扩散模型学习笔记 —— 理论

一. 预备：1. 评价指标：IS / FID

2. ELBO + VAE + HVAE

3. VDM： x_0 / ε / $\nabla \log p(x)$

二. DDPM

三. DDIM

四. SDE 统一框架

五. Analyse-DPM (exand-)

六. 概率流 ODE

七. 条件生成 { classifier guidance
classifier-free

一、预备：评价：1) Inception Score：对于生成图片x，使用 Inception-V3 分类网络得到类预测 $y \in R^{1000}$

$\begin{cases} p(y|x): 单一样本类别从原来 \to 由 \downarrow 给出的类预测，质量高的图应该集中 \\ p(y): 生成样本的多样性 \to 分类后统计得到，应分布均匀 \end{cases}$

熵高

熵低

$\to IS(G) = \exp\left(E_{x\sim p(x)} D_{KL}(p(y|x)\|p(y))\right)$

$= -p(y|x)\log p(y) - (-p(y|x)\log p(y|x))$

交叉熵 $+p(y|x)$ 熵

2) FID：为了衡量生成样本集与真实分布的距离。

将图片输入 Inception 得到分类层前的激活向量 $\in R^{2048}$，$p(y)$ 熵↓

再用高斯建模分布得到 μ 和 Σ.

$d^2((\mu,\Sigma),(\mu_w,\Sigma_w)) = \|\mu-\mu_w\|^2 + tr(\Sigma+\Sigma_w-2(\Sigma\Sigma_w)^{\frac{1}{2}})$

生成分布　真实分布

生成模型：建模 $p(x)$ -> 可以采样+新样本

2. ELBO:
(Evidence Lower Bound)

$\log p(x)$

对于一个真实样本 x，真实分布为 p^*。借助隐变量 z 建模低维表示。

直接极大化似然不现实：$p(x)=\int f(x,z)dz=$ 积分 ✗ $\frac{p(x,z)}{p(z|x)}$ 未知 gt 编码器

$$\log p(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] \rightarrow \text{无 ELBO}$$

学习，使 $q_\phi(z|x)$ 近似 后验 $p(z|x)$ <=> 最大化 ELBO

（最小化 KL 散度 <=> 最大化对数似然：$E_{x\sim p^*(x)}[\ln p_\theta(x)] = -H(p^*) - D_{KL}(p^*(x)||p_\theta(x))$

从真实分布中采样　p_θ 与 p^* 的相似度

$$\log p(x) = \log p(x)\int q_\phi(z|x)dz = \int q_\phi(z|x)(\log p(x))dz = E_{q_\phi(z|x)}[\log p(x)]$$

$$= E_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{p(z|x)}\right] = E_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{p(z|x)}\frac{q_\phi(z|x)}{q_\phi(z|x)}\right]$$

$$= E_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] + E_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(z|x)}\right]$$

$$\geq E_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] \qquad D_{KL}(q_\phi(z|x)||p(z|x)) \geq 0$$

最大化 ELBO？：① 是 $\log p(x)$ 的变分下界

② 对于给定 ϕ，ELBO 与 KL 散度 = 常数，因此 max ELBO → min D_{KL} → $q_\phi(z|x) \Rightarrow p(z|x)$

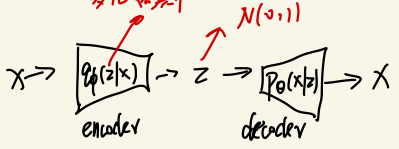
③ 训好后，ELBO 可以用来估计 $p(x)$

2. VAE：变分？求最优化 $q_\phi(z|x)$ | 自编码器？训练时输入 = 输出。

$$E_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] = E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right] = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))$$

重建 $z \rightarrow x$

训练时保证 $x \rightarrow z$ 的编码符合 $p(z)$ 分布，防止拟合成 δ 函数。



① encoder：预测 μ, σ^2 → $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x) I)$ → 计算 $D_{KL}(q_\phi(z|x); p(z))$ 解析解

② 从 $q_\phi(z|x)$ 中采样 L个样本 $z^{(l)}$ → $E_{q_\phi}[\log p_\theta(x|z)] = \frac{1}{L}\sum_{l}^{L} \log p_\theta(x|z^{(l)})$

(为了使采样过程可传梯度，用重参数技巧)

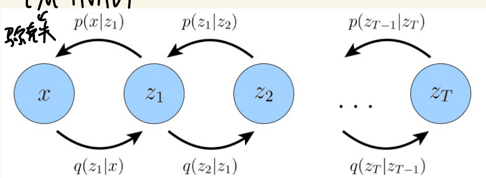
$z^{(l)} = \mu_\phi(x) + \sigma^2_{\phi(x)} \varepsilon, \quad \varepsilon \sim N(0,1)$ → μ_ϕ 和 σ_ϕ 可反传.

③ 目标：对于每个x：$\arg\max_{\phi,\theta} \sum_{l=1}^{L} \log p_\theta(x|z^{(l)}) - D_{KL}(q_\phi(z|x) \| p(z))$

[PS]：$p_\theta(x|z)$ 一般也建模为高斯，经推导可得 $-\log p_\theta(x|z^{(l)}) = \sum_w \sum_h (x^{u,v} - \mu_\theta(z)^{u,v})^2$

4. 层次VAE. (M-HVAE)

$p(x, z_{1:T}) = p(z_T) p_\theta(x|z_1) \prod_{t=2}^{T} p_\theta(z_{t-1}|z_t)$.



