Derivation Steps

1) Define Modeling objective
2) Introduce known forward process to get ELBO
3) Start decomposing to get $\mathcal{L}_{ELBO}(\theta)$
4) Final decomposition of $\mathcal{L}_\epsilon$
5) weighted MSE / l2-norm term

1) Define Modeling objective

Training objective

$$\max_\theta \mathbb{E}_{x_0 \sim D}\left[\lg p_\theta(x_0)\right], \text{ where } D = \text{data}$$

$p_\theta(x_0) = \int p_\theta(x_{T:0})\, dx_{T:1}$ the marginalization over all intermediate latents $x_T \longrightarrow x_1$, iow all latent trajectories that could lead back to x_0.

$$= \int p(x_T) \prod \underbrace{p(x_{t-1} | x_t)}_{\text{unknown}} dx_{T:1}$$

2) We introduce known forward process

$$q(x_{1:T} | x_0) = \prod \underbrace{q(x_t | x_{t-1})}_{N(\sqrt{\alpha_t} x_{t-1}, (1-\bar{\alpha}_t) I)}$$

a tractable way to sample latent trajectories

$x_1 \longrightarrow x_t | x_0$

$$\Rightarrow \lg p(x_0) = \lg \int \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} q(x_{1:T}|x_0)\, dx_{T:1}$$

$$= \lg E_{q(x_{1:T}|x_0)}\left[\frac{p(x_T)\prod p(x_{t-1}|x_t)}{\prod q(x_t|x_{t-1})}\right]$$

$f(E[x]) \geq E[f(x)]$ for f concave

$$\geq E_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T)\prod p(x_{t-1}|x_t)}{\prod q(x_t|x_{t-1})}\right]$$

$$\Rightarrow \boxed{\mathcal{L}_E(\theta) = E_q\Big[\log p(x_T) + \underbrace{\sum \log p(x_{t-1}|x_t)}_{B} - \underbrace{\sum \log q(x_t|x_{t-1})}_{C}\Big]}$$

3) Start decomposing $\mathcal{L}_E$

a) Group terms by 't': isolate $t=1$; $1<t<T$

$$B - C = \sum \log p(x_{t-1}|x_t) - \log q(x_t|x_{t-1})$$

$$= \log p(x_0|x_1) - \log q(x_1|x_0) + \sum_{t=2}^{T} \log p(x_{t-1}|x_t) - \log q(x_t|x_{t-1})$$

$$\Rightarrow \boxed{\begin{aligned}\mathcal{L}_E(\theta) = {} & E_q \log p(x_T) + E_q\left[\log p(x_1|x_1) - \log q(x_1|x_0)\right] \\ & + \sum_{t=2}^{T} \underbrace{E_q\left[\log p(x_{t-1}|x_t) - \log q(x_t|x_{t-1})\right]}_{G.1} \quad (G)_{q=q(x_{1:T}|x_0)}\end{aligned}}$$

b) Insert posterior identity $\forall t \in \{2 \dots T\}$ add and subtract $\log q(x_{t-1}|x_t, x_0)$ in (G.1)

$$\log p_\theta(x_{t-1}|x_t) - \log q(x_t|x_{t-1})$$

$$= \underbrace{\left[\log p_\theta(x_{t-1}|x_t) - \log q(x_{t-1}|x_t, x_0)\right]}_{b.1} + \underbrace{\left[\log q(x_{t-1}|x_t, x_0) - \log q(x_t|x_{t-1})\right]}_{b.2}$$

b.1) Bring Expectation and get KL Divergence

$$E_{q(x_{1:T}|x_0)}[(b.1)] = E_{q(x_t|x_0)}\Big[-KL\left(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)\right)\Big]$$

$$KL(P(x)\|Q(x)) = E_{x\sim p}\left[\log \frac{P(x)}{Q(x)}\right]$$

How does the expectation turn into KL?

