arXiv:2603.02452v1 [cs.LG] 2 Mar 2026

流形感知去噪分数匹配（MAD）

Alona Levy-Jurgenson [1] **Alvaro Prat** [1] **James Cuin** [2] **Yee Whye Teh** [1]

摘要

设计用于学习定义在流形上的分布的方法时，一个主要关注点是减轻隐式学习流形的需求，使学习能够专注于流形内的数据分布。然而，实现这一目标往往会导致计算密集型方案。在这项工作中，我们对环境空间中的去噪得分匹配作出简单修改，以隐式考虑流形，从而减轻学习流形的负担，同时保持计算效率。具体而言，我们提出将得分函数简单分解为一个已知分量 s^{base} 和一个剩余分量 $s - s^{base}$ （学习目标）；前者隐式包含有关数据流形所在位置的信息。我们以解析形式推导了若干重要情形下的已知分量 s^{base}，包括旋转矩阵上的分布和离散分布，并用这些情形展示了此方法的实用价值。

1. 引言

使用去噪得分匹配（DSM）训练的基于得分的生成模型（SGM）为生成建模提供了一种高效且可扩展的方法（Song & Ermon，2019；Song 等人，2020）。然而，许多具有实际应用价值的数据模态都支撑在低维流形上。这包括 SO(3) 中的旋转，这类旋转与药物设计和机器人学相关（Senanayake & Ramos，2018；Zhang 等人，2024；Prat 等人，2025），也包括地质学和气候数据（Karpatne 等人，2018）。此外，离散数据也值得关注；这类数据在文本生成（Shi 等人，2024；Amin 等人，2025；Feng 等人，2025）和生物信息学（Huang 等人，2024）中十分常见，其支撑集位于嵌入在高维空间中的低维集合上。在这类情形下，假设数据在环境空间中具有全支撑集的标准 DSM，实际上需要同时学习流形及其上的分布。与目标分布在整个环境空间中均有支撑集的情形相比，这会带来更具挑战性的学习问题。

为此，一些方法构建了能够显式考虑流形结构的生成模型，例如定义流形上的生成过程（Rozen 等人，2021；Huang 等人，2022年；De Bortoli 等，2022；Prat 等人，2025），其中分数函数通过流形的切空间定义；或者在低维设置中应用欧几里得生成模型，使用图册等方法使维数与流形的内禀维数相匹配，再映射回流形（Gemici 等人，2016；Kalatzis 等人，2021）。前一种方法虽然有效，但训练时计算开销可能很大，所采用的采样方案通常需要进行精细离散化，以便在曲率较高的流形上移动（De Bortoli 等，2022；Lou 等人，2023）。另一方面，后一种方法往往需要多次映射，这也可能引入失真，或导致结果依赖于所选映射（Gemici 等人，2016；Brehmer & Cranmer，2020；Kalatzis 等人，2021）。

相比之下，在环境空间中进行 DSM 计算效率高，且易于部署。此外，在欧几里得空间中操作放宽了路径的几何约束：轨迹不必受限于沿着流形上的弯曲测地线前进，而是可以走更直接的“弦状”路线。值得注意的是，对于嵌入流形 $d_{\mathcal{M}}(x,\ y)\ \geq ||x-y||_2$ (Lee，2012）。然而，环境空间扩散仅通过训练数据隐式地考虑流形结构。因此，许多环境空间概率流模型一直难以生成连贯的流形内结构，原因正是这一问题（例如（Buttenschoen 等人，2024））。对流形结构信息的缺失会影响学习动力学——近期研究发现，扩散模型，尤其是 DSM，必须先恢复数据分布的支撑集，才能学习任何关于密度的信息（Li 等人，2025）。综合来看，这些发现引出了一个自然的问题：能否改进环境空间中的 DSM，使其考虑支撑结构，同时保留其简洁性并提高学习效率?

