A Collection of Frequently Asked Questions about Denoising Diffusion Probabilistic Models*

Kaibo Tang

November 21, 2024

1 Does the forward process tend to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$?

Proposition 1.1. *Suppose* $a_n \in (0,1)\ \forall n \in \mathbb{N}$. *Then,*

$$\prod_{n=1}^{\infty}(1-a_n) = 0 \iff \sum_{n=1}^{\infty} a_n \text{ diverges.} \tag{1}$$

Proof.

1. ($\Longrightarrow$). For the sake of contradiction, assume $\sum_{n=1}^{\infty}$ converges, in which case, we know $a_n \to 0$ as $n \to \infty$. In addition, since

$$\prod_{n=1}^{\infty}(1-a_n) = 0, \tag{2}$$

we have

$$\sum_{n=1}^{\infty} -\ln(1-a_n) = \infty. \tag{3}$$

By L'Hopital's rule, we see that

$$\lim_{n\to\infty} -\frac{a_n}{\ln(1-a_n)} = \lim_{n\to\infty}(1-a_n) = 1, \tag{4}$$

and limit comparison test applies, in which case, we know $\sum_{n=1}^{\infty}$ also diverges, which is contradiction.

*https://arxiv.org/abs/2006.11239

2. ($\Longleftarrow$). Notice that

$$\sum_{n=1}^{\infty} -\ln(1-a_n) \geq \sum_{n=1}^{\infty} a_n. \tag{5}$$

By comparison test, we see that

$$\sum_{n=1}^{\infty} -\ln(1-a_n) = \infty, \tag{6}$$

i.e.,

$$\prod_{n=1}^{\infty}(1-a_n) = 0. \tag{7}$$

□

Proposition 1.2. *Given $\beta_t \in (0,1)$ such that $\sum_{t=1}^{\infty} \beta_t$ diverges, a Markov chain with Gaussian transitions*

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{8}$$

satisfies

$$p(\boldsymbol{x}_t) \to \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{as } t \to \infty. \tag{9}$$

Proof.

1. For each $t \geq 1$, let

$$\boldsymbol{y}_t = \frac{\boldsymbol{x}_t - \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}}{\sqrt{\beta_t}}. \tag{10}$$

We see that $\boldsymbol{y}_t$ is a standard normal random variable, given $\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{t-1}$. Hence, we see that $\boldsymbol{y}_t$ is independent on $\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{t-1}$. Similarly, we see that $\boldsymbol{x}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_t$ are independent.

2. Now, notice that

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{y}_t && (11)\\
&= \sqrt{1-\beta_t}\sqrt{1-\beta_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\beta_t}\sqrt{\beta_{t-1}}\boldsymbol{y}_{t-1} + \sqrt{\beta_t}\boldsymbol{y}_t && (12)\\
&= \cdots && (13)\\
&= \prod_{i=1}^{t}\sqrt{1-\beta_i}\boldsymbol{x}_0 + \sum_{i=1}^{t-1}\left[\left(\prod_{j=i+1}^{t}\sqrt{1-\beta_j}\right)\sqrt{\beta_i}\boldsymbol{y}_i\right] + \sqrt{\beta_t}\boldsymbol{y}_t. && (14)
\end{aligned}$$

We see that the distribution of $\boldsymbol{x}_t$ given $\boldsymbol{x}_0$ is normal with expectation

$$\boldsymbol{\mu} = \prod_{i=1}^{t} \sqrt{1-\beta_i}\boldsymbol{x}_0, \tag{15}$$

and covariance matrix

$$\boldsymbol{\Sigma} = \sum_{i=1}^{t-1} \left[\left(\prod_{j=i+1}^{t} (1-\beta_j) \right) \beta_i \boldsymbol{I} \right] + \beta_t \boldsymbol{I}. \tag{16}$$

