A Review of Unify Diffusion Model Theory Based on EDM

Created By Zeng Zhiwen
Reference："Elucidating the Design Space of Diffusion-Based Generative Models"

前向：通用加噪公式与SDE的相互转化

从一个分布 $\mathbf{x_0}$ 到另一个分布 $\mathbf{x_t}$ 的桥梁，也即流（Flow）：

$$p(\boldsymbol{x_t}|\boldsymbol{x_0}) = \mathcal{N}(\boldsymbol{x_t}; s(t)\boldsymbol{x_0}, s^2(t)\sigma^2(t)\boldsymbol{I}) \tag{1}$$

将公式（1）进行重参数化采样，这里的 $\sigma(t)$为噪声强度的相对系数，$s(t)\sigma(t)$ 为绝对噪声强度（即噪声实际标准差）：

$$\mathbf{x_t} = s(t)\mathbf{x_0} + s(t)\sigma(t)\epsilon \tag{2}$$

- $s(t)\mathbf{x_0}$ **为信息部分:**
  信号功率 $= \mathbb{E}\left[\|s(t)\mathbf{x}_0\|^2\right] = s(t)^2 \cdot \mathbb{E}[\|\mathbf{x}_0\|^2] = s(t)^2 \cdot \alpha \cdot n$
- $s(t)\sigma(t)\epsilon$**为噪声部分:**
  噪声功率 $= \mathbb{E}\left[\|s(t)\sigma(t)\boldsymbol{\epsilon}\|^2\right] = s(t)^2\sigma(t)^2 \cdot \mathbb{E}[\|\boldsymbol{\epsilon}\|^2] = s(t)^2\sigma(t)^2 \cdot n$
- **信噪比:**
  $SNR(t) = \frac{\alpha}{\sigma^2(t)}$

$$\hat{\mathbf{x}}_\mathbf{t} = \frac{\mathbf{x_t}}{s(t)} = \mathbf{x_0} + \sigma(t)\epsilon$$

最终单步递推公式（见 #证明1 ）：

$$\mathbf{x}_t = \frac{s(t)}{s(t-1)}\mathbf{x}_{t-1} + s(t)\sqrt{\sigma(t)^2 - \sigma(t-1)^2}\epsilon_t$$

具体地，可以得到各个版本扩散模型加噪公式，例如 DDPM ：

$$\mathbf{x_t} = \sqrt{1-\beta_t}\mathbf{x_{t-1}} + \sqrt{\beta_t}\epsilon$$

其中，$s(t) = \sqrt{\bar{\alpha}_t}, \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s), \quad \sigma(t) = \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}$

对应的随机微分方程（SDE），形式如下：

$$d\mathbf{x_t} = f(t)\mathbf{x}_t dt + g(t)dw_t \tag{3}$$

反之，递推式可通过求取极限（见 #证明2 ）得到公式（3），对应系数如下：

$$s(t) = e^{\int_0^t f(r)dr} \tag{4}$$
$$f(t) = \frac{s'(t)}{s(t)}$$
$$\sigma^2(t) = \int_0^t \frac{g^2(r)}{s^2(r)} dr \tag{5}$$
$$g(t) = s(t)\sqrt{2 \cdot \sigma(t)\sigma'(t)}$$

公式（2）与公式（3）是等价的，公式（4）和（5）是沟通二者的桥梁，即起点分布相同条件下（即原图像 $\boldsymbol{x_0}$），针对特定加噪时间点 $\boldsymbol{t}$，公式（3）SDE的解 $\boldsymbol{x_t}$ 与公式（1）采样获得的 $\boldsymbol{x_t}$ 在分布上是一致的。特别地，EDM框架令 $s(t) = 1$, $\delta(t) = t$，保证了时间和噪声水平完全等价。

模型：通用模型框架

EDM概括了所有扩散模型中，神经网络部分的模型框架：

$$D_\theta(\hat{\mathbf{x}}; \sigma) = C_{skip}(\sigma)\hat{\mathbf{x}} + C_{out}(\sigma)F_\theta(C_{in}(\sigma)\hat{\mathbf{x}}; C_{nosie}(\sigma)) \tag{6}$$

- $D_\theta(\hat{\boldsymbol{x}}; \sigma)$ 是接收一个规范化的噪声图片（即原始图片直接添加 σ 水平的噪声，不进行尺度缩放），以及我们为其指定的噪声水平 σ ，输出降噪后的“纯净图像”，但是直接训练一个纯净网络效果不佳，因此 F_θ （残差）才是真正的网络组成。
  - $\hat{\mathbf{x}} = \frac{\mathbf{x}}{s(t)}$，用于统一所有模型的输入尺度
- C_{skip} , C_{out}： 纯净去噪网络 $D_\theta(\hat{\boldsymbol{x}}; \sigma)$ 的输出由噪声图片 $\hat{\boldsymbol{x}}$ 和模型 F_θ 输出加权组成。
- C_{in}： 用于适配不同网络对标准输入$\hat{\boldsymbol{x}}$ 的系数，如 $s(t)$。
- C_{noise}：EDM要求模型框架$D_\theta(\hat{\boldsymbol{x}}; \sigma)$ 的输入为σ ，但不同模型真正输入 F_θ的参数可能为σ 的函数，因此需要作变换。

VP

$$D_\theta(\hat{\mathbf{x}}; \sigma) = \underbrace{1}_{cskip} \cdot \hat{\mathbf{x}} \underbrace{- \sigma}_{cout} \cdot F_\theta\Big(\underbrace{\tfrac{1}{\sqrt{\sigma^2+1}}}_{cin} \cdot \hat{\mathbf{x}};\ \underbrace{(M-1)\ \sigma^{-1}(\sigma)}_{cnoise}\Big)$$

