Diffusion Models

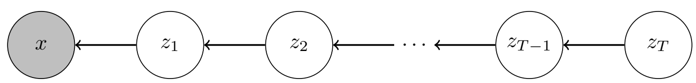
Karl Stratos

1 Framework

Let $p_\theta(x, z) = \pi_\theta(z) \times \kappa_\theta(x|z)$ denote a latent-variable generative model defining a joint distribution over an observed image $x \in \mathbb{R}^d$ and an unobserved "latent image" $z \in \mathbb{R}^d$. Given an image x and a choice of approximate posterior $q(z|x)$, a variational autoencoder (VAE) maximizes the evidence lower bound (ELBO) on the marginal log-likelihood

$$\log \left(\int_z p_\theta(x, z) dz \right) \geq \mathop{\mathbf{E}}_{z \sim q(\cdot|x)} [\log \kappa_\theta(x|z)] - \mathrm{KL}\left(q(\cdot|x) || \pi_\theta\right)$$

A diffusion model is a VAE that assumes the latent is a sequence $z_1 \dots z_T \in \mathbb{R}^d$ for some fixed number of steps (e.g., $T = 1000$). It first draws a completely random image $z_T \in \mathbb{R}^d$ and repeatedly refines it through a *backward* Markov chain, so-called **backward (denoising) process**. More formally,



$$\begin{aligned} p_\theta(x, z_1 \dots z_T) &= \overleftarrow{p}_\theta(z_T|\varnothing, T+1) \times \overleftarrow{p}_\theta(z_{T-1}|z_T, T) \times \cdots \times \overleftarrow{p}_\theta(z_1|z_2, 2) \times \overleftarrow{p}_\theta(x|z_1, 1) \\ &= \prod_{t=1}^{T+1} \overleftarrow{p}_\theta(z_{t-1}|z_t, t) \end{aligned}$$

where we have defined $z_{T+1} = \varnothing$ and $z_0 = x$. A key assumption in diffusion models is that the approximate posterior has a matching form (but conditioning on x): $q(z_1 \dots z_T|x) = \prod_{t=2}^{T+1} \overleftarrow{q}(z_{t-1}|x, z_t, t)$. With this, the ELBO is

$$\max_\theta \underbrace{\mathop{\mathbf{E}}_{z_1 \dots z_T \sim q(\cdot|x)} [\log \overleftarrow{p}_\theta(x|z_1, 1)]}_{\text{reconstruction term}} - \mathop{\mathbf{E}}_{z_1 \dots z_T \sim q(\cdot|x)} \left[\sum_{t=2}^{T+1} \underbrace{\mathrm{KL}\left(\overleftarrow{q}(\cdot|x, z_t, t) || \overleftarrow{p}_\theta(\cdot|z_t, t)\right)}_{\text{stepwise KL term}} \right] \tag{1}$$

1.1 Gaussian Paramaterization

A natural definition of the model is $\overleftarrow{p}_\theta(\cdot|\varnothing, T+1) = \mathcal{N}(0_d,\ I_{d\times d})$ and for $t = T \dots 1$

$$\overleftarrow{p}_\theta(z_{t-1}|z_t, t) = \mathcal{N}(\overleftarrow{\mu}_\theta(z_t, t),\ \sigma_t^2 I_{d\times d})(z_{t-1}) \tag{2}$$

where $\sigma_T^2 > \ldots > \sigma_1^2 > 0$ is some fixed decreasing variance schedule. Here, $\overleftarrow{\mu}_\theta(z, t) \in \mathbb{R}^d$ is a mean regressor. The reconstruction term becomes

$$\mathop{\mathbf{E}}_{z_1 \dots z_T \sim q(\cdot|x)} [\log \overleftarrow{p}_\theta(x|z_1, 1)] = \mathop{\mathbf{E}}_{z_1 \dots z_T \sim q(\cdot|x)} \left[-\frac{1}{2\sigma_1^2} \left|\left| x - \overleftarrow{\mu}_\theta(z_1, 1) \right|\right|^2 \right] + C$$

