

## Derivation Steps

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

### 1) Define Modeling objective

Training objective

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

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

$$= \int p(x_t) \underbrace{\prod p(x_{t+1}|x_t) dx_{T:1}}$$

unknown

### 2) We introduce known forward process

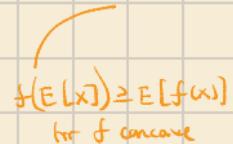
$$g(x_{1:T}|x_0) = \underbrace{\prod g(x_t|x_{t-1})}_{N(\sum_{k=t}^T x_{k+1}, (1-\bar{\alpha}_t)^2)}$$

a tractable way to sample latent trajectories

$$x_1 \rightarrow x_t | x_0$$

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

$$= \lg \mathbb{E}_{g(x_{1:T}|x_0)} \left[ \frac{p(x_T) \prod p(x_{t+1}|x_t)}{\prod g(x_t|x_{t-1})} \right]$$



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

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

C

$$\Rightarrow L_E(\theta) = E_g \left[ \lg p(x_T) + \sum \lg p(x_{t+1} | x_t) - \sum \lg g(x_t | x_{t-1}) \right]$$

3) Start decomposing  $L_E$

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

$$B - C = \sum \lg p(x_{t+1} | x_t) - \lg g(x_t | x_{t-1})$$

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

$$\begin{aligned} L_E(\theta) &= E_g \lg p(x_T) + E_g [\lg p(x_1 | x_0) - \lg g(x_1 | x_0)] \\ &\quad + \sum_{t=2}^T E_g [\lg p(x_{t+1} | x_t) - \lg g(x_t | x_{t-1})] \end{aligned} \quad (6)$$

→

b) Insert posterior identity  $\forall t \in \{2, \dots, T\}$  add and subtract  $\lg g(x_{t-1} | x_t, x_0)$  in (6.1)

$$\begin{aligned} &\lg p_\theta(x_{t-1} | x_t) - \lg g(x_t | x_{t-1}) \\ &= [\lg p_\theta(x_{t-1} | x_t) - \lg g(x_{t-1} | x_t, x_0)] \quad b.1 \\ &\quad + [\lg g(x_{t-1} | x_t, x_0) - \lg g(x_t | x_{t-1})] \quad b.2 \end{aligned}$$

b.1) Bring Expectation and get KL Divergence

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

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

How does the expectation turn into KL?

$$\text{Let } f(x_{t-1}, x_t) = \lg p_0(x_{t-1} | x_t) - \lg g(x_{t-1} | x_t, x_0) \quad (\text{b.12})$$

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

$f$  only depends on  $x_{t-1}$  &  $x_t \Rightarrow$  all other variables  
integrate out

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

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

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

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

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

$$\text{KL}(p(x) || q(x)) = \mathbb{E}_{x \sim p} \left[ \log \frac{p(x)}{q(x)} \right]$$

$$\Rightarrow \mathbb{E}_{g(x_t | x_0)} \left[ \lg \frac{p_0(x_{t-1} | x_t)}{g(x_{t-1} | x_t, x_0)} \right]$$

$$= -\text{KL}(p_0(x_{t-1} | x_t) || g(x_{t-1} | x_t, x_0))$$

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

$$= \int g(x_t | x_0) \left[ -\text{KL}(p_0(x_{t-1} | x_t) || g(x_{t-1} | x_t, x_0)) \right] dx_t$$

$$= \mathbb{E}_{g(x_t | x_0)} [-\text{KL}(p_0(x_{t-1} | x_t) || g(x_{t-1} | x_t, x_0))] \quad (\text{b.13})$$

$$(b.2) \quad \lg g(x_{t-1} | x_t, x_0) - \lg g(x_t | x_{t-1}) \xrightarrow{\text{by Markov}}$$

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

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

$$\Rightarrow \lg \frac{g(x_t | x_{t-1}) g(x_{t-1} | x_0)}{g(x_t | x_0)} - \lg g(x_t | x_{t-1}) \\ = \cancel{\lg g(x_t | x_{t-1})} + \lg g(x_{t-1} | x_0) - \lg g(x_t | x_0) - \cancel{\lg g(x_t | x_{t-1})}$$

