Assignment 1: Normalizing Flows & Diffusion Models

Deep Generative Models Course 361.2.2370

December 10, 2024

Contents

1 Diffusion Models

In this exercise, we will drill down into the probabilistic framework underlying denoising diffusion models.

1.1 Conditional Diffusion Distribution

We defined the conditional probability $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ as the mixing process. To reverse this process, we apply Bayes' rule:

$$q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \frac{q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})q(\boldsymbol{z}_{t-1})}{q(\boldsymbol{z}_t)}$$

This is intractable since we cannot compute the marginal distribution $q(\boldsymbol{z}_{t-1})$. However, if we know the starting variable $\boldsymbol{x}$, we do know the distribution $q(\boldsymbol{z}_{t-1}|\boldsymbol{x})$.

1.1.1 (a)

Consider two multivariate statistically independent normal distributions,

$$P_1(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad P_2(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2).$$

Answer

The multivariate Gaussian distribution is given by:

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

The product of $P_1(\boldsymbol{x})$ and $P_2(\boldsymbol{x})$ is:

$$P_1(\boldsymbol{x})\cdot P_2(\boldsymbol{x}) = \frac{1}{(2\pi)^{d}|\boldsymbol{\Sigma}_1|^{1/2}|\boldsymbol{\Sigma}_2|^{1/2}} \exp\left(-\frac{1}{2}\Big[(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\Big]\right).$$

Expanding the quadratic terms:

$$(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) = \boldsymbol{x}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1,$$

$$(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) = \boldsymbol{x}^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2.$$

Combining the terms:

$$P_1(\boldsymbol{x})\cdot P_2(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\Big[\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - 2\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\Big]\right).$$

We will compare the resulting expression to the expression for a normal distribution:

$$\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - 2\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

And we get:

$$\boldsymbol{\Sigma}_{\text{new}}^{-1} = \boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1},$$

$$\boldsymbol{\Sigma}_{\text{new}}^{-1}\boldsymbol{\mu}_{\text{new}} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2.$$

Multiplying $\boldsymbol{\Sigma}_{\text{new}}$ to get $\boldsymbol{\mu}_{\text{new}}$:

$$\boldsymbol{\mu}_{\text{new}} = \boldsymbol{\Sigma}_{\text{new}}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).$$

Therefore, the covariance and mean of the resulting distribution are:

$$\boldsymbol{\Sigma}_{\text{new}} = (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1},$$

$$\boldsymbol{\mu}_{\text{new}} = (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).$$

Thus, the product is proportional to:

$$\mathcal{N}\big(\boldsymbol{x}; \boldsymbol{\mu}_{\text{new}}, \boldsymbol{\Sigma}_{\text{new}}\big).$$

Substituting these results yields the desired form:

$$P_1(\boldsymbol{x})\cdot P_2(\boldsymbol{x}) \propto \mathcal{N}\big(\boldsymbol{x}; (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2), (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}\big).$$

Proves the relation.

1.1.2 (b)

Consider a multivariate normal distribution $\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B})$. Prove the following change of variable identity:

$$\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B}) \propto \mathcal{N}\big(\boldsymbol{y}; (A^T \boldsymbol{B}^{-1} A)^{-1} A^T \boldsymbol{B}^{-1} \boldsymbol{x}, (A^T \boldsymbol{B}^{-1} A)^{-1}\big).$$

Proof

The probability density function of the multivariate Gaussian $\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B})$ is given by:

$$\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{B}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - A \cdot \boldsymbol{y})^T \boldsymbol{B}^{-1} (\boldsymbol{x} - A \cdot \boldsymbol{y}) \right).$$

Expanding the quadratic term in the exponent:

$$(\boldsymbol{x} - A \cdot \boldsymbol{y})^T \boldsymbol{B}^{-1} (\boldsymbol{x} - A \cdot \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{B}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^T \boldsymbol{B}^{-1} A \cdot \boldsymbol{y} + \boldsymbol{y}^T A^T \boldsymbol{B}^{-1} A \cdot \boldsymbol{y}.$$

The first term, $\boldsymbol{x}^T \boldsymbol{B}^{-1} \boldsymbol{x}$, is a constant with respect to $\boldsymbol{y}$, so we can write:

$$\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B}) \propto \exp\left(-\frac{1}{2} \big(-2\boldsymbol{x}^T \boldsymbol{B}^{-1} A \cdot \boldsymbol{y} + \boldsymbol{y}^T A^T \boldsymbol{B}^{-1} A \cdot \boldsymbol{y} \big) \right).$$