for some constant C. To match (2), we want an approximate posterior of the form: for $t = T \dots 2$

$$\overleftarrow{q}(z_{t-1}|x, z_t, t) = \mathcal{N}(\tilde{\mu}_t,\ \tilde{\sigma}_t^2 I_{d\times d})(z_{t-1}) \tag{3}$$

where $\tilde{\mu}_t \in \mathbb{R}^d$ and $\tilde{\sigma}_t^2 > 0$ are some functions of x and z_t (thus random variables themselves). The KL term in (1) is, for $t = 2 \dots T$ (ignoring $t = T+1$ which is constant)

$$\mathrm{KL}\left(\overleftarrow{q}(\cdot|x, z_t, t) || \overleftarrow{p}_\theta(\cdot|z_t, t)\right) = \frac{1}{2\sigma_t^2} \left|\left| \tilde{\mu}_t - \overleftarrow{\mu}_\theta(z_t, t) \right|\right|^2 + C'$$

for some constant C'. Note that $\tilde{\sigma}_t^2$ is ignored. Defining $\tilde{\mu}_1 = x$, we see that (1) is equivalent to

$$\min_{\theta} \mathop{\mathbf{E}}_{z_1 \ldots z_T \sim q(\cdot|x)} \left[\sum_{t=1}^{T} \frac{1}{2\sigma_t^2} \left|\left|\tilde{\mu}_t - \overleftarrow{\mu}_\theta(z_t, t)\right|\right|^2 \right] \tag{4}$$

This is a weighted regression problem $\overleftarrow{\mu}_\theta(z_t, t) \approx \tilde{\mu}_t$. Since σ_t^2 is decreasing, the prediction at small t is counted (substantially) more than at large t. To avoid sampling an entire sequence, we assume that the marginal distribution

$$\bar{q}(z|x, t) = \int_{z_1 \ldots z_T :\, z_t = z} q(z_1 \ldots z_T|x)\, d(z_1 \ldots z_T) \tag{5}$$

is easy to sample from (hint: Gaussian). Then (4) is equivalent to

$$\min_{\theta} \mathop{\mathbf{E}}_{t \sim \mathrm{Unif}\{1 \ldots T\},\, z_t \sim \bar{q}(\cdot|x,t)} \left[\frac{1}{2\sigma_t^2} \left|\left|\tilde{\mu}_t - \overleftarrow{\mu}_\theta(z_t, t)\right|\right|^2 \right] \tag{6}$$

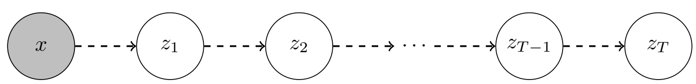
In summary, assuming the Gaussian Markov backward process $\overleftarrow{p}_\theta(\cdot|z_t, t) = \mathcal{N}(\overleftarrow{\mu}_\theta(z_t, t),\ \sigma_t^2 I_{d\times d})$, if we design an approximate posterior $q(z_1 \ldots z_T|x)$ such that

1. Its backward form is also Gaussian Markov: $\overleftarrow{q}(\cdot|x, z_t, t) = \mathcal{N}(\tilde{\mu}_t,\ \tilde{\sigma}_t^2 I_{d\times d})$
2. The stepwise marginal distribution $\bar{q}(\cdot|x, t)$ is easy to sample from (e.g., Gaussian)

then optimizing the ELBO (1) is equivalent to optimizing the samplable weighted regression problem (6).

2 DDPM

A denoising diffusion probabilistic model (DDPM) (Ho *et al.*, 2020) satisfies Condition 1 and 2 by defining the approximate posterior to be a *forward* Gaussian Markov chain, so-called **forward (noising) process**. More formally,



$$q(z_1 \ldots z_T|x) = \prod_{t=1}^{T} \mathcal{N}(\sqrt{1-\beta_t} z_{t-1},\ \beta_t I_{d\times d})(z_t) \tag{7}$$

where $0 < \beta_1 < \cdots < \beta_T < 1$ is some fixed increasing variance schedule (recall $z_0 = x$).