$\therefore$

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

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

$$\begin{aligned} \mathcal{L}_E(\theta) &= \mathbb{E}_g \lg p(x_T) + \mathbb{E}_g [\lg p(x_t | x_{t-1}) - \lg g(x_t | x_0)] \\ &\quad + \sum_{t=2}^T \mathbb{E}_g [\lg p(x_{t-1} | x_t) - \lg g(x_{t-1} | x_{t-1})], \text{ where } g = g(x_{t-1} | x_0) \\ &= \mathbb{E}_g \lg p(x_T) + \mathbb{E}_g [\lg p(x_t | x_{t-1}) - \lg g(x_t | x_0)] \\ &\quad + \sum_{t=2}^T \mathbb{E}_{g(x_t | x_{t-1})} [-\text{KL}(g(x_{t-1} | x_{t-1}, x_0) || p_\theta(x_{t-1} | x_t))] \\ &\quad + \underbrace{\sum_{t=2}^T \mathbb{E}_g [\lg g(x_{t-1} | x_0) - \lg g(x_t | x_0)]}_{\text{telescoping sum}} \end{aligned}$$

telescoping sum =  $E_g[\lg g(x_1 | x_0)] - E_g[\lg g(x_T | x_0)]$   
middle cancels

$$\begin{aligned}
&= E_g \lg p(x_T) + E_g [\lg p(x_0 | x_0)] - \cancel{E_g \lg g(x_0 | x_0)} \\
&\quad + \sum_{t=2}^T E_{g(x_t | x_0)} [-\text{KL}(g(x_{t-1} | x_{t-2}, x_0) \| p_0(x_{t-1} | x_{t-2}))] \\
&\quad + \cancel{E_g [\lg g(x_0 | x_0)]} - E_g [\lg g(x_T | x_0)] \\
&= E_g \lg p(x_0 | x_0) + \sum_{t=2}^T E_{g(x_t | x_0)} [-\text{KL}(g(x_{t-1} | x_{t-2}, x_0) \| p_0(x_{t-1} | x_{t-2}))] \\
&\quad + E_{g(x_T | x_0)} [\lg p(x_T) - \lg g(x_T | x_0)]
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{L}_{E(\Theta)}(\Theta) &= \sum_{t=2}^T E_{g(x_t | x_0)} [-\text{KL}(g(x_{t-1} | x_{t-2}, x_0) \| p_0(x_{t-1} | x_{t-2}))] \\
&\quad + E_{g(x_T | x_0)} [-\text{KL}(g(x_T | x_0) \| p(x_T))] + E_g \lg p(x_0 | x_0)
\end{aligned}$$

5) How to get weighted MSE / l2-norm term

Look at  $\sum_{t=2}^T E_{g(x_t | x_0)} [-\text{KL}(g(x_{t-1} | x_{t-2}, x_0) \| p_0(x_{t-1} | x_{t-2}))]$

i) Both conditionals are Gaussians with same covariance

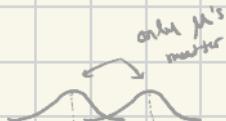
i) True posterior

$$g(x_{t-1} | x_t, x_0) = N(\mu_{\text{post}}(x_t, x_0), \tilde{\beta}_t I)$$

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

ii) Learned reverse kernel

$$p_0(x_{t-1} | x_t) = N(\mu_0(x_t, t), \tilde{\beta}_t I)$$



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

KEY: If cov same  $\rightarrow \text{KL} = \frac{1}{2\tilde{\beta}_t} \|\mu_{\text{post}} - \mu_0\|^2$

