Notes: Stochastic Differential Equations

Guan Luo [*]

Tsinghua University

该文章是看 Stochastic Differential Equations and Diffusion Models [1] 的学习笔记。

1 Discrete-Time Markov Processes

离散马尔可夫过程由一系列的随机变量定义，每个随机变量的分布只依赖于上一时刻的随机变量的分布，因而给定当前状态下，过去和未来是相互独立。

$$X_0, X_1, ..., X_{i-1} \perp X_{i+1}, X_{i+2}, ...|X_i \tag{1}$$

这些随机变量的联合分布可以拆解为初始状态的分布和每一时刻的转移分布的乘积。

$$P(X_0 = x_0, X_1 = x_1, X_2 = x_2, ...) = P(X_0 = x_0) \prod_{t\in\{1,2,...\}} P(X_t = x_t|X_{t-1} = x_{t-1}) \tag{2}$$

我们着重关注任意时刻的条件分布 (条件概率质量) 及边缘分布 (概率质量)，在离散马尔可夫过程的范畴中如下。

$$P(X_t = x_t|X_s = x_s) = \sum_k P(X_t = x_t|X_m = k)P(X_m = k|X_s = x_s), t \le m \le s \tag{3}$$

$$P(X_t = x_t) = \sum_k P(X_t = x_t|X_m = k)P(X_m = k) \tag{4}$$

需要注意的是 $X_0, X_1, ..., X_T$ 是马尔可夫过程，其反向过程 $X_T, X_{T-1}, ..., X_0$ 同样是马尔可夫过程。

2 Continuous-Time Markov Processes

让离散马尔可夫过程中相邻的状态的时间间隔趋于 0，离散马尔可夫过程变换成连续状态的马尔可夫过程，我们同样可以定义转移分布 (条件概率质量) 和边缘分布 (概率质量)。

$$p(x;t|y;s) = \int_{-\infty}^{\infty} p(x;t|k;m)p(k;m|y;s)dk \tag{5}$$

$$p(x;t) = \int_{-\infty}^{\infty} p(x;t|k;m)p(k;m)dk \tag{6}$$

3 The Kolmogorov Equation

3.1 The Forward Kolmogorov Equation

我们考虑条件概率质量函数在时间上推进了非常小的间隔 dt，可得条件概率质量函数如下。

$$p(x;t+dt|y;s) = \int_{-\infty}^{\infty} p(x;t+dt|m;t)p(m;t|y;s)dm \tag{7}$$

[*] lg22@mails.tsinghua.edu.cn

不妨简记推进 dt 时间的条件概率为

$$\phi_t(\Delta; z) = p(z + \Delta; t + dt|z; t) \tag{8}$$

则可得 $m = x - \Delta$，带入 Eq. 7 式可得

$$p(x; t + dt|y; s) = \int_{-\infty}^{\infty} \phi_t(\Delta; m)p(m; t|y; s)d\Delta \tag{9}$$

我们对 Eq. 9 积分内做泰勒展开可得

$$\begin{aligned} p(x; t + dt|y; s) &= \int_{-\infty}^{\infty} \phi_t(\Delta; x)p(x; t|y; s)d\Delta \quad (10)\\ &- \int_{-\infty}^{\infty} \Delta \frac{\partial}{\partial x}\phi_t(\Delta; x)p(x; t|y; s)d\Delta \quad (11)\\ &+ \int_{-\infty}^{\infty} \frac{\Delta^2}{2}\frac{\partial^2}{\partial x^2}\phi_t(\Delta; x)p(x; t|y; s)d\Delta + \cdots \quad (12)\end{aligned}$$