2.1 Marginals

Lemma 2.1. Under (7), the marginal probability (5) is

$$\bar{q}(z|x, t) = \mathcal{N}(\sqrt{\alpha_t} x, (1 - \alpha_t)\, I_{d\times d})(z) \tag{8}$$

where $\alpha_t = \prod_{s=1}^{t} (1 - \beta_s)$.

Proof. We first note that the forward process (7) implies

$$\bar{q}(z|x, t) = \mathop{\mathbf{E}}_{z_1 \ldots z_{t-1} \sim q(\cdot|x)} \left[\mathcal{N}(\sqrt{1-\beta_t} z_{t-1},\ \beta_t I_{d\times d})(z) \right]$$

The base case $z_1 \sim \mathcal{N}(\sqrt{1-\beta_1} z_0,\ \beta_1 I_{d\times d}) = \mathcal{N}(\sqrt{\alpha_1} x, (1 - \alpha_1) I_{d\times d})$ holds by premise. By the reparameterization trick, for $t > 1$,

$$\begin{aligned}
z_t &= \sqrt{1-\beta_t} z_{t-1} + \sqrt{\beta_t} \epsilon_t && (7) \\
&= \sqrt{1-\beta_t} \left(\sqrt{\alpha_{t-1}} x + \sqrt{1-\alpha_{t-1}} \epsilon_{t-1} \right) + \sqrt{\beta_t} \epsilon_t && \text{(inductive step)} \\
&= \sqrt{\alpha_t} x + \sqrt{(1-\beta_t)(1-\alpha_{t-1})} \epsilon_{t-1} + \sqrt{\beta_t} \epsilon_t
\end{aligned}$$

where $\epsilon_{t-1}, \epsilon_t \sim \mathcal{N}(0_d, I_{d\times d})$. The last two terms are independently normally distributed with mean 0_d and covariances $(1-\beta_t)(1-\alpha_{t-1})I_{d\times d}$ and $\beta_t I_{d\times d}$. Thus their sum is distributed as $\mathcal{N}(0_d, \nu^2 I_{d\times d})$ where

$$\begin{aligned}
\nu^2 &= (1-\beta_t)(1-\alpha_{t-1}) + \beta_t \\
&= 1 - (1-\beta_t)\alpha_{t-1} \\
&= 1 - \alpha_t
\end{aligned} \tag{9}$$

This shows that $z_t \sim \mathcal{N}(\sqrt{\alpha_t}x, (1-\alpha_t)I_{d\times d})$. □

Note that $\alpha_t = \prod_{s=1}^{t}(1-\beta_s) = (1-\beta_t)\alpha_{t-1}$ is mapped to β_t by $1-\beta_t = \frac{\alpha_t}{\alpha_{t-1}}$, which is used frequently in derivations. The quantity $1-\alpha_t$ is a variance schedule for $\bar{q}(\cdot|x,t) = \mathcal{N}(\sqrt{\alpha_t}x, (1-\alpha_t)\, I_{d\times d})$, increasing since

$$0 < \beta_1 < \cdots < \beta_T < 1 \qquad \Rightarrow \qquad \begin{aligned} &1 = \alpha_0 > \alpha_1 > \cdots > \alpha_T > 0 \\ &0 = (1-\alpha_0) < (1-\alpha_1) < \cdots < (1-\alpha_T) < 1 \end{aligned}$$

where we have defined $\alpha_0 = 1$. The marginals are "consistent at the extremes" in the following sense. At $t=0$, the marginal becomes a point-mass density on x,

$$\bar{q}(z|x,0) = \mathcal{N}(x, 0_{d\times d})(z) = \begin{cases} 1 & \text{if } z = x \\ 0 & \text{otherwise} \end{cases}$$

As $t \to \infty$, the marginal converges to a standard Gaussian,

$$\lim_{t\to\infty} \bar{q}(\cdot|x,t) = \lim_{t\to\infty} \mathcal{N}(\sqrt{\alpha_t}x, (1-\alpha_t)\, I_{d\times d}) = \mathcal{N}(0_d, I_{d\times d})$$