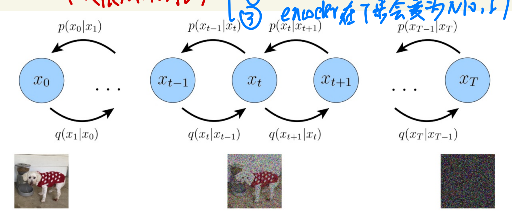
$q_\phi(z_{1:T}|x) = q_\phi(z_1|x) \prod_{t=2}^{T} q_\phi(z_t|z_{t-1})$

$$\therefore ELBO = E_{q_\phi(z_{1:T}|x)}\left[\log \frac{p(x, z_{1:T})}{q_\phi(z_{1:T}|x)}\right]$$

$$= E_{q_\phi}\left[\log \frac{p(z_T) p_\theta(x|z_1) \prod_{t=2}^{T} p_\theta(z_{t-1}|z_t)}{q_\phi(z_1|x) \prod_{t=2}^{T} q_\phi(z_t|z_{t-1})}\right]$$

5. Variational Diffusion Models. $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$

(受限MHVAE)
- ① z与x维度一致
- ② encoder不用学，x_t 总是以 x_{t-1} 为中心的高斯模型.
- ③ encoder在T步会变为 $N(0, I)$



$q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t) I)$ 加噪过程完全由高斯建模

保方差

$p(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$ 去噪训练θ

$p(x_T) = N(x_T; 0, I)$

(1) ELBO

$$\log p(x_0) = \log \int p(x_{0:T}) dx_{1:T} = \log \int \frac{p(x_{0:T}) q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} dx_{1:T} = \log E_{q(x_{1:T}|x_0)}\left[\frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$\geq E_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right] = E_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\right]$$

$$= E_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T) p_\theta(x_0|x_1)}{q(x_T|x_{T-1})}\right] + E_q\left[\log \prod_{t=1}^{T-1} \frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right]$$

$= E_{q(x_{1:T}|x_0)}\left[\log p_\theta(x_0|x_1)\right] + E_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T)}{q(x_T|x_{T-1})}\right] + \sum_{t=1}^{T-1} E_{q(x_{t:T}|x_0)}\left[\log \frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right]$

$q(x_T, x_{T-1}|x_0)$　　$q(x_{t+1}, x_t, x_{t-1}|x_0)$

$= E_{q(x_1|x_0)}\left[\log p_\theta(x_0|x_1)\right] - E_{q(x_{T-1}|x_0)}\left[D_{KL}(q(x_T|x_{T-1})\|p(x_T))\right] - \sum_{t=1}^{T-1} E_{q(x_{t+1}, x_t|x_0)}\left[D_{KL}(q(x_t|x_{t+1})\|p_\theta(x_t|x_{t+1}))\right]$

$q(x_T|x_{T-1})$　　$N(0, I)$

重建项，与VAE一致　　先验匹配项　　一致项（使 x_t 在前/后向保持一致）

（不需优化，无参数，当时间足够大时，$q(x_T|x_0)$ 为0）

训练时会高效估计

用蒙特卡洛估计期望时有两个随机变量 x_{t+1}, x_t → 方差大，不准

另一种推导：

$\log p(x) \geq E_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right] = E_q\left[\log \frac{p(x_T)\, p_\theta(x_0|x_1)}{q(x_1|x_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right]$

$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$

$= E_q\left[\log \frac{p(x_T)\, p_\theta(x_0|x_1)}{q(x_1|x_0)} + \log \frac{q(x_1|x_0)}{q(x_T|x_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right] = \frac{q(x_{t-1}|x_t, x_0)\, q(x_t|x_0)}{q(x_{t-1}|x_0)}$

$= E_q\left[\log p_\theta(x_0|x_1)\right] + E_{q(x_T|x_0)}\left[\log \frac{p(x_T)}{p(x_T|x_0)}\right] + \sum_{t=2}^{T} E_{q(x_t, x_{t-1}|x_0)}\left[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right]$

$= E_{q(x_1|x_0)}\left[\log p_\theta(x_0|x_1)\right] - D_{KL}(p(x_T|x_0)\|p(x_T)) - \sum_{t=2}^{T} E_{q(x_t|x_0)}\left[D_{KL}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t))\right]$

重建　　先验　　CT　　匹配

$\left[q(x_{t-1}, x_t|x_0) = q(x_t|x_0)\cdot q(x_{t-1}|x_t, x_0)\right]$

去噪匹配项

（当T=1时，VDM → VAE）

(2)

优化：重建项如 VAE．去噪匹配项：$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)\, q(x_{t-1}|x_0)}{q(x_t|x_0)}$

① $q(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, \sqrt{1-\alpha_t}I)$

② $q(x_t|x_0) = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t^*$

$= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\varepsilon_{t-1}^* + \sqrt{1-\alpha_t}\varepsilon_t^*$　　$N(0, I)$

$= \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t}$

$= \sqrt{1-\alpha_t\alpha_{t-1}}\,\varepsilon_{t-1}$

$= \cdots$

$= \sqrt{\prod_{i=1}^{t}\alpha_i}\, x_0 + \sqrt{1-\prod_{i=1}^{t}\alpha_i}\,\varepsilon_0 = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon_0$

