Derivation of KL Divergence for Multivariate Gaussian Distributions

Chieh Wu

Jan/6/2025

1 Introduction

The KL divergence between 2 multivariate Gaussian distribution is a classic derivation that is used in many algorithms. The motivation of this derivation is to apply it to the Variational Autoencoder.

2 Definitions

Let P and Q be two multivariate Gaussian distributions defined as follows:

$$P(x) = \mathcal{N}(x; \mu_1, \Sigma_1) \tag{1}$$

$$Q(x) = \mathcal{N}(x; \mu_2, \Sigma_2) \tag{2}$$

where:

- $x \in \mathbb{R}^d$ is a d-dimensional random vector.
- $\mu_1, \mu_2 \in \mathbb{R}^d$ are the mean vectors.
- $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$ are the covariance matrices.

3 KL Divergence Formula

The KL divergence from Q to P is given by:

$$D_{KL}(P \| Q) = \int P(x) \log \frac{P(x)}{Q(x)} \, dx \tag{3}$$

Step-by-Step Derivation

1. Express the Probability Density Functions (PDFs)

The PDF of a multivariate Gaussian distribution is:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \tag{4}$$

Therefore, the PDFs for P and Q are:

$$P(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \right) \tag{5}$$

$$Q(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_2|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \right) \tag{6}$$

2. Compute the Logarithm of the Ratio

$$\log \frac{P(x)}{Q(x)} = \log P(x) - \log Q(x) \tag{7}$$

Substituting the PDFs:

$$\log P(x) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_1| - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \tag{8}$$

$$\log Q(x) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_2| - \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \tag{9}$$

Therefore:

$$\log \frac{P(x)}{Q(x)} = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} \left[(x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) - (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right] \tag{10}$$

3. Compute the Expectation with Respect to $P(x)$

The KL divergence is the expectation of $\log \frac{P(x)}{Q(x)}$ with respect to $P(x)$:

$$D_{KL}(P \| Q) = \mathbb{E}_{x \sim P} \left[\log \frac{P(x)}{Q(x)} \right] \tag{11}$$

Substituting the expression for $\log \frac{P(x)}{Q(x)}$:

$$D_{KL}(P \| Q) = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} \mathbb{E}_{x \sim P} \left[(x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) - (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right] \tag{12}$$

4. Simplify the Expectation Terms

We now simplify the expectation terms in Equation (12). Let's break it down step by step.

4.1 Simplify $\mathbb{E}_{x \sim P} \left[(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right]$

This term represents the expected value of the quadratic form $(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)$ under the distribution P. Let's derive this step by step.

Step 1: Rewrite the Quadratic Form

The quadratic form can be rewritten using the trace operator:

$$(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) = \mathrm{Tr} \left(\Sigma_1^{-1} (x-\mu_1)(x-\mu_1)^T \right) \tag{13a}$$

Step 2: Take the Expectation

Take the expectation of both sides with respect to P:

$$\mathbb{E}_{x \sim P} \left[(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right] = \mathbb{E}_{x \sim P} \left[\mathrm{Tr} \left(\Sigma_1^{-1} (x-\mu_1)(x-\mu_1)^T \right) \right] \tag{13b}$$

Move the expectation inside the trace (linearity of trace):

$$\mathbb{E}_{x \sim P} \left[\mathrm{Tr} \left(\Sigma_1^{-1} (x-\mu_1)(x-\mu_1)^T \right) \right] = \mathrm{Tr} \left(\Sigma_1^{-1} \mathbb{E}_{x \sim P} \left[(x-\mu_1)(x-\mu_1)^T \right] \right) \tag{13c}$$

Step 3: Compute the Expectation $\mathbb{E}_{x \sim P} \left[(x-\mu_1)(x-\mu_1)^T \right]$

The term $\mathbb{E}_{x \sim P} \left[(x-\mu_1)(x-\mu_1)^T \right]$ is the covariance matrix of x under P, which is Σ_1:

$$\mathbb{E}_{x \sim P} \left[(x-\mu_1)(x-\mu_1)^T \right] = \Sigma_1 \tag{13d}$$

Step 4: Substitute Back into the Trace

Substitute Equation (13d) into Equation (13c):

$$\mathrm{Tr} \left(\Sigma_1^{-1} \mathbb{E}_{x \sim P} \left[(x-\mu_1)(x-\mu_1)^T \right] \right) = \mathrm{Tr} \left(\Sigma_1^{-1} \Sigma_1 \right) \tag{13e}$$

Step 5: Simplify the Trace

The product $\Sigma_1^{-1} \Sigma_1$ is the identity matrix I:

$$\Sigma_1^{-1} \Sigma_1 = I \tag{13f}$$

The trace of the identity matrix I of size $d \times d$ is equal to d:

$$\mathrm{Tr}(I) = d \tag{13g}$$

Step 6: Final Result

Combining Equations (13e), (13f), and (13g), we arrive at the final result:

$$\mathbb{E}_{x \sim P} \left[(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right] = \mathrm{Tr}(\Sigma_1^{-1} \Sigma_1) = d$$

4.2 Simplify $\mathbb{E}_{x\sim P}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$

This term is more complex and requires expanding the quadratic form. Let's rewrite $x-\mu_2$ as $(x-\mu_1)+(\mu_1-\mu_2)$:

$$(x-\mu_2) = (x-\mu_1)+(\mu_1-\mu_2) \tag{14}$$

Substituting into the quadratic form:

$$(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) = \left[(x-\mu_1)+(\mu_1-\mu_2)\right]^T\Sigma_2^{-1}\left[(x-\mu_1)+(\mu_1-\mu_2)\right] \tag{15}$$

Expanding the quadratic form:

$$= (x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)+2(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) \tag{16}$$

Now, take the expectation with respect to P:

$$\mathbb{E}_{x\sim P}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] = \mathbb{E}_{x\sim P}\left[(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)\right]+2\mathbb{E}_{x\sim P}\left[(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) \tag{17}$$

4.3 Simplify Each Term in Equation (17)

- **First Term:** $\mathbb{E}_{x\sim P}\left[(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)\right]$
This is the expected value of a quadratic form under P. It can be computed as:

$$\mathbb{E}_{x\sim P}\left[(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)\right] = \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) \tag{18}$$

- **Second Term:** $2\mathbb{E}_{x\sim P}\left[(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$
Since $\mathbb{E}_{x\sim P}[x-\mu_1]=0$, this term vanishes:

$$2\mathbb{E}_{x\sim P}\left[(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right] = 0 \tag{19}$$

- **Third Term:** $(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)$
This is a constant term and remains unchanged.

4.4 Combine the Results

Substituting Equations (18) and (19) into Equation (17):

$$\mathbb{E}_{x\sim P}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] = \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) \tag{20}$$

5. Combine the Results

Substituting Equations (13) and (20) into Equation (12):

$$D_{KL}(P\|Q) = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}+\frac{1}{2}\left[\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)-d\right] \tag{21}$$

Simplifying further:

$$D_{KL}(P\|Q) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|}+\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)-d\right) \tag{22}$$

Final Expression

The KL divergence between two multivariate Gaussian distributions P and Q is:

$$D_{KL}(P\|Q) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|}+\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)-d\right) \tag{23}$$