2.2 Backward Form

A highlight of the Gaussian parameterization is the linear-Gaussian Bayes' rule:

$$\begin{aligned} &\mu \sim \mathcal{N}(\mu_0, \gamma_0 I_{d\times d}) \\ &z|\mu \sim \mathcal{N}(c\mu + b, \gamma I_{d\times d}) \end{aligned} \quad \Rightarrow \quad \begin{aligned} &z \sim \mathcal{N}\left(c\mu_0 + b, (\gamma + c^2\gamma_0) I_{d\times d}\right) \\ &\mu|z \sim \mathcal{N}\left(\left(\frac{\gamma}{\gamma + c^2\gamma_0}\right)\mu_0 + \left(\frac{c\gamma_0}{\gamma + c^2\gamma_0}\right)(z-b),\ \left(\frac{\gamma_0\gamma}{\gamma + c^2\gamma_0}\right) I_{d\times d}\right) \end{aligned} \tag{10}$$

Using the fact that the marginals (8) are Gaussian and the forward noising process (7) is linear-Gaussian,

$$\begin{aligned} z_{t-1}|x &\sim \mathcal{N}(\sqrt{\alpha_{t-1}}x, (1-\alpha_{t-1})\, I_{d\times d}) \\ z_t|x, z_{t-1} &\sim \mathcal{N}(\sqrt{1-\beta_t}z_{t-1},\ \beta_t I_{d\times d}) \end{aligned} \quad \Rightarrow \quad z_{t-1}|x, z_t \sim \underbrace{\mathcal{N}(\tilde{\mu}_t(x, z_t),\ \tilde{\sigma}_t^2 I_{d\times d})}_{\overleftarrow{q}(\cdot|x,z_t,t)} \tag{11}$$

where

$$\tilde{\mu}_t(x, z_t) = \frac{\beta_t\sqrt{\alpha_{t-1}}}{1-\alpha_t}x + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}z_t \tag{12}$$

$$\tilde{\sigma}_t^2 = \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t} \tag{13}$$

2.3 Noise Predictive Formulation

Plugging (8) and (12) in the ELBO (6), we have

$$\min_\theta \mathop{\mathbf{E}}_{\substack{t\sim\text{Unif}\{1\ldots T\},\ \epsilon_t\sim\mathcal{N}(0_d, I_{d\times d}): \\ z_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t}} \left[\frac{1}{2\sigma_t^2}\left|\left|\tilde{\mu}_t(x, z_t) - \overleftarrow{\mu}_\theta(z_t, t)\right|\right|^2\right] \tag{14}$$

To avoid directly modeling $\tilde{\mu}_t(x, z_t) \in \mathbb{R}^d$ which is high-variance for random x, note that $z_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t$ or equivalently

$$x = \frac{z_t - \sqrt{1-\alpha_t}\epsilon_t}{\sqrt{\alpha_t}} \tag{15}$$

where $\epsilon_t \sim \mathcal{N}(0_d, I_{d\times d})$. This allows us to express $\tilde{\mu}_t(x, z_t)$ a function of only z_t and ϵ_t. While not necessary, it can be simplified as

$$\begin{aligned}
\tilde{\mu}_t(x, z_t) &= \frac{\beta_t \sqrt{\alpha_{t-1}}}{1-\alpha_t}\left(\sqrt{\frac{1}{\alpha_t}} z_t - \sqrt{\frac{1-\alpha_t}{\alpha_t}} \epsilon_t\right) + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t} z_t \\
&= \sqrt{\frac{1}{1-\beta_t}}\left(\frac{\beta_t}{1-\alpha_t} z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_t + \frac{(1-\beta_t)(1-\alpha_{t-1})}{1-\alpha_t} z_t\right) \\
&= \sqrt{\frac{1}{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_t\right) \qquad (16)
\end{aligned}$$