注意 ϕ_t 对于 Δ 积分为 1，截断到展开的二次项可得

$$\begin{aligned} p(x; t + dt|y; s) - p(x; t|y; s) &= -\frac{\partial}{\partial x}(\mathbb{E}_{\Delta\sim\phi_t(;x)}[\Delta]p(x; t|y; s)) \quad (13)\\ &+ \frac{1}{2}\frac{\partial^2}{\partial x^2}(\mathbb{E}_{\Delta\sim\phi_t(;x)}[\Delta^2]p(x; t|y; s)) \quad (14)\end{aligned}$$

我们设定展开的前两项都是 dt 阶矩，则有

$$\mathbb{E}_{\Delta\sim\phi_t(;x)}[\Delta] := f(x, t)dt \tag{15}$$

$$\mathbb{E}_{\Delta\sim\phi_t(;x)}[\Delta^2] := g^2(x, t)dt \tag{16}$$

将 Eq. 14 两边除以 dt 即得 Kolmogorov Forward Equation，也被称为 Fokker-Planck Equation。

$$\frac{\partial}{\partial t}p(x; t|y; s) = -\frac{\partial}{\partial x}(f(x, t)p(x; t|y; s)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(g^2(x, t)p(x; t|y; s)) \tag{17}$$

以相同的方式，可以得到边缘分布的 Kolmogorov Forward Equation 形式。

$$\frac{\partial}{\partial t}p(x; t) = -\frac{\partial}{\partial x}(f(x, t)p(x; t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(g^2(x, t)p(x; t)) \tag{18}$$

截断的原因：假定在概率分布变化较为平滑，则 ϕ_t 集中在非常小的区域，则有 $\mathbb{E}[\Delta^3] \ll O(\mathbb{E}[\Delta]) = O(dt)$，$\mathbb{E}[\Delta^4] \ll O(\mathbb{E}[\Delta^2]) = O(dt)$，因此只考虑 dt 阶矩时可以忽略后续展开项。

伊藤过程：ϕ_t 表示在 dt 时间内随机变量的变化量的分布，即

$$dX \sim \phi_t(; X) \tag{19}$$

$$\mathbb{E}[dX] = f(X, t)dt \tag{20}$$

$$\mathrm{Var}(dX) = g^2(X, t)dt - O(dt^2) \approx g^2(X, t)dt \tag{21}$$

其中 f 称为扩散过程的 drift coefficient，g 称为扩散过程的 diffusion coefficient，且该过程重参数化如下，其中 dw 服从均值为 0，方差为 dt 的正态分布。

$$dX = f(X, t)dt + g(X, t)dw \tag{22}$$

数值求解：将时间序列离散化，并通过欧拉法做数值积分。假设我们已知 $t = 0$ 时刻的值 X_0，需要计算 $t = 1$ 时刻的值，将时间等分为 N 份，则有

$$\begin{aligned} \int_{X_0}^{X_1} dX &\approx \sum_{i=1}^{N} f_i \int_{\frac{i-1}{N}}^{\frac{i}{N}} dt + \sum_{i=1}^{N} g_i \int_{\frac{i-1}{N}}^{\frac{i}{N}} dw \quad (23)\\ &= \sum_{i=1}^{N} \frac{f_i}{N} + \sum_{i=1}^{N} g_i \int_{\frac{i-1}{N}}^{\frac{i}{N}} dw \quad (24)\end{aligned}$$

注意到第二项的各个积分相互独立，由于 dw 之间相互独立，因此以下我们只考虑第一项，第一个积分项进一步划分成 δt 的间隔有

$$\int_0^{\frac{1}{N}} dw = \lim_{\Delta t \to 0} \sum_{j=1}^{\frac{1}{N\Delta t}} \Delta w_j \tag{25}$$

且这些 Δw_j 之间相互独立，都服从均值为 0，方差为 Δt，根据中心极限定理，我们知道这些随机变量的和服从正态分布，且均值为随机变量的均值的和，方差为随机变量的方差的和，即有

$$\int_0^{\frac{1}{N}} dw \sim \mathcal{N}(0, \frac{1}{N}) \tag{26}$$

因此认为每个积分项为该分布的一个采样，带入 Eq. 24 得

$$\int_{X_0}^{X_1} dX \approx \sum_{i=1}^{N} \frac{f_i}{N} + \sum_{i=1}^{N} g_i \Delta w_i \tag{27}$$