To simplify the covariance matrix, notice that

$$\left(\prod_{j=i+1}^{t} (1-\beta_j) \right) \beta_i = \left(\prod_{j=i+1}^{t} (1-\beta_j) \right) (1-(1-\beta_i)) \tag{17}$$

$$= \prod_{j=i+1}^{t} (1-\beta_j) - \prod_{j=i}^{t} (1-\beta_j). \tag{18}$$

Hence, we have

$$\boldsymbol{\Sigma} = \sum_{i=1}^{t-1} \left[\prod_{j=i+1}^{t} (1-\beta_j)\boldsymbol{I} - \prod_{j=i}^{t} (1-\beta_j)\boldsymbol{I} \right] + \beta_t \boldsymbol{I} \tag{19}$$

$$= (1-\beta_t)\boldsymbol{I} - \prod_{j=1}^{t} (1-\beta_j)\boldsymbol{I} + \beta_t \boldsymbol{I} \tag{20}$$

$$= \left[1 - \prod_{j=1}^{t} (1-\beta_j) \right] \boldsymbol{I}. \tag{21}$$

3. Finally, since $\sum_{j=1}^{\infty} \beta_j$ diverges, we see that $\prod_{j=1}^{\infty}(1-\beta_j) = 0$ from Proposition 1.1. Hence, we see that as $t \to \infty$, we have

$$\boldsymbol{\mu} = \prod_{i=1}^{t} \sqrt{1-\beta_i}\boldsymbol{x}_0 \to \boldsymbol{0}, \quad \boldsymbol{\Sigma} = \left[1 - \prod_{j=1}^{t} (1-\beta_j) \right] \boldsymbol{I} \to \boldsymbol{I} \tag{22}$$

for any $\boldsymbol{x}_0$ finite.

□

Remark 1.1. *Let $\alpha_j := 1 - \beta_j$ and $\bar{\alpha}_t := \prod_{j=1}^{t} \alpha_j$. We see that*

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}). \tag{23}$$

2 How do you get from NLL to L_{simple}?

Starting with the negative log likelihood,

$$-\ln p_\theta(\boldsymbol{x}_0), \tag{24}$$

the authors manage to reduce this objective to

$$L_{\text{simple}}(\theta) = E_{t,\boldsymbol{\epsilon}}\left(\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right). \tag{25}$$

This is done in three steps, which I will outline below.

2.1 ELBO

We first find a variational bound for the negative log likelihood term, i.e.,

$$-\ln p_\theta(\boldsymbol{x}_0) \le E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left(-\ln \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right). \tag{26}$$

Proof of this inequality involves noticing that $\ln(\cdot)$ is a convex function and invoking Jensen's inequality.

2.2 KL divergence

Using the definition of p_θ and q, we see that

$$-\ln \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} = -\ln p(\boldsymbol{x}_T) - \sum_{t=1}^{T} \ln \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \tag{27}$$

$$= -\ln p(\boldsymbol{x}_T) - \sum_{t=2}^{T} \ln \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} - \ln \frac{p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}, \tag{28}$$

where the second equality follows from taking the first term out of the summation. Notice that

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t|\boldsymbol{x}_0) = q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0). \tag{29}$$

Hence, we have

$$-\ln \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} = -\ln p(\boldsymbol{x}_T) - \sum_{t=2}^{T} \ln \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \cdot \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} - \ln \frac{p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \tag{30}$$

$$= -\ln \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} - \sum_{t=2}^{T} \ln \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} - \ln p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1). \tag{31}$$

Now, we notice that

$$L_T := E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left(-\ln \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right) \tag{32}$$

$$= E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[E_{\boldsymbol{x}_T\sim q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \left(-\ln \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right) \right] \tag{33}$$

$$= E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\int q(\boldsymbol{x}_T|\boldsymbol{x}_0) \ln \frac{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}{p(\boldsymbol{x}_T)} \; d\boldsymbol{x}_T \right] \tag{34}$$

$$= E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T)) \right]. \tag{35}$$

Remark 2.1. *Given non-learnable variance schedule β_t, the term L_T has no learnable parameter and is thus constant during training.*

Notice that

$$E_{\boldsymbol{x}_{t-1},\boldsymbol{x}_t\sim q(\boldsymbol{x}_{t-1}\boldsymbol{x}_t|\boldsymbol{x}_0)} \left(-\ln \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right) \tag{36}$$

$$= \iint q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t|\boldsymbol{x}_0) \ln \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)} \; d\boldsymbol{x}_{t-1} d\boldsymbol{x}_t \tag{37}$$

$$= \int q(\boldsymbol{x}_t|\boldsymbol{x}_0) \int q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \ln \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)} \; d\boldsymbol{x}_{t-1} d\boldsymbol{x}_t \tag{38}$$

$$= E_{\boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \right], \tag{39}$$

where we use the fact that

$$q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t|\boldsymbol{x}_0) = q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t|\boldsymbol{x}_0) \tag{40}$$

for the second equality. Hence, we have

$$L_{t-1} := E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left(-\ln\frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right) \tag{41}$$

$$= E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[E_{\boldsymbol{x}_{t-1},\boldsymbol{x}_t\sim q(\boldsymbol{x}_{t-1}\boldsymbol{x}_t|\boldsymbol{x}_0)}\left(-\ln\frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right)\right] \tag{42}$$

$$= E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right] \tag{43}$$

Finally, let

$$L_0 := E_{\boldsymbol{x}_{1:T}\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left(-\ln p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right). \tag{44}$$

