3-12周报

> 本周工作：
>
> 1. Foundation Model 调研任务；
> 2. 点云（和 Mesh）数据集收集；
> 3. 寻找梯度估计的灵感：收集传统点云降噪论文和与3D梯度有关的深度论文；
> 4. 完善公式推导，~~找到了一个可尝试的修正了权重的新loss。~~ **证明了原loss更可靠。**
> 5. 对模型部分代码优化，在不影响效率下大大降低了GPU显存占用；
>
> 周报内容是对 工作4 的阐述。
>
> 关于上周提出的 **对于 $x_a^{(t)}$ 生成策略的修改**，训练实验结果显示不合理，且基于 $Nearest$ 生成的结果显然不通过 Jarque-Bera 检验。（注：原方法的中间变量均通过JB检验）

1. 公式推导与分析

> 对部分符号进行重定义，为了降低公式编写难度，例如 $x^{(t)}$ 重定义为了 x^t。

过去的 Loss 的依据是 Score-based，但 Diffusion的随机过程和 Langevin 过程存在区别，因此具体内容的计算上存在不同。

1.1. SBD Loss推导

定义 Diffusion Process 的分布描述：

$$q(x^{1:T}|x^0) = \prod_{t=1}^{T} q(x^t|x^{t-1}),\ q(x^t|x^{t-1}) = \mathcal{N}(x^t;\ \sqrt{1-\beta_t}x^{t-1},\ \beta_t\mathrm{I})$$

定义 Sampling Process 的分布描述：（带 θ 的为需要训练的

$$p_\theta(x^{0:T}|F_T) = p(x^T)\prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, F_T)$$

$$where\ p_\theta(x^{t-1}|x^t, F_T) = \mathcal{N}(x^t;\ \sqrt{1-\beta_t}x^{t-1},\ \beta_t\mathrm{I}),\ p(x^T) = \mathcal{N}(x^T; x^0, L_{noise}\mathrm{I}^3)$$

其中，$F_T = EdgeFeature(x^T)$

训练方法 p_θ 使用它的负对数似然估计的变分上界：（**绿色部分**为相比于上行公式**修改的部分**，目的是降低阅读难度）

$$
\begin{aligned}
\mathbb{E}[-\log p_\theta(x^0)] &\le \mathbb{E}_q\left[-\log\frac{p_\theta(x^{0:T}|F_T)}{q(x^{1:T}|x^0)}\right] \\
&= E_q\left[-\log p(x^T) - \sum_{t\ge 1}\log\frac{p_\theta(x^{t-1}|x^t,F_T)}{q(x^t|x^{t-1})}\right] \\
&= E_q\left[-\log p(x^T) - \sum_{t>1}\log\frac{p_\theta(x^{t-1}|x^t,F_T)}{q(x^t|x^{t-1})} - \log\frac{p_\theta(x^0|x^1,F_T)}{q(x^1|x^0)}\right]
\end{aligned}
$$

根据贝叶斯定理: $q(x^t|x^{t-1}) = \dfrac{q(x^{t-1}|x^t)q(x^t)}{q(x^{t-1})}$，但是 $q(x^{t-1}|x^t)$ 是无法直接处理的。又因 $q(x^{1:T}|x^0)$ 令 $q(x^t|x^{t-1},x^0) = q(x^t|x^{t-1})$ 满足，因此引入 x^0 作为条件 $q(x^t|x^{t-1}) = q(x^t|x^{t-1},x^0) = \dfrac{q(x^{t-1}|x^t,x^0)q(x^t|x^0)}{q(x^{t-1}|x^0)}$。