[1]牛津大学统计系 [2]伦敦帝国理工学院数学系，英国伦敦。通讯作者：Alona Levy- Jurgenson <alona.jurgenson@stats.ox.ac.uk>、Yee Whye Teh<y.w.teh@stats.ox.ac.uk>。

预印本。*2026*年*3*月*4*日。

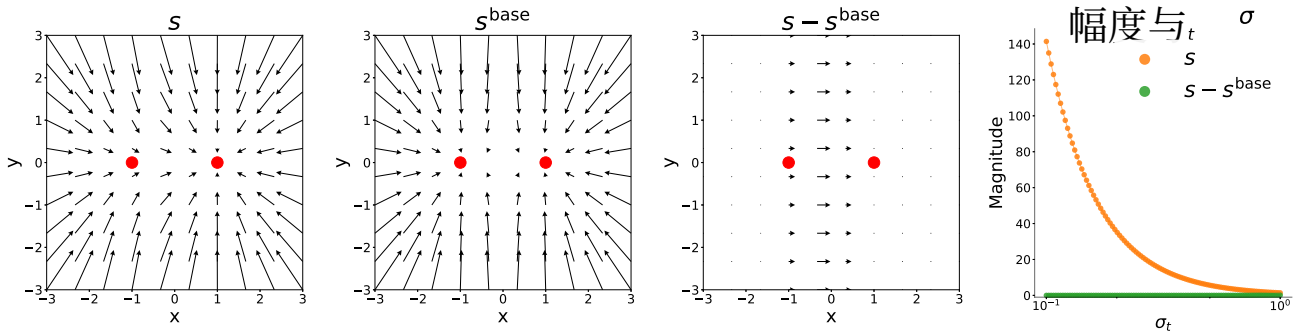


图 *1.* 通过 $\mathcal{M} = \{(-1,0),(1,0)\}$（红色）上的玩具离散分布介绍MAD，分别带有 $p = \{0.1, 0.9\}$以及 $p^{base} = \{0.5, 0.5\}$。图中展示了实际分数 s、基准分数 s^{base} 以及差值 $s - s^{base}$ 在 $\sigma_t = 0.8$下的向量场。 s（DSM的学习目标）与 $s - s^{base}$（MAD的学习目标）的幅度随 σ_t 变化的关系针对 $x =$(1,0)给出。

在这项工作中，我们通过将分数分解为两个分量来回答这一问题：一个已知的基准分数，用于捕捉流形结构；以及一个未知的残差项，它取决于目标分布。我们将这种方法称为流形感知去噪分数匹配 (MAD)。图1展示了我们的方法。我们对一系列重要流形推导出解析的基准分数，包括旋转、n维球面和离散数据。由于基准分数已知，学习过程可以专注于残差这一依赖分布的分量。我们表明，MAD 能够加快收敛并提高分布保真度，同时达到与计算量更大的替代方法相当、在某些情况下甚至更好的性能，且不牺牲标准 DSM 的效率和简洁性。我们还展示了这种分解如何在离散情形下为得分匹配带来更稳定的学习目标，从而无需学习分数函数；随着噪声水平降低，该函数会发散。

2. 方法

形式上，我们希望学习一个扩散模型，其目标分布支撑在某个 $n \in \mathbb{N}_{>0}$的子集 $\mathcal{M} \subseteq \mathbb{R}^n$上。我们假设在 $\mathcal{M}$ 上有一个简单且易于处理的基准概率测度 μ，而目标分布 p （存在一定记号重叠）关于 μ绝对连续。例如，我们可以取 $\mathcal{M}$为一个已知的紧致黎曼流形， μ 为 $\mathcal{M}$ 上归一化的均匀Haar测度；或者取一个有限集， μ 为 $\mathcal{M}$上的均匀计数测度。

2.1. 背景：扩散模型

扩散模型由一个向样本添加噪声的前向过程和一个对样本去噪以恢复原始分布的反向过程定义，其中样本是按照 p 抽取的。设 x_t 为时间 t、噪声方差为 σ_t^2时的含噪样本。