where the second equality uses $\frac{\alpha_{t-1}}{\alpha_t} = \frac{1}{1-\beta_t}$ and the final equality makes the same observation in (9). We now define the mean regressor $\overleftarrow{\mu}_\theta(z_t, t)$ in matching form:

$$\overleftarrow{\mu}_\theta(z_t, t) = \sqrt{\frac{1}{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(z_t, t)\right) \qquad (17)$$

where $\epsilon_\theta : \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}^d$ is a noise predictor (e.g., U-Net with sinusoidal step embeddings). Plugging (17) and (16) in (14), we have

$$\min_\theta \underset{\substack{t\sim \mathrm{Unif}\{1\ldots T\},\ \epsilon_t \sim \mathcal{N}(0_d, I_{d\times d}): \\ z_t = \sqrt{\alpha_t} x + \sqrt{1-\alpha_t}\epsilon_t}}{\mathbf{E}} \left[\lambda_t^{\mathrm{DDPM}} \left|\left|\epsilon_t - \epsilon_\theta(z_t, t)\right|\right|^2\right] \qquad (18)$$

for the stepwise weights $\lambda_t^{\mathrm{DDPM}} = \frac{\beta_t^2}{2\sigma_t^2(1-\beta_t)(1-\alpha_t)}$, again larger for small t. Ho *et al.* (2020) overwrite $\lambda_t^{\mathrm{DDPM}} \leftarrow 1$. This "surrogate objective" is no longer the true ELBO and corresponds to upweighting large t (i.e., focus more on the noisy phase).[1]

2.4 Generation

Once the noise predictor ϵ_θ is trained, we can sample $x, z_1 \ldots z_T \sim p_\theta$ by the backward process (2) as

1. Sample $z_T \sim \mathcal{N}(0_d, I_{d\times d})$.
2. For $t = T \ldots 1$, sample $z_{t-1} \sim \mathcal{N}(\sqrt{\frac{1}{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(z_t, t)\right),\ \sigma_t^2 I_{d\times d})$ (see (17)).
3. Return $x = z_0, z_1 \ldots z_T$.

Note that the model variance σ_t^2 does not affect training with the surrogate objective, but still affects generation.

3 DDIM

It is tempting to speed up the stepwise generation 2 by "skipping" some steps. But this requires marginalizing over the skipped steps. Denoising diffusion implicit models (DDIMs) (Song *et al.*, 2021) get around this difficulty by

1. *Defining* the approximate posterior to be a Markov backward chain over any *subsequence* of the latent images

$$q_\tau(z_{\tau_1} \ldots z_{\tau_m}|x) = \prod_{l=2}^{m+1} \overleftarrow{q}(z_{\tau_{l-1}}|x, z_{\tau_l}, \tau_l, \tau_{l-1}) \qquad (19)$$

   where $\tau = (\tau_1 \ldots \tau_m)$ satisfies $\tau_1 < \cdots < \tau_m = T$ and $\tau_{m+1} = \varnothing$ is a dummy step

2. *Defining* the model to match the approximate posterior, except for replacing x with a prediction

$$p_\theta(x = z_{\tau_0}, z_{\tau_1} \ldots z_{\tau_m}) = \prod_{l=1}^{m+1} \overleftarrow{q}\left(z_{\tau_{l-1}}|f_\theta(z_{\tau_l}, \tau_l), z_{\tau_l}, \tau_l, \tau_{l-1}\right) \qquad (20)$$

A neat technical trick is that we can define (19) to have the same marginals in (8). Since the ELBO (6) only depends on samples from the marginals and the means of (20) and (19) (matched by construction), DDIMs with the full sequence $\tau = (1 \ldots T)$ will have the same surrogate objective as DDPMs.

[1] This will also hold if the model $\epsilon_\theta(z_t, t)$ is not shared across time steps (not the case in practice).