VE

$$D_\theta(\hat{\mathbf{x}}; \sigma) = \underbrace{1}_{cskip} \cdot \hat{\mathbf{x}} + \underbrace{\sigma}_{cout} \cdot F_\theta\Big(\underbrace{1}_{cin} \cdot \hat{\mathbf{x}};\ \underbrace{\log\left(\tfrac{1}{2}\sigma\right)}_{cnoise}\Big)$$

> 注意：这里的 `VE` $\hat{x} = x$

训练：通用训练框架

训练的过程，是针对从原始图像集合中采样得到的真实图像 $\mathbf{x}_0$ ，进行一次 $\boldsymbol{\sigma}$ 级别的噪声添加，得到 $\mathbf{x}_0 + \boldsymbol{n}$ ，随后可构造损失函数：

$$\mathcal{L}_{diff} = \mathbb{E}_{\sigma,n,\mathbf{x}_0}\left[\lambda(\sigma)||D_\theta(\hat{\mathbf{x}}; \sigma) - \mathbf{x_0}||_2^2\right]$$
$$\mathcal{L}_{diff} = \underbrace{\mathbb{E}_{\sigma,n,x_0}}_{p_{train}}\Big[\underbrace{\lambda(\sigma)\ C_{out}^2(\sigma)}_{\text{损失权重w}(\sigma)}\Big|\Big|\underbrace{F_\theta\big(C_{in}(\sigma)\cdot(\boldsymbol{x_0}+\boldsymbol{n}); C_{noise}(\sigma)\big)}_{\text{模型输出}} - \underbrace{\tfrac{1}{C_{out}(\sigma)}\big(\boldsymbol{x_0} - C_{skip}(\sigma)\cdot(\boldsymbol{x_0}+\boldsymbol{n})\big)}_{\text{训练目标}}\Big|\Big|_2^2\Big] \tag{7}$$

- $\sigma \sim P_{train}$ ，即前向噪声采样分布，由各个模型决定
- $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, $\mathbf{x}_0 \sim P_{data}$
- $Var(\mathbf{x}_0) = \sigma_{data}^2$
- C_{in}：保证神经网络的输入保持单位方差（式117）

$$c_{in}(\sigma)=\frac{1}{\sqrt{\sigma^2+\sigma_{data}^2}}$$

- C_{out},C_{skip}：保证训练目标保持方差恒为1，同时让C_{out}^2被最小化（式138、131）:

$$C_{skip}(\sigma)=\frac{\sigma_{data}^2}{\sigma_{data}^2+\sigma^2}$$
$$C_{out}(\sigma)=\frac{\sigma\cdot\sigma_{data}}{\sqrt{\sigma^2+\sigma_{data}^2}}$$

- $\lambda(\sigma)$：保证损失权重 $w(t)=1$ （式 144）：

$$\lambda=\frac{\sigma^2+\sigma_{data}^2}{(\sigma\cdot\sigma_{data})^2}$$

- 当初始化神经网络权重为0（即输出恒0），方差暂时固定某个初始值时，有:

$$\mathbb{E}(\mathcal{L})=1$$

- σ：损失函数在加噪水平很低或很高情况下，损失函数均难以下降，因此损失（时间步）的选择如下：

$$ln(\sigma)\sim\mathcal{N}(P_{mean},P_{std}^2)$$

  其中$P_{mean}=-1.2,P_{std}=1.2$
- $\sigma_{data}=0.5$

VP

$$\underbrace{\mathbb{E}_{\sigma^{-1}(\sigma)\sim\mathcal{U}(\epsilon_t,1)}}_{p_{train}}\mathbb{E}_{\mathbf{x_0},\mathbf{n}}\left[\underbrace{\frac{1}{\sigma^2}}_{\text{损失权重}}\left\|D_\theta\left(\mathbf{x_0}+\mathbf{n};\sigma\right)-\mathbf{x_0}\right\|_2^2\right]$$

VE

$$\underbrace{\mathbb{E}_{\ln(\sigma)\sim\mathcal{U}(\ln(\sigma_{min}),\ln(\sigma_{max}))}}_{p_{train}}\mathbb{E}_{\mathbf{x_0},\mathbf{n}}\left[\underbrace{\frac{1}{\sigma^2}}_{\text{损失权重}}\left\|D_\theta\left(\mathbf{x_0}+\mathbf{n};\sigma\right)-\mathbf{x_0}\right\|_2^2\right]$$

$\hat{\mathbf{x}}=\mathbf{x_0}+\mathbf{n}=\frac{\mathbf{x}}{s(t)}$

损失类型	数学形式	适用场景
残差损失	$\|D_\theta(\hat{\mathbf{x}})-\mathbf{x}_0\|^2$	直接去噪
噪声损失	$\|F_\theta(\mathbf{x})-\boldsymbol{\epsilon}\|^2$	DDPM 类模型
分数匹配	$\|N_\theta(\mathbf{x})-\nabla\log p(\mathbf{x})\|^2$	基于分数的生成模型

反向：通用推理过程

确定性过程

通用 概率流常微分方差 PFODE

🚩 通用反向：对于任意一个扩散模型加噪SDE（公式(3)），通过福克普朗克方程，可进一步推导出一个常微分方程（ODE），也叫概率流常微分方程（PFODE）：

$$d\boldsymbol{x_t}=\left[f(t)\boldsymbol{x_t}-\frac{1}{2}g^2(t)\bigtriangledown_{\boldsymbol{x_t}}logp_t(\boldsymbol{x_t})\right]dt \tag{8}$$

注意，这里的 $p_t(\mathbf{x}_t)$ 可以描述为：

$$
\begin{aligned}
p_t(x) &= \int_{\mathcal{R}^d} p_{0t}(x|x_0)p_{data}(x_0)dx_0 \\
&= s(t)^{-d}[p_{data} * \mathcal{N}(0, \sigma^2(t)\mathbf{I})](\mathbf{x}/s(t)) \\
&= s(t)^{-d}p(\mathbf{x}/s(t); \sigma(t))
\end{aligned}
$$

