 1, $\int q(z|x)\, dz = 1$, as a valid probability density function (PDF).
- Role: Represents likely latent codes for a given input.

3.2 $p(z)$: Prior

- Form: Standard normal, $p(z) = \mathcal{N}(0, I)$ (mean 0, variance 1 per dimension).
- Fixed: Not learned, defines the latent space structure.
- Normalization: Integrates to 1, $\int p(z)\,dz = 1$.
- Why Chosen: Easy sampling, analytical KL-divergence, smooth latent space.

4. VAE Loss Function

The VAE minimizes:

$$\text{Loss} = \text{Reconstruction Loss} + \text{KL-Divergence}$$

4.1 Reconstruction Loss

- Purpose: Measures how well the decoder reconstructs x from z.
- Form: Often mean squared error or binary cross-entropy:

$$L(x, \hat{x}) = \mathbb{E}_{q(z|x)}[\log p(x|z)]$$

- Example: Ensures pixel values of reconstructed image $\hat{x}$ match input x.

4.2 KL-Divergence

- Purpose: Measures how close $q(z|x)$ is to $p(z)$, regularizing the latent space.
- Form: For normal distributions:

$$\text{KL}(q(z|x)||p(z)) = \frac{1}{2}\sum_{i=1}^{k}\left[\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1\right]$$

  where k is the latent space dimensionality.
- Components: Penalizes $\mu_i \neq 0$, $\sigma_i^2 \neq 1$, and small variances.
- Example: For 1D, $\mu = 0.5$, $\sigma = 0.8$:

$$\text{KL} = \frac{1}{2}[0.5^2 + 0.8^2 - \log(0.8^2) - 1] \approx 0.173$$

- Note: Not always 1; varies with $\mu(x)$ and $\sigma(x)$.

5. Reparameterization Trick

Enables differentiable sampling from $q(z|x)$ for training.

- Problem: Direct sampling from $\mathcal{N}(\mu(x), \sigma(x))$ is not differentiable.
- Solution: Reparameterize:

$$z = \mu(x) + \sigma(x) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- Steps:
  1. Encoder outputs $\mu(x)$ and $\log \sigma(x)$ (for stability).
  2. Compute $\sigma(x) = \exp(\log \sigma(x))$.
  3. Sample $\epsilon \sim \mathcal{N}(0, I)$.
  4. Compute z, pass to decoder.
- Why It Works: Randomness in ϵ; operations with $\mu(x)$ and $\sigma(x)$ are differentiable.
- Example: For $\mu(x) = 2$, $\sigma(x) = 0.5$, $\epsilon = 1$:

$$z = 2 + 0.5 \cdot 1 = 2.5$$

6. Why Normal Distributions?

- Tractability: Closed-form KL-divergence.
- Sampling: Easy with reparameterization trick.
- Latent Space: $p(z) = \mathcal{N}(0, I)$ ensures a smooth, centered space.

7. Latent Space

- Space of all z values, structured by $p(z) = \mathcal{N}(0, I)$.
- $q(z|x)$ maps inputs to regions in this space.
- KL-divergence ensures $q(z|x) \approx p(z)$, enabling generation by sampling $z \sim p(z)$.

8. Visualization of Reparameterization Trick

- Concept: Transform a sample from $\mathcal{N}(0, 1)$ to $q(z|x)$.
- Example: For $q(z|x) = \mathcal{N}(2, 0.5^2)$, sample $\epsilon = 1$:

$$z = 2 + 0.5 \cdot 1 = 2.5$$

- Visualized as shifting/scaling a point on the $\mathcal{N}(0, 1)$ curve to the $q(z|x)$ curve.

This summary covers the VAE's mathematical foundation about distributions, KL-divergence, and the reparameterization trick.