Let $x_0 \in \mathbb{R}^d$ be an image and D be our image data

$p_\theta(x)$ is our data distribution & $p(x_0)$ the probability of a single image

Training objective is maximize $p_\theta(x_0)$

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

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T}$$ the marginalization over all intermediate latents $x_1, \dots, x_T$

intractable

We observe x_0 an image, all intermediate $x_{1:T}$ are latent.

"The probability of seeing x_0 is the total probability of all possible reverse trajectories that end at x_0, weighted by how likely each trajectory is under the model"

What is $p_\theta(x_{0:T})$?

$$p_\theta(x_{0:T}) = p(x_T)\, p(x_{T-1} \mid x_T)\, p(x_{T-2} \mid x_{T-1}, x_T) \cdots p(x_0 \mid x_{1:T})$$

by chain Rule

$$= p(x_T) \prod_{t=1}^{T} p(x_{t-1} \mid x_t)$$ by Markov Assumption

$$\Rightarrow p_\theta(x_0) = \int p(x_T) \prod p_\theta(x_{t-1} \mid x_t)\, dx_{1:T}$$

The true reverse kernel $p(x_{t-1} \mid x_t)$ is unknown

"Given a noisy latent x_t, what is the distribution of the slightly less noisy latent x_{t-1} in the true data generating process?" - we don't know. So we introduce known $q(x_{1:T} \mid x_0)$, the forward process

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1})$$

we don't know this yet

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}\, x_0, (1-\bar{\alpha}_t) I)$$

Why do we introduce $q(x_{1:T} | x_0)$?

We need a tractable way to sample latent trajectories $(x_1, \dots, x_T)$ given a real data sample x_0. That's what the forward noising process $q(x_{1:T} | x_0)$ is for

$q(x_{1:T} | x_0) = \prod_{t=1}^{T} q(x_t | x_{t-1})$. It is known Gaussian kernel we design ourselves.

Deriving the forward process

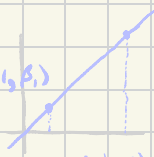
How do we get the parameters of the conditional distribution $q(x_t | x_0)$?

Let's start with $q(x_t | x_{t-1}) = N(a_t x_{t-1}, \beta_t I)$

where mean $= a_t x_{t-1}$ the a_t is scaling the signal from the previous latent

What about β_t?

DDPM: ww the corruption to progress linearly



Slope $m = \frac{\beta_T - \beta_1}{T - 1}$

Line $y = mx + b$

point slope form at $t = 1$: $\beta_t = \beta_1 + m(t-1)$

$$\therefore \beta_t = \beta_1 + \frac{t-1}{T-1}(\beta_T - \beta_1)$$

Known $\mu = a_t x_{t-1}$ & $\sigma^2 = \beta_t I$

and using the sampling form of a Gaussian

$x = \mu + \sigma\epsilon, \epsilon \sim N(0,1) \Rightarrow x_t = a_t x_{t-1} + \sqrt{\beta_t}\,\epsilon_t, \epsilon_t \sim N(0, I)$

What is α_t?

Now, Let $Var(X_{t-1}) = 1$ and we choose to preserve unit variance at each timestep t

That means now $Var(X_t) = 1$

$$\Rightarrow Var(X_t) = Var(\alpha_t X_{t-1} + \sqrt{\beta_t}\,\epsilon_t)$$
$$= \alpha_t^2\, Var(X_{t-1}) + \beta_t\, Var(\epsilon_t)$$
$$= \alpha_t^2 + \beta_t = 1$$

Now $\alpha_t^2 + \beta_t = 1$

$\Rightarrow \alpha_t = \sqrt{1-\beta_t}$ $\Rightarrow$

Define: $\alpha_t := 1 - \beta_t$

Then $\alpha_t = \sqrt{\alpha_t}$

$$\therefore q(X_t \mid X_{t-1}) = N(\sqrt{\alpha_t}\, X_{t-1}, \beta_t I)$$

Deriving closed form $q(X_t \mid X_0)$

We've just defined the one-step forward kernel $q(X_t \mid X_{t-1})$ which tells us: if we know the image at $t-1$ how do we sample the next noisy image X_t

That's useful for simulation, but during training we do something different. we sample a t and jump straight from X_0 to X_t in one go. why? so we can avoid all intermediate steps

In the training objective we'll draw $X_t \mid X_0$ and schedule $\beta_1 \longrightarrow \beta_t$ then feed that to the network

Thus, mathematically we need $q(X_t \mid X_0)$

Luckily $q(x_t \mid x_0)$ comes directly from $q(x_t \mid x_{t-1})$ by unrolling the recursion and tracking how the scale and noise accumulate.