$\sim N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$

$\therefore q(x_{t-1}|x_t, x_0) \propto N\left(x_{t-1}; \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}I\right)$

$\mu_q(x_t, x_0)$ $\Sigma_q(t)$

∵ $q(x_{t-1}|x_t, x_0)$也是正态分布，因此 $p_\theta(x_{t-1}|x_t)$直接建模，得到 $\mu_\theta(t)$, $\sigma_\theta^2(t)$

时刻t的x_0未知，必须建模。 时刻t预先确定，$\sigma_\theta^2(t) = \Sigma_q(t)$

$$\arg\min_\theta D_{KL}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t))$$

$$= \arg\min_\theta D_{KL}(N(x_{t-1}; \mu_q, \Sigma_q(t)) \| N(x_{t-1}; \mu_\theta, \Sigma_q(t)))$$

预测 $x_{t-1} \to x_0$

解析解

$$= \arg\min_\theta \frac{1}{2\sigma_q^2(t)} [\| \mu_\theta - \mu_q \|_2^2]$$

$\mu_\theta(x_t, t)$, $\mu_q(x_t, x_0)$

将μ_θ写成μ_q形式：$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{x}_\theta(x_t, t)}{1-\bar{\alpha}_t}$

$$= \arg\min_\theta \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} [\| \hat{x}_\theta(x_t, t) - x_0 \|_2^2]$$

相当于，训一个NN，可以从任意步certain预测原始图片。

沿时间求和 近似为 $\arg\min_\theta E_{t\sim U\{2,T\}}[E_{q(x_t|x_0)}[D_{KL}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t))]]$

→ 可以沿时间步随机采样来优化。

(3) 噪音参数的优化： 用NN建模 α_t 是低效的，因为推理时需要计算 $\bar{\alpha}_t$

目标：$\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} [\| \hat{x}_\theta(x_t, t) - x_0 \|_2^2] = \frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}} - \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right) [\| \hat{x}_\theta(x_t, t) - x_0 \|_2^2]$

$$= \frac{1}{2}(SNR(t-1) - SNR(t)) [\| \hat{x}_\theta(x_t, t) - x_0 \|_2^2]$$

$SNR := \frac{\mu^2}{\sigma^2}$, $q(x_t|x_0) = (x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$

signal to noise ratio 表示原始信号(μ)与噪声量的比率

联合优化目标

→ 直接用NN建模 SNR.

$SNR(t) = e^{-w_y(t)}$ → 递增 (时间t推进，SNR变小，直到 $\to N(0, I)$)

→ $\bar{\alpha}_t = sigmoid(-w_y(t))$, $1-\bar{\alpha}_t = sigmoid(w_y(t))$

$$\frac{1}{\frac{\alpha_{t-1}}{1-\alpha_{t-1}}} - \frac{1-\alpha_t}{1-\bar{\alpha}_t} = \frac{1}{-1+\frac{1}{1-\alpha_t}} - \frac{1}{\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}}$$

(4) 其它解释： ① VDM <=> 从任意噪声步骤 x_t 中预测原始图片 x_0

② <=> 预测噪声，DDPM便是如此，效果更好！

预测 $x_{t-1} \to x_0 \to \varepsilon$

$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$ → $x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon}{\sqrt{\bar{\alpha}_t}}$

→ $\mu_q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}$

$$= \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \varepsilon_t$$

写成用网络 $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \hat{\varepsilon}_\theta(x_t, t)$

$\therefore$ 目标 $\Rightarrow \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \left[\| \varepsilon_\theta - \hat{\varepsilon}_\theta(x_t, t) \|_2^2 \right]$

③ $\Rightarrow$ 预测得分函数（在数据空间的梯度），为什么？噪声水平，$\nabla \log p(x_t)$

从 Tweedie's Formula 角度：对个指数族的均值，可以采样样本，再经验平均值 + 涉及分数的校正项来估计（若采样 1 则方差为 0）

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

$\rightarrow E[\mu_{x_t} | x_t] = x_t + (1-\bar{\alpha}_t) \nabla_{x_t} \log p(x_t)$

作为分数，会将 x_t 往 p 增大的修正

$= \sqrt{\bar{\alpha}_t} x_0$

$\therefore x_0 = \frac{x_t + (1-\bar{\alpha}_t) \nabla \log p(x_t)}{\sqrt{\bar{\alpha}_t}} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\, \varepsilon_0}{\sqrt{\bar{\alpha}_t}} \Rightarrow \nabla \log p(x_t) = \frac{-1}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_0$

使 x_t 的 $\log p$ 增大的方向，即“噪声”的反方向

代入 $\mu_q(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}} \nabla \log p(x_t)$

将 $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}} s_\theta(x_t, t)$ 预测分数

目标：$\arg\min_\theta \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{\alpha_t} \left[\| s_\theta(x_t, t) - \nabla \log p(x_t) \|_2^2 \right]$

6. Score-based Generative Models：建模 score function，并使用郎可夫蒂特卡罗法（如朗之万动力学）生成样本

任何概率分布均可以写成 $p_\theta(x) = \frac{1}{Z_\theta} e^{-f_\theta(x)}$，$f_\theta$ 能量函数，Z_θ 归一化系数

由于 f_θ 复杂 $\rightarrow Z_\theta$ 复杂 $\rightarrow$ 无法极大似然学 $p(x)$

$\rightarrow \nabla_x \log p_\theta(x) = \nabla_x \log \frac{1}{Z_\theta} e^{-f_\theta(x)}$

