Diffusion Model

Pin-Jing, Li

ouo.ee11@nycu.edu.tw
National Yang Ming Chiao Tung University

September 22, 2025

Reference (1)

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning*, 2015, pp. 2256–2265.

[2] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[3] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

Reference (2)

[4] C. K. Sonderby, T. Raiko, L. Maaloe, S. K. Sonderby, and O. Winther, "Ladder variational autoencoders," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015.

[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," in *International Conference on Learning Representations (ICLR)*, 2021.

Goal of Generative Models

Given training samples $x_0 \sim p_{\text{data}}(x)$, the objective is to train a model such that

$$p_\theta(x_0) \;\approx\; p_{\text{data}}(x),$$

where

- p_{data}: the true (unknown) data distribution,
- p_θ: the learned distribution parameterized by the model.

Intuition: learn to generate new samples $\hat{x}_0 \sim p_\theta$ that are indistinguishable from real data.

Variational Autoencoders (1)

VAE is one of the canonical examples of a generative model. In here our task is,

- for a set of observed data $x \sim p_X$
- VAE defines a parameterized model $p_\theta(x)$ that aims to match the true distribution $p_X(x)$

Directly parameterizing $p_\theta(x)$ in high-dimentional space is not computationally-friendly.

Variational Autoencoders (2)

To get access to the parametric, learned distribution $p_\theta(x)$, we introduce the latent space. Supposed there is a prior of the latent variable with lower dimensionality $z \sim p(z)$, we can now rewrite the learned distribution as if it's induced by marginalizing over the latent space,

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz$$

where

- z the latent variable
- $p(z)$ the prior distribution of the latent variable.
- $p_\theta(x|z)$ the parameterized decoder that decodes the latent space variable back into the data.

We can now generate x by sampling z and then decoding with the decoder $p_\theta(x|z)$.
But how to train the decoder model $p_\theta(x|z)$?

Variational Autoencoders (3)

An idea is to use the posterior $p_\theta(z|x)$ to generate the latent variable given a set of data x. to obtain this, we need

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)}$$

but this return us back to the problem that we don't have the direct access to $p_\theta(x)$ and $p_\theta(x|z)$, and this term is intractable. To get around this, we introduce the encoder $q_\phi(z|x)$ that aims to parameterize the true posterior $p_\theta(z|x)$.

Variational Autoencoders (4)

Now, to train the parameterized models, we aim to maximize the log likelihood of the distribution:

$$\log p_\theta(\mathbf{x})$$

In the following we show how this term is equivalent to

$$\log p_\theta(x) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))$$

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log p_\theta(\mathbf{x}) \int q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} \qquad \left[\int q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} = 1\right] && (1)\\
&= \int \log p_\theta(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} && (2)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] && (3)
\end{aligned}$$

Variational Autoencoders (5)

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \qquad \text{[by conditional probability]} \tag{4}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \qquad \left[\text{multiply by } 1 = \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \tag{5}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\right] \tag{6}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \tag{7}$$

Variational Autoencoders (6)

Since D_{KL} is a non-negative metric, we now have the term as the lower bound of the evidance (ELBO)

$$\text{ELBO} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right]$$

By maximizing ELBO, we are minimizing the KL divergence between the parameterized encoder $q_\phi(z|x)$ and the true, intractable encoder $p_\theta(z|x)$, and also maximizing the log likelihood of the parameterized distribution $p_\theta(x)$

Variational Autoencoders (7)

If we further expand the ELBO

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] \quad \text{[conditional prob.]} \tag{8}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] \tag{9}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})) \tag{10}$$

Variational Autoencoders (8)

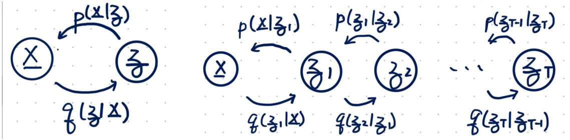
$$\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})) \tag{11}$$

From the expansion on the ELBO, we can now see that, by maximizing ELBO, we are

- **Reconstruction:** maximizing $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]$
  the likelihood of data given latent variable (the decoder).
- **Regularization:** minimizing $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}))$
  push the parametric prior to the true prior (the encoder).

Hierarchical Variational Autoencoders (1)

the latent variable of VAE is extended to multiple hierarchies T to form a HVAE, where each level of the latent variables are modeled by a higher level latents.