在我们的设置中，我们使用方差爆炸型随机微分方程（Song 等人，2020年）。记 $g_t^2 := \frac{d\sigma_t^2}{dt}$ 和 dW_t 为$\mathbb{R}^n$中的标准维纳过程。前向过程为布朗运动：

$$x_0 \sim p \tag{1}$$

$$dx_t = g_t dW_t \tag{2}$$

反向过程由一个随机微分方程（SDE）给出，

$$dx_t = g_t^2 \nabla_{x_t} \log p_t(x_t) dt + g_t dW_t, \tag{3}$$

其中分数函数为，

$$\begin{aligned} s(x_t, t) &:= \nabla_{x_t} \log p_t(x_t) \\ &= \nabla_{x_t} \log \int N_{\sigma_t}(x_t - x_0) p(x_0) \mu(dx_0) \\ &= \frac{\mathbb{E}[x_0 | x_t] - x_t}{\sigma_t^2} \end{aligned} \tag{4}$$

$$N_{\sigma_t}(x_t - x_0) := \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{\|x_t - x_0\|^2}{2\sigma_t^2}\right) \tag{5}$$

通常使用神经网络来近似分数函数 $s(x_t, t) \approx s_\theta(x_t, t)$，其中 θ 通常通过得分匹配进行学习：

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, x_t, t}\left[\left\| \sigma_t s_\theta(x_t, t) - \frac{x_0 - x_t}{\sigma_t} \right\|_2^2\right]. \tag{6}$$

其中，根据（Song & Ermon，2019，章节 4.2），出于数值方面的原因，得分匹配损失由 σ_t^2 加权。

2.2. 分数分解

我们方法背后的核心思想是，将定义在环境空间中的时间相关分数函数 $s(x_t,\ t)$ 分解为两个分量：

$$s(x_t, t) = s^{base}(x_t, t) + \delta(x_t, t) \tag{7}$$

第一项 s^{base} 是简单基分布 μ 的分数函数，而第二项 δ 是未知分量，也是我们的学习目标。我们假设 s^{base} 已知且可进行解析处理，并将在下文通过若干示例明确推导该项。由于 μ 是均匀分布，且支撑集为 $\mathcal{M}$，因此 s^{base} 实际上捕捉了 $\mathcal{M}$ 的几何结构。因此，δ 实际上捕捉了支撑在 $\mathcal{M}$ 上的目标分布 p 的性质，而无需隐式编码 $\mathcal{M}$ 的几何结构。

一个自然的问题是，为什么学习 δ 可能比学习原始分数函数 s 更容易；我们的实验结果也证实了这一点。直观来说，分数函数 s 同时捕捉了 $\mathcal{M}$ 的几何结构和目标分布 p，而 δ 只捕捉 p，因为几何结构已由 s^{base} 捕捉，因此学习 δ 应该更容易，因为需要学习的内容更少。

另一个有用的直觉如下：$s(x_t, t)$在其定义域的大部分范围内都应表现良好，因为对于 $t > 0$，加噪数据的分布是光滑的。例外情况是，当 $t \approx 0$（等价地，$\sigma_t \approx 0$）时，$s(x_t, t)$可能取很大的值。这是因为在反向随机微分方程中，分数函数需要迫使 x_t 收敛到 $\mathcal{M}$，因为 $t \to 0$。对于 s^{base}，情况也一样；事实上，$s(x_t, t)$和 $s^{base}(x_t, t)$都应发挥类似作用，迫使 x_t 收敛到 $\mathcal{M}$ 上同一个最近点，前提是 p 在局部看起来与 μ 相似。因此，当 $t \approx 0$ 时，$\delta(x_t, t)$——即 $s(x_t, t)$ 与 $s^{base}(x_t, t)$ 之差——其取值应远小于二者各自的取值。

我们可以在离散分布这一简单情形下用数学方式证明这一点：

定理 2.1. 设 $N \in \mathbb{N}$，$N > 1$且 $\mathcal{M} = \{u_i\}_{i=1}^N \subset \mathbb{R}^n$ 对于某个 $n \in \mathbb{N}$。设 μ 为均匀归一化的[1]在 $\mathcal{M}$ 上的计数测度，且 p 是关于 μ 的密度函数，其支撑集为 $\mathcal{M}$。设 $x \in \mathbb{R}^n$ 使得 x 与 $\mathcal{M}$ 中的所有点都不等距。当 $t \to 0$（等价地，$\sigma_t \to 0$）时：