$$
\begin{aligned}
&= E_q\left[-\log p(x^T) - \sum_{t>1}\log\frac{p_\theta(x^{t-1}|x^t,F_T)}{q(x^{t-1}|x^t,x^0)}\cdot\frac{q(x^{t-1}|x^0)}{q(x^t|x^0)} - \log\frac{p_\theta(x^0|x^1,F_T)}{q(x^1|x^0)}\right] \\
\because\ &\sum_{t>1}\log\left[\frac{q(x^{t-1}|x^0)}{q(x^t|x^0)}\right] = \log q(x^1|x^0) - \log q(x^T|x^0) \\
\therefore\ &= E_q\left[-\log\frac{p(x^T)}{q(x^T|x^0)} - \sum_{t>1}\log\frac{p_\theta(x^{t-1}|x^t,F_T)}{q(x^{t-1}|x^t,x^0)} - \log p_\theta(x^0|x^1,F_T)\right] \\
&= E_q\left[D_{KL}(q(x^T|x^0)\,||\,p(x^T)) + \sum_{t>1}D_{KL}(q(x^{t-1}|x^t,x^0)\,||\,p_\theta(x^{t-1}|x^t,F_T)) - \log p_\theta(x^0|x^1,F_T)\right] \\
&= D_{KL}(q(x^T|x^0)\,||\,p(x^T)) + E_q\left[\sum_{t>1}D_{KL}(q(x^{t-1}|x^t,x^0)\,||\,p_\theta(x^{t-1}|x^t,F_T))\right] - \log p_\theta(x^0|x^1,F_T) \\
&=: loss
\end{aligned}
$$

其中，红色项显然是常数项，对Loss的下降并不会起到任何作用，因此带 F_T 的 Diffusion 最优化问题可描述为:

$$
\begin{aligned}
Simplify \Rightarrow loss &= E_q\left[\sum_{t>1}D_{KL}(q(x^{t-1}|x^t,x^0)\,||\,p_\theta(x^{t-1}|x^t,F_T))\right] \\
&\Leftrightarrow \arg\min_\theta E_q\left[\sum_{t>1}D_{KL}(q(x^{t-1}|x^t,x^0)\,||\,p_\theta(x^{t-1}|x^t,F_T))\right]
\end{aligned}
$$

结论:

- 引入 F_T 只影响 $q(x^{t-1}|x^t,x^0)$ 和 $p_\theta(x^{t-1}|x^t,F_T)$ 的相对熵;

1.1.1. 引入 Score-based 得到目标 Loss

对于 Diffusion process 来说:

$$
q(x^{t-1}|x^t,x^0) = \mathcal{N}(x^{t-1};\tilde{\mu}(x^t,x^0),\tilde{\beta}_t)
$$

分解得到: $(note: \bar{\alpha}_t = \prod_{i=1}^T \alpha_i, \alpha_t = 1-\beta_t)$

$$
\begin{aligned}
q(x^{t-1}|x^t,x^0) &= q(x^t|x^{t-1},x^0)\frac{q(x^{t-1}|x^0)}{q(x^t|x^0)} \\
&= \mathcal{N}(x^t;\sqrt{\alpha_t}x^{t-1},\beta_t\mathrm{I})\frac{\mathcal{N}(x^{t-1};\sqrt{\bar{\alpha}_{t-1}}x^0,(1-\bar{\alpha}_{t-1})\mathrm{I})}{\mathcal{N}(x^t;\sqrt{\bar{\alpha}_t}x^0,(1-\bar{\alpha}_t)\mathrm{I})} \\
&\propto \exp\left(-\frac{1}{2}(\frac{(x^t-\sqrt{\alpha_t}x^{t-1})^2}{\beta_t}+\frac{(x^{t-1}-\sqrt{\bar{\alpha}_{t-1}}x^0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(x^t-\sqrt{\bar{\alpha}_t}x^0)^2}{1-\bar{\alpha}_t})\right) \\
&= \exp\left(-\frac{1}{2}((\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})(x^{t-1})^2-2(\frac{\sqrt{\alpha_t}}{\beta_t}x^t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x^0)x^{t-1}+C(x^t,x^0))\right)
\end{aligned}
$$

由此可以解得：

$$
\begin{aligned}
\tilde{\beta}_t &= \frac{1}{\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \\
\tilde{\mu}(x^t,x^0) &= \frac{\frac{\sqrt{\alpha_t}}{\beta_t}x^t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x^0}{\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}} \\
&= \frac{\sqrt{\alpha_t}1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}x^t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x^0 \\
Reparameterizing &\Rightarrow \frac{1}{\sqrt{\alpha_t}}(x^t(x^0,z)-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z), z\sim\mathcal{N}(0,\mathrm{I})
\end{aligned}
$$