3.1 Approximate Posterior

Lemma 3.1. Let $\alpha, \sigma \in \mathbb{R}^T_{\geq 0}$ and $\tau = (\tau_1 \dots \tau_m)$ a subsequence of $(1 \dots T)$ such that $\tau_1 < \dots < \tau_m = T$. Let $\tau_{m+1} = \varnothing$ denote a dummy step. Given $x \in \mathbb{R}^d$, define

$$q_{\alpha,\sigma,\tau}(z_{\tau_1} \dots z_{\tau_m} | x) = \prod_{l=2}^{m+1} \overleftarrow{q}_{\alpha,\sigma}(z_{\tau_{l-1}} | x, z_{\tau_l}, \tau_l, \tau_{l-1}) \tag{21}$$

where $\overleftarrow{q}_{\alpha,\sigma}(\cdot | x, z_\varnothing, \varnothing, \tau_m) = \mathcal{N}(\sqrt{\alpha_{\tau_m}} x, (1 - \alpha_{\tau_m}) I_{d \times d})$ and for $l = m \dots 2$,

$$\overleftarrow{q}_{\alpha,\sigma}(\cdot | x, z_{\tau_l}, \tau_l, \tau_{l-1}) = \mathcal{N}\left(\sqrt{\alpha_{\tau_{l-1}}} x + \sqrt{\frac{1 - \alpha_{\tau_{l-1}} - \sigma^2_{\tau_l}}{1 - \alpha_{\tau_l}}} \left(z_{\tau_l} - \sqrt{\alpha_{\tau_l}} x\right),\ \sigma^2_{\tau_l} I_{d \times d}\right) \tag{22}$$

Then (22) is the only distribution of the linear-Gaussian form $\mathcal{N}(c z_{\tau_l} + b, \sigma^2_{\tau_l} I_{d \times d})$ for some $c \in \mathbb{R}$ and $b \in \mathbb{R}^d$ such that the marginals of (21) (as defined in (5)) satisfy $\bar{q}(\cdot | x, \tau_l) = \mathcal{N}(\sqrt{\alpha_{\tau_l}} x, (1 - \alpha_{\tau_l}) I_{d \times d})$ for $l = 1 \dots m$.

Proof. At $l = m$, the statement is true by definition. We now give a constructive proof by induction. Let $t = \tau_l$ and $s = \tau_{l-1}$ for $l \leq m$. Using (i) the inductive step, (ii) the Markov assumption in (21), and (iii) Bayes rule' (10),

$$\begin{aligned} z_t | x &\sim \mathcal{N}(\sqrt{\alpha_t} x, (1 - \alpha_t) I_{d \times d}) && \Rightarrow && z_s | x \sim \mathcal{N}\left(c \sqrt{\alpha_t} x + b, (\sigma_t^2 + c^2 (1 - \alpha_t)) I_{d \times d}\right) \\ z_s | x, z_t &\sim \mathcal{N}(c z_t + b, \sigma_t^2 I_{d \times d}) \end{aligned} \tag{23}$$

We want $c \sqrt{\alpha_t} x + b = \sqrt{\alpha_s} x$ and $\sigma_t^2 + c^2 (1 - \alpha_t) = 1 - \alpha_s$. Solving for c in the latter and then b in the former gives

$$c = \sqrt{\frac{1 - \alpha_s - \sigma_t^2}{1 - \alpha_t}} \qquad\qquad b = \sqrt{\alpha_s} x - \sqrt{\frac{1 - \alpha_s - \sigma_t^2}{1 - \alpha_t}} \sqrt{\alpha_t} x$$

We conclude that to have the marginal $\bar{q}(z | x, s) = \mathcal{N}(\sqrt{\alpha_s} x, (1 - \alpha_s) I_{d \times d})$, the distribution $z_s | x, z_t \sim \mathcal{N}(c z_t + b, \sigma_t^2 I_{d \times d})$ must have the form

$$z_s | x, z_t \sim \mathcal{N}\left(\sqrt{\alpha_s} x + \sqrt{\frac{1 - \alpha_s - \sigma_t^2}{1 - \alpha_t}} (z_t - \sqrt{\alpha_t} x), \sigma_t^2 I_{d \times d}\right)$$

□

Corollary 3.2. The DDPM approximate posterior (7), with the associated $\beta, \alpha \in \mathbb{R}^T_{\geq 0}$, is a special case of the DDIM approximate posterior (21) using the full sequence $\tau = (1 \dots T)$ and the variance $\sigma_t^2 = \frac{\beta_t (1 - \alpha_{t-1})}{1 - \alpha_t}$.