$= \nabla_x \log \frac{1}{Z_\theta} + \nabla_x \log e^{-f_\theta(x)} = -\nabla_x f_\theta(x) \propto s_\theta(x)$

$\rightarrow$ 优化 $E_{p(x)} \left[\| s_\theta(x) - \nabla \log p(x) \|_2^2 \right]$ 时避开 Z_θ $\rightarrow$ 得分匹配 loss，通过对 $\nabla \log p(x)$ 的需求而最小化似然度

得分函数 $\nabla \log p(x)$ 含义：给定 x 的 $\log$ 似然的梯度 $\Rightarrow$ 给出使似然增大的方向

从而，通过学习 s_θ，可以从样本空间任意点开始，迭代地根据 score 更新位置来产生样本。

$x_{i+1} \leftarrow x_i + \epsilon \nabla \log p(x_i) + \sqrt{2\epsilon}\, z$ 朗之万动力学

二、DDPM：

$$L = E_q\left[D_{KL}(q(x_T|x_0)\|p(x_T)) + \sum_{t>1} D_{KL}(q(x_{t-1}|x_t,x_0)\|p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1)\right]$$

L_T　　L_{t-1}　　L_0

1. 前向：α_t/β_t 固定，无可学习参数，L_T 忽略。

2. 反向：$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$

将 $\mu_\theta(x_t,t)$ 参数化为 $\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_\theta(x_t,t)\right)$

Σ_θ set to $\sigma_t^2 I$ 无需训练（t小时的权重下降，应该得把此代码里化简一下），从而使模型将注意力放到t大时。

$$\Rightarrow L_{t-1} \overset{t\ge 2}{=} \frac{1}{2\sigma_t^2}\frac{\beta_t^2}{\alpha_t(1-\bar\alpha_t)}\|\varepsilon - \varepsilon_\theta(x_t,t)\|^2$$

（$x_t = \sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\varepsilon$）

文中实验证明，预测 ε 比 x_0 和 μ_θ 效果好。

3. L_0：简化，与 L_{t-1} 同一形式

简化后

$$= E_{t,x_0,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(\sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\varepsilon, t)\|^2\right]$$

$t\in[1,T]$　把 $\|\|$ 前的对 t 的权重省了，文中说，这种简化使 t 小时的权重下降

train：$\forall x_0$，采样 $t\sim U[1,T]$，$\varepsilon\sim\mathcal{N}(0,I)$

$\Rightarrow x_t = \sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\varepsilon$

$\Rightarrow L = \|\varepsilon - \varepsilon_\theta(x_t,t)\|^2$ 反向传播。

sample (infer)：$\forall x_T\sim\mathcal{N}(0,I)$，

for $t=T,\cdots,1$:

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

$\mu_\theta(x_t,t)$　$z\sim\mathcal{N}(0,I)$，$t>1$；$=0$，$t=1$

前向过程不再引入噪声，而均与 x_{t-1}，x_0 解

三、DDIM：

$$q_\sigma(x_{1:T}|x_0) := q_\sigma(x_T|x_0)\prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t,x_0)$$

① 从更高角度，将 DDPM 的训练过程由马尔科夫推广到非马尔科夫（并未改变目标，仅是另种角度/解释）

② 由此，提出一种加速采样方法。

$$q(x_{t-1}|x_t,x_0) = \frac{q(x_t|x_{t-1},x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)}$$（贝叶斯）

$$\Downarrow$$

$$q_\sigma(x_{t-1}|x_0) = \int q_\sigma(x_{t-1}|x_t,x_0)q_\sigma(x_t|x_0)dx_t$$

$\mathcal{N}(\sqrt{\bar\alpha_{t-1}}x_0, (1-\bar\alpha_{t-1})I)$　$\Rightarrow \mathcal{N}(\sqrt{\bar\alpha_t}x_0, (1-\bar\alpha_t)I)$

分析：DDPM：$q(x_t|x_{t-1})\to q(x_t|x_0)$

$\Rightarrow q(x_{t-1}|x_t,x_0)$ ← 匹配 $p_\theta(x_{t-1}|x_t)$

损失只依赖 $q(x_t|x_0)$

采样只依赖 $p(x_{t-1}|x_t)$ → 省去 $q(x_t|x_{t-1})$

从假设 $q(x_t|x_{t-1})\to q(x_t|x_0)$

待定系数可求解：$x_{t-1} = k_t(\sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\varepsilon_1) + \lambda_t x_0 + \sigma_t\varepsilon_2$

$N(x_{t-1}; k_t x_t + \lambda_t x_0, \sigma_t^2 I)$

$= (k_t\sqrt{\bar\alpha_t} + \lambda_t)x_0 + (k_t\sqrt{1-\bar\alpha_t}\varepsilon_1 + \sigma_t\varepsilon_2)$

得到的 x_{t-1} 相当于采样到 x_t，再根据 x_t 和 x_0 采样到 x_{t-1}，此时再只关注 x_{t-1} 的分布，即相当于通过积分剔去 x_t。

$\therefore k_t\sqrt{\bar\alpha_t} + \lambda_t = \sqrt{\bar\alpha_{t-1}}$，$\sqrt{k_t^2(1-\bar\alpha_t) + \sigma_t^2} = \sqrt{1-\bar\alpha_{t-1}}$

k_t, λ_t, σ_t 有两个方程

$\therefore q_\sigma(x_{t-1}|x_t,x_0) = \mathcal{N}\left(\sqrt{\bar\alpha_{t-1}}x_0 + \sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t - \sqrt{\bar\alpha_t}x_0}{\sqrt{1-\bar\alpha_t}}, \sigma_t^2 I\right)$