Let's unroll the recursion

$$x_3 = \sqrt{\alpha_3}\, x_2 + \sqrt{\beta_3}\, \epsilon_3$$
$$x_2 = \sqrt{\alpha_2}\, x_1 + \sqrt{\beta_2}\, \epsilon_2$$
$$x_1 = \sqrt{\alpha_1}\, x_0 + \sqrt{\beta_1}\, \epsilon_1$$

$$x_3 = \sqrt{\alpha_3}\left(\sqrt{\alpha_2}\, x_1 + \sqrt{\beta_2}\, \epsilon_2\right) + \sqrt{\beta_3}\, \epsilon_3$$
$$= \sqrt{\alpha_3 \alpha_2}\, x_1 + \sqrt{\alpha_3}\sqrt{\beta_2}\, \epsilon_2 + \sqrt{\beta_3}\, \epsilon_3$$
$$= \sqrt{\alpha_3 \alpha_2}\left(\sqrt{\alpha_1}\, x_0 + \sqrt{\beta_1}\, \epsilon_1\right) + \sqrt{\alpha_3}\sqrt{\beta_2}\, \epsilon_2 + \sqrt{\beta_3}\, \epsilon_3$$
$$= \sqrt{\prod_{s=1}^{3} \alpha_s}\, x_0 + \sqrt{\alpha_3 \alpha_2 \beta_1}\, \epsilon_1 + \sqrt{\alpha_3}\sqrt{\beta_2}\, \epsilon_2 + \sqrt{\beta_3}\, \epsilon_3$$

$$= \sqrt{\prod_{s=1}^{t} \alpha_s}\, x_0 + \sum_{s=1}^{t} \sqrt{\beta_s}\left(\prod_{m=s+1}^{t} \alpha_m\right) \epsilon_s$$

(P) (Q)

P, our mean → $\sqrt{\prod_{s=1}^{t} \alpha_s} = \sqrt{\prod_{s=1}^{t} \alpha_s} \equiv \sqrt{\bar{\alpha}_s}$

What about Q?

Property:
Linear combination of independent Gaussians is Gaussian

Since the $\epsilon_s \sim N(0, I)$ (independent Gaussians)

$$\Rightarrow \sum_{s=1}^{t} c_s \epsilon_s, \text{ with } c_s = \sqrt{\beta_s \prod_{m=s+1}^{t} \alpha_m} \sim N\left(0, \sum_{s=1}^{t} c_s^2\right)$$

$$\text{Var}(x_t \mid x_0) = \sum_{s=1}^{t} c_s^2 = \sum_{s=1}^{t} \beta_s \prod_{m=s+1}^{t} \alpha_m$$

Can we get a better closed form for the variance?

SDXL → $q(x_t \mid x_0) \sim N\left(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I\right)$

How do we get $\text{var} = 1 - \bar{\alpha}_t$?

Recall

$$\alpha_t = 1 - \beta_t \rightarrow \beta_t = 1 - \alpha_t$$

and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$

$$\Rightarrow \sum_{s=1}^{t} \beta_s \prod_{u=s+1}^{t} \alpha_u = \sum_{s=1}^{t} (1-\alpha_s) \prod_{u=s+1}^{t} \alpha_u$$

$$(1-\alpha_s) \prod_{u=s+1}^{t} \alpha_u = \prod_{u=s+1}^{t} \alpha_u - \alpha_s \prod_{u=s+1}^{t} \alpha_u = \prod_{u=s+1}^{t} \alpha_u - \prod_{u=s}^{t} \alpha_u$$

$$\Rightarrow \sum_{s=1}^{t} \beta_s \prod_{u=s+1}^{t} \alpha_u = \sum_{s=1}^{t} \left(\prod_{u=s+1}^{t} \alpha_u - \prod_{u=s}^{t} \alpha_u \right)$$

$$\prod_{u=2}^{t} \alpha_u - \prod_{u=1}^{t} \alpha_u + \prod_{u=3}^{t} \alpha_u - \prod_{u=2}^{t} \alpha_u + \cdots + \prod_{u=t+1}^{t} \alpha_u - \prod_{u=t}^{t} \alpha_u = 1 - \bar{\alpha}_t$$

$t=3 \rightarrow$ $\prod_{u=1}^{t} \alpha_u$ $\quad 1 \quad \rightarrow 1 - \bar{\alpha}_t$

$$\therefore \text{var}(x_t | x_0) = \sum_{s=1}^{t} \beta_s \prod_{u=s+1}^{t} \alpha_u = 1 - \bar{\alpha}_t = 1 - \prod_{s=1}^{t} (1-\beta_s)$$

$$\therefore q(x_t | x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I\right)$$