其中，$x^t(x^0,z)=\sqrt{\bar{\alpha}_t}x^0+\sqrt{1-\bar{\alpha}_t}z$

这里开始，我**引入 Score-based** 作为 p_θ 中计算梯度的模型，把上面的最优化问题进行限制。定义 Score-based 计算梯度：

$$
\sqrt{\bar{\alpha}_t}\nabla_x log[s_\theta(x_a^{t-1}|x_a^t,F_T)] \approx -z_\theta(x^t,F_T) \propto \min\{||x_i^0-x_a^t||_2^2\ |x_i^0\in x^0\},\ x_a^t=\frac{x^t}{\sqrt{\bar{\alpha}_t}}
$$

Sampling Process 同理得到：

$$
\begin{aligned}
p_\theta(x^{t-1}|x^t,F_T) &= \mathcal{N}(x^{t-1};\mu_\theta(x^t,t,F_T),\Sigma_\theta(x^t,t)),\Sigma_\theta(x^t,t)=\sigma^2\mathrm{I}=\tilde{\beta}_t^2\mathrm{I} \\
\Rightarrow \mu_\theta(x^t,t,F_T) &= \frac{1}{\sqrt{\alpha_t}}(x^t(x^0,z)-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z_\theta(x^t,F_T))
\end{aligned}
$$

对于两个高斯分布的 KL 散度来说（借鉴VAE的推导），我们可以得到如下推导：

$$
\begin{aligned}
D_{KL}(q\,||\,p_\theta) &= D_{KL}(\mathcal{N}(x^{t-1};\tilde{\mu}(x^t,x^0),\tilde{\beta}_t)||\mathcal{N}(x^{t-1};\mu_\theta(x^t,t,F_T),\Sigma_\theta(x^t,t))) \\
&= \frac{1}{2}(n+\frac{1}{\tilde{\beta}_t^2}||\tilde{\mu}(x^t,x^0)-\mu_\theta(x^t,t,F_T)||^2-n+\log 1) \\
&= \frac{1}{2\tilde{\beta}_t^2}||\tilde{\mu}(x^t,x^0)-\mu_\theta(x^t,t,F_T)||^2 \\
&= \frac{1}{2\tilde{\beta}_t^2}\left|\left|\left(\frac{1}{\sqrt{\alpha_t}}(x^t(x^0,z)-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z)\right)-\left(\frac{1}{\sqrt{\alpha_t}}(x^t(x^0,z)-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z_\theta(x^t,F_T))\right)\right|\right|_2^2 \\
&= \frac{1}{2\tilde{\beta}_t^2}\left|\left|\frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\left(z_\theta(x^t,F_T)-z\right)\right|\right|_2^2 \\
\because p(x;\mu,\Sigma) &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}} \propto p(x;\mu,\sigma^2)=\frac{1}{\sqrt{(2\pi)^n}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, when\ \Sigma=\sigma^2\mathrm{I} \\
\therefore \quad &\Rightarrow \frac{1}{2\tilde{\beta}_t^2}\left|\left|\frac{\beta_t\sqrt{\bar{\alpha}_t}}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\left(\nabla_x log[s_\theta(x_a^{t-1}|x_a^t,F_T)]-\nabla_x q(x_a^t)\right)\right|\right|_2^2
\end{aligned}
$$

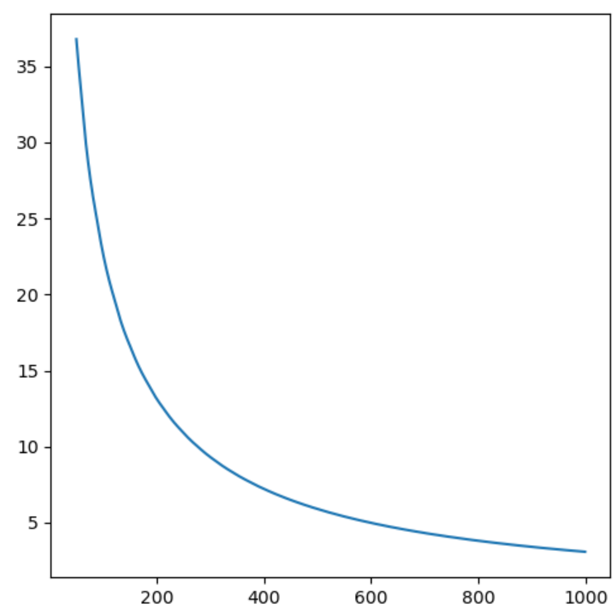
其中，$\nabla_x q(x_a^t)$ 为我们通过算法估计的梯度方向，目前是个待改善内容。

