

1 Definitions

1. **Variational Autoencoder:** A VAE is a generative model that encodes data into a probability distribution over a latent space, then decodes samples from that distribution to reconstruct the input.
2. **ELBO:** The Evidence Lower Bound (ELBO) is a tractable lower bound on the log-likelihood of the data, used to optimize the Variational Autoencoder (VAE). Since the true marginal likelihood $\log p(x)$ is intractable, we instead maximize:

$$\text{ELBO}(x) = \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x) \| p(z))$$

Maximizing the ELBO encourages the model to reconstruct the data well while keeping the learned latent distribution $q(z|x)$ close to the prior $p(z)$.

3. **Reconstruction Loss:** Reconstruction loss measures how well the VAE decoder can reconstruct the original input x from the latent representation z . This term ensures that the decoder preserves important information from the input.
4. **KL Divergence:** KL divergence is a measure of how much a distribution deviates from another:

$$D_{\text{KL}}(p(x) \| q(x)) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

For VAEs, we use this to quantify how much the approximate posterior $q(z|x)$ deviates from the prior distribution $p(z)$. Assuming $p(z)$ is a standard normal and $q(z|x) \sim \mathcal{N}(\mu, \sigma^2)$, for example, gives us:

$$D_{\text{KL}}(q(z|x) \| p(z)) = \frac{1}{2} \sum_j (\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1)$$

Minimizing this term regularizes the latent space by encouraging $q(z|x)$ to remain close to a known prior, typically $\mathcal{N}(0, I)$.

5. **Reparameterization Trick:** The reparameterization trick enables backpropagation by expressing sampling in a differentiable way. Rather than sampling directly from $z \sim \mathcal{N}(\mu, \sigma^2)$, we write:

$$z = \mu + \sigma \cdot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, 1)$$

While these are functionally equivalent, this formulation isolates the randomness in a separate variable ϵ , allowing gradients to propagate through μ and σ during training.

6. **Jensen's Inequality:** Jensen's inequality states that for any convex function $f(x)$, the secant line should always lie above the graph. Mathematically,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

for $t \in [0, 1]$. This can be extended to probability theory, where

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

when φ is convex. As an example, $-\log$ is convex, so

$$-\log(\mathbb{E}[X]) \leq -\mathbb{E}[\log(X)] \implies \log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$$

2 KL Divergence

2.1 Experimenting with Distributions

The associated [Colab notebook](#) visualizes two univariate Gaussian distributions and computes the KL divergence between them for varying values of mean and variance. Use this interactive tool to explore the behavior of KL divergence, and answer the following questions based on your observations and understanding.

Given two univariate Gaussian distributions:

$$P = \mathcal{N}(\mu_1, \sigma_1^2), \quad Q = \mathcal{N}(\mu_2, \sigma_2^2)$$

The KL divergence from P to Q is given by:

$$D_{\text{KL}}(P \parallel Q) = \frac{1}{2} \left[\frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_2 - \mu_1)^2}{\sigma_2^2} - 1 + \log \left(\frac{\sigma_2^2}{\sigma_1^2} \right) \right]$$

1. Under what conditions does $D_{\text{KL}}(P \parallel Q) = D_{\text{KL}}(Q \parallel P)$? What can we infer about the symmetry of KL divergence from this? Do the results we observe for Gaussians extend to other distributions?

2. When does the KL divergence become exactly zero? Is this condition unique, or can multiple different pairs of distributions yield zero KL divergence?

3. Is it possible for KL divergence to be negative for these distributions?

2.2 Divergence vs Distance

For a function $d(P, Q)$ to qualify as a distance metric, it must satisfy the following properties:

- **Non-negativity:** $d(P, Q) \geq 0$
- **Identity of indiscernibles:** $d(P, Q) = 0$ if and only if $P = Q$
- **Symmetry:** $d(P, Q) = d(Q, P)$
- **Triangle inequality:** $d(P, R) \leq d(P, Q) + d(Q, R)$

1. KL divergence is **not** considered a distance metric, because it does not satisfy some of the required properties. Which of these are satisfied by KL divergence?