Rewriting the exponent:

$$\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B}) \propto \exp\left(-\frac{1}{2} \big(\boldsymbol{y}^T A^T \boldsymbol{B}^{-1} A \cdot \boldsymbol{y} - 2\boldsymbol{y}^T A^T \boldsymbol{B}^{-1} \boldsymbol{x} \big) \right).$$

We will compare the resulting expression

$$\boldsymbol{y}^T A^T \boldsymbol{B}^{-1} A \cdot \boldsymbol{y} - 2\boldsymbol{y}^T A^T \boldsymbol{B}^{-1} \boldsymbol{x},$$

to the expression for a normal distribution:

$$\boldsymbol{y}^T \Sigma^{-1} \boldsymbol{y} - 2\boldsymbol{y}^T \Sigma^{-1} \mu$$

And we get:

$$\boldsymbol{\Sigma}_y^{-1} = A^T \boldsymbol{B}^{-1} A,$$

$$\boldsymbol{\Sigma}_y^{-1} \boldsymbol{\mu}_y = A^T \boldsymbol{B}^{-1} \boldsymbol{x}.$$

Multiplying $\boldsymbol{\Sigma}_y$ to get $\boldsymbol{\mu}_y$:

$$\boldsymbol{\mu}_y = \boldsymbol{\Sigma}_y (A^T \boldsymbol{B}^{-1} \boldsymbol{x}).$$

Therefore, the covariance and mean of the resulting distribution are:

$$\boldsymbol{\mu}_y = (A^T \boldsymbol{B}^{-1} A)^{-1} A^T \boldsymbol{B}^{-1} \boldsymbol{x},$$

$$\boldsymbol{\Sigma}_y = (A^T \boldsymbol{B}^{-1} A)^{-1}.$$

Thus, we have shown:

$$\mathcal{N}(\boldsymbol{x}; A \cdot \boldsymbol{y}, \boldsymbol{B}) \propto \mathcal{N}\big(\boldsymbol{y}; (A^T \boldsymbol{B}^{-1} A)^{-1} A^T \boldsymbol{B}^{-1} \boldsymbol{x}, (A^T \boldsymbol{B}^{-1} A)^{-1}\big).$$

1.1.3 (c)

Show that:

$$q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{x}) \propto q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) \cdot q(\boldsymbol{z}_{t-1}|\boldsymbol{x}).$$

Answer

Using Bayes rule:

$$q(z_{t-1} \mid z_t, x) = \frac{q(z_t \mid z_{t-1}, x) \cdot q(z_{t-1} \mid x)}{q(z_t \mid x)}.$$

Assuming conditional independence of z_t and x given z_{t-1}:

$$q(z_t \mid z_{t-1}, x) = q(z_t \mid z_{t-1}),$$

We substitute this into the equation:

$$q(z_{t-1} \mid z_t, x) = \frac{q(z_t \mid z_{t-1}) \cdot q(z_{t-1} \mid x)}{q(z_t \mid x)}.$$

Here, $q(z_t \mid x)$ integrates over all possible values of z_{t-1}, thus becoming a normalization constant that does not depend on z_{t-1}. Therefore, we can express the proportionality:

$$q(z_{t-1} \mid z_t, x) \propto q(z_t \mid z_{t-1}) \cdot q(z_{t-1} \mid x).$$

1.1.4 (d)

Use (a), (b), and (c) to find a proportional term for $q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{x})$, where:

- $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t} \cdot \boldsymbol{z}_{t-1}, \beta_t \cdot \boldsymbol{I})$,
- $q(\boldsymbol{z}_{t-1}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_{t-1}; \sqrt{\alpha_{t-1}} \cdot \boldsymbol{x}, (1-\alpha_{t-1}) \cdot \boldsymbol{I})$.

Hint: Note that the relation in (a) requires the normal distributions to have the same support.