3.2 The Backward Kolmogorov Equation

我们考虑时间向前推进 ds，通过类似的泰勒展开和截断过程有

$$\begin{aligned}
p(x;t|y;s-ds) &= \int_{-\infty}^{\infty} \phi_{s-ds}(\Delta;y) p(x;t|y+\Delta;s) d\Delta &(28)\\
&= \int_{-\infty}^{\infty} \phi_{s-ds}(\Delta;y) p(x;t|y;s) d\Delta &(29)\\
&+ \int_{-\infty}^{\infty} \phi_{s-ds}(\Delta;y) \Delta \frac{\partial}{\partial y} p(x;t|y;s) d\Delta &(30)\\
&+ \int_{-\infty}^{\infty} \phi_{s-ds}(\Delta;y) \frac{\Delta^2}{2} \frac{\partial^2}{\partial y^2} p(x;t|y;s) d\Delta + \cdots &(31)
\end{aligned}$$

令 ds 趋于零，则可得 Kolmogorov Backward Equation。

$$-\frac{\partial}{\partial s} p(x;t|y;s) = f(y,s) \frac{\partial}{\partial y} p(x;t|y;s) + \frac{g^2(y;s)}{2} \frac{\partial^2}{\partial y^2} p(x;t|y;s) \tag{32}$$

4 Reversing Time

给定初始分布 $p(x;0)$ 和 drift 及 diffusion coefficients f 和 g，我们可以描述分布随着时间的变化轨迹，假设在 T 时刻，边缘分布为 $p(x;T)$，我们是否能够描述分布 q，使得 $q(x;0) := p(x;T)$，且轨迹和 p 相同，但时间是逆流的。考虑伊藤积分形式

$$dX = f(X,t)dt + g(X,t)dw \tag{33}$$

即在给定当前时刻的状态 X 时，经过很短的时间间隔 dt，状态的变化量为 dX，且其中分布项 dw 只依赖于当前状态，但当反向从 t 时刻经过短时间 $-dt$ 时，该 dw 不依赖于 X，而是依赖于 $X_s, s < t$，即对于轨迹 q 而言，当前时刻的变化量依赖于其未来的状态，因此不能直接对 Eq. 33 直接取负号。

考虑联合分布 $p(x;t,y;s)$，我们关注在反向过程的条件分布 $p(y;s|x;t), s \leq t$，有

$$\begin{aligned}
&p(x;t,y;s) = p(y;s)p(x;t|y;s) &(34)\\
\Rightarrow\ &p(x;t,y;s) = p(y;s)\frac{\partial}{\partial s}p(x;t|y;s) + p(x;t|y;s)\frac{\partial}{\partial s}p(y;s) &(35)\\
&= -p(y;s)\left[f(y;s)\frac{\partial}{\partial y}\frac{p(x;t,y;s)}{p(y;s)} + \frac{g^2(y;s)}{2}\frac{\partial^2}{\partial y^2}\frac{p(x;t,y;s)}{p(y;s)}\right] &(36)\\
&+ \frac{p(x;t,y;s)}{p(y;s)}\left[-\frac{\partial}{\partial y}f(y;s)p(y;s) + \frac{1}{2}\frac{\partial^2}{\partial y^2}g^2(y;s)p(y;s)\right] &(37)
\end{aligned}$$