- * 表示卷积操作
- $\mathbf{x}/s(t)$ 表示分布在此处的取值

$$
\begin{aligned}
\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) &= \nabla_{\mathbf{x}} \log s(t)^{-d} + \nabla_{\mathbf{x}} \log[p_t(\frac{\mathbf{x}_t}{s(t)}; \sigma(t))] = \nabla_{\mathbf{x}} \log[p_t(\frac{\mathbf{x}_t}{s(t)}; \sigma(t))] \\
&= \frac{1}{s(t)} \nabla_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}}; \sigma(t)) = \frac{1}{s(t)\sigma^2(t)} (D_\theta(\hat{\mathbf{x}}; \sigma(t)) - \hat{\mathbf{x}})
\end{aligned} \tag{10}
$$

> 分数函数的方向由当前噪声图谱 $\hat{\mathbf{x}}$ 指向神经网络预测的真实的分布 $\mathcal{D}_\theta(\hat{\mathbf{x}}; \sigma(t))$

🚩通用反向：因此，在确定起点 $\mathbf{x_0}$ （前向）或 $\mathbf{x_N}$ （逆向）前提下，式（8）解的分布 $p(\mathbf{x_t})$，即$\mathbf{x_t}$的边缘概率密度与加噪过程SDE求解得到的分布是完全相同：

$$
d\boldsymbol{x}_t = \left[\left(\frac{\dot{\sigma}(t)}{\sigma(t)} + \frac{\dot{s}(t)}{s(t)} \right) \boldsymbol{x}_t - \frac{\dot{\sigma}(t)s(t)}{\sigma(t)} D_\theta(\frac{\boldsymbol{x}_t}{s(t)}; \sigma(t)) \right] dt \tag{11}
$$

> 说明：针对PFODE，当dt取反时，便可实现前向加噪和后向加噪的切换，因此式（8）、式（9）、式（11）都可以兼顾前向和反向的描述，但**不可用PFODE**实现图像加噪，因为PFODE在给定起点时，其终点是确定的，于是变形成非分布的一对一输入输出样本匹配对。扩散模型建模的是真实分布与完全噪声分布之间的关系，必须通过随机采样配对实现，因此不能用PFODE实现图像加噪，而是仅能用于反向采样去噪

通用确定性采样

公式（11）的 $\mathcal{D}_\theta$ 可用神经网络模拟，具体为公式（6），随后通过使用ODE求解器，如一阶Euler，二阶Heun，在给定起点 $\boldsymbol{X_N}$ 下，逐步采样获得生成图像。**注意：训练过程的时间步和采样过程的时间步定义不同**，EDM采样过程的噪声水平定义为：

$$
\sigma_{i<N} = (\sigma_{max}{}^{\frac{1}{\rho}} + \frac{i}{N-1}(\sigma_{min}{}^{\frac{1}{\rho}} - \sigma_{max}{}^{\frac{1}{\rho}}))^\rho \; and \; \sigma_N = 0 \tag{12}
$$

Algorithm 1 Deterministic sampling using Heun's 2nd order method with arbitrary $\sigma(t)$ and $s(t)$.

procedure HEUNSAMPLER($D_\theta(\boldsymbol{x}; \sigma)$, $\sigma(t)$, $s(t)$, $t_{i\in\{0,\ldots,N\}}$)
  sample $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \sigma^2(t_0)\, s^2(t_0)\, \mathbf{I})$ ▷ Generate initial sample at t_0
  for $i \in \{0, \ldots, N-1\}$ **do** ▷ Solve Eq. 4 over N time steps
    当前图像的噪声水平
    $\boldsymbol{d}_i \leftarrow \left(\frac{\dot{\sigma}(t_i)}{\sigma(t_i)} + \frac{\dot{s}(t_i)}{s(t_i)} \right) \boldsymbol{x}_i - \frac{\dot{\sigma}(t_i)s(t_i)}{\sigma(t_i)} D_\theta\left(\frac{\boldsymbol{x}_i}{s(t_i)}; \sigma(t_i) \right)$ ▷ Evaluate $d\boldsymbol{x}/dt$ at t_i
    $\boxed{\boldsymbol{x}_{i+1}} \leftarrow \boldsymbol{x}_i + (t_{i+1} - t_i)\boldsymbol{d}_i$ ▷ Take Euler step from t_i to t_{i+1}
    if $\sigma(t_{i+1}) \neq 0$ **then** 取反 $x_i \rightarrow D_\theta$ ▷ Apply 2nd order correction unless σ goes to zero
      $\boldsymbol{d}'_i \leftarrow \left(\frac{\dot{\sigma}(t_{i+1})}{\sigma(t_{i+1})} + \frac{\dot{s}(t_{i+1})}{s(t_{i+1})} \right) \boxed{\boldsymbol{x}_{i+1}} - \frac{\dot{\sigma}(t_{i+1})s(t_{i+1})}{\sigma(t_{i+1})} D_\theta\left(\frac{\boldsymbol{x}_{i+1}}{s(t_{i+1})}; \sigma(t_{i+1}) \right)$ ▷ Eval. $d\boldsymbol{x}/dt$ at t_{i+1}
      $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + (t_{i+1} - t_i)\left(\frac{1}{2}\boldsymbol{d}_i + \frac{1}{2}\boldsymbol{d}'_i \right)$ ▷ Explicit trapezoidal rule at t_{i+1}
  return $\boldsymbol{x}_N$ ▷ Return noise-free sample at t_N

通用 随机微分方程

🚩 通用：**逆向**随机形式 SDE 为：

$$
d\mathbf{x} = \left[\mathbf{f}(t)\mathbf{x}_t - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt + g(t) d\bar{\mathbf{w}}, \tag{a}
$$

