

Denoising Diffusion Probabilistic Models (DDPM): A Complete Derivation

These notes follow the structure of the video *Diffusion Models: DDPM | Generative AI Animated* and fill in the derivations it skips: the closed-form forward marginal, the ELBO decomposition (Appendix A of Ho et al., 2020), the true posterior, the Gaussian KL divergence, and the noise reparameterization.

Contents

1	The naive forward process (variance exploding)	1
2	The variance-preserving forward process	2
3	Forward and reverse joint distributions	2
4	From negative log-likelihood to the ELBO	3
5	Rewriting the ELBO as KL divergences	3
6	The true posterior $q(x_{t-1} x_t, x_0)$	4
7	The Gaussian KL reduces to a squared distance	4
8	Reparameterizing in terms of noise	5
9	Random time steps and the simplified loss	5
10	Training and sampling	5

1 The naive forward process (variance exploding)

Start from a clean sample $x_0 \sim p(x)$ and add Gaussian noise of variance β at each step:

$$x_t = x_{t-1} + \sqrt{\beta} \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \mathbf{I}), \quad (1)$$

equivalently $q(x_t | x_{t-1}) = \mathcal{N}(x_t; x_{t-1}, \beta \mathbf{I})$. Unrolling the recursion,

$$x_t = x_0 + \sqrt{\beta} (\varepsilon_1 + \cdots + \varepsilon_t). \quad (2)$$

A sum of t independent $\mathcal{N}(0, \beta \mathbf{I})$ variables is $\mathcal{N}(0, t\beta \mathbf{I})$, hence

$$q(x_t | x_0) = \mathcal{N}(x_t; x_0, t\beta \mathbf{I}). \quad (3)$$

The mean stays at x_0 and the variance $t\beta \rightarrow \infty$, so the process never converges to $\mathcal{N}(0, \mathbf{I})$. This is the *variance exploding* process.

2 The variance-preserving forward process

To drive the mean to 0 and the variance to 1, DDPM scales the previous state before adding noise:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}). \quad (4)$$

Define

$$\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s. \quad (5)$$

With the reparameterization trick, one step reads

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \varepsilon_t. \quad (6)$$

Why $\sqrt{1 - \beta_t}$? If $\text{Var}(x_{t-1}) = 1$, then $\text{Var}(x_t) = (1 - \beta_t) \cdot 1 + \beta_t = 1$. The coefficient is exactly what keeps unit variance fixed, hence “variance preserving”.

Closed form for $q(x_t | x_0)$. We show by induction that

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \bar{\varepsilon}, \quad \bar{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (7)$$

For $t = 1$ this holds since $\bar{\alpha}_1 = \alpha_1$. Assuming it for $t - 1$,

$$\begin{aligned} x_t &= \sqrt{\alpha_t} \left(\sqrt{\bar{\alpha}_{t-1}} x_0 + \sqrt{1 - \bar{\alpha}_{t-1}} \bar{\varepsilon}' \right) + \sqrt{1 - \alpha_t} \varepsilon_t \\ &= \sqrt{\bar{\alpha}_t} x_0 + \underbrace{\sqrt{\alpha_t(1 - \bar{\alpha}_{t-1})} \bar{\varepsilon}' + \sqrt{1 - \alpha_t} \varepsilon_t}_{\text{independent zero-mean Gaussians}}. \end{aligned} \quad (8)$$

The noise term has variance

$$\alpha_t - \alpha_t \bar{\alpha}_{t-1} + 1 - \alpha_t = 1 - \bar{\alpha}_t, \quad (9)$$

so it equals $\sqrt{1 - \bar{\alpha}_t} \bar{\varepsilon}$. Therefore

$$\boxed{q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I})} \quad (10)$$

For a constant $\beta \in (0, 1)$, $\bar{\alpha}_t = (1 - \beta)^t \rightarrow 0$, so the mean tends to 0 and the variance to 1: $q(x_t | x_0) \rightarrow \mathcal{N}(0, \mathbf{I})$. With a schedule β_t , $\bar{\alpha}_t = \prod_s (1 - \beta_s)$, as in the paper.