(a) VAE (b) HVAE

Hierarchical Variational Autoencoders (2)

We usually assume the transition along hierarchy is Markovian, then the joint distribution of data $\mathbf{x}$ and all latents $\mathbf{z}_{1:T}$ can therefore be written as

$$p_\theta(\mathbf{x}, \mathbf{z}_{1:T}) = p(\mathbf{z}_T)p_\theta(\mathbf{x}|\mathbf{z}_1)\prod_{t=2}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t),$$

and the posterior is

$$q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}) = q_{\boldsymbol{\phi}}(\mathbf{z}_1|\mathbf{x})\prod_{t=2}^{T} q_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{z}_{t-1})$$

Hierarchical Variational Autoencoders (3)

The ELBO just add the KL divergence terms for the latent trajectories in HVAE,

$$
\begin{aligned}
\log p_\theta(\mathbf{x}_0) &\geq \text{ELBO} \\
&= \mathbb{E}_q\Big[\log p_\theta(\mathbf{x}_0 \mid \mathbf{z}_1)\Big] - \sum_{t=2}^{T} D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t)\big) \\
&\quad - D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_T \mid \mathbf{x}_0) \,\|\, p(\mathbf{z}_T)\big).
\end{aligned}
$$

we include the ELBO of the VAE here for an easy comparison.

$$
\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})) \tag{12}
$$

from HVAE to DDPM

Diffusion models extend to a hierarchy of latents $x_{1:T}$, with fixed forward process:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\right).$$

The hierarchical ELBO becomes:

$$\mathcal{L}_{\mathrm{DDPM}} = \mathbb{E}_q[\log p_\theta(x_0 \mid x_1)] - \sum_{t=2}^{T} D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)) \\ - D_{\mathrm{KL}}(q(x_T \mid x_0) \,\|\, p(x_T))$$

- Same variational structure as VAE, but encoder $q(x_t \mid x_{t-1})$ is **fixed Gaussian**.
- Training = denoising; learning only the reverse process.

from ELBO to MSE

from the above loss function of DDPM, we know that maximizing the ELBO is equivalent to minimizing the

$$D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big)$$

with some arithmetics the conditional distribution

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$$

can be rewritten as a Gaussian with

$$\sim \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\bar{\alpha}_{t-1}}}_{\mu_q(\mathbf{x}_t, \mathbf{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_{t-1}}\mathbf{I}}_{\mathbf{\Sigma}_q(t)=\sigma_q^2(t)\mathbf{I}})$$

Objective with Gaussian Assumption (1)

to fit the model better into predicting the noise distribution, we design our model $p_{\boldsymbol{\theta}}$ as a Gaussian. It is characterized by

- for $\boldsymbol{\Sigma}_p$, since variance in forward process q is given, we set $\boldsymbol{\Sigma}_p(t) = \boldsymbol{\Sigma}_q(t) = \sigma_q^2(t)\mathbf{I}$
- we parameterize the mean with $\boldsymbol{\mu}_p = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$

since now the forward q and the reverse $p_{\boldsymbol{\theta}}$ are assumed to be Gaussian, the KL divergence

$$D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t))$$

can be simplified with the equation

$$D_{\mathrm{KL}}(\mathcal{N}(\mathbf{x}_1; \boldsymbol{\mu}_{x_1}, \boldsymbol{\Sigma}_{x_1})||\mathcal{N}(\mathbf{x}_2; \boldsymbol{\mu}_{x_2}, \boldsymbol{\Sigma}_{x_2}))$$

$$= \frac{1}{2}\left[\log \frac{|\boldsymbol{\Sigma}_{x_2}|}{|\boldsymbol{\Sigma}_{x_1}|} - \dim(\mathbf{x}) + \mathrm{tr}(\boldsymbol{\Sigma}_{x_2}^{-1}\boldsymbol{\Sigma}_{x_1}) + (\boldsymbol{\mu}_{\mathbf{x}_2} - \boldsymbol{\mu}_{\mathbf{x}_1})^T \boldsymbol{\Sigma}_{x_2}^{-1}(\boldsymbol{\mu}_{\mathbf{x}_2} - \boldsymbol{\mu}_{\mathbf{x}_1})\right]$$

Objective with Gaussian Assumption (2)

with again, some tedious arithmetics, minimizing the KL divergence can be shown to be equal to minimizing the MSE of the estimated mean.

$$\begin{aligned}&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t))\\ =&\arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[||(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)||^2\right]\end{aligned}$$

We further show in the following that this can be rewritten to be equivalent to minimizing the MSE of the reconstructed noise or samples