→ 自由参数

与与 DDPM 类似，代入 $x_t = \sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\varepsilon$

$$\therefore \mu_q(x_{t-1}|x_t,x_0) = \sqrt{\bar\alpha_{t-1}}\frac{x_t - \sqrt{1-\bar\alpha_t}\varepsilon}{\sqrt{\bar\alpha_t}} + \sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\,\varepsilon$$

$$= \frac{x_t}{\sqrt{\alpha_t}} + \left(\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}}\right)\varepsilon$$

$$\therefore \mu_\theta(x_{t-1}|x_t) = \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1}{\sqrt{\alpha_t}}\left(\sqrt{\alpha_t(1-\bar{\alpha}_{t-1}-\sigma_t^2)} - \sqrt{1-\bar{\alpha}_t}\right)\varepsilon_\theta(x_t, t)$$

$\sigma_\theta = \sigma_t$

σ_t 决定了前向过程的随机程度，当 $\sigma_t \to 0$ 时，只要 x_t 和 x_0 和 t 确定，x_{t-1} 就确定。

$$p_\theta(x_{t-1}|x_t) = \begin{cases} \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_\theta),\ \sigma_1 I\right) & , t=1 \\ q_\sigma\left(x_{t-1}|x_t, x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_\theta)\right) & , t>1 \end{cases}$$

当 $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\sqrt{1-\alpha_t}$ 时，等价于马尔科夫的 DDPM。

当 $\sigma_t = 0$，DDIM（一个以 DDPM 训练目标的隐含概率模型 / implicit

加速采样：由于前向过程可以视为非马尔科夫过程。

对于一个训好的采样时，可以将其作为 dim τ 步训好的模型，$\tau = [\tau_1, \cdots, \tau_{\dim(\tau)}]$ 为 $[1, T]$ 的任意子序列。

→ 生成也只需 dim τ 步，

$$p(x_{\tau_{i-1}}|x_{\tau_i}) = \mathcal{N}\left(x_{\tau_{i-1}}; \frac{\bar{\alpha}_{\tau_{i-1}}}{\alpha_{\tau_i}}\left(x_{\tau_i} - \left(\bar{\beta}_{\tau_i} - \frac{\bar{\alpha}_{\tau_i}}{\bar{\alpha}_{\tau_{i-1}}}\sqrt{\bar{\beta}_{\tau_{i-1}}^2 - \tilde{\sigma}_{\tau_i}^2}\right)\varepsilon_\theta(x_{\tau_i}, \tau_i)\right), \tilde{\sigma}_{\tau_i} I\right)$$

论文给出了两种方法：线性采样 / 二次采样

DDIM 没改变 DDPM 的训练，而仅改变了采样方法 { τ：加速；σ：σ=0 时为确定性 DDIM

四、Score-based：DDPM：离散的 T 步 → SDE：时间上的连续微分方程

drift 系数　diffusion 系数

统一 SDE 框架：前向过程：$dx = f_t(x)\,dt + g_t\,dw$　布朗运动

（自定义的，由比例和 g_t 配设置）

可视作离散过程：$x_{t+\Delta t} - x_t = f_t(x_t)\Delta t + g_t\sqrt{\Delta t}\,\varepsilon$，当 $\Delta t \to 0$ 时的极限　↑ SDE 的设置

∴ 越小的 Δt，代表对 SDE 越好的逼近　类似于 $\sqrt{dt}\,x_t$？　→ dw

DDPM 中不同的 T，可视为对 SDE 不同离散化程度的体现。

→ 引入 SDE 可以“在分析时借助连续性 SDE”，而在实践时逐步的离散化。

逆向过程：$dx = [f_t(x) - g_t^2 \nabla_x \log p_t(x)]\,dt + g_t\,dw$　（推导省去）

未知，“Score”

→ 离散：$x_t - x_{t+\Delta t} = -[f_{t+\Delta t}(x_{t+\Delta t}) - g_{t+\Delta t}^2 \nabla_{x_{t+\Delta t}} \log p(x_{t+\Delta t})]\Delta t + g_{t+\Delta t}\sqrt{\Delta t}\,\varepsilon$

$p(x_t) = \int p(x_t|x_0)\tilde{p}(x_0)dx_0 = E_{x_0}[p(x_t|x_0)]$

$\rightarrow \nabla_{x_t}\log p(x_t) = \dfrac{\nabla_{x_t} p(x_t)}{p(x_t)} = \dfrac{E_{x_0}[\nabla_{x_t} p(x_t|x_0)]}{E_{x_0}[p(x_t|x_0)]} = \dfrac{E_{x_0}[p(x_t|x_0)\nabla_{x_t}\log p(x_t|x_0)]}{E_{x_0}[p(x_t|x_0)]}$ 这步是对的

To 使模型预测 $s_\theta(x_t, t) \approx \nabla_{x_t}\log p(x_t)$

$$\min \frac{E_{x_0}[p(x_t|x_0)\|s_\theta(x_t,t) - \nabla_{x_t}\log p(x_t|x_0)\|^2]}{E_{x_0}[p(x_t|x_0)]}$$