最后等式分别带入 Kolmogorov Backward Equation 32和 Kolmogorov Forward Equation 的边缘分布 18。两边除以 $p(x;t)$ 得

$$\frac{\partial}{\partial s}p(y;s|x;t) = -p(y;s)\left[f(y;s)\frac{\partial}{\partial y}\frac{p(y;s|x;t)}{p(y;s)} + \frac{g^2(y;s)}{2}\frac{\partial^2}{\partial y^2}\frac{p(y;s|x;t)}{p(y;s)}\right] \tag{38}$$

$$+\frac{p(y;s|x;t)}{p(y;s)}\left[-\frac{\partial}{\partial y}f(y;s)p(y;s) + \frac{1}{2}\frac{\partial^2}{\partial y^2}g^2(y;s)p(y;s)\right] \tag{39}$$

以下我们简化记 f_y 表示 $\frac{\partial f}{\partial y}$，$q$ 和 p 表示 $p(y;s|x;t)$ 和 $p(y;s)$，则上式简化如下

$$q_s = -pf(\frac{q}{p})_y - \frac{pg^2}{2}\left(\frac{q}{p}\right)_{yy} - \frac{q}{p}(fp)_y + \frac{q}{2p}(g^2p)_{yy} \tag{40}$$

将第一项和第三项结合，并加和减去 $\frac{(pg^2)_y}{2}(\frac{q}{p})_y$ 结合第二项和第四项得

$$q_s = -(fq)_y - \left(\frac{pg^2}{2}\left(\frac{q}{p}\right)_y\right)_y + \left(\frac{q}{2p}(g^2p)_y\right)_y \tag{41}$$

最后加和减去第三项，综合得到

$$q_s = -\left(\left(f - \frac{1}{p}(g^2p)_y\right)q\right)_y - \frac{(g^2q)_{yy}}{2} \tag{42}$$

将该式子和 Eq. 17 对比，就可得 drift 和 diffusion coefficient 项，逆时间的伊藤积分表示为

$$dX = \left(f(X,t) - \frac{1}{p(X,t)}\frac{\partial}{\partial X}g^2(X,t)p(X,t)\right)dt + g(X,t)d\tilde{w} \tag{43}$$

此处的由于时间是从 T 逆流到 0，则 dt 是负的，$d\tilde{w}$ 服从均值为 0，方差为 $-dt$ 的高斯分布。

5 Diffusion Model

Score-Based Generative Modeling Using Stochastic Differential Equation [6] 采用如下 SDE 作为统一的框架描述 Diffusion Model，Score-based [5, 6] 的模型和 Diffusion-based [2, 4] 的模型采用不同的 drift 和 diffusion 系数，其中将 diffusion 项简化成只和时间相关。

$$dX = f(X,t)dt + g(t)dw \tag{44}$$

其逆向过程伊藤积分如下

$$dX = (f(X,t) - \frac{g^2(t)}{p(X,t)}\nabla_X p(X,t))dt + g(t)dw \tag{45}$$

$$= (f(X,t) - g^2(t)\nabla_X \log p(X,t))dt + g(t)dw \tag{46}$$

注意 Eq. 44 和 Eq. 46 的 dw 不相同。

[6] 指出如果 $f(X,t)$ 是 affine 形式即 $f(X,t) = f(t)X$，则这个过程始终是高斯分布，对于更泛化的 SDE 的讨论可以参考 [3, 6]。

Variance Exploding(VE) [5] 的前向过程为

$$dX = \sqrt{\frac{d[\sigma^2(t)]}{dt}}dw \tag{47}$$

对应的条件概率分布为

$$p(x_t|x_0) = \mathcal{N}(x_t|x_0, [\sigma_t^2 - \sigma_0^2]I) \tag{48}$$

Variance Preserving(VP) [2] 的前向过程为

$$dX = -\frac{1}{2}\beta(t)Xdt + \sqrt{\beta(t)}dw \tag{49}$$

对应的条件概率分布为

$$p(x_t|x_0) = \mathcal{N}(x_t|x_0 e^{-\frac{1}{2}\int_0^t \beta(s)ds}, [1 - e^{-\int_0^t \beta(s)ds}]I) \tag{50}$$