$\mathbf{x}_t$ reparameterization (1)

From the Gaussian assumption on the forward process we can write $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_{t-1})$ the latent with

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}$$

recursively substitute each step of latent

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}
\end{aligned}$$

Here since $\boldsymbol{\epsilon}$ are sampled from a standard Gaussian Distribution, we treat

$$\sqrt{\alpha_t-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} \sim \mathcal{N}(\mathbf{0}, (\alpha_t-\alpha_{t-1})\mathbf{I})$$

and

$\mathbf{x}_t$ reparameterization (2)

$$\sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \sim \mathcal{N}(\mathbf{0}, (1-\alpha_t)\mathbf{I})$$

therefore the sum of two independent gaussian can be treated as

$$(\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}) \sim \mathcal{N}(\mathbf{0}, ((\alpha_t - \alpha_t\alpha_{t-1}) + (1-\alpha_t))\mathbf{I})$$

thus we rewrite the sum as

$$\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} = \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}$$

with $\boldsymbol{\epsilon}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\mathbf{x}_t$ reparameterization (3)

therefore we can now keep expand $\mathbf{x}_t$ as

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \boldsymbol{\epsilon}^*_{t-2} \\
&= \sqrt{\prod_{i=1}^{t} \alpha_i \mathbf{x}_0} + \sqrt{1 - \prod_{i=1}^{t} \alpha_i \boldsymbol{\epsilon}^*_0} \\
&= \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}^*_0 \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, 1 - \bar{\alpha}_t \mathbf{I})
\end{aligned}$$

so we now have

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

that is, with $\boldsymbol{\epsilon}$ from a standard Gaussian $\mathcal{N}(0, \mathbf{I})$,

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$$

further MSE objectives (1)

we know from derivation,

$$\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\bar{\alpha}_{t-1}}$$

since $\mathbf{x}_t$ and all the α are known at the tth timestep, if we set our model as

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)}{1-\bar{\alpha}_{t-1}}$$

further MSE objectives (2)

our objective can be further simplified with

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[||(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)||_2^2 \right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \Bigg[|| \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)}{1-\bar{\alpha}_{t-1}} \\
&\qquad - \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\bar{\alpha}_{t-1}} ||_2^2 \Bigg] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[|| \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)}{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\bar{\alpha}_{t-1}} ||_2^2 \right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_{t-1}} \left[||\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{x}_0||_2^2 \right]
\end{aligned}
$$

further MSE objectives (3)

another interpretation is by exploiting the reparameterization done earlier again,

$$\mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

then the mean $\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0)$ can be further expanded by

$$\begin{aligned}
\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) &= \frac{1}{1-\bar{\alpha}_t}[\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0] \\
&= \frac{1}{1-\bar{\alpha}_t}[\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}] \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t}{1-\bar{\alpha}_t} + \frac{(1-\alpha_{t-1})\mathbf{x}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} + \frac{(1-\alpha_{t-1}\sqrt{1-\bar{\alpha}_t})\boldsymbol{\epsilon}_0}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}
\end{aligned}$$

further MSE objectives (4)

and simplified to

$$\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0$$

thus we further parameterize our model as

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)$$

further MSE objectives (5)

the KL divergence of the trajectories now become

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)||p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t))||\mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_q(t))) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}[||\frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t,t) \\
&\qquad - \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0||_2^2] \\
&\qquad = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\alpha_t)\alpha_t}[||\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t,t)||_2^2]
\end{aligned}
$$

Diffusion Process in an SDE perspective

the Diffusion process and the Stochastical differential equation has an intuitive perspective given by physics.

$$dX_t = \underbrace{f(X_t, t)}_{\text{drift}}\, dt + \underbrace{g(t)}_{\text{diffusion}}\, dW_t, \qquad t \in [0,1], \quad X_0 \sim p_{\text{data}}.$$

think of x_t as the state of a particle.