$$\|s(x,t) - s^{base}(x,t)\| \to 0 \tag{8}$$

证明见附录 B.1。定理 2.1表明，学习目标 $\delta(x, t)$ 按 $o(1)$ 的尺度变化 $\sigma_t \to 0$。结合已知的离散基准分数（见下文第 2.3 章节），这一结果从理论上使得我们能够将给定 x 对应的分数函数恢复到误差在 $o(1)$ 以内，且该误差随 $\sigma_t \to 0$ 变化。这一点很重要，因为（Li 等人，2025）指出，要恢复数据密度，学习到的分数函数必须与真实分数函数的误差在 $o(1)$ 以内，且误差随 $\sigma_t \to 0$变化。

我们现在可以通过参数化来对 $s(x_t, t)$ 进行参数化 $\delta(x_t, t) \approx \delta_\theta(x_t, t)$，得到：

$$s(x_t,t) \approx s^{base}(x_t,t) + \delta_\theta(x_t,t) \tag{9}$$

为了学习 θ，我们可以调整之前的损失函数（6），得到如下损失函数：

$$\mathcal{L}(\theta) = \tag{10}$$
$$\mathbb{E}_{x_0,x_t,t}\left[\left\|\sigma_t\delta_\theta(x_t,t) - \left(\frac{x_0 - x_t}{\sigma_t} - \sigma_t s^{base}(x_t,t)\right)\right\|_2^2\right]$$

实际上，环境空间 DSM 学习会得到略带噪声的 p（Song &Ermon，2019，章节 3.1），通常记为 $p_{\sigma_{\min}}$，也就是说，我们学习的是 $p_{\sigma_{\min}}$ 而不是p_0。因此，作为生成样本的最后一步，为了确保样本位于流形上，除非另有说明（例如图 7），我们会将环境空间中的生成样本投影到流形上。

2.3. 示例

下面我们推导几个基准分数函数，这些函数也将在结果章节中使用。所有证明均见附录 B。附录图 8展示了用于基准分数的分布的实证结果，其中的样本完全使用基准分数生成。

2.3.1. 离散分布

命题 2.2.（离散分布的基准分数）对于支撑在有限点集 $\mathcal{M} = \{u_i\}_{i=1}^N$上的均匀离散分布，

$$s^{base}(x_t,t) = \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x_t - u_i)}{\sum_{i=1}^N N_{\sigma_t}(x_t - u_i)} \tag{11}$$

2.3.2. 球面上的分布

下面我们推导一般 n球面，特别是二维球面上的基准分数。这也将用于推导三维旋转上分布的基准分数。

命题 2.3.（n球面的基准分数）假设 $n \in \mathbb{N}_{>0}$。令 $\mathcal{M} = \mathcal{S}^n(1) := \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$，且$\mu$ 为 $\mathcal{M}$ 上的归一化球面测度（通常记作 σ^n）。那么对于 $x_t \neq 0$,

$$s^{base}(x_t,t) = -\frac{x_t}{\sigma_t^2} + \frac{1-n}{2}\frac{x_t}{\|x_t\|^2} + \tag{12}$$
$$\frac{1}{2I_{\frac{n-1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)}\left(I_{\frac{n-3}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right) + I_{\frac{n+1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)\right)\frac{x_t}{\sigma_t^2\|x_t\|}$$

其中，I_k 表示第一类修正贝塞尔函数，其阶数为 k。

[1]请注意，在图 1的图注中使用了非归一化计数测度，但对其归一化对应形式也可以得到一种等价但不太直观的形式。

根据上述结果，我们可以得到二维球面上分布的以下简化表达式。

推论 2.4. （2-球面的基准分数）对于 $\mathcal{M} = \mathcal{S}^2(1) = \{x \in \mathbb{R}^3 : \ ||x|| = 1\}$ 以及 μ 在 $\mathcal{S}^2(1)$ 上的通常归一化球面测度，

$$s^{base}(x_t, t) = x_t \left(-\frac{1}{\sigma_t^2} - \frac{1}{\|x_t\|^2} + \frac{1}{\sigma_t^2 \|x_t\| \tanh \frac{\|x_t\|}{\sigma_t^2}} \right) \tag{13}$$