Answer

From (c) we know:

$$q(z_{t-1} \mid z_t, x) \propto q(z_t \mid z_{t-1}) \cdot q(z_{t-1} \mid x)$$

and we know that $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t} \cdot \boldsymbol{z}_{t-1}, \beta_t \cdot \boldsymbol{I})$.

we will denote $A = \sqrt{1-\beta_t}$, $B = \beta_t \cdot I$, $x = z_t$, $y - z_{t-1}$, and we will calculate the expressions from (b):

$$\mu_{z-1} = (A^T \boldsymbol{B}^{-1} A)^{-1} A^T \boldsymbol{B}^{-1} \boldsymbol{z_t} = \frac{\sqrt{1-\beta_t}(\beta_t \cdot I)^{-1}}{\sqrt{1-\beta_t}(\beta_t \cdot I)^{-1}\sqrt{1-\beta_t}} z_t = \frac{1}{\sqrt{1-\beta_t}} z_t$$

$$\Sigma_{z-1} = (A^T B^{-1} A)^{-1} = \frac{1}{\sqrt{1-\beta_t}(\beta_t \cdot I)^{-1}\sqrt{1-\beta_t}} = \frac{\beta_t}{1-\beta_t} \cdot I$$

Therefore:

$$q(z_{t-1} \mid z_t) \propto \mathcal{N}\left(z_{t-1}; \frac{1}{\sqrt{1-\beta_t}} z_t, \frac{\beta_t}{1-\beta_t} I\right)$$

We also know:

$$q(z_{t-1} \mid x) = \mathcal{N}\left(z_{t-1}; \sqrt{\alpha_{t-1}} \cdot x, (1-\alpha_{t-1}) \cdot I\right)$$

Combining $q(z_{t-1} \mid z_t)$ and $q(z_{t-1} \mid x)$, both expressed as Gaussian distributions in z_{t-1}, we have:

$$q(z_{t-1} \mid z_t, x) \propto \mathcal{N}\left(z_{t-1}; \mu_1, \Sigma_1\right) \cdot \mathcal{N}\left(z_{t-1}; \mu_2, \Sigma_2\right)$$

where:

$$\mu_1 = \frac{1}{\sqrt{1-\beta_t}} z_t, \quad \Sigma_1 = \frac{\beta_t}{1-\beta_t} I$$

$$\mu_2 = \sqrt{\alpha_{t-1}} \cdot x, \quad \Sigma_2 = (1-\alpha_{t-1}) \cdot I$$

From (a) we know that the covariance of the resulting Gaussian is:

$$\Sigma = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$$

Substitute:

$$\Sigma_1^{-1} = \frac{1-\beta_t}{\beta_t} I, \quad \Sigma_2^{-1} = \frac{1}{1-\alpha_{t-1}} I$$

$$\Sigma = \left(\frac{1-\beta_t}{\beta_t} + \frac{1}{1-\alpha_{t-1}}\right)^{-1} I$$

Simplify:

$$\begin{aligned}
\frac{1-\beta_t}{\beta_t} + \frac{1}{1-\alpha_{t-1}} &= \frac{(1-\beta_t)(1-\alpha_{t-1}) + \beta_t}{\beta_t(1-\alpha_{t-1})} \\
&= \frac{1-\alpha_{t-1} - \beta_t - \beta_t \alpha_{t-1} + \beta_t}{\beta_t(1-\alpha_{t-1})} \\
&= \frac{1-\alpha_{t-1}(1-\beta_t)}{\beta_t(1-\alpha_{t-1})} \\
&= \frac{1-\alpha_t}{\beta_t(1-\alpha_{t-1})}
\end{aligned}$$

Where The last equality follows from the definition of α_t:

$$\alpha_t = \alpha_{t-1}(1-\beta_t)$$

Thus:

$$\Sigma = \left(\frac{1-\alpha_t}{\beta_t(1-\alpha_{t-1})}\right)^{-1} = \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t} I$$

The mean of the resulting Gaussian is:

$$\mu = \Sigma \left(\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2 \right)$$

Substitute:

$$\Sigma_1^{-1} \mu_1 = \frac{1 - \beta_t}{\beta_t} \cdot \frac{1}{\sqrt{1 - \beta_t}} z_t = \frac{\sqrt{1 - \beta_t}}{\beta_t} z_t$$

$$\Sigma_2^{-1} \mu_2 = \frac{1}{1 - \alpha_{t-1}} \cdot \sqrt{\alpha_{t-1}} x = \frac{\sqrt{\alpha_{t-1}}}{1 - \alpha_{t-1}} x$$

$$\mu = \frac{\beta_t (1 - \alpha_{t-1})}{1 - \alpha_t} \left(\frac{\sqrt{1 - \beta_t}}{\beta_t} z_t + \frac{\sqrt{\alpha_{t-1}}}{1 - \alpha_{t-1}} x \right)$$