- the drift term is the deterministic force field. (i.e. gravity, friction) taking only this term give us the classical machenics in an ODE form $dx = f(x,t)dt$
- the diffusion term describe the thermal noise acting on the particle.
- dW_t is the Wiener Process (the Brownian noise) following $dW_t \coloneqq W_{t+dt} - W_t \sim \mathcal{N}(0, dt)$

if we can try to formulate the corresponding reverse process as an SDE, we will be able to reconstruct X_0 from a X_t

from Foward to Reverse SDE (1)

We start from the Forward SDE

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t.$$

If the marginal of X_t (the distribution of X at time t) is known

$$q_t(x) = \Pr\{X_t = x\},$$

then the Fokker–Planck equation gives the PDE for the marginal:

$$\partial_t q_t(x) = -\nabla\cdot\big(f(x,t)\,q_t(x)\big) + \tfrac{1}{2}g^2(t)\,\Delta q_t(x).$$

from Foward to Reverse SDE (2)

Reverse process: Let $Y_\tau = X_{T-\tau}$ with marginal $q_{T-\tau}(x)$. Then

$$\partial_\tau q_{T-\tau}(x) = -\partial_t q_t(x)\Big|_{t=T-\tau}.$$

Fokker–Planck becomes:

$$\partial_\tau q_{T-\tau}(x) = \nabla\cdot\big(f(x,t)\, q_{T-\tau}(x)\big) - \tfrac{1}{2} g^2(t)\, \Delta q_{T-\tau}(x).$$

Suppose a reverse SDE exists with this given marginal,

$$dX_t = f_{\mathrm{rev}}(X_t,t)\, dt + g(t)\, d\widetilde{W}_t,$$

according to Fokker-Planck again, the $q_{T-\tau}(x)$ satisfies

$$\partial_\tau q_{T-\tau}(x) = -\nabla\cdot\big(f_{\mathrm{rev}}(x,t)\, q_{T-\tau}(x)\big) + \tfrac{1}{2} g^2(t)\, \Delta q_{T-\tau}(x).$$

Thus, solving for $f_{\mathrm{rev}}(x,t)$ gives the drift of the reverse process.

Reverse SDE (Anderson, 1982) (1)

From equating the Fokker–Planck equations, we obtain

$$\nabla\cdot\big(f(x,t)q_t(x)\big) - \tfrac{1}{2}g^2(t)\Delta q_t(x) = -\nabla\cdot\big(f_{\text{rev}}(x,t)q_t(x)\big) + \tfrac{1}{2}g^2(t)\Delta q_t(x).$$

Therefore,

$$\nabla\cdot\big(f_{\text{rev}}(x,t)q_t(x)\big) = \nabla\cdot\big(f(x,t)q_t(x)\big) - g^2(t)\,\Delta q_t(x).$$

using the identity

$$\Delta q_t(x) = \nabla\cdot\nabla q_t(x) = \nabla\cdot\big(q_t(x)\nabla\log q_t(x)\big).$$

We can see that if we let the reverse drift function to be

$$f_{\text{rev}}(x,t) = f(x,t) - g^2(t)\,\nabla_x \log q_t(x).$$

then the divergence would be matching.

Reverse SDE (Anderson, 1982) (2)

We conclude our **Reverse-time SDE** to be

$$dX_t = \Big(f(X_t, t) - g^2(t)\,\nabla_x \log q_t(X_t)\Big)\,dt + g(t)\,d\widetilde{W}_t.$$

- time is now running reversely as $t = T - \tau$
- Drift contains the original $f(x,t)$ plus a correction term.
- The correction involves the **score** $\nabla_x \log q_t(x)$.
- This is the mathematical foundation for score-based generative models.

Score function

As we just derived the Reverse-time SDE

$$dX_t = \big(f(X_t, t) - g(t)^2 \nabla_x \log q_t(X_t)\big)\, dt + g(t)\, d\widetilde{W}_t.$$

has an unknown score term $\nabla_x \log q_t(x)$. We parameterize $s_\theta(x, t)$ to learn $\nabla_x \log q_t(x)$ to get rid of solving it explicitly. [1]

This is the idea of the score-based generative model.

[1] [5]

DDPM as a discretized SDE (1)

We want to express a discrete Markov chain (DDPM) as a discretized forward SDE.

$$dX_t = f(X_t, t)\, dt + g(t)\, dW_t.$$

Starts from our DDPM of one timestep,

$$x_i = \sqrt{1-\beta_i}\, x_{i-1} + \sqrt{\beta_i}\, \epsilon_{i-1}, \qquad \epsilon_{i-1} \sim \mathcal{N}(0, I).$$

Let $\tilde{\beta}_i = N \cdot \beta_i$ and rewrite:

$$x_i = \sqrt{1-\frac{\tilde{\beta}_i}{N}}\, x_{i-1} + \sqrt{\frac{\tilde{\beta}_i}{N}}\, \epsilon_{i-1}.$$

DDPM as a discretized SDE (2)