这也提供了一些直观理解：通过重新排列各项，我们可以看到：

$$s^{base}(x_t, t) = \frac{-x_t + \frac{x_t}{\|x_t\| \tanh \frac{\|x_t\|}{\sigma_t^2}} - \frac{\sigma_t^2 x_t}{\|x_t\|^2}}{\sigma_t^2}$$

因此根据方程4 我们可以推出：

$$\mathbb{E}^{base}[x_0 | x_t] = \frac{x_t}{\|x_t\| \tanh \frac{\|x_t\|}{\sigma_t^2}} - \frac{\sigma_t^2 x_t}{\|x_t\|^2} \tag{14}$$

因此，随着 $\sigma_t \to 0$，期望值会简单地收敛到投影到 $\mathcal{S}^2$上的扰动样本，这对于均匀分布而言是合理的。

2.3.3. 三维旋转上的分布

另一个值得关注的情形是 SO(3) 中的旋转，它们可以表示为 $\mathcal{S}^3$ 中的单位向量（严格来说，是$\mathbb{R}^4$ 中的四元数——参见附录 A）。在这种情况下，我们可以使用定理 2.3，令 $n = 3$，并进一步简化表达式，使其仅包含阶数为0和 1的第一类修正贝塞尔函数；这两种函数在深度学习框架中的实现更为常见：

推论 2.5。 （3-球面的基准分数）对于 $\mathcal{M} = \mathcal{S}^3(1) = \{x \in \mathbb{R}^4 : \ ||x|| = 1\}$ 以及 μ 在 $\mathcal{S}^3(1)$ 上的通常归一化球面测度，

$$s^{base}(x_t, t) = -\frac{x_t}{\sigma_t^2} + \frac{1-n}{2} \frac{x_t}{\|x_t\|^2} + \left(\frac{I_0\left(\frac{\|x_t\|}{\sigma_t^2}\right)}{I_1\left(\frac{\|x_t\|}{\sigma_t^2}\right)} - \frac{\sigma_t^2}{\|x_t\|} \right) \frac{x_t}{\sigma_t^2 \|x_t\|} \tag{15}$$

请注意，这种表示是通过从 $\mathcal{S}^3$ 到 SO(3) 的二对一映射实现的，其中两个对跖点 x 和 $-x$ 对应于同一个旋转。

奇偶等变性。为了考虑旋转群的双重覆盖几何，$\mathcal{S}^3$ 的分布必须在对跖对称性下保持不变。因此，对于所有 $x \in \mathcal{S}^3$，$p(x) = p(-x)$。由此，分数函数 $s(x_t, t)$ 必须满足奇偶等变性：$s(-x_t, t) = -s(x_t, t)$，也就是说，它关于对跖点是奇函数。为了从结构上确保这一点，我们可以通过反对称化（δ_θ，利用奇偶等变函数的通用参数化，来参数化残差（De Bortoli 等人，2022年）：

$$\delta(x_t, t) \approx \frac{1}{2} \Big(f_\theta(x_t, t) - f_\theta(-x_t, t) \Big) \tag{16}$$

其中，$f_\theta(x_t, t)$ 是一个神经网络。最后，很容易证明，式 (15) 中的基准分数具有奇偶等变性，即 $s^{\text{base}}(-x_t, t) = -s^{\text{base}}(x_t, t)$。因此，通过组合 $s_\theta(x_t, t) := \delta_\theta(x_t, t) + s^{\text{base}}(x_t, t)$ 定义的分数场具有严格的奇偶等变性。

通过规范化实现商空间扩散。 在存在旋转对称性的情况下，学习三维旋转的条件分布会遇到一个关键问题：**不可辨识性**。具体而言，后验分布 $p(R \mid c)$ 通常具有多峰性，其中 c 是任意条件变量（例如图像或分子图）：多个不同的旋转对应同一物理状态，因此也对应同一观测。形式上，对于具有离散对称群 $G \subset \text{SO(3)}$ 的物体，观测映射在 G 的右作用下保持不变，即对于所有 $g \in G$，$c(R) = c(Rg)$。在理想化的无噪声极限下，后验分布集中于真实姿态的对称轨道上，并可写成狄拉克测度的均匀混合，

$$p(R \mid c) = \frac{1}{|G|} \sum_{g \in G} \delta\big(R_{\text{gt}}\, g\big), \tag{17}$$