Let

$$p(x) = N(x; m_1, \Sigma) \quad \left| \begin{array}{l} \text{same } \Sigma \in \mathbb{R}^{\text{dxd}} \\ \text{pos def} \end{array} \right.$$

$$g(x) = N(x; m_2, \Sigma)$$

$$\log p(x) - \log g(x) =$$

$$\cancel{-\frac{1}{2}(\lg(2\pi) + \lg|\Sigma| + (x-m_1)^T \Sigma^{-1} (x-m_1))}$$

$$\cancel{+\frac{1}{2}(\lg(2\pi) + \lg|\Sigma| + (x-m_2)^T \Sigma^{-1} (x-m_2))}$$

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

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

$$\text{KL}(p||g) = -\frac{1}{2} \mathbb{E}_{x \sim N(m_1, \Sigma)=p} [A - B]$$

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

$\text{E} \circ \text{tr}$  both linear

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

Squared Mahalanobis Distance

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

$$\textcircled{1} \quad \mathbb{E}_{x \sim p} [A] = \mathbb{E}_{x \sim p} [(x-m_1)^T \Sigma^{-1} (x-m_1)]$$

$$= \mathbb{E}_{x \sim p} [\text{tr}(\Sigma^{-1} (x-m_1)(x-m_1)^T)]$$

$$= \text{tr} (\Sigma^{-1} \mathbb{E}_{x \sim p} [(x-m_1)(x-m_1)^T]) = \text{tr}(\Sigma^{-1} \Sigma) = \text{tr}(I) = d$$

PDF Multivariate Gaussian

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

log pdf Multivariate Gaussian

$$\log N(x; m, \Sigma) =$$

$$-\frac{1}{2} [\lg(2\pi) + \log|\Sigma| + (x-m)^T \Sigma^{-1} (x-m)]$$

$$\text{KL}(p||g) =$$

$$\mathbb{E}_p [\log p(x) - \log g(x)]$$

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

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

Now, what about  $E_p[\text{B}]$ , the second term?

$$E_{x \sim N(\mu_1, \Sigma)} [(x - \mu_2)^T \Sigma^{-1} (x - \mu_2)] \quad (\text{J})$$

Let  $r := x - \mu_1 \sim N(0, \Sigma)$ ,

$$\delta := \mu_1 - \mu_2$$

$$\Rightarrow x - \mu_2 = (x - \mu_1) + (\mu_1 - \mu_2) = r + \delta$$

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

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

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

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

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

Plug into expectation, 3 cases

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

$$+ \text{tr}(\Sigma^{-1} E_p[r r^T]) = \text{tr}(\Sigma^{-1} \Sigma) = \text{tr}(\text{I}) = d$$

ii)  $E_p[2\delta^T \Sigma^{-1} r] \rightarrow$  only ' $r$ ' depends on  $x$   
 $\delta = x - \mu_1$ , removing mean  
 $\text{so } E_{\Omega}[r] = 0$

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

$(xy)^T = y^T x^T$
$(a+b)^T = a^T + b^T$
$M(cab) = Mca + Mcb$

$M$ symmetric $\rightarrow M^T = M$
$a, b$ scalars
$a^T M b = (a^T M b)^T = b^T M^T a = b^T M a$

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

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

Nwh  $E_p[Q] = d + \delta + \delta^\top \Sigma^{-1} \delta$  where  $\delta = m_1 - m_2$

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

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

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

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

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

In diffusion, our case

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

$$m_1 = \mu_{post}(x_t, x_0) \rightarrow \text{KL}(q(x_{t-1}|x_t, x) \| p_0(x_{t-1}|x))$$

$$m_2 = \mu_0(x_t, t)$$

$$= \frac{1}{2\tilde{\beta}_t} \|\mu_{post} - \mu_0\|_2^2$$

Substitute the  $\epsilon$ -parameterization

$\Rightarrow$

$$\|\epsilon - \epsilon_0\|_2^2$$

"plug in"

From (IM.1)

$$\mu_{post}/\phi = \frac{1}{1-\alpha_t} \left( x_t - \frac{\beta_t}{1-\alpha_t} \{ \epsilon, \epsilon_0(x_t, t) \} \right)$$

$$\therefore \underline{\epsilon}_{\phi}(0) = \|\epsilon - \epsilon_0\|_2^2 \quad (\text{a.2})$$

Finish!