> SDE适用于前向和反向过程，因此这里特别分别给出了其在扩散模型中的前向、反向过程定义。而PFODE在扩散模型领域只适用于反向过程，见式（8）。
>
> 当 $g(t)=0$ 时，方程变为 $d\mathbf{x}=\mathbf{f}(\mathbf{x},t)dt$ ，此时不能称为 扩散模型的确定性采样过程，因为扩散项以及被消除了，不再是扩散模型的范畴。真正确定性逆向采样参见式（8）。
>
> 思考：既然 $-g(t)^2\bigtriangledown_{\mathbf{x}}\log p_t(\mathbf{x})$ 已经能够为逆向采样过程提供降噪方向正确的、降噪强度确定的保证了，为什么还要后面的随机项？保持生成多样性，避免坍缩到单一模式。

#猜想：变成随机后二分之一没了，但多了一个随机项

#注：下面可以忽略

🚩 一般通用（结合前向、逆向）：结合热方程偏微分方程和福克普朗克方程：

$$dx=\left(\frac{1}{2}\,g(t)^2-\dot{\sigma}(t)\,\sigma(t)\right)\,\bigtriangledown_x\log p\big(\boldsymbol{x};\sigma(t)\big)dt+g(t)dw_t \tag{b}$$

其中$g(t)$和$\sigma(t)$随便取值。

非通用：
注意，此处s(t) = 1，令g(t)=0，则转化为PFODE（式(9)）；$g(t)=\sqrt{2\beta(t)}\sigma(t)$ 则转化为：

$$d\boldsymbol{x}_{\pm}=\underbrace{-\dot{\sigma}(t)\sigma(t)\bigtriangledown_x\log p\big(\boldsymbol{x};\sigma(t)\big)dt}_{\text{PFODE}}\pm\underbrace{\underbrace{\beta(t)\sigma(t)^2\bigtriangledown_x\log p\big(\boldsymbol{x};\sigma(t)\big)\,dt}_{\text{deterministic noise decay}}+\underbrace{\sqrt{2\beta(t)}\sigma(t)dw_t}_{\text{noise injection}}}_{\text{Langevin diffusion SDE}} \tag{13}$$

正负分别表示前向SDE和**逆向SDE**过程，后者为随机性采样所使用的形式。

- PFODE: 这一部分的出现预示着基于逆向SDE的随机性采样过程也一定包含与确定性采样类似的过程
- deterministic noise decay：带入式（10）可得：
- $$\pm\beta(t)(\mathcal{D}_\theta(\boldsymbol{x};\sigma(t))-\boldsymbol{x})dt\approx\pm\beta(t)(\boldsymbol{y}-\boldsymbol{x})dt=\mp\beta(t)\boldsymbol{n}dt \tag{14}$$

  说明此部分为确定性噪声衰减项，其值与该时间步所提供的噪声水平成正比。
- noise injection：转化为:
- $$\sqrt{2\beta(t)}\sigma(t)\epsilon\sqrt{dt}=\sqrt{2\beta(t)}\boldsymbol{n}'\sqrt{dt} \tag{15}$$

  反向过程中，式（14）与式（15）分别进行着相同噪声水平的去噪和加噪过程，$\beta(t)$控制二者相对速率。

非通用 随机性采样

随机性采样过程方法众多，甚至和逆向SDE公式本身“关系不大”。EDM论文也表示它设计的机性采样过程不是一种通用的SDE求解器，而是一种面向扩散模型问题的垂类SDE求解器。EDM设计的随机性采样过程非常简单，其核心就是在确定性采样的基础上增加了 **“回退”** 操作，也即先对样本额外加噪，再采用ODE求解器采样获得下一个时间点的图像。这种回退操作可以有效修正前面迭代步骤产生的误差，所以通常相比PFODE的生成效果更好，但同时也要花

费更多的采样步数。EDM提出的SDE采样器(求解器)基本算法流程如图所示:

最大化噪声水平（完全高斯随机噪声）

告诉模型每一轮降噪的力度

Algorithm 2 Our stochastic sampler with $\sigma(t) = t$ and $s(t) = 1$.

procedure StochasticSampler$(D_\theta(\boldsymbol{x};\sigma), t_{i\in\{0,\dots,N\}}, \gamma_{i\in\{0,\dots,N-1\}}, S_{\text{noise}})$

  sample $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0},\ t_0^2\ \mathbf{I})$

  for $i \in \{0,\dots,N-1\}$ **do** ▷ $\gamma_i = \begin{cases} \min\left(\frac{S_{\text{churn}}}{N}, \sqrt{2}-1\right) & \text{if } t_i \in [S_{\text{tmin}}, S_{\text{tmax}}] \\ 0 & \text{otherwise} \end{cases}$

    sample $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0},\ S_{\text{noise}}^2\ \mathbf{I})$

    $\hat{t}_i \leftarrow t_i + \gamma_i t_i$ ▷ Select temporarily increased noise level $\hat{t}_i$

    $\hat{\boldsymbol{x}}_i \leftarrow \boldsymbol{x}_i + \boxed{\sqrt{\hat{t}_i^2 - t_i^2}\ \boldsymbol{\epsilon}_i}$ ▷ Add new noise to move from t_i to $\hat{t}_i$

    $\boldsymbol{d}_i \leftarrow (\hat{\boldsymbol{x}}_i - D_\theta(\hat{\boldsymbol{x}}_i; \hat{t}_i))/\hat{t}_i$ ▷ Evaluate $\mathrm{d}\boldsymbol{x}/\mathrm{d}t$ at $\hat{t}_i$