We see that now the variational bound is given by

$$-\ln p_\theta(\boldsymbol{x}_0) \le L_T + \sum_{t=2}^{T} L_{t-1} + L_0. \tag{45}$$

2.3 Further simplification

We state (without proof) the expression for the KL divergence between two multivariate normal distributions.[1]

Lemma 2.1. *Let P and Q denote two multivariate normal distributions with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^n$, respectively, and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{M}_{n\times n}$, respectively. Then, the KL divergence of P from Q is given by*

$$D_{\mathrm{KL}}(P\parallel Q) = \frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - \ln\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_2} - n\right]. \tag{46}$$

Corollary 2.1. *In the previous case, let $\boldsymbol{\Sigma}_1 = \sigma_1\boldsymbol{I}$ and $\boldsymbol{\Sigma}_2 = \sigma_2\boldsymbol{I}$. Then,*

$$D_{\mathrm{KL}}(P\parallel Q) = \frac{1}{2}\left[\frac{\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2}{\sigma_2^2} + \frac{n\sigma_1^2}{\sigma_2^2} - \ln\frac{\sigma_1}{\sigma_2} - n\right]. \tag{47}$$

Now, consider the reverse process at time t given by

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t,t)) \quad \text{for } t = 2,3,\cdots,T. \tag{48}$$

[1] A detailed proof can be found at https://statproofbook.github.io/P/mvn-kl.html

The following simplification assumes the covariance matrix at each step to be non-learnable and dependent only on time t, i.e.,

$$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \sigma_t^2 \boldsymbol{I}, \tag{49}$$

in which case, we have

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I}) \tag{50}$$

Recall that for the forward process posterior, we have

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I}), \tag{51}$$

$$\text{where } \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t \tag{52}$$

$$\text{and } \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{53}$$

From Corollary 2.1, we see that

$$D_{\mathrm{KL}}(P \parallel Q) = \frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)\|^2 + \text{const.}, \tag{54}$$

since σ_t and $\tilde{\beta}_t$ are fixed.

Now, we recall equation (23), i.e.,

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}). \tag{55}$$

Notice that we can reparameterize $\boldsymbol{x}_t$ as

$$\boldsymbol{x}_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{56}$$

Substituting (52), (56) into (54), we see that

$$L_{t-1} = E_{\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}), t)\right\|^2\right] + \text{const.} \tag{57}$$

Instead of directly parameterizing $\boldsymbol{\mu}_\theta$, consider

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right), \tag{58}$$

where $\boldsymbol{\epsilon}_\theta$ predicts $\boldsymbol{\epsilon}$ from $\boldsymbol{x}_t$.

Remark 2.2. *During inference, we need to sample* $\boldsymbol{x}_{t-1} \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. *Reparameterizing again, we see that*

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \boldsymbol{z} \quad \textit{where } \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{59}$$

Finally, we arrive at

$$L_{t-1} = E_{\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right] \tag{60}$$

3 Disclaimer

There are still several loose ends that are yet to be tied. For example, we have not demonstrated how to deal with L_0, which is nontrivial. But I believe the most important stuff are well-covered in this note. In addition, I would like to also use this opportunity to highlight some caveats.

1. Although the authors do not make this explicit, in Proposition 1.2, we see that the divergence of $\sum_{t=1}^{\infty} \beta_t$ is essential for DDPM to work.

2. In the original paper, the authors start with $E_{\boldsymbol{x}_0}(-\ln p_\theta(\boldsymbol{x}_0))$. However, I think taking the expectation over $\boldsymbol{x}_0$ creates confusion. Hence, I started with $-\ln p_\theta(\boldsymbol{x}_0)$.

3. The authors use E_q ambiguously. Since I dropped $E_{\boldsymbol{x}_0}$, we only need to deal with $E_{\boldsymbol{x}_{1:T} \sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}$ and later on $E_{\boldsymbol{\epsilon}}$, which makes the derivation a lot clearer.

4. Proposition 1.1, Proposition 1.2, Lemma 2.1, and Corollary 2.1 are great conclusions to remember.