2. Prove that KL divergence does not satisfy the triangle inequality. Show the following claim: you can choose variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$, such that the distributions $p = N(0, \sigma_1^2), q = N(0, \sigma_2^2), r = N(0, \sigma_3^2)$ violate the triangle inequality, i.e. show that:

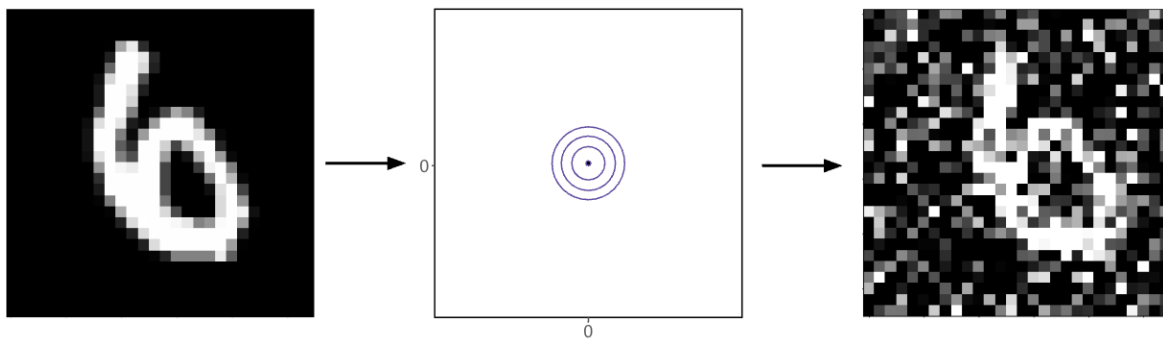
$$D_{\text{KL}}(p \parallel r) > D_{\text{KL}}(p \parallel q) + D_{\text{KL}}(q \parallel r)$$

3 VAEs and Loss

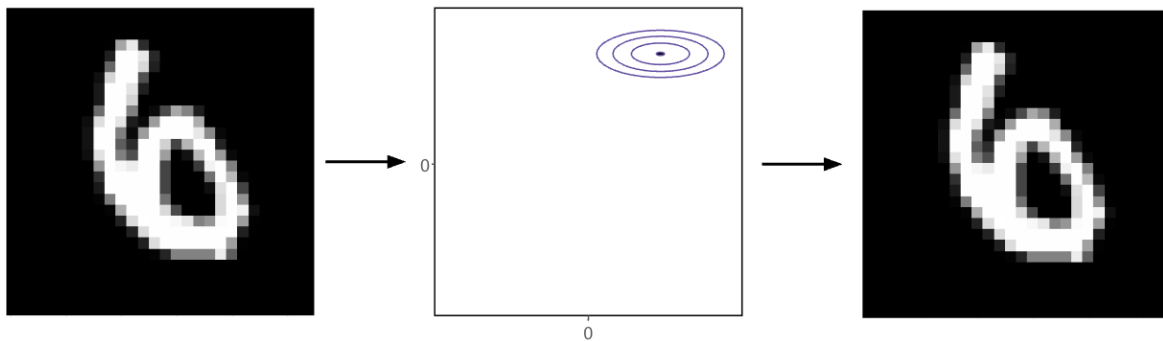
1. What is the difference between a standard autoencoder and a variational autoencoder (VAE)? Consider the differences in the latent space, loss functions, and generative capability. Why is the latent space in VAEs more suitable for generative tasks?

2. Show that the Evidence Lower Bound (ELBO) is a lower bound on the log-likelihood of the data.

3. Each row in the figures below shows the output of a Variational Autoencoder (VAE) for the same input image of the digit **6**.



VAE 1



VAE 2

- The **left image** is the original input x .
- The **middle plot** is a visualization of the encoder's latent embedding in the 2D space.
- The **right image** is the decoder's reconstruction.

Answer the following questions based on these examples.

- (a) Which example (1 or 2) has the higher **reconstruction error**? How can you tell?

- (b) Which example (1 or 2) has the higher **KL divergence** between $q(z|x)$ and the prior $p(z)$ (the prior is a standard normal distribution)? Why?

- (c) Assuming we calculate the KL divergence and the reconstruction error for both of these data points, how do we get the total VAE loss for this point?