    $\boldsymbol{x}_{i+1} \leftarrow \hat{\boldsymbol{x}}_i + \underbrace{(t_{i+1} - \hat{t}_i)}_{<0}\boldsymbol{d}_i$ ▷ Take Euler step from $\hat{t}_i$ to t_{i+1}

    if $t_{i+1} \neq 0$ **then**

      $\boldsymbol{d}'_i \leftarrow (\boldsymbol{x}_{i+1} - D_\theta(\boldsymbol{x}_{i+1}; t_{i+1}))/t_{i+1}$ ▷ Apply 2nd order correction

      $\boldsymbol{x}_{i+1} \leftarrow \hat{\boldsymbol{x}}_i + (t_{i+1} - \hat{t}_i)(\tfrac{1}{2}\boldsymbol{d}_i + \tfrac{1}{2}\boldsymbol{d}'_i)$

  return $\boldsymbol{x}_N$

N轮降噪，每一轮对应一种降噪的力度

扩散模型的采样过程本质是在噪声空间和数据空间之间插值。若新增噪声不足（Snoise=1.0），去噪器可能因噪声过强而丢失低频信息，可能过早“确信”某些区域已干净，导致信息丢失。

加噪回退，在不考虑S_noise情况下，所添加的噪声水平不超过当前图像噪声（当前指定模型降噪的力度）

令 $\varepsilon_i \sim N(0, I)$ max $\boxed{\sqrt{2}-1}$

$\hat{t}_i \leftarrow t_i + \gamma_i t_i$

$\hat{t}_i = t_i + (\sqrt{2}-1)t_i = \sqrt{2}t_i$

$\hat{x}_i = x_i + \sqrt{2t_i^2 - t_i^2}\ \varepsilon_i$

$= x_i + t_i\varepsilon_i$ 加噪不能大于 x_i 原有噪声

$N(x, t_i^2 I)$ $N(0, t_i I)$

其间涉及多个超参数，均为实验性、经验性取值。

VP (DDPM / DDIM)

VP(Variance Perserving)，噪声调度满足 **信号与噪声的方差总和恒定，前向**公式：

$$
\begin{aligned}
\mathbf{x}_t &= \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon \\
\mathbf{x}_t &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \\
\hat{\mathbf{x}}_\mathbf{t} &= \frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} = \mathbf{x}_\mathbf{0} + \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon \\
d\mathbf{x} &= -\frac{1}{2}\beta(t)\mathbf{x}dt + \sqrt{\beta(t)}d\mathbf{w}
\end{aligned} \tag{16}
$$

对离散递推式取极限可以直接导出连续式（16)。
VP一步加噪式中的 $\hat{\sigma} = \sqrt{1-\bar{\alpha}_t}$ 不是绝对噪声方差，下面的 $\sigma(t)$ 才是：

$$
\begin{aligned}
\bar{\sigma}(t) &= \sqrt{1-\bar{\alpha}_t} \\
s(t) &= \sqrt{\bar{\alpha}_t} \\
\sigma(t) &= \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}
\end{aligned}
$$

满足

$$
s(t) = \frac{1}{\sqrt{\sigma^2(t)+1}}
$$

上面的符号与式（2）对应。
漂移项$f(t)$:

$$
f(t) = -\frac{1}{2}\beta(t)
$$

扩散项 $g(t)$:

$$
g(t) = \sqrt{\beta(t)} \tag{18}
$$

式（17)、式（18）带入式（b）得到**VP逆向SDE**过程:

$$
\begin{aligned}
d\mathbf{x} &= \frac{1}{2}\beta(t)\sigma(t)^2\nabla_\mathbf{x}\log p(\mathbf{x})dt + \sqrt{\beta(t)}d\mathbf{w} \\
&= \left[-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_\mathbf{x}\log p(\mathbf{x})\right]dt + \sqrt{\beta(t)}d\mathbf{w}
\end{aligned} \tag{19}
$$

VE (SMLD)

VE(Variance Exploding)，VE过程的噪声调度允许 **噪声方差无限增长，前向**公式：

$$
\begin{aligned}
\mathbf{x}_t &= \mathbf{x}_{t-1} + \sqrt{\sigma^2(t) - \sigma^2(t-1)}\epsilon \\
\mathbf{x}_t &= \mathbf{x_0} + \sigma(t)\epsilon \\
d\mathbf{x} &= \sqrt{\frac{d\bar{\sigma}^2(t)}{dt}} d\mathbf{w}
\end{aligned} \tag{20}
$$

其中，$\bar{\sigma}(t) = s(t) \cdot \sigma(t) = \sigma(t)$

扩散项：

$$
g(t) = \sqrt{\frac{d\bar{\sigma}^2(t)}{dt}} = \sqrt{\frac{d\sigma^2(t)}{dt}} \tag{22}
$$

漂移项:

$$
f(t) = 0
$$

带入式（a）得到与song等人定义一致的**VE逆向**过程：

$$
\begin{aligned}
d\mathbf{x} &= [-g(t)^2 \bigtriangledown_x \log p_t(\mathbf{x})]dt + g(t)d\mathbf{w} \\
&= [-\frac{d\sigma^2(t)}{dt} \bigtriangledown_x \log p_t(\mathbf{x})]dt + \sqrt{\frac{d\sigma^2(t)}{dt}} d\mathbf{w}
\end{aligned} \tag{23}
$$

附录

【补充1】如何理解 VP 方差保持与 VE 方差爆炸

考虑**VP**递推：

$$
x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon
$$

由方差性质：

$$
Var(x_t) = \alpha_t Var(x_{t-1}) + (1-\alpha_t) \cdot 1
$$

由数学归纳法，假设 $Var(x_{t-1}) = 1$，有：

$$
Var(x_t) = \alpha_t + 1 - \alpha_t = 1
$$

因此方差是保持的。
而VE中，$Var(x_t) = 1 - \alpha_t = \beta_t$ ，由于噪声水平逐步增大，因此方差是爆炸式增大的。

【补充2】DDPM与VP的关系

DDPM是VP的离散化形式

模型	前向加噪公式
VP SDE	$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}dt + \sqrt{\beta(t)}d\mathbf{w}$
DDPM	$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon$ $\mathbf{x_t} = \sqrt{\bar{\alpha}_t}\mathbf{x_0} + \sqrt{1-\bar{\alpha}_t}\epsilon$