General Form [3] EDM 得出在 affine 的形式下，即前向过程为

$$dX = f(t)Xdt + g(t)dw \tag{51}$$

其逆向过程对应的概率流 ODE 为

$$dX = (f(t)X - g^2(t)\nabla_x \log p(X,t))dt \tag{52}$$

其条件概率分布有着统一的形式。

$$p(x_t|x_0) = \mathcal{N}(x_t|s(t)x_0, s^2(t)\sigma^2(t)I) \tag{53}$$

其中，

$$s(t) = \exp\left(\int_0^t f(\xi)d\xi\right), \sigma(t) = \sqrt{\int_0^t \frac{g^2(\xi)}{s^2(\xi)}d\xi} \tag{54}$$

从而其边缘分布为

$$\begin{aligned}
p(x_t) &= \int_{\mathbb{R}^d} p(x_t|x_0)p_{\text{data}}(x_0)dx_0 && (55)\\
&= \int_{\mathbb{R}^d} p_{\text{data}}(x_0)\left[\mathcal{N}(x; s(t)x_0, s^2(t)\sigma^2(t)I)\right]dx_0 && (56)\\
&= \int_{\mathbb{R}^d} p_{\text{data}}(x_0)\left[s(t)^{-d}\mathcal{N}(x/s(t); x_0, \sigma^2(t)I)\right]dx_0 && (57)\\
&= s(t)^{-d}\int_{\mathbb{R}^d} p_{\text{data}}(x_0)\mathcal{N}(x/s(t); x_0, \sigma^2(t)I)dx_0 && (58)\\
&= s(t)^{-d}\left[p_{\text{data}} * \mathcal{N}(0, \sigma^2(t)I)\right](x/s(t)) && (59)
\end{aligned}$$

记被扰动后的分布为 $p(x;\sigma)$，则有

$$p(x;\sigma) = p_{\text{data}} * \mathcal{N}(0, \sigma^2(t)I), p(x_t) = s(t)^{-d}p(x/s(t);\sigma(t)) \tag{60}$$

我们可以重参数化用 $s(t)$ 和 $\sigma(t)$ 表示，则有

$$f(t) = \frac{\dot{s}(t)}{s(t)}, g(t) = s(t)\sqrt{2\dot{\sigma}(t)\sigma(t)} \tag{61}$$

代入得到的逆向过程对应的概率流 ODE 为

$$dX = \left[\frac{\dot{s}(t)}{s(t)}X - s^2(t)\dot{\sigma}(t)\sigma(t)\nabla_x \log p\left(\frac{X}{s(t)};\sigma(t)\right)\right]dt \tag{62}$$