综上所述：

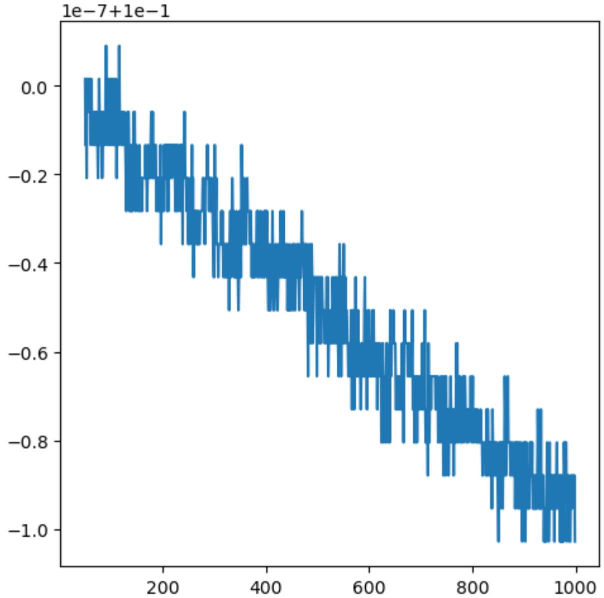
$$
\mathcal{L}(x_a^{0:T},\{\beta_i\}_{i=1}^T)=\sum_{t>1}\frac{1}{2\tilde{\beta}_t^2}\mathbb{E}_q\left[\left|\left|\frac{\beta_t\sqrt{\bar{\alpha}_t}}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\left(\nabla_x log[s_\theta(x_a^{t-1}|x_a^t,F_T)]-\nabla_x q(x_a^t)\right)\right|\right|_2^2\right]
$$

结论：

- loss形式上和当前使用的loss一致，但增加了跟噪声方差有关的权重项。
- loss形式上和DDPM的差不多。



可视化权重项可见（上图），这个权重大小和噪声方差负相关。将这个权重与真实噪声方差相乘（下图），得到的结果是稳定在 0.1 ± 10^{-7} 的一个数值。由此不难侧面证明这个权重项的作用是把噪声项归一化。



1.2. 缺陷

通过梯度估计算法提取的“噪声样本”可以理解为一个真实梯度混合了错误噪声的样本。那么对于噪声程度较小的样本来说，错误噪声在噪声样本中的占比的偏高，若把它归一化，那么这个错误就会被很大程度地放大，最终干扰到loss下降结果。

解决方法有两个：

- 使用更高精度的梯度估计算法；
- 训练时避免训练噪声程度过小的样本。

1.3. 进一步优化

权重的作用是让loss更加关注细节，但对于点云来说，局部Patch的点数不足以支持梯度估计算法对细节的微小噪声进行提取（或许这里我可以尝试通过假设检验算出一个置信度）。但是我们可以在 $\mathcal{L}(x_a^{0:T}, \{\beta_i\}_{i=1}^{T})$ 的基础上继续得到一个变分上界。

$$
\begin{aligned}
\mathcal{L}(x_a^{0:T}, \{\beta_i\}_{i=1}^{T}) &\leq \sum_{t>1} \mathbb{E}_q\left[\left|\left|\nabla_x log[s_\theta(x_a^{t-1}|x_a^t, F_T)] - \nabla_x q(x_a^t)\right|\right|_2^2\right] \\
&\leq \mathbb{E}_{t,q}\left[\left|\left|\nabla_x log[s_\theta(x_a^{t-1}|x_a^t, F_T)] - \nabla_x q(x_a^t)\right|\right|_2^2\right] \\
&=: \mathcal{L}_{simple}(x_a^{0:T}, \{\beta_i\}_{i=1}^{T})
\end{aligned}
$$

最后这个公式就是原先我们训练使用的公式。当 $\mathcal{L}_{simple}(x_a^{0:T}, \{\beta_i\}_{i=1}^{T}) \to 0$，则 $\mathcal{L}(x_a^{0:T}, \{\beta_i\}_{i=1}^{T}) \to 0$ 显然成立。

2. 总结

丰富了Loss推导的理论过程，并从数学上证明了原Loss的正确性与可靠性。