# 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:

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

   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:

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

   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:

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

   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 $\epsilon$, allowing gradients to propagate through $\mu$ and $\sigma$ 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 $\varphi$ 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:

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

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

   We can see that these are only equal when the two distributions are equal, or when the variances are the same but the means differ. This tells us that KL divergence is not symmetric in general. For other distributions, the two values will be the same when the distributions are equal, but may not necessarily be equal when the variances are equal.

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

   KL divergence only becomes zero when the two distributions are exactly equal. This is unique, and no possible pair of different distributions can yield a value of zero.

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

   No, it is not possible for KL divergence to be negative.

## 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?

   Non-negativity and identity of indiscernibles. In general, KL divergence does not satisfy symmetry or triangle inequality.

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:

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

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel r) &= \mathbb{E}_p\left[\log\left(\frac{\sigma_3}{\sigma_1} \cdot e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_1^2} - \frac{x^2}{\sigma_3^2}\right)}\right)\right] \\ &= \mathbb{E}_p\left[\log\left(\frac{\sigma_3}{\sigma_1}\right) - \frac{1}{2}\left(\frac{x^2}{\sigma_1^2} - \frac{x^2}{\sigma_3^2}\right)\right] \\ &= \log\left(\frac{\sigma_3}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_3^2} - 1\right) \quad (\text{since } \mathbb{E}_p[x^2] = 1) \end{aligned}$$

$$\text{LHS: } D_{\mathrm{KL}}(p \parallel r) = \log\left(\frac{\sigma_3}{\sigma_1}\right) + \frac{\sigma_1^2}{2\sigma_3^2} - \frac{1}{2}$$

$$\begin{aligned} \text{RHS: } D_{\mathrm{KL}}(p \parallel q) + D_{\mathrm{KL}}(q \parallel r) &= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2} + \log\left(\frac{\sigma_3}{\sigma_2}\right) + \frac{\sigma_2^2}{2\sigma_3^2} - \frac{1}{2} \\ &= \log\left(\frac{\sigma_2}{\sigma_1} \cdot \frac{\sigma_3}{\sigma_2}\right) + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{\sigma_2^2}{2\sigma_3^2} - 1 \\ &= \log\left(\frac{\sigma_3}{\sigma_1}\right) + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{\sigma_2^2}{2\sigma_3^2} - 1 \end{aligned}$$

Hence we want to find $\sigma_1, \sigma_2, \sigma_3$ such that
$\frac{\sigma_1^2}{\sigma_3^2} > \frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_3^2} - 1$

One possibility: $\sigma_1^2 = 1, \sigma_2^2 = 2, \sigma_3^2 = 4$

# 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?

   - **Latent Representation:** Autoencoders map inputs to points; VAEs map inputs to distributions.
   - **Loss Function:** Autoencoders minimize reconstruction error; VAEs add a KL divergence term to align the latent distribution with a prior (typically $\mathcal{N}(0, I)$).
   - **Generative Modeling:** VAEs can generate new, coherent data by sampling from the latent prior. Autoencoders cannot do this in a principled way.

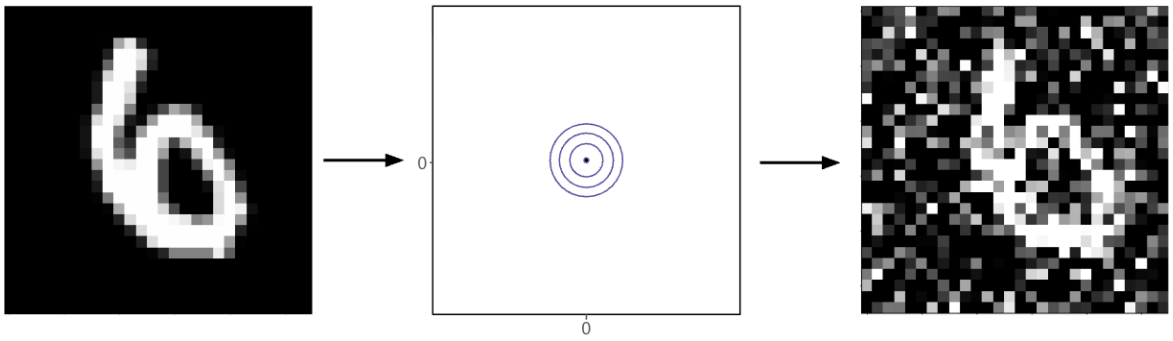
   Because the VAE forces the latent space to follow a continuous, smooth distribution, sampling becomes well-defined. The KL divergence term ensures that similar inputs are encoded to nearby regions, and the decoder learns to handle points throughout the latent space — not just sparse, discrete encodings.

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

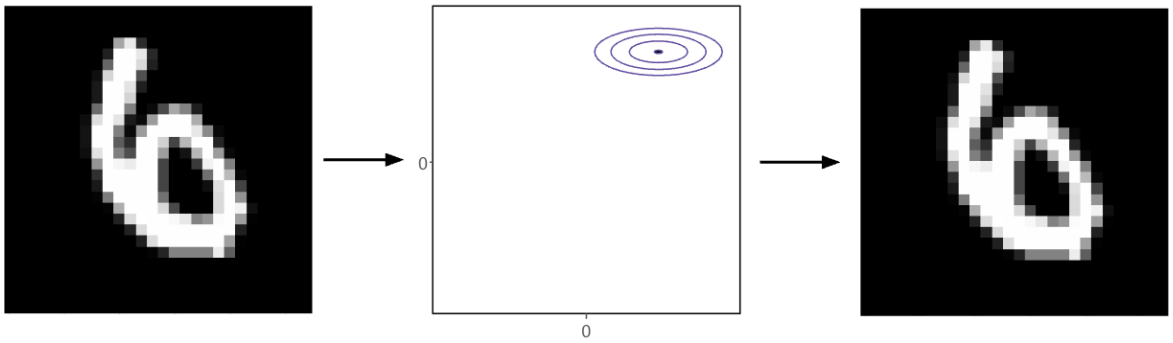
   $\phi$ parametrizes the encoder, $\theta$ parametrizes the decoder. $x$ represents the inputs to the model and $z$ is the latent/compressed representation produced by the encoder.

$$
\begin{aligned}
\text{MLE:} \quad & \arg\max_{\theta,\phi} \log p(x) \\
& = \log \int_z p_\theta(x \mid z)\, p(z)\, dz \\
& = \log \int_z p_\theta(x \mid z)\, p(z)\, \frac{q_\phi(z \mid x)}{q_\phi(z \mid x)}\, dz \\
& = \log \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \frac{p_\theta(x \mid z)\, p(z)}{q_\phi(z \mid x)} \right] \\
& \text{Using Jensens Inequality} \\
& \geq \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log \left( \frac{p_\theta(x \mid z)\, p(z)}{q_\phi(z \mid x)} \right) \right] \\
& = \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log p_\theta(x \mid z) \right] + \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log \frac{p(z)}{q_\phi(z \mid x)} \right] \\
& = \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log p_\theta(x \mid z) \right] - D_{\mathrm{KL}} \left( q_\phi(z \mid x) \,\|\, p(z) \right)
\end{aligned}
$$

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?

VAE 1 has higher reconstruction error, since the resulting image is not as similar to the input image.

(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?

VAE 2 has higher KL divergence, since the latent distribution of VAE 1 more closely resembles a standard Gaussian centered at 0, as compared to VAE 2.

(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?

Taking the sum of the KL divergence and the reconstruction error will give us the total VAE loss for this point.