Proof. The DDPM has the Markov backward form $\mathcal{N}(c z_t + b, \frac{\beta_t (1 - \alpha_{t-1})}{1 - \alpha_t} I_{d \times d})$ (11) with $\mathcal{N}(\sqrt{\alpha_t} x, (1 - \alpha_t) I_{d \times d})$ as the marginals. By Lemma 3.1, this is the same distribution as (22) using $\tau = (1 \dots T)$ and $\sigma_t^2 = \frac{\beta_t (1 - \alpha_{t-1})}{1 - \alpha_t}$. It also implies that the DDIM now has the same Markov forward noising process (7) since they have the same marginals and likelihoods. □

3.2 Model

Lemma 3.3. Let $\alpha, \sigma \in \mathbb{R}^T_{\geq 0}$ and $\tau = (\tau_0, \tau_1 \dots \tau_m)$ a subsequence of $(0, 1 \dots T)$ such that $\tau_0 = 0 < \tau_1 < \dots < \tau_m = T$. Pick $f_\theta : \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}^d$ and define

$$p_{\alpha,\sigma,\tau,\theta}(x = z_{\tau_0}, z_{\tau_1} \dots z_{\tau_m}) = \mathcal{N}(0_d, I_{d \times d})(z_T) \times \prod_{l=1}^{m} \overleftarrow{q}_{\alpha,\sigma}(\cdot | f_\theta(z_{\tau_l}, \tau_l), z_{\tau_l}, \tau_l, \tau_{l-1}) \tag{24}$$

where $\overleftarrow{q}_{\alpha,\sigma}$ is defined in (22). If $\tau = (0, 1 \dots T)$ and $f_\theta(z, t) = \frac{z - \sqrt{1 - \alpha_t} \epsilon_\theta(z, t)}{\sqrt{\alpha_t}}$ for some $\epsilon_\theta : \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}^d$, the ELBO (6) using the approximate posterior in Lemma 3.1 is equivalent to

$$\min_\theta \mathop{\mathbf{E}}_{\substack{t \sim \mathrm{Unif}\{1 \dots T\},\ \epsilon_t \sim \mathcal{N}(0_d, I_{d \times d}): \\ z_t = \sqrt{\alpha_t} x + \sqrt{1 - \alpha_t} \epsilon_t}} \left[\lambda_t^{\mathrm{DDIM}} \left|\left|\epsilon_t - \epsilon_\theta(z_t, t)\right|\right|^2\right] \tag{25}$$

for the stepwise weights $\lambda_t^{\mathrm{DDIM}} = \frac{\beta_t + (1 - \beta_1) \sigma_t^2}{2 \sigma_t^2 (1 - \beta_t)}$.

Proof. In the ELBO (6), the mean of the backward Markov approximate posterior is $\tilde{\mu}_t = (\sqrt{\alpha_{t-1}} - c\sqrt{\alpha_t})x + cz_t$ and the mean of the model is $\overleftarrow{\mu}_\theta(z_t, t) = (\sqrt{\alpha_{t-1}} - c\sqrt{\alpha_t})f_\theta(z_t, t) + cz_t$ where $c = \sqrt{(1 - \alpha_{t-1} - \sigma_t^2)/(1 - \alpha_t)}$. Thus it becomes

$$\min_\theta \mathop{\mathbf{E}}_{t \sim \text{Unif}\{1 \ldots T\},\ z_t \sim \bar{q}(\cdot|x,t)} \left[\frac{(\sqrt{\alpha_{t-1}} - c\sqrt{\alpha_t})^2}{2\sigma_t^2} \left|\left| x - f_\theta(z_t, t) \right|\right|^2 \right]$$