Let $f(x_{t-1}, x_t) = \lg p_\theta(x_{t-1}|x_t) - \lg q(x_{t-1}|x_t, x_0)$ (b.12)

$$\mathbb{E}_{q(x_{1:T}|x_0)}[f(x_{t-1}, x_t)] = \int \cdots \int f(x_{t-1}, x_t) q(x_{1:T}|x_0) dx_{1:T}$$

f only depends on x_{t-1} & x_t ⇒ ade other variables integrate out

$$= \iint f(x_{t-1}, x_t) q(x_{t-1}, x_t|x_0) dx_{t-1} dx_t$$

$$q(x_{t-1}, x_t|x_0) = q(x_t|x_0) q(x_{t-1}|x_t, x_0)$$

$$\mathbb{E}_{q(x_{1:T}|x_0)}[f(x_{t-1}, x_t)]$$

$$= \int q(x_t|x_0) \underbrace{\left(\int f(x_{t-1}, x_t) q(x_{t-1}|x_t, x_0) dx_{t-1} \right)}_{\mathbb{E}_{q(x_{t-1}|x_t, x_0)}[f(x_{t-1}, x_t)]} dx_t$$

Swap in (b.12) for $f(x_{t-1}, x_t)$

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

$$\Rightarrow \mathbb{E}_{q(x_{t-1}|x_t, x_0)}\left[\lg \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right]$$

$$= -KL\left(q(x_{t-1}|x_t, x_0) \,||\, p_\theta(x_{t-1}|x_t)\right)$$

$$\Rightarrow \mathbb{E}_{q(x_{1:T}|x_0)}[f(x_{t-1}, x_t)]$$

$$= \int q(x_t|x_0) \left[-KL\left(q(x_{t-1}|x_t, x_0) \,||\, p_\theta(x_{t-1}|x_t)\right)\right] dx_t$$

$$= \mathbb{E}_{q(x_t|x_0)}\left[-KL\left(q(x_{t-1}|x_t, x_0) \,||\, p_\theta(x_{t-1}|x_t)\right)\right] \quad (b.13)$$

(b.2) $\lg q(x_{t-1}|x_t, x_0) - \lg q(x_t|x_{t-1})$ — by Markov

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)\, q(x_{t-1}|x_0)}{q(x_t|x_0)}$$ by Bayes

$$= \frac{q(x_t|x_{t-1})\, q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$\Rightarrow \lg \frac{q(x_t|x_{t-1})\, q(x_{t-1}|x_0)}{q(x_t|x_0)} - \lg q(x_t|x_{t-1})$$

$$= \lg q(x_t|x_{t-1}) + \lg q(x_{t-1}|x_0) - \lg q(x_t|x_0) - \lg q(x_t|x_{t-1})$$

$\therefore$

$$\lg q(x_{t-1}|x_t, x_0) - \lg q(x_t|x_{t-1}) = \lg q(x_{t-1}|x_0) - \lg q(x_t|x_0) \quad (b.21)$$

Now, back to $\mathcal{L}_{ELBO}(\theta)$. We swap b.13 and b.21 back into the ELBO loss

$$\mathcal{L}_E(\theta) = \mathbb{E}_q \lg p(x_T) + \mathbb{E}_q\left[\lg p(x_0|x_1) - \lg q(x_1|x_0)\right]$$
$$+ \sum_{t=2}^{T} \mathbb{E}_q\left[\lg p(x_{t-1}|x_t) - \lg q(x_t|x_{t-1})\right], \text{ where } q = q(x_{1:T}|x_0)$$

$$= \mathbb{E}_q \lg p(x_T) + \mathbb{E}_q\left[\lg p(x_0|x_1) - \lg q(x_1|x_0)\right]$$
$$+ \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[-KL\left(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)\right)\right]$$
$$+ \sum_{t=2}^{T} \mathbb{E}_q\left[\lg q(x_{t-1}|x_0) - \lg q(x_t|x_0)\right]$$

telescoping sum $= \mathbb{E}_q[\lg q(x_1|x_0)] \cdots - \mathbb{E}_q[\lg q(x_T|x_0)]$

middle cancels

$$= E_q \lg p(x_T) + E_q\left[\lg p(x_0|x_1)\right] - \cancel{E_q \lg q(x_1|x_0)}$$

$$+ \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[-KL(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))\right]$$

$$+ \cancel{E_q\left[\lg q(x_1|x_0)\right]} - E_q\left[\lg q(x_T|x_0)\right]$$

$$= E_q \lg p(x_0|x_1) + \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[-KL(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))\right]$$

$$+ E_{q(x_{1:T}|x_0)}\left[\lg p(x_T) - \lg q(x_T|x_0)\right]$$

4) Final decomposition of $\mathcal{L}_E$

$$\mathcal{L}_{ELBO}(\theta) = \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[-KL(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))\right]$$

$$+ E_{q(x_T|x_0)}\left[-KL(q(x_T|x_0)\,\|\,p(x_T))\right] + E_q \lg p(x_0|x_1)$$