其中 $R_{\text{gt}} \in \text{SO(3)}$ 是与 c一致的任意代表元，而具有物理意义的姿态是等价类$[R_{\text{gt}}] := \{R_{\text{gt}} g : \ g \in G\}$。

标准去噪分数匹配训练网络来近似条件分数 $s(R_t, t \mid c) = \nabla_{R_t} \log\, p_t(R_t \mid c)$。当$p(R \mid c)$呈多模态时，来自不同对称相关模态的扩散轨迹会在高噪声水平下重叠，因此同一噪声状态 R_t 可能对应多个彼此不兼容的去噪目标。在环境空间参数化中（例如 $\mathbb{R}^4$ 中的四元数），这会引发梯度冲突，并推动模型趋向各模态的欧几里得平均；此类平均通常落在低密度区域，且未必对应有效旋转，从而产生退化的**幽灵旋转**。

问题的根源在于，具有物理意义的变量并非 $R \in \text{SO}(3)$，而是等价类 $[R] \in \text{SO}(3)/G$。因此，我们在商空间$\mathcal{M}_q = \text{SO}(3)/G$上重新表述这个问题。在该空间中，每个对称轨道都收缩为一个点，条件分布也因此变为单峰分布。实际操作中，我们会在注入噪声之前，对每个真实姿态进行规范化，也就是从轨道 $\{Rg$ 中 $g \in G\}$在固定基本域内选取一个确定性代表。附录 A.1给出了具体的基于四元数的规范化规则及相关提升过程。

3. 结果

我们在多个基准测试上评估所提出的方法，并将其与流形方法和环境空间方法进行比较。为确保评估条件可控，我们在每项任务中使用相同的训练样本集，并让每种方法以相同的批次大小、相同（或相近）的神经网络架构运行相同的步数。我们采用一种适用于流形的、基于样本的指标——最大均值差异（MMD）（Mehraban& Pich-ler，2025），并采用与（De Bortoli 等人，2022年）相同的方法，以避免比较不同的流形对数似然近似方法，并避免环境空间方法中相关的对数似然病态问题（参见（Loaiza-Ganem 等人，2024））。完整的数据和训练规格见附录D。

结果概要。我们的结果揭示了三个主要发现：（1）与其他方法相比，我们的方法和 DSM 的最大均值差异相近，有时甚至更优，同时收敛和采样速度更快（表1、2、3），这表明环境空间方法在这一领域具有很强的竞争力；（2）与 DSM 相比，MAD 收敛更快（图4），且损失函数始终更低（附录图 9），同时保持了简洁性。这表明，环境空间 DSM 确实可以通过有效改造来纳入支撑集几何信息，从而使学习能够聚焦于其中的分布（例如图 2g）；（3）在离散实验中，正如定理 2.1所预示的，MAD 能够恢复真实分布，而 DSM 则难以做到，常常生成分布外样本。这些发现凸显了 MAD 在低维设置中的实际优势。

3.1. $\mathcal{S}^2$中的地球数据

我们从以点的形式表示于 $\mathcal{S}^2$ 中的地球数据开始（附录C.1）。结果如图2 和表1所示。从表1 中我们观察到，所有方法要么具有相当的最大均值差异（例如火山，最小的数据集），要么MAD和DSM表现略好（例如火灾，最大的数据集）。总体而言，地球实验证实，与流形上方法相比，MAD可以获得更好的最大均值差异，并且能够在捕捉更清晰的分布细节的同时与DSM相当（图2g-2i）。