上式为基于 $s(t)$ 和 $\sigma(t)$ 的一般表示形式，当简化设定 $s(t) = 1$ 时，我们得到了 EDM [3] 的 Eq.1 如下

$$dX = -\dot{\sigma}(t)\sigma(t)\nabla_X \log p(X;\sigma(t))dt \tag{63}$$

直观地根据 Eq. 53可知，改变 $\sigma(t)$ 为重参数化时间维度 t，改变 $s(t)$ 为重参数化 X。

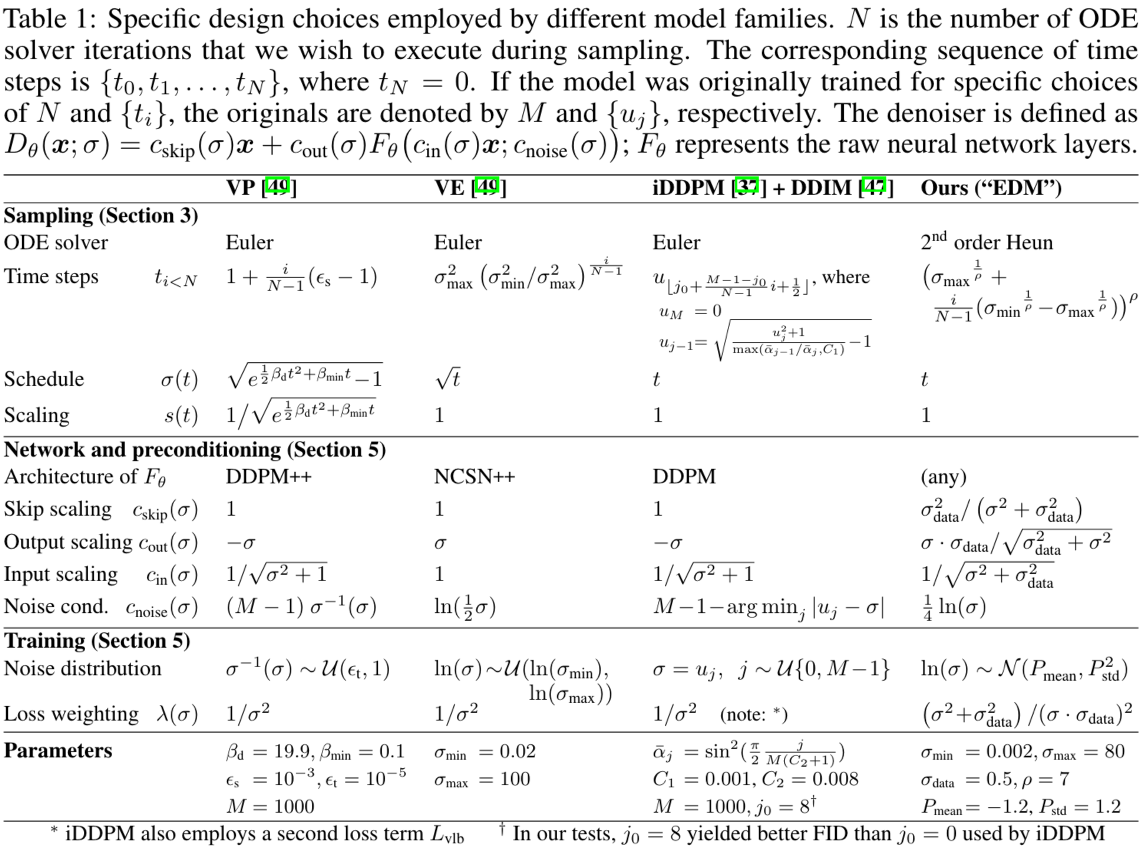

Table 1: Specific design choices employed by different model families. N is the number of ODE solver iterations that we wish to execute during sampling. The corresponding sequence of time steps is $\{t_0, t_1, \dots, t_N\}$, where $t_N = 0$. If the model was originally trained for specific choices of N and $\{t_i\}$, the originals are denoted by M and $\{u_j\}$, respectively. The denoiser is defined as $D_\theta(\boldsymbol{x}; \sigma) = c_{\text{skip}}(\sigma)\boldsymbol{x} + c_{\text{out}}(\sigma)F_\theta\big(c_{\text{in}}(\sigma)\boldsymbol{x}; c_{\text{noise}}(\sigma)\big)$; F_θ represents the raw neural network layers.