In continuous time we consider

$$\tilde{\beta}_i \to \tilde{\beta}(t = \frac{i}{N})$$

$$\epsilon_i \to z(t = \frac{i}{N})$$

$$x_t \to X(t = \frac{i}{N})$$

as $N \to \infty$, and by letting $t = \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}$, $\Delta t = \frac{1}{N}$

$$X(t + \Delta t) = \sqrt{1 - \tilde{\beta}(t)\Delta t}\, X(t) + \sqrt{\tilde{\beta}(t)\Delta t}\, z(t), \quad z(t) \sim \mathcal{N}(0, I).$$

DDPM as a discretized SDE (3)

using the approximation with small x,

$$\sqrt{1-x} \approx 1 - \tfrac{1}{2}x.$$

and for small δt we have $\tilde{\beta}(t+\delta t) = \tilde{\beta}(t)$, therefore we have,

$$X(t+\Delta t) - X(t) \;\approx\; -\tfrac{1}{2}\tilde{\beta}(t)\, X(t)\, \Delta t + \sqrt{\tilde{\beta}(t)}\, z(t)\sqrt{\Delta t}.$$

Taking the limit,

$$dX_t = -\tfrac{1}{2}\beta(t)\, X_t\, dt + \sqrt{\beta(t)}\, dW_t.$$

This is exactly an SDE with affine drift $f(X_t, t) = -\tfrac{1}{2}\beta(t)\, X_t$ and diffusion $g(t) = \sqrt{\beta(t)}$ coefficient.

training of DDPM in an SDE perspective

the training process of DDPM can be seen as trying to solve a reverse problem,
Forward (known): corrupt $x_0 \sim p_{\text{data}}$ into noisy x_t.

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

Inverse (learned): undo corruption.

$$p_\theta(x_{t-1} \mid x_t) \ \approx \ q(x_{t-1} \mid x_t, x_0)$$

The following shows how the training process of DDPM can be seen as a score-matching process.

From noise MSE to score matching (1)

Starting from our noise MSE objective of DDPM,

$$\arg\min_\theta D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t))$$
$$= \arg\min_\theta \; \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \|\epsilon_0 - \hat{\epsilon}_\theta(x_t, t)\|^2.$$

Score of $q(x_t \mid x_0)$ is the gradient of the log likelihood

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon_0, \quad q(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t) I).$$

$$\nabla_{x_t} \log q(x_t \mid x_0) = -\tfrac{1}{\sqrt{1-\bar{\alpha}_t}} \epsilon_0.$$

therefore the model predicting the noise can also be seen as a score predictor

$$s_\theta(x_t, t) := -\tfrac{1}{\sqrt{1-\bar{\alpha}_t}}\, \hat{\epsilon}_\theta(x_t, t).$$

From noise MSE to score matching (2)

MSE can therefore be interpreted as score-matching

$$\|\epsilon_0 - \hat{\epsilon}_\theta(x_t, t)\|^2 = (1 - \bar{\alpha}_t)\,\|\nabla_{x_t} \log q(x_t \mid x_0) - s_\theta(x_t, t)\|^2.$$

Thus DDPM noise MSE $\Leftrightarrow$ *denoising score matching.*

Inverse Problem as reverse SDE (1)

Generally speaking, an inverse problem is trying to recover X from the output of a corruption model

$$Y = \mathcal{A}(X) + \sigma_y Z, \qquad Z \sim \mathcal{N}(0, I).$$

- X: clean signal (unknown).
- $\mathcal{A}$: forward operator (blur, subsampling, general map).
- $\sigma_y Z$: Gaussian measurement noise.

Inverse Problem as reverse SDE (2)

if we describe the corruption in an SDE perspective,

$$\underbrace{X_0}_{\text{clean measurement}} := \mathcal{A}(X), \qquad \boxed{dX_t = \sigma_y \, dW_t, \quad t \in [0,1]}$$

- Start at the noiseless observation $X_0 = \mathcal{A}(X)$.
- Evolve with Brownian diffusion
- Each tiny step adds a small Gaussian jitter accumulating over time.

then the Endpoint at unit time = original corruption

$$X_1 = \mathcal{A}(X) + \sigma_y W_1 \overset{d}{=} \mathcal{A}(X) + \sigma_y Z \equiv Y.$$

Inverse Problem as reverse SDE (3)

- Gives a continuous dial from clean ($t = 0$) to fully corrupted ($t = 1$).
- Makes corruption Markovian. the "noise-injection" becomes a forward diffusion.
- solving Inversion = reverse diffusion: remove a little noise per step using the score $\nabla_x \log q_t(x)$.