模型	反向采用公式
VP SDE	$d\mathbf{x} = [-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_x \log p_t(x)]dt + \sqrt{\beta(t)}d\bar{\mathbf{w}}$
DDPM	$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \beta_t \frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}}\right) + \sqrt{\beta_t}\epsilon_t$$ $$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t + \beta_t \frac{D_\theta(\frac{\mathbf{x}_t}{s(t)}; \sigma(t)) - \frac{\mathbf{x}_t}{s(t)}}{s(t)\sigma^2(t)}\right) + \sqrt{\beta_t}\epsilon_t$$

VP中的变量定义与转化

1. $\alpha_t = 1 - \beta_t$
2. $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$
3. $\bar{\sigma}(t) = s(t)\sigma(t) = \sqrt{1-\bar{\alpha}_t}$ ，即 `t` 时刻的图像噪声标准差，取值范围由$0 \to 1$ 。
4. $s(t) = \sqrt{\bar{\alpha}_t}$，与式（2）对应
5. $\sigma(t) = \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}$，与（2）对应

【其他】

1. EDM前向

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \sqrt{\sigma^2(t) - \sigma^2(t-1)}\epsilon$$
$$\mathbf{x}_t = \mathbf{x_0} + \sigma(t)\epsilon$$
$$d\mathbf{x} = \sqrt{\frac{d\bar{\sigma}^2(t)}{dt}}d\mathbf{w}$$

与VE一致，但噪声分布不同：

$$\ln(\sigma(t)) = \ln(t) \sim \mathcal{N}(P_{mean}, P_{std}^2)$$

2. Wiener过程

$w_t \sim \mathcal{N}(0, t)$ 是一个布朗运动（Wiener） 过程

- 独立增量性：$w_{t+\triangle t} - w_t$ 与 w_t 独立
- 布朗运动的增量服从正态分布：$w_{t+\triangle t} - w_t \sim \mathcal{N}(0, \triangle t)$
- 令 $\triangle_t \to 0, d_w \sim \mathcal{N}(0, d_t)$
  - 布朗运动无穷小增量的平方 $d_w^2 = d_t$ 为确定性量
  - 重参数化展开：$d_w = \sqrt{d_t} \cdot \epsilon, \epsilon \sim \mathcal{N}(0, 1)$

3. EDM论文相关

> Song et al. present a stochastic differential equation (SDE) that **maintains** the desired distributionp as sample x evolves over time

若一个SDE的解 $\mathbf{x}_t$ 的边际分布 p_t 满足：

$$\lim_{t\to\infty} p_t(\mathbf{x}) = p(\mathbf{x}) \quad 且 \quad 一旦达到p后，分布不再随时间变化$$

则称 p_t 是该SDE的**不变分布**（或稳态分布）。此时，SDE“保持”了分布 p_t。

- **正向过程**：
从数据分布 p_{data} 出发，通过SDE逐渐将数据破坏为噪声分布纯粹的高斯分布 ϵ 。

- **逆向过程**：
从噪声 ϵ 出发，通过SDE将样本演化回 p_{data}。

> To specify the ODE, we must first choose a schedule $\sigma(t)$ that defines the desired noise level at time t.

在PFODE中，$\sigma(t)$ **直接表示 t 时刻数据的噪声水平（累积结果）**，而非单步添加量。这样一来，在前向加噪训练时，针对某一时刻 t 噪声水平 $\sigma(t)$，直接向 $\mathbf{x_0}$ 添加 $\mathcal{N} \sim (0, \sigma^2(t))$ 的高斯随机噪声即可。在反向降噪采样时，也可以直接告诉神经网络当前图像的噪声水平 $\sigma(t)$，从而做出相应力度的降噪操作。

> The score function has the remarkable property that it does not depend on the generally intractable normalization constant of the underlying density function $p(\mathbf{x};\sigma)$

Score Function它不依赖于概率密度函数 p(x;σ) 的归一化常数（normalization constant）。
假设概率密度函数可以分解为：

$$p(x;\sigma) = \frac{1}{Z(\sigma)}\tilde{p}(x;\sigma)$$

其中：
- $\tilde{p}(x;\sigma)$ 是未归一化的概率密度（可能难以计算积分）。
- $Z(\sigma)$ 是归一化常数（通常难以计算，尤其是高维数据）。

取对数后：

$$\log p(x;\sigma) = \log \tilde{p}(x;\sigma) - \log Z(\sigma)$$

计算梯度时：

$$\nabla_x \log p(x;\sigma) = \nabla_x \log \tilde{p}(x;\sigma) - \underbrace{\nabla_x \log Z(\sigma)}_{=0}$$

由于 $Z(\sigma)$ 不依赖 x，其梯度为零，因此：

$$\nabla_x \log p(x;\sigma) = \nabla_x \log \tilde{p}(x;\sigma)$$

结论：Score 函数仅依赖于未归一化的 $\tilde{p}(x;\sigma)$，与 $Z(\sigma)$ 无关！

> Excessive Langevin-like addition and removal of noise results in gradual **loss of detail** in the generated images, There is also a drift toward oversaturated colors at very low and high noise levels. We suspect that practical denoisers induce a slightly nonconservative vector field.