可视为以 $p(x_t|x_0)p(x_0)$ 为权重对 $\nabla_{x_t}\log p(x_t|x_0)$ 加权平均的结果

对 x_t 求期望以对每个 x_t 最小化 → 可忽略成加权 loss → 可能等，若 $f_t(x)$ 为仿射变换，则 $p(x_t|x_0)$ 也是高斯

$\rightarrow \min E_{x_0, x_t\sim p(x_t|x_0)p(x_0)}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p(x_t|x_0)\|\right]^2$ 得分匹配的损失函数

2. 解释 DDPM：$p(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I) \rightarrow x_t = \sqrt{\alpha_t}x_{t-1} + (1-\alpha_t)\varepsilon$

$\Rightarrow$ SDE: $dx = -\frac{1}{2}\beta(t)x\,dt + \sqrt{\beta(t)}\,dw$

↓

[推导：定义一个新的序列：$\bar{\beta}_i = N\beta_i$ $\{\bar{\beta}_i\}_{i=1}^N$ $(\beta_i = 1-\alpha_i)$

$\rightarrow x_t = \sqrt{1-\frac{\bar{\beta}_i}{N}}x_{t-1} + \sqrt{\frac{\bar{\beta}_i}{N}}\varepsilon$ ，当 $N\to\infty$ 时，$\{\bar{\beta}_i\}_{i=1}^N \to \beta(t)$, $t\in[0,1]$ $\beta(\frac{i}{N}) = \bar{\beta}_i$ ，$x(\frac{i}{N}) = x_i$

令 $\Delta t = \frac{1}{N}$ →

$x(t+\Delta t) = \sqrt{1-\beta(t+\Delta t)\Delta t}\,x(t) + \sqrt{\beta(t+\Delta t)}\Delta t\,\varepsilon$

$\overset{\text{泰勒}}{\approx} (1-\frac{1}{2}\beta(t+\Delta t)\Delta t)x(t) + \sqrt{\beta(t+\Delta t)}\sqrt{\Delta t}\,\varepsilon$

$\approx (1-\frac{1}{2}\beta(t)\Delta t)x(t) + \sqrt{\beta(t)}\sqrt{\Delta t}\,\varepsilon$]

$= x(t) - \underbrace{\frac{1}{2}\beta(t)x(t)\Delta t}_{f_t}\,t + \underbrace{\sqrt{\beta(t)}}_{g_t}\sqrt{\Delta t}\,\varepsilon$

设：$q(x_t|x_0) = N(x_t; \alpha_t x_0, \sigma_t^2 I)$

$\Leftrightarrow f_t = \dfrac{d\log\alpha_t}{dt}$

$g_t^2 = \dfrac{d\sigma_t^2}{dt} - 2\dfrac{d\log\alpha_t}{dt}\sigma_t^2$

五. Analytic-DPM：在 DDIM 的基础上推出了方差的解析解.

DDIM: $p(x_{t-1}|x_t, x_0) = N\left(x_{t-1}; \dfrac{\sqrt{\bar{\sigma}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}x_t + \underbrace{(\bar{\alpha}_{t-1} - \dfrac{\bar{\alpha}_t\sqrt{\bar{\sigma}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t})}_{\gamma_t}x_0,\ \sigma_t^2 I\right)$

$\rightarrow p(x_{t-1}|x_t) \approx p(x_{t-1}|x_t, x_0 = \bar{\mu}(x_t))$ → 不准确！时从 x_t 预测 x_0 具有不确定性，应用分布描述

$p(x_{t-1}|x_t) = \int p(x_{t-1}|x_t, x_0)p(x_0|x_t)dx_0$

→ 用 $N(x_0; \bar{\mu}(x_t), \bar{\sigma}_t^2 I)$ 近似 $p(x_0|x_t)$，则

$p(x_{t-1}|x_t) = N\left(x_{t-1}; \dfrac{\sqrt{\bar{\sigma}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}x_t + \gamma_t\bar{\mu}(x_t),\ (\sigma_t^2 + \bar{\sigma}_t^2\gamma_t^2)I\right)$

数据的假设

与过去一致

$+\sigma_t^2$，$\bar{\sigma}_t^2\gamma_t^2$ 即最优方差的修正项.

优化：待预测 $\bar{\mu}(x_t)$，$\bar{\sigma}_t$

均值优化不变，将 $\bar{\mu}(x_t)$ 参数化后得到 $\|\epsilon-\epsilon_\theta(x_t,t)\|^2$

$$\bar{\mu}(x_t)=\frac{1}{\bar{\alpha}_t}(x_t-\bar{\beta}_t\epsilon_\theta(x_t,t))$$

证：$\Sigma(x_t)=E_{x_0\sim p(x_0|x_t)}[(x_0-\bar{\mu}(x_t))(x_0-\bar{\mu}(x_t))^T]$

$$=E_{x_0}\left[\left(\left(x_0-\frac{x_t}{\bar{\alpha}_t}\right)+\frac{\bar{\beta}_t}{\bar{\alpha}_t}\epsilon_\theta(x_t,t)\right)\left(\left(x_0-\frac{x_t}{\sqrt{\bar{\alpha}_t}}\right)+\frac{\bar{\beta}_t}{\bar{\alpha}_t}\epsilon_\theta(x_t,t)\right)^T\right]$$

$$=E_{x_0}\left[\left(x_0-\frac{x_t}{\bar{\alpha}_t}\right)\left(x_0-\frac{x_t}{\bar{\alpha}_t}\right)^T\right]-\frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\epsilon_\theta(x_t,t)\,\epsilon_\theta(x_t,t)^T$$

$$=\frac{1}{\bar{\alpha}_t^2}E_{x_0\sim p(x_0|x_t)}[(x_t-\bar{\alpha}_t x_0)(x_t-\bar{\alpha}_t x_0)^T]-\frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\epsilon_\theta(x_t,t)\epsilon_\theta(x_t,t)^T$$