Simplify:

$$\mu = \frac{(1 - \alpha_{t-1}) \sqrt{1 - \beta_t}}{1 - \alpha_t} z_t + \frac{\beta_t \sqrt{\alpha_{t-1}}}{1 - \alpha_t} x$$

Final Expression

$$q(z_{t-1} \mid z_t, x) = \mathcal{N} \left(z_{t-1}; \mu, \Sigma \right)$$

Where:

$$\mu = \frac{(1 - \alpha_{t-1}) \sqrt{1 - \beta_t}}{1 - \alpha_t} z_t + \frac{\beta_t \sqrt{\alpha_{t-1}}}{1 - \alpha_t} x$$

$$\Sigma = \frac{\beta_t (1 - \alpha_{t-1})}{1 - \alpha_t} I$$

1.2 Evidence Lower Bound (ELBO)

1.2.1 (a)

Consider two D-dimensional multivariate statistically independent normal distributions, $N(x; \mu_1, \Sigma_1)$ and $N(x; \mu_2, \Sigma_2)$. Derive a concise closed-form formula for the Kullback-Leibler (KL) divergence:

$$D_{KL}(N(x; \mu_1, \Sigma_1) \parallel N(x; \mu_2, \Sigma_2)).$$

Answer

Let

$$P := x \sim \mathcal{N}(\mu_1, \Sigma_1), \quad Q := x \sim \mathcal{N}(\mu_2, \Sigma_2)$$

The KL divergence for a continuous random variable is defined as:

$$D_{KL}[P \parallel Q] = \int_X p(x) \log \frac{p(x)}{q(x)} \, dx.$$

Substituting the probability density functions of the multivariate normal distributions, we get:

$$D_{KL}[P \parallel Q] = \int_{\mathbb{R}^n} \mathcal{N}(x;\mu_1,\Sigma_1) \log \frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)}\, dx$$
$$= \mathbb{E}_P \left[\log \frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)} \right]$$

The log term can be expanded as:

$$\log \frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)} = \log \frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_1|}}}{\frac{1}{\sqrt{(2\pi)^n|\Sigma_2|}}} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$$
$$= \log \frac{\sqrt{(2\pi)^n}}{\sqrt{(2\pi)^n}} + \log \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$$
$$= \frac{1}{2}\log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$$

Simplifying:

$$\log \frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)} = \frac{1}{2}\left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right].$$

We will use the following properties:

1. Since the terms $(x-\mu)\Sigma^{-1}(x-\mu)$ are scalars, $(x-\mu)\Sigma^{-1}(x-\mu) = \mathbf{tr}[(x-\mu)\Sigma^{-1}(x-\mu)]$.
2. $\mathbf{tr}[ABC] = \mathbf{tr}[BCA]$.
3. $\mathbb{E}[X^TX] = \mu^T\mu + \mathbf{tr}(\Sigma)$.

And we get:

$$D_{KL}[P \parallel Q] = \mathbb{E}_P \left[\log \frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)}\right] =$$