3 Forward and reverse joint distributions

Since the forward process is a Markov chain,

$$q(x_{1:T} | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}). \quad (11)$$

The learned reverse process starts from pure noise $p(x_T) = \mathcal{N}(0, \mathbf{I})$ and is unconditional:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t). \quad (12)$$

4 From negative log-likelihood to the ELBO

We want to minimize $-\log p_\theta(x_0)$. Marginalizing,

$$p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T}, \quad (13)$$

which sums over every path from noise to x_0 and is intractable. Multiplying and dividing by the forward process,

$$p_\theta(x_0) = \int q(x_{1:T} | x_0) \frac{p_\theta(x_{0:T})}{q(x_{1:T} | x_0)} dx_{1:T} = \mathbb{E}_{q(x_{1:T} | x_0)} \left[\frac{p_\theta(x_{0:T})}{q(x_{1:T} | x_0)} \right]. \quad (14)$$

Since $-\log$ is convex, Jensen's inequality gives $-\log \mathbb{E}[Y] \leq \mathbb{E}[-\log Y]$, so

$$-\log p_\theta(x_0) \leq \mathbb{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} | x_0)} \right] =: L. \quad (15)$$

L is an upper bound on the negative log-likelihood; equivalently $-L$ is a lower bound on the log-likelihood, the *evidence lower bound* (ELBO).

5 Rewriting the ELBO as KL divergences

Step 1: expand. Plugging in the factorizations,

$$L = \mathbb{E}_q \left[-\log p(x_T) - \sum_{t=1}^T \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} \right]. \quad (16)$$

Separating the $t = 1$ term,

$$L = \mathbb{E}_q \left[-\log p(x_T) - \sum_{t=2}^T \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} - \log \frac{p_\theta(x_0 | x_1)}{q(x_1 | x_0)} \right]. \quad (17)$$

Step 2: flip the forward step with Bayes' rule. By the Markov property $q(x_t | x_{t-1}) = q(x_t | x_{t-1}, x_0)$, and by Bayes' rule

$$q(x_t | x_{t-1}, x_0) = \frac{q(x_{t-1} | x_t, x_0) q(x_t | x_0)}{q(x_{t-1} | x_0)}. \quad (18)$$

The sum over $t \geq 2$ therefore splits into

$$-\sum_{t=2}^T \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_{t-1} | x_t, x_0)} - \sum_{t=2}^T \log \frac{q(x_{t-1} | x_0)}{q(x_t | x_0)}. \quad (19)$$

Step 3: telescope. The second sum telescopes:

$$\sum_{t=2}^T [\log q(x_{t-1} | x_0) - \log q(x_t | x_0)] = \log q(x_1 | x_0) - \log q(x_T | x_0). \quad (20)$$

Combining with the remaining terms, the $q(x_1 | x_0)$ factors cancel:

$$L = \mathbb{E}_q \left[-\log \frac{p(x_T)}{q(x_T | x_0)} - \sum_{t=2}^T \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_{t-1} | x_t, x_0)} - \log p_\theta(x_0 | x_1) \right]. \quad (21)$$

Step 4: recognize KL divergences. Each log-ratio involves only a few variables; averaging over the rest yields

$$L = \mathbb{E}_q \left[\underbrace{D_{\text{KL}}(q(x_T | x_0) \| p(x_T))}_{L_T} + \sum_{t=2}^T \underbrace{D_{\text{KL}}(q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t))}_{L_{t-1}} \underbrace{- \log p_\theta(x_0 | x_1)}_{L_0} \right]. \quad (22)$$

Dropping terms. L_T contains no trainable parameters. L_0 concerns only the final, nearly noise-free step and contributes little learning signal; it is dropped in practice (the paper models it with a discrete decoder). We are left with the sum of L_{t-1} terms.

6 The true posterior $q(x_{t-1} | x_t, x_0)$

By Bayes' rule, $q(x_{t-1} | x_t, x_0) \propto q(x_t | x_{t-1}) q(x_{t-1} | x_0)$. Up to a constant, its log-density is

$$-\frac{1}{2} \left[\frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|^2}{\beta_t} + \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0\|^2}{1 - \bar{\alpha}_{t-1}} \right]. \quad (23)$$

This is quadratic in x_{t-1} , so the posterior is Gaussian, $\mathcal{N}(\tilde{\mu}_t, \tilde{\beta}_t \mathbf{I})$. Matching with $-\frac{1}{2\tilde{\beta}_t} \|x_{t-1} - \tilde{\mu}_t\|^2$:

Variance. From the coefficient of $\|x_{t-1}\|^2$,

$$\frac{1}{\tilde{\beta}_t} = \frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} = \frac{\alpha_t(1 - \bar{\alpha}_{t-1}) + \beta_t}{\beta_t(1 - \bar{\alpha}_{t-1})} = \frac{1 - \bar{\alpha}_t}{\beta_t(1 - \bar{\alpha}_{t-1})}, \quad (24)$$

using $\alpha_t - \bar{\alpha}_t + 1 - \alpha_t = 1 - \bar{\alpha}_t$. Hence

$$\boxed{\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t} \quad (25)$$

Mean. From the linear term in x_{t-1} ,

$$\tilde{\mu}_t = \tilde{\beta}_t \left(\frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0 \right) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} x_0. \quad (26)$$

The mean is a weighted combination of x_t and x_0 .