我们希望 $\bar{\sigma}_t$ 与 x_t 无关 → 对 $x_t\sim p(x_t)$ 求均值.

$$E_{x_t\sim p(x_t)}E_{x_0\sim p(x_0|x_t)}[(x_t-\bar{\alpha}_t x_0)(x_t-\bar{\alpha}_t x_0)^T]$$

$$=E_{x_0\sim\tilde{p}(x_0)}E_{x_t\sim p(x_t|x_0)}[(x_t-\bar{\alpha}_t x_0)(x_t-\bar{\alpha}_t x_0)^T]$$ 关键

$p(x_t|x_0)$ 的 μ　　$p(x_t|x_0)$ 的 Σ

$$=E_{x_0\sim\tilde{p}(x_0)}[\bar{\beta}_t^2 I]=\bar{\beta}_t^2 I$$

$$\therefore \Sigma_t=E_{x_t\sim p(x_t)}[\Sigma(x_t)]=\frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}(I-E_{x_t\sim p(x_t)}[\epsilon_\theta(x_t,t)\epsilon_\theta(x_t,t)^T])$$

取迹降dim

转为标量 $\bar{\sigma}_t^2=\frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(1-\frac{1}{d}E_{x_t\sim p(x_t)}[\|\epsilon_\theta(x_t,t)\|^2]\right)\le\frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}$ （原文是用得分函数推的）

蒙特卡洛估计 $\frac{1}{M}\sum_{m=1}^{M}\|\epsilon_\theta(x_{n,m},t)\|^2$ ← 随机步

Extended-Analytic-DPM：之前假设 $p(x_0|x_t)$ 的协为 $\bar{\sigma}_t^2 I$，即对角阵且元素相等

四、extend：概率流ODE：　DDIM的连续形式　　SDE离散时步长不能太大

1. F-P方程：对于分布 $p(x)=\int\Delta(x-y)p(y)dy=E_y[\Delta(x-y)]$ 转为期望

$$\to p(x)f(x)=\int\Delta(x-y)p(y)f(y)dy=E_y[\Delta(x-y)f(y)]$$

$$\to\nabla_x[p(x)f(x)]=E_y[\nabla_x\Delta(x-y)f(y)]=E_y[f(y)\nabla_x\Delta(x-y)]$$ ①

由SDE：$dx=f_t(x)dt+g_t dw$ → $x_{t+\Delta t}=x_t+f_t(x_t)\Delta t+g_t\sqrt{\Delta t}\,\varepsilon$，$\varepsilon\sim N(0,1)$

$$\therefore\Delta(x-x_{t+\Delta t})=\Delta(x-x_t-f_t(x_t)\Delta t-g_t\sqrt{\Delta t}\,\varepsilon)$$

展开 $\approx\Delta(x-x_t)-(f_t(x_t)\Delta t+g\sqrt{\Delta t}\,\varepsilon)\nabla_x\Delta(x-x_t)+\frac{1}{2}(g_t\sqrt{\Delta t}\,\varepsilon)^2\nabla_x^2\Delta(x-x_t)+O(\Delta t)$

左右求期望可得：

$$p_{t+\Delta t}(x) = \mathbb{E}_{x_{t+\Delta t}}\left(\Delta(x - x_{t+\Delta t})\right) \approx \mathbb{E}_{x_t,\xi}\left[\Delta(x-x_t) - \left(f_t(x_t)\Delta t + g_t\sqrt{\Delta t}\,\xi\right)\nabla_x \Delta(x-x_t) + \frac{1}{2}\left(g_t\sqrt{\Delta t}\,\xi\right)^2 \nabla_x^2 \Delta(x-x_t)\right]$$

$\mathbb{E}(\xi)=0$　　$\mathbb{E}(\xi^2)=1$

$$= \mathbb{E}_{x_t}\left[\Delta(x-x_t) - f_t(x_t)\Delta t \cdot \nabla_x \Delta(x-x_t) + \frac{1}{2} g_t^2 \Delta t\, \nabla_x^2 \Delta(x-x_t)\right]$$

由①

$$= p_t(x) - \nabla_x\left[f_t(x)\Delta t\, p_t(x)\right] + \frac{1}{2} g_t^2 \Delta t\, \nabla_x^2 p_t(x)$$

除以Δt并$\Delta t \to 0$

$$\Rightarrow \frac{\partial p_t(x)}{\partial t} = -\nabla_x\left[f_t(x)\, p_t(x)\right] + \frac{1}{2} g_t^2 \nabla_x^2 p_t(x)$$

原SDE的F-P方程，描述转移概率分布的偏微分方程.