5) How to get weighted MSE / $\ell 2$-norm term

Look at $\sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[-KL(q(x_{t-1}|x_t,x_0)\,\|\,p_\theta(x_{t-1}|x_t))\right]$

i) Both conditionals are Gaussians with same covariance

i) True posterior

$$q(x_{t-1}|x_t,x_0) = \mathcal{N}(\mu_{post}(x_t,x_0), \tilde{\beta}_t I) \quad (m_1)$$

with $\tilde{\beta}_t = \beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$

ii) Learned reverse kernel (m_2)

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t,\hat{x}_0), \tilde{\beta}_t I)$$

Because the covariances are Equal, the KL simplifies to a squared distance between the means

KEY: If cov same $\Rightarrow$ KL $= \frac{1}{2\tilde{\beta}_t}\|\mu_{post} - \mu_\theta\|^2$

PDF Multivariate Gaussian

$$\frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(x-\hat{x})^T\Sigma^{-1}(x-\hat{x})$$

log pdf Multivariate Gaussian

$$\log \mathcal{N}(x; m, \Sigma) = -\frac{1}{2}\left[d \log(2\pi) + \log|\Sigma| + (x-m)^T\Sigma^{-1}(x-m)\right]$$

Let

$$r(x) = \mathcal{N}(x; m_1, \Sigma)$$
$$s(x) = \mathcal{N}(x; m_2, \Sigma)$$

same $\Sigma \in \mathbb{R}^{d \times d}$ pos def

$$\log r(x) - \log s(x) =$$

$$-\frac{1}{2}\left(d \log(2\pi) + \log|\Sigma| + (x-m_1)^T\Sigma^{-1}(x-m_1)\right)$$
$$+\frac{1}{2}\left(d \log(2\pi) + \log|\Sigma| + (x-m_2)^T\Sigma^{-1}(x-m_2)\right)$$

$$= -\frac{1}{2}A + \frac{1}{2}B = -\frac{1}{2}(A-B) = -\frac{1}{2}\left[(x-m_1)^T\Sigma^{-1}(x-m_1) - (x-m_2)^T\Sigma^{-1}(x-m_2)\right]$$

$$KL(r\|s) = E_r[\log r(x) - \log s(x)]$$

Now take expectation under $x \sim p = \mathcal{N}(m_1, \Sigma)$ our learned model since $KL(p\|q)$

$$KL(p\|q) = -\frac{1}{2} E_{x \sim \mathcal{N}(m_1, \Sigma) = p}[A - B]$$

$E[\mathrm{tr}(A)] = \mathrm{tr}[E(A)]$

E & tr both linear

$$a^T M a = \mathrm{tr}(a^T M a) = \mathrm{tr}(M a a^T)$$

$a \in \mathbb{R}^n \quad M \in \mathbb{R}^{n \times n}$

Squared Mahalanobis Distance

$$D_M^2(x) = (x-m)^T\Sigma^{-1}(x-m)$$

① $$E_{x \sim p}[A] = E_{x \sim p}\left[(x-m_1)^T\Sigma^{-1}(x-m_1)\right]$$

$$= E_{x \sim p}\left[\mathrm{tr}\left(\Sigma^{-1}(x-m_1)(x-m_1)^T\right)\right]$$

$$= \mathrm{tr}\left(\Sigma^{-1} \underbrace{E_{x \sim p}\left[(x-m_1)(x-m_1)^T\right]}_{\Sigma}\right) = \mathrm{tr}(\Sigma^{-1}\Sigma) = \mathrm{tr}(I) = d$$

So $KL(p\|q) = -\frac{1}{2} E_{x\sim N(m_1,\Sigma)=p}[A-B]$

$= -\frac{1}{2}\left(E_p[A] - E_p[B]\right) = -\frac{1}{2}\left(d - E_p[B]\right)$

Now, what about $E_p[B]$, the second term?

$$E_{x\sim N(m_2,\Sigma)}\underbrace{\left[(x-m_2)^T\Sigma^{-1}(x-m_2)\right]}_{(J)}$$