引入随机性（SDE，朗之万噪声步骤）虽然能修正早期采样误差，但会导致**细节丢失**和在极端噪声水平下的颜色过饱和。
原因可能在于：

- Denoiser的过渡去噪移除了比理论值更多的噪声，破坏了朗之万扩散所需的保守向量场；
- $\mathcal{L}^2$ 损失使得模型倾向于预测均值，忽略极端边缘细节
  解决方案在于：
- 限制噪声添加的时机范围 $t_i \in [S_{t_{min}}, S_{t_{max}}]$
- 使得每次添加随机噪声的水平S_{noise} 略微大于1抵消细节损失
- 确保每次新增噪声的强度不超过当前图像的噪声水平，防止过度破坏结构

【证明1】前向离散一步加噪式（2）转化为离散单步递推式

离散一步加噪形式：

$$\mathbf{x_t} = s(t)\mathbf{x_0} + s(t)\sigma(t)\epsilon$$

假设：

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

1. **期望匹配**：

$$\mathbb{E}[\mathbf{x}_t] = s(t)\mathbf{x}_0 = \alpha_t \mathbb{E}[\mathbf{x}_{t-1}] = \alpha_t s(t-1)\mathbf{x}_0$$

因此：

$$\alpha_t = \frac{s(t)}{s(t-1)}$$

2. **方差匹配**：

$$\mathrm{Var}(\mathbf{x}_t) = s(t)^2\sigma(t)^2 = \alpha_t^2 \mathrm{Var}(\mathbf{x}_{t-1}) + \beta_t^2$$

代入 $\mathrm{Var}(\mathbf{x}_{t-1}) = s(t-1)^2\sigma(t-1)^2$：

$$\beta_t^2 = s(t)^2\sigma(t)^2 - \left(\frac{s(t)}{s(t-1)}\right)^2 s(t-1)^2\sigma(t-1)^2 = s(t)^2\left(\sigma(t)^2 - \sigma(t-1)^2\right)$$

因此：

$$\beta_t = s(t)\sqrt{\sigma(t)^2 - \sigma(t-1)^2}$$

3. **最终单步递推公式**：

$$\mathbf{x}_t = \frac{s(t)}{s(t-1)}\mathbf{x}_{t-1} + s(t)\sqrt{\sigma(t)^2 - \sigma(t-1)^2}\boldsymbol{\epsilon}_t \tag{24}$$

【证明2】前向加噪离散形式到连续形式的转化

原始递推式如下：

$$\mathbf{x}_t = \frac{s(t)}{s(t-1)}\mathbf{x}_{t-1} + s(t)\sqrt{\sigma(t)^2 - \sigma(t-1)^2}\boldsymbol{\epsilon}_t$$

两边减去 $\mathbf{x}_{t-1}$，并写为极限形式：

$$\mathbf{x}_t - \mathbf{x}_{t-\Delta t} = \left(\frac{s(t)}{s(t-\Delta t)} - 1\right)\mathbf{x}_{t-\Delta t} + s(t)\sqrt{\sigma(t)^2 - \sigma(t-\Delta t)^2}\boldsymbol{\epsilon}_t.$$

当时间步长 $\Delta t = t-(t-1) \to 0$ 时:

(1) 对 $s(t-\Delta t)$ 进行泰勒展开

$$s(t-\Delta t) \approx s(t) - s'(t)\Delta t + \mathcal{O}(\Delta t^2)$$

$$\frac{s(t)}{s(t-\Delta t)} \approx \frac{s(t)}{s(t) - s'(t)\Delta t} \approx 1 + \frac{s'(t)}{s(t)}\Delta t$$

$$\left(\frac{s(t)}{s(t-\Delta t)} - 1\right) \approx \frac{s'(t)}{s(t)}\Delta t$$

当$\Delta t \to 1$有：

$$f(t)=\left(\frac{s(t)}{s(t-1)}-1\right)\approx\frac{s'(t)}{s(t)}$$

(2)对 $\sigma^2(t-\Delta t)$ 泰勒展开

$$\sigma(t-\Delta t)^2\approx\sigma(t)^2-\frac{d}{dt}[\sigma(t)^2]\Delta t$$

因此

$$\sigma(t)^2-\sigma(t-\Delta t)^2\approx 2\sigma(t)\sigma'(t)\Delta t$$

当$\Delta t\to 1$

$$\sigma(t)^2-\sigma(t-1)^2\approx 2\sigma(t)\sigma'(t)$$

因此

$$g(t)=s(t)\sqrt{\sigma(t)^2-\sigma(t-1)^2}\boldsymbol{\epsilon}_t=s(t)\sqrt{2\sigma(t)\sigma'(t)}$$

【证明3】VE 前向离散形式推导

$$dx=\sqrt{\frac{d^2\sigma(t)}{dt}}dw_t$$

由欧拉离散化 $x_t=x_{t-1}+dx(t-1)$：

$$x_t=x_{t-1}+\sqrt{\frac{\sigma^2(t)-\sigma^2(t-\Delta t)}{\Delta t}}\sqrt{\Delta t}\epsilon$$

令$\Delta t=1$得：

$$x_t=x_{t-1}+\sqrt{\sigma^2(t)-\sigma^2(t-1)}\epsilon$$

由迭代求和可得：

$$x_t=x_0+\sigma(t)\epsilon$$

特别地，$\sigma(t)=\sigma_{min}(\frac{\sigma_{max}}{\sigma_{min}})^t$ ，其中$t\sim\mathcal{U}(0,1)$
也可由式（24）通式带入相关项得到。

【证明4】VP 前向离散形式推导

$$d\mathbf{x}_t=-\frac{1}{2}\beta(t)\mathbf{x}_t dt+\sqrt{\beta(t)}dw_t$$

欧拉离散化：

$$\mathbf{x}_t=\mathbf{x}_{t-1}+(-\frac{1}{2}\beta(t)\mathbf{x}_{t-1}\Delta t+\sqrt{\beta(t)}\sqrt{\Delta t}\epsilon)$$