Using the fact that $\bar{q}(\cdot|x, t) = \mathcal{N}(\sqrt{\alpha_t}x, (1 - \alpha_t)\, I_{d \times d})$ (Lemma 3.1), the Gaussian parameterization trick (15), and the parameterization $f_\theta(z, t) = \frac{z - \sqrt{1-\alpha_t}\epsilon_\theta(z,t)}{\sqrt{\alpha_t}}$, it is equivalent to

$$\min_\theta \mathop{\mathbf{E}}_{\substack{t \sim \text{Unif}\{1 \ldots T\},\ \epsilon_t \sim \mathcal{N}(0_d, I_{d \times d}): \\ z_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t}} \left[\frac{(\sqrt{\alpha_{t-1}} - c\sqrt{\alpha_t})^2 \frac{1-\alpha_t}{\alpha_t}}{2\sigma_t^2} \left|\left| \epsilon_t - \epsilon_\theta(z_t, t) \right|\right|^2 \right]$$

Simplifying the coefficient gives the statement.[2] □

Training. The ELBO under DDIMs (25) and the ELBO under DDPMs (18) are the same except for slightly different stepwise coefficients (λ_t^{DDIM} vs λ_t^{DDPM}). In particular, under the surrogate objective that overwrites the coefficients to be 1, training a DDIM over the full sequence $\tau = (0, 1 \ldots T)$ using any $\alpha, \sigma \in \mathbb{R}_{\geq 0}^T$ is equivalent to training a DDPM using those $\alpha_1 \ldots \alpha_T$.

3.3 Generation

We take a trained DDPM noise predictor ϵ_θ associated with $\beta, \alpha \in \mathbb{R}_{\geq 0}^T$. We choose a subsequence $\tau = (\tau_0, \tau_1 \ldots \tau_m)$ of $(0, 1 \ldots T)$ where $\tau_0 = 0 < \tau_1 < \cdots < \tau_m = T$ and a variance schedule $\sigma^2 \in \mathbb{R}_{\geq 0}^T$. We then sample $x, z_{\tau_1} \ldots z_{\tau_m} \sim p_{\alpha,\sigma,\tau,\theta}$ by the backward process (24) as

1. Sample $z_T \sim \mathcal{N}(0_d, I_{d \times d})$.
2. For $l = m \ldots 1$, sample $z_{\tau_{l-1}} \sim \mathcal{N}\left(\sqrt{\alpha_{\tau_{l-1}}} \left(\frac{z_{\tau_l} - \sqrt{1-\alpha_{\tau_l}}\epsilon_\theta(z_{\tau_l}, \tau_l)}{\sqrt{\alpha_{\tau_l}}} \right) + \sqrt{1 - \alpha_{\tau_{l-1}} - \sigma_{\tau_l}^2}\epsilon_\theta(z_{\tau_l}, \tau_l),\ \sigma_{\tau_l}^2 I_{d \times d} \right)$.
3. Return $x = z_{\tau_0}, z_{\tau_1} \ldots z_{\tau_m}$.

Like DDPMs, the variance schedule σ^2 only affects generation. If we choose $\tau = (0, 1 \ldots T)$ and $\sigma_t^2 = \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t}$, by Corollary 3.2 we recover the DDPM sampling (Section 2.4).

References

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, **33**, 6840–6851.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

[2] $\sqrt{\alpha_{t-1}} - c\sqrt{\alpha_t} = \sqrt{\frac{\alpha_{t-1}(1-\alpha_t) - (1-\alpha_{t-1}-\sigma_t^2)\alpha_t}{1-\alpha_t}} = \sqrt{\frac{\alpha_{t-1} - \alpha_t(1-\sigma_t^2)}{1-\alpha_t}} \Rightarrow \text{numerator} = \frac{1}{1-\beta_t} - 1 + \sigma_t^2 \Rightarrow \text{coeff} = \frac{\beta_t + (1-\beta_t)\sigma_t^2}{2\sigma_t^2(1-\beta_t)}$