Let $r := x - m_1 \sim N(0,\Sigma)$,

$\delta := m_1 - m_2$

$\Rightarrow x - m_2 = (x-m_1) + (m_1-m_2) = r+\delta$

$\Rightarrow (J) = (r+\delta)^T\Sigma^{-1}(r+\delta)$

$(r^T+\delta^T)(\Sigma^{-1}r + \Sigma^{-1}\delta)$

$= r^T\Sigma^{-1}r + \delta^T\Sigma^{-1}r + r^T\Sigma^{-1}\delta + \delta^T\Sigma^{-1}\delta$

$= r^T\Sigma^{-1}r + \delta^T\Sigma^{-1}r + \delta^T\Sigma^{-1}r + \delta^T\Sigma^{-1}\delta$

$= r^T\Sigma^{-1}r + 2\delta^T\Sigma^{-1}r + \delta^T\Sigma^{-1}\delta$ (Q)

$(xy)^T = y^Tx^T$

$(a+b)^T = a^T+b^T$

$M(c+d) = Mc + Md$

M symmetric $\Rightarrow M^T = M$

a, b scalars

$a^TMb = (a^TMb)^T = b^TM^Ta = b^TMa$

Plug into expectation, 3 cases

i) $E_p[r^T\Sigma^{-1}r]$, where $p = x\sim N(m_1,\Sigma)$

$tr\left(\Sigma^{-1}E_p[rr^T]\right) = tr(\Sigma^{-1}\Sigma) = tr(I) = d$

ii) $E_p[2\delta^T\Sigma^{-1}r]$ → only 'r' depends on x

$\dot\alpha$ $r = x - m_1$, removing mean

so $E_p[r] = 0$

$\Rightarrow 2\delta^T\Sigma^{-1}E_p[r] = 0$

iii) $E_p[\delta^T \Sigma^{-1} \delta]$ is constant

$\rightarrow \delta^T \Sigma^{-1} \delta \, E[1] = \delta^T \Sigma^{-1} \delta$

Now $E_p[Q] = d + 0 + \delta^T \Sigma^{-1} \delta$ where $\delta = m_1 - m_2$

$\rightarrow E_p[Q] = d + (m_1 - m_2)^T \Sigma^{-1} (m_1 - m_2)$

and if $\Sigma = \sigma^2 I \Rightarrow (m_1 - m_2)^T \Sigma^{-1} (m_1 - m_2)$

$= \frac{1}{\sigma^2} \| m_1 - m_2 \|_2^2$

$\therefore \frac{1}{2} E_{x \sim N(m_1, \Sigma)} [(x - m_1)^T \Sigma^{-1} (x - m_1) - (x - m_2)^T \Sigma^{-1} (x - m_2)]$

$= -\frac{1}{2} (d - d - \frac{1}{\sigma^2} \| m_1 - m_2 \|_2^2) = \boxed{\frac{1}{2\sigma^2} \| m_1 - m_2 \|_2^2}$

In diffusion, our case

$\Sigma = \tilde{\beta}_t I$

$m_1 = \mu_{post}(x_t, x_0)$ $\Rightarrow KL(q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t))$

$m_2 = \mu_\theta(x_t, t)$ $= \frac{1}{2\tilde{\beta}_t} \| \mu_{post} - \mu_\theta \|_2^2$

Recall $\leftarrow p(x_t | x_0)$ one-shot kernel

$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim N(0, 1)$

solve for x_0 and get

$x_0 = \frac{1}{\sqrt{\bar{\alpha}}} (x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon)$

$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}}} (x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t))$ our prediction

$$\mu_{post} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

$$\mu_{post} - \mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}(\epsilon_\theta(x_t, t) - \epsilon)\right)$$

$$= \frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}(\epsilon_\theta(x_t, t) - \epsilon)$$

Can drop!

$$\Rightarrow \frac{1}{2\tilde{\beta}_t}\|\mu_{post} - \mu_\theta\| = \frac{1}{2\tilde{\beta}_t}\frac{\beta_t^2}{\alpha_t(1-\bar{\alpha}_t)}\|\epsilon_\theta(x_t, t) - \epsilon\|^2$$

timestep dependent weight

$$\therefore L_t(\theta) = \|\epsilon_{pred} - \epsilon_\theta\|_2^2 \quad (a.2)$$

Finish!