令$\Delta t=1$，得：

$$\mathbf{x}_t=(1-\frac{1}{2}\beta(t))\mathbf{x}_{t-1}+\sqrt{\beta(t)}\epsilon$$

泰勒展开近似有：$\sqrt{1-\beta(t)}=1-\frac{1}{2}\beta(t)$ 得:

$$\mathbf{x}_t=\sqrt{1-\beta(t)}\mathbf{x}_{t-1}+\sqrt{\beta(t)}\epsilon$$

递推求和得，$\sigma(t)^2=1-e^{-\int_0^t\beta(s)ds}$：

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

特别地，$\beta(t) = (\beta_{max} - \beta_{min})t + \beta_{min}$ ，其中，$t \sim \mathcal{U}(\epsilon_t, 1)$

【证明5】VE 反向形式离散化推导过程

由VE反向连续SDE形式：

$$d\mathbf{x} = \sqrt{\frac{d\sigma^2(t)}{dt}}dw_t$$

可知

$$g(t) = \sqrt{\frac{d\sigma^2(t)}{dt}}$$

对VE反向SDE连续形式欧拉离散化：

$$d\mathbf{x} = [-g(t)^2 \bigtriangledown_x \log p_t(\mathbf{x})]dt + g(t)d\mathbf{w}$$

$$\mathbf{x}_{t-1} = \mathbf{x}_t + d\mathbf{x}_t = \mathbf{x}_t + \frac{d^2\sigma_t}{dt} \cdot \frac{\epsilon_\theta(\mathbf{x}_t, t)}{\bar{\sigma}_t} + \sqrt{\frac{d^2\sigma_t}{dt}}\sqrt{dt}\epsilon_t$$

进一步，令$\Delta t = 1$：

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \frac{\sigma_t^2 - \sigma_{t-1}^2}{\bar{\sigma}_t}\epsilon_\theta(\mathbf{x}_t, t) + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\epsilon_t$$

由于 $\bar{\sigma}(t) = \sigma(t)$ ，因此：

噪声预测形式：

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \frac{\sigma_t^2 - \sigma_{t-1}^2}{\sigma_t}\epsilon_\theta(\mathbf{x}_t, t) + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\epsilon_t$$

残差预测形式：

$$\mathbf{x}_{t-1} = \mathbf{x}_t - (\sigma_t^2 - \sigma_{t-1}^2)\frac{1}{s(t)\sigma^2(t)}(D_\theta(\frac{\mathbf{x}_t}{s(t)}; \sigma(t)) - \frac{\mathbf{x}_t}{s(t)}) + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\epsilon_t$$

由于 $s(t) = 1$，有:

$$\mathbf{x}_{t-1} = \mathbf{x}_t - \frac{(\sigma_t^2 - \sigma_{t-1}^2)}{\sigma_t^2}(D_\theta(\mathbf{x}_t; \sigma_t) - \mathbf{x}_t) + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\epsilon_t$$

【证明6】VP 反向形式离散化为DDPM推导过程

给定**VP逆向SDE**：

$$d\mathbf{x} = \left[-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_x \log p_t(\mathbf{x})\right]dt + \sqrt{\beta(t)}d\bar{\mathbf{w}}$$

（1）忽略扩散项

$$d\mathbf{x} = \left[-\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_x \log p_t(\mathbf{x})\right]dt$$

（2）欧拉离散化

对时间 (t) 离散化，步长 Δt，逆向时间从 (t) 到 (t-1)：

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \left[-\frac{1}{2}\beta_t\mathbf{x}_t - \beta_t\nabla_x \log p_t(\mathbf{x}_t)\right](-\Delta t)$$

令$\Delta t = 1$

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \frac{1}{2}\beta_t\mathbf{x}_t + \beta_t\nabla_x \log p_t(\mathbf{x}_t)$$

(3）得分函数替换（残差形式）:

$$\nabla_x \log p_t(\mathbf{x}_t) \approx -\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}} = \frac{1}{s(t)\sigma^2(t)}(D_\theta(\frac{\mathbf{x}_t}{s(t)};\sigma(t)) - \frac{\mathbf{x}_t}{s(t)})$$

- 这里 ϵ_θ 表示噪声方向
- 前者为噪声预测形式，后者为残差形式。
- $\sqrt{1-\bar{\alpha}_t} = \bar{\sigma}_t$
- $\mathbf{x}_t$ 表示原模型输入 F_θ 的形式

> 直观理解：分母 $\sqrt{1-\bar{\alpha}_t} = s(t)\sigma(t)$，分子量纲为 $\sigma(t)\cdot\epsilon_\theta$，因此需要再除一份 $\sigma(t)$

(3')得分函数替换（噪声形式）:

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \frac{1}{2}\beta_t\mathbf{x}_t - \beta_t\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}}$$

$$\mathbf{x}_{t-1} = \left(1 + \frac{1}{2}\beta_t\right)\mathbf{x}_t - \beta_t\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}}$$

此处 $1+\frac{1}{2}\beta_t$ 泰勒展开近似

$$\mathbf{x}_{t-1} \approx \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \sqrt{\alpha_t}\beta_t\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}}\right)$$

由于 β_t 近似0，$\alpha_t = 1-\beta_t$ 近似1，因此 $\sqrt{\alpha_t}\beta_t \approx \beta_t$

(4）加入随机噪声

噪声预测形式：

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \beta_t\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}}\right) + \sqrt{\beta_t}\epsilon_t$$

残差预测形式：

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t + \beta_t\frac{D_\theta(\frac{\mathbf{x}_t}{s(t)};\sigma(t)) - \frac{\mathbf{x}_t}{s(t)}}{s(t)\sigma^2(t)}\right) + \sqrt{\beta_t}\epsilon_t$$