7 The Gaussian KL reduces to a squared distance

Choose $p_\theta(x_{t-1} | x_t) = \mathcal{N}(\mu_\theta(x_t, t), \sigma_t^2 \mathbf{I})$ with fixed σ_t^2 (the paper uses $\sigma_t^2 = \beta_t$ or $\tilde{\beta}_t$). For isotropic Gaussians in d dimensions,

$$D_{\text{KL}}(\mathcal{N}(\mu_1, s_1^2 \mathbf{I}) \| \mathcal{N}(\mu_2, s_2^2 \mathbf{I})) = \frac{d}{2} \left(\frac{s_1^2}{s_2^2} - 1 - \log \frac{s_1^2}{s_2^2} \right) + \frac{\|\mu_1 - \mu_2\|^2}{2s_2^2}. \quad (27)$$

Both variances are fixed, so only the mean term depends on θ :

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] + C. \quad (28)$$

8 Reparameterizing in terms of noise

From (7), $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$, so

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (x_t - \sqrt{1 - \bar{\alpha}_t} \varepsilon). \quad (29)$$

Substituting into $\tilde{\mu}_t$ and using $\sqrt{\bar{\alpha}_{t-1}}/\sqrt{\bar{\alpha}_t} = 1/\sqrt{\alpha_t}$:

$$\text{coefficient of } x_t : \quad \frac{\alpha_t(1 - \bar{\alpha}_{t-1}) + \beta_t}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)} = \frac{1}{\sqrt{\alpha_t}}, \quad (30)$$

$$\text{coefficient of } \varepsilon : \quad -\frac{\beta_t \sqrt{1 - \bar{\alpha}_t}}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)} = -\frac{\beta_t}{\sqrt{\alpha_t} \sqrt{1 - \bar{\alpha}_t}}. \quad (31)$$

Therefore

$$\boxed{\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon \right)} \quad (32)$$

The network sees x_t but not ε , so we give its mean the same form and let it predict the noise:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta(x_t, t) \right). \quad (33)$$

Subtracting, the x_t terms cancel:

$$\tilde{\mu}_t - \mu_\theta = \frac{\beta_t}{\sqrt{\alpha_t} \sqrt{1 - \bar{\alpha}_t}} (\varepsilon_\theta(x_t, t) - \varepsilon), \quad (34)$$

and so

$$L_{t-1} = \mathbb{E}_{x_0, \varepsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \left\| \varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon, t) \right\|^2 \right]. \quad (35)$$

9 Random time steps and the simplified loss

Summing over all T steps per sample is expensive. Since $\sum_{t=1}^T f(t) = T \mathbb{E}_{t \sim \mathcal{U}\{1, \dots, T\}}[f(t)]$, drawing one random t per sample gives an unbiased estimate of the full objective. DDPM additionally sets the time-dependent weight to 1, which often trains better:

$$\boxed{L_{\text{simple}}(\theta) = \mathbb{E}_{t, x_0, \varepsilon} \left[\left\| \varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon, t) \right\|^2 \right]} \quad (36)$$

10 Training and sampling

Training. Repeat until converged:

1. Sample $x_0 \sim p(x)$, $t \sim \mathcal{U}\{1, \dots, T\}$, $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$.
2. Form $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$.
3. Take a gradient step on $\nabla_\theta \left\| \varepsilon - \varepsilon_\theta(x_t, t) \right\|^2$.

Sampling. Start from $x_T \sim \mathcal{N}(0, \mathbf{I})$. For $t = T, \dots, 1$:

1. Draw $z \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else set $z = 0$.
2. Update

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_{\theta}(x_t, t) \right) + \sigma_t z. \quad (37)$$

Return x_0 . Each update is the posterior mean from the previous section plus a small amount of fresh noise; it is typically run for $T = 1000$ steps.

References

- [1] J. Ho, A. Jain, P. Abbeel. *Denoising Diffusion Probabilistic Models*. NeurIPS, 2020.
- [2] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. ICML, 2015.