2. 等价变换：对于函数 $\sigma_t^2 = g_t^2$，（完全等价）$\frac{\partial p_t(x)}{\partial t} = -\nabla_x\left[f_t(x) p_t(x) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_x p_t(x)\right] + \frac{1}{2}\sigma_t^2 \nabla_x^2 p_t(x)$

$$= -\nabla_x\left[\left(f_t(x) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_x \log p_t(x)\right) p_t(x)\right] + \frac{1}{2}\sigma_t^2 \nabla_x^2 p_t(x)$$

相当于 $g_t \to \sigma_t$，而 $f_t \to f_t(x) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_x \log p_t(x)$

→ 将F-P方程对应回SDE：

$$dx = \left(f_t(x) - \frac{1}{2}(g_t^2 - \sigma_t^2)\nabla_x \log p_t(x)\right)dt + \sigma_t dw$$

完全等价于原SDE

→ 存在 不同方差的前向过程，产生的边际分布是一样的.　→ DDIM可统一进来

→ 反向SDE：$dx = \left(f_t(x) - \frac{1}{2}(g_t^2 + \sigma_t^2)\nabla_x \log p_t(x)\right)dt + \sigma_t dw$

3. 神经ODE：当取 $\sigma_t = 0$ 时，SDE退化为ODE（常微分）

$$dx = \left(f_t(x) - \frac{1}{2} g_t^2 \nabla_x \log p_t(x)\right)dt$$

概率流 ODE

用 $s_\theta(x,t)$ 近似 → 对应一个"神经ODE"

此时，反向与前向ODE完全相同，可逆变换.

且 前向与反向均为确定性变换.

没有随机性，从而可以在离散化时取更大步长

六. 条件生成：

$$\nabla \log p(x_t|y) = \nabla \log\left(\frac{p(x_t)p(y|x_t)}{p(y)}\right) = \nabla \log p(x_t) + \nabla \log p(y|x_t)$$

无条件　　条件

1. Classifier-Guidance：（只需训classifier）

(1) $p(x_{t-1}|x_t) \to p(x_{t-1}|x_t, y) = \frac{p(x_{t-1}|x_t)\, p(y|x_{t-1}, x_t)}{p(y|x_t)} = p(y|x_{t-1})$

$p(x_{t-1}|y) = \frac{p(x_{t-1})\, p(y|x_{t-1})}{p(y)}$　（添加条件）

$= p(x_{t-1}|x_t)\, e^{\log p(y|x_{t-1}) - \log p(y|x_t)}$

近似后可得

$$p(x_{t-1}|x_t, y) = \mathcal{N}\left(x_{t-1};\ \mu(x_t) + \sigma_t^2 \nabla_{x_t} \log p(y|x_t),\ \sigma_t^2 I\right)$$

采样: $x_{t-1} = \mu(x_t) + \sigma_t^2 \nabla_{x_t} \log p(y|x_t)\big|_{x_t=\mu(x_t)} + \sigma_t \varepsilon$

梯度缩放: 向分类器梯度加个缩放参数 γ, $\sigma_t^2 \gamma \nabla_{x_t} \log p(y|x_t)$, 当 $\gamma > 1$ 时, 会提高结果与 y 的相关性, 降低多样性.

(2) 分类 → 文本…: $p(x_{t+1}|x_t, y) = \frac{p(x_{t+1}|x_t)\, e^{\gamma \mathrm{sim}(x_{t+1}, y)}}{Z_t}$, Z_t 是归一化系数.

近似 → $p(x_{t-1}|x_t, y) = N(x_{t-1}; \mu(x_t) + \sigma_t^2 \gamma \nabla_{x_t} \mathrm{sim}(x_t, y), \sigma_t^2 I)$

$\sigma_t^2 = 0$? → SDE; $\mathrm{sim}(x_t, y)$: 余弦空间相似度

(2) 连续视角: SDE: 反向

$$dx = (f_t(x) - \frac{1}{2}(g_t^2 + \sigma_t^2) \nabla_x \log p(x)) dt + \sigma_t dw$$

$\nabla_x \log p(x|y)$

$$\therefore \nabla_x \log p_t(x|y) = -\frac{\varepsilon_\theta(x_t, t) - \bar{\beta}_t \nabla_x \log p_t(y|x)}{\bar{\beta}_t} = \nabla_x \log p(x) + \nabla_x \log (y|x)$$

$\frac{\varepsilon_\theta(x_t, t)}{\bar{\beta}_t}$

↓

只需将 $\varepsilon_\theta(x_t, t) \Rightarrow \varepsilon_\theta(x_t, t) - \bar{\beta}_t \nabla_x \log p_t(y|x)$

2. classifier-free: 定义 $p(x_{t-1}|x_t, y) = N(x_{t-1}; \mu(x_t, y), \sigma_t^2 I)$

$$\rightarrow \mu(x_t, y) = \frac{1}{\alpha_t}(x_t - \frac{\beta_t^2}{\bar{\beta}_t} \varepsilon_\theta(x_t, y, t))$$

$$\rightarrow L = E_{x_0, y \sim \tilde{p}(x_0, y), \varepsilon \sim N(0,1)} [\|\varepsilon - \varepsilon_\theta(x_t, y, t)\|^2]$$

为了平衡相关性和多样性, 引入 $w = \gamma - 1$

$$\hat{\varepsilon}_\theta(x_t, y, t) = (1+w)\varepsilon_\theta(x_t, y, t) - w\varepsilon_\theta(x_t, t)$$

新引入一个特定输入 ϕ, 对应图像为全体图像 (条件)

$$= (1+w)\varepsilon_\theta(x_t, y, t) - w\varepsilon_\theta(x_t, \phi, t)$$