		VP [49]	**VE [49]**	**iDDPM [37] + DDIM [47]**	**Ours ("EDM")**
Sampling (Section 3)					
ODE solver		Euler	Euler	Euler	2nd order Heun
Time steps	$t_{i<N}$	$1 + \frac{i}{N-1}(\epsilon_s - 1)$	$\sigma_{\max}^2 \left(\sigma_{\min}^2/\sigma_{\max}^2\right)^{\frac{i}{N-1}}$	$u_{\lfloor j_0 + \frac{M-1-j_0}{N-1} i + \frac{1}{2} \rfloor}$, where $u_M = 0$ $u_{j-1} = \sqrt{\frac{u_j^2+1}{\max(\bar{\alpha}_{j-1}/\bar{\alpha}_j, C_1)} - 1}$	$\left(\sigma_{\max}{}^{\frac{1}{\rho}} + \frac{i}{N-1}(\sigma_{\min}{}^{\frac{1}{\rho}} - \sigma_{\max}{}^{\frac{1}{\rho}})\right)^\rho$
Schedule	$\sigma(t)$	$\sqrt{e^{\frac{1}{2}\beta_d t^2 + \beta_{\min} t} - 1}$	$\sqrt{t}$	t	t
Scaling	$s(t)$	$1/\sqrt{e^{\frac{1}{2}\beta_d t^2 + \beta_{\min} t}}$	1	1	1
Network and preconditioning (Section 5)					
Architecture of F_θ		DDPM++	NCSN++	DDPM	(any)
Skip scaling	$c_{\text{skip}}(\sigma)$	1	1	1	$\sigma_{\text{data}}^2 / \left(\sigma^2 + \sigma_{\text{data}}^2\right)$
Output scaling	$c_{\text{out}}(\sigma)$	$-\sigma$	σ	$-\sigma$	$\sigma \cdot \sigma_{\text{data}} / \sqrt{\sigma_{\text{data}}^2 + \sigma^2}$
Input scaling	$c_{\text{in}}(\sigma)$	$1/\sqrt{\sigma^2+1}$	1	$1/\sqrt{\sigma^2+1}$	$1/\sqrt{\sigma^2 + \sigma_{\text{data}}^2}$
Noise cond.	$c_{\text{noise}}(\sigma)$	$(M-1)\,\sigma^{-1}(\sigma)$	$\ln(\frac{1}{2}\sigma)$	$M-1-\arg\min_j \lvert u_j - \sigma \rvert$	$\frac{1}{4}\ln(\sigma)$
Training (Section 5)					
Noise distribution		$\sigma^{-1}(\sigma) \sim \mathcal{U}(\epsilon_t, 1)$	$\ln(\sigma) \sim \mathcal{U}(\ln(\sigma_{\min}), \ln(\sigma_{\max}))$	$\sigma = u_j, \quad j \sim \mathcal{U}\{0, M-1\}$	$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$
Loss weighting	$\lambda(\sigma)$	$1/\sigma^2$	$1/\sigma^2$	$1/\sigma^2$ (note: *)	$\left(\sigma^2 + \sigma_{\text{data}}^2\right) / (\sigma \cdot \sigma_{\text{data}})^2$
Parameters		$\beta_d = 19.9, \beta_{\min} = 0.1$ $\epsilon_s = 10^{-3}, \epsilon_t = 10^{-5}$ $M = 1000$	$\sigma_{\min} = 0.02$ $\sigma_{\max} = 100$	$\bar{\alpha}_j = \sin^2(\frac{\pi}{2} \frac{j}{M(C_2+1)})$ $C_1 = 0.001, C_2 = 0.008$ $M = 1000, j_0 = 8^\dagger$	$\sigma_{\min} = 0.002, \sigma_{\max} = 80$ $\sigma_{\text{data}} = 0.5, \rho = 7$ $P_{\text{mean}} = -1.2, P_{\text{std}} = 1.2$

* iDDPM also employs a second loss term L_{vlb} † In our tests, $j_0 = 8$ yielded better FID than $j_0 = 0$ used by iDDPM

图 1: EDM [3] 框架的总结和整理，详细的推导参考 EDM [3] 的 Appendix C

参考文献

[1] Tanmaya Shekhar Dabral. Stochastic differential equations and diffusion models. URL: https://www.vanillabug.com/posts/sde/.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[3] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

[4] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. URL: http://proceedings.mlr.press/v37/sohl-dickstein15.html.

[5] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019. URL: `https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html`.

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020. URL: `https://arxiv.org/abs/2011.13456`, `arXiv:2011.13456`.