$$= \mathbb{E}_P \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$$
$$= \mathbb{E}_P \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - \mathbf{tr}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right] + \mathbf{tr}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]\right]$$
$$= \mathbb{E}_P \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - \mathbf{tr}\left[\Sigma_1^{-1}(x-\mu_1)^T(x-\mu_1)\right] + \mathbf{tr}\left[\Sigma_2^{-1}(x-\mu_2)^T(x-\mu_2)\right]\right]$$
$$= \mathbb{E}_P \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbf{tr}\left[\Sigma_1^{-1}[-xx^T + 2x\mu_1 - \mu_1^T\mu_1)]\right] + \mathbf{tr}\left[\Sigma_2^{-1}[xx^T - 2x\mu_2 + \mu_2^T\mu_2)]\right]\right]$$
$$= \mathbb{E}_P \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbf{tr}\left[[-\Sigma_1^{-1} + \Sigma_2^{-1}]xx^T + 2x[\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)] - \Sigma_1^{-1}\mu_1^T\mu_1 + \Sigma_2^{-1}\mu_2^T\mu_2]\right]\right]$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbf{tr}[-\Sigma_1^{-1} + \Sigma_2^{-1}]\mathbb{E}_P[xx^T] + \mathbb{E}_P[2x]\mathbf{tr}[\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)] - \mathbf{tr}[\Sigma_1^{-1}\mu_1^T\mu_1 + \Sigma_2^{-1}\mu_2^T\mu_2]\right]$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbf{tr}[-\Sigma_1^{-1} + \Sigma_2^{-1}][\mu_1\mu_1^T + \mathbf{tr}(\Sigma_1)] + 2\mu_1\mathbf{tr}[\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)] - \mathbf{tr}[\Sigma_1^{-1}\mu_1^T\mu_1 + \Sigma_2^{-1}\mu_2^T\mu_2]\right]$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbf{tr}[-\Sigma_1^{-1}\Sigma_1 + \Sigma_2^{-1}\Sigma_1] + [-\Sigma_1^{-1} + \Sigma_2^{-1}]\mu_1\mu_1^T + 2\mu_1\mathbf{tr}[\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)] - \mathbf{tr}[\Sigma_1^{-1}\mu_1^T\mu_1 + \Sigma_2^{-1}\mu\right.$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathbf{tr}\left[\Sigma_2^{-1}\Sigma_1 - \Sigma_1^{-1}\mu_1\mu_1^T + \Sigma_2^{-1}\mu_1\mu_1^T + \Sigma_1^{-1}2\mu_1\mu_1^T - \Sigma_2^{-1}2\mu_1\mu_2 - \Sigma_1^{-1}\mu_1^T\mu_1 + \Sigma_2^{-1}\mu_2^T\mu_2\right]\right]$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathbf{tr}(\Sigma_2^{-1}\Sigma_1) + \mathbf{tr}\left[\Sigma_2^{-1}\mu_1\mu_1^T - \Sigma_2^{-1}2\mu_1\mu_2^T + \Sigma_2^{-1}\mu_2\mu_2^T\right]\right]$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathbf{tr}(\Sigma_2^{-1}\Sigma_1) + \mathbf{tr}\left[\mu_1^T\Sigma_2^{-1}\mu_1 - 2\mu_2^T\Sigma_2^{-1}\mu_1 + \mu_2^T\Sigma_2^{-1}\mu_2\right]\right]$$
$$= \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathbf{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2(\mu_2-\mu_1)\right]$$

Final Expression:

$$D_{KL}[P \parallel Q] = \frac{1}{2}\left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right].$$

1.2.2 (b)

Use (a) to derive an expression for:

$$D_{KL}(q(z_{t-1} \mid z_t, x) \parallel P(z_{t-1} \mid z_t; \phi_t)),$$

where:

- $q(z_{t-1} \mid z_t, x) = N\left(z_{t-1}; \frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot \sqrt{1-\beta_t} \cdot z_t + \frac{\beta_t\sqrt{\alpha_{t-1}}}{1-\alpha_t} \cdot x, \beta_t \cdot \frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot I\right)$
- $P(z_{t-1} \mid z_t; \phi_t) = N\left(z_{t-1}; f_{\phi_t}(z_t), \sigma_t^2 \cdot I\right)$

Answer

The goal is to derive an expression for:

$$D_{KL}[q \parallel p] = \frac{1}{2}\left[\log\frac{|\Sigma_1|}{|\Sigma_2|} - n + \mathrm{tr}(\Sigma_1^{-1}\Sigma_2) + (\mu_1-\mu_2)^T\Sigma_1^{-1}(\mu_1-\mu_2)\right]$$

Where:

- $\mu_1 = \frac{1-\alpha_{t-1}}{1-\alpha_t}\cdot\sqrt{1-\beta_t}\cdot z_t + \frac{\beta_t\sqrt{\alpha_{t-1}}}{1-\alpha_t}\cdot x$
- $\mu_2 = f_{\phi_t}(z_t)$
- $\Sigma_1 = \beta_t\cdot\frac{1-\alpha_{t-1}}{1-\alpha_t}\cdot I$
- $\Sigma_2 = \sigma_t^2\cdot I$

Based on (a), the KL divergence between two multivariate Gaussian distributions $q \sim \mathcal{N}(\mu_1,\Sigma_1)$ and $p\sim\mathcal{N}(\mu_2,\Sigma_2)$ is given by:

$$D_{KL}[q \parallel p] = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

where n is the dimensionality of the distributions. The covariance matrices are diagonal (scaled identity matrices), so their determinants are the product of their diagonal entries:

$$|\Sigma_1| = \left(\beta_t\cdot\frac{1-\alpha_{t-1}}{1-\alpha_t}\right)^n, \quad |\Sigma_2| = (\sigma_t^2)^n$$

Thus, the log term is:

$$\log\frac{|\Sigma_2|}{|\Sigma_1|} = \log(\sigma_t^2)^n - \log\left(\beta_t\cdot\frac{1-\alpha_{t-1}}{1-\alpha_t}\right)^n$$

$$= n\log\sigma_t^2 - n\log\left(\beta_t\cdot\frac{1-\alpha_{t-1}}{1-\alpha_t}\right)$$

To derive the trace term:

$$\mathrm{tr}(\Sigma_2^{-1}\Sigma_1),$$

Since both Σ_1 and Σ_2 are diagonal:

$$\Sigma_2^{-1} = \frac{1}{\sigma_t^2}I$$

$$\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) = \mathrm{tr}\left(\frac{1}{\sigma_t^2}\cdot\beta_t\cdot\frac{1-\alpha_{t-1}}{1-\alpha_t}\cdot I\right)$$

$$= \frac{\beta_t\cdot(1-\alpha_{t-1})}{\sigma_t^2(1-\alpha_t)}\cdot n$$

The quadratic term is:

$$(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)$$

Substitute $\mu_1 = \frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot \sqrt{1-\beta_t} \cdot z_t + \frac{\beta_t\sqrt{\alpha_{t-1}}}{1-\alpha_t} \cdot x$ and $\mu_2 = f_{\phi_t}(z_t)$:

$$\mu_1 - \mu_2 = \frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot \sqrt{1-\beta_t} \cdot z_t + \frac{\beta_t\sqrt{\alpha_{t-1}}}{1-\alpha_t} \cdot x - f_{\phi_t}(z_t)$$

Let $\delta = \mu_1 - \mu_2$. Then:

$$\delta^T \Sigma_2^{-1} \delta = \delta^T \cdot \frac{1}{\sigma_t^2} \cdot \delta = \frac{1}{\sigma_t^2} \delta^T \cdot \delta = \frac{1}{\sigma_t^2} \cdot \|\delta\|^2$$

Where $\|\delta\|^2$:

$$\|\delta\|^2 = \left\| \frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot \sqrt{1-\beta_t} \cdot z_t + \frac{\beta_t\sqrt{\alpha_{t-1}}}{1-\alpha_t} \cdot x - f_{\phi_t}(z_t) \right\|^2$$

Substitute all terms into the KL divergence expression:

$$D_{KL}[q \parallel p] = \frac{1}{2} \left[n \log \sigma_t^2 - n \log \left(\beta_t \cdot \frac{1-\alpha_{t-1}}{1-\alpha_t} \right) - n + \frac{\beta_t \cdot \frac{1-\alpha_{t-1}}{1-\alpha_t}}{\sigma_t^2} \cdot n + \frac{1}{\sigma_t^2} \cdot \|\delta\|^2 \right]$$

1.2.3 (c)

Answer

In the KL divergence $D_{KL}(q(z_{t-1} \mid z_t, x) \parallel P(z_{t-1} \mid z_t; \phi_t))$:

The term $q(z_{t-1} \mid z_t, x)$ represents the desired distribution for the denoising step. It is derived from the known data-generating process and incorporates the observed data x as well as the latent relationship between z_t and z_{t-1}. Specifically:

- $q(z_{t-1} \mid z_t, x)$ is a posterior distribution that quantifies the true process of denoising z_t to obtain z_{t-1}.
- It is referred to as the target distribution, as it is the distribution we aim to match through the model's predictions.

The term $P(z_{t-1} \mid z_t; \phi_t)$ represents the approximated distribution predicted by the model during the denoising process. This distribution is parameterized by ϕ_t, which are the learnable parameters of the model. Specifically

- $P(z_{t-1} \mid z_t; \phi_t)$ is the model's attempt to approximate $q(z_{t-1} \mid z_t, x)$.
- The parameters ϕ_t are optimized to minimize the divergence between the target $q(z_{t-1} \mid z_t, x)$ and the prediction $P(z_{t-1} \mid z_t; \phi_t)$, thereby improving the accuracy of the denoising steps.