3.2. SO(3)中的旋转

接下来，我们将探讨一个更复杂的任务——生成SO(3)旋转矩阵，并以四元数表示S^3。我们测试了SO(3) 上的高斯混合，实验设置与（De Bortoli 等人，2022年）中的实验相同，分量数量逐渐增加，从$K = 16$ 个分量（最简单）到 $K = 64$ 个分量（最难）。如表2所示，对于 $K = 16$，RSGM 的表现优于所有方法。不过，随着分布复杂度增加，其最大均值差异也随之增大，直至

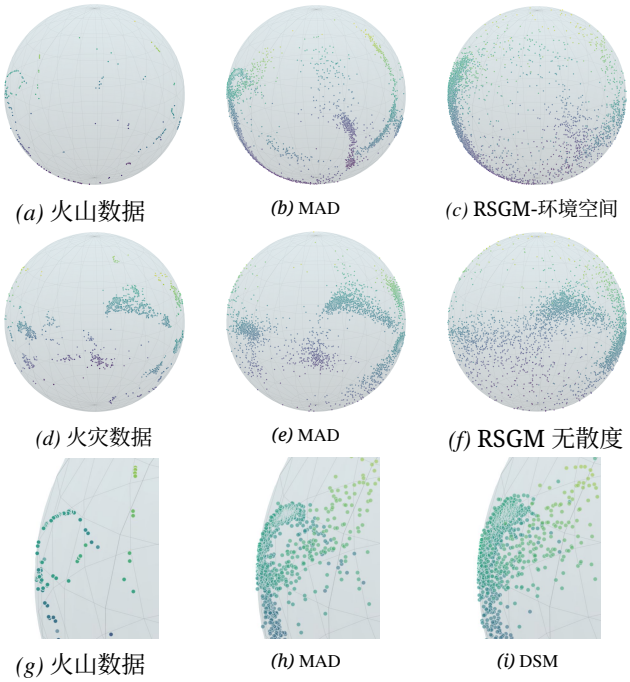


图2. 地球数据（左图）与不同方法的生成样本的对比。由极轴定义的色图仅保留用于可视化。

分布复杂度增加，直到所有方法在 $K = 64$时表现相当。值得注意的是，MAD 的收敛速度最快（图4），且与 DSM 相比几乎没有增加采样成本（表2）。这表明，引入基准分数能够改善优化效果，同时不牺牲环境空间采样的简便性。图3 展示了不同方法之间的定性差异：与 DSM 相比，MAD 对不同分量的区分略好；MAD 与 RSGM 的区分效果则相近。

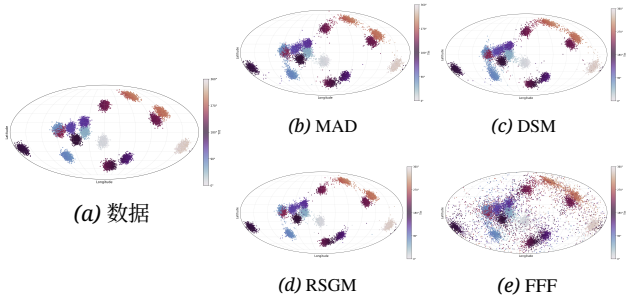


图 3. 不同方法和数据下，$K = 16$ 的 SO(3) 结果以摩尔魏德投影的形式呈现。所有图均来自随机种子 0。点的颜色根据角度而非分量编码。

SymSol：SO(3) 及其对称性。我们研究具有离散

表 *1.* 地球 MMD 结果　　（均值 ± 标准差）根据 1000 个样本与可用测试集大小之间的最小值计算。

方法	火山	地震	洪水	FIRE
RSGM-DIVFREE	0.1272 ± 0.0289	0.0602 ± 0.0101	0.0579 ± 0.0094	0.0779 ± 0.0172
RSGM-AMBIENT	0.1114 ± 0.0090	0.0803 ± 0.0067	0.0751 ± 0.0087	0.0830 ± 0.0152
DSM	0.1082 ± 0.0217	0.0488 ± 0.0096	0.0503 ± 0.0092	0.0456 ± 0.0165
MAD（我们的方法）	0.1030 ± 0.0185	0.0526 ± 0.0108	0.0500 ± 0.0100	0.0452 ± 0.0131
数据规模	827	6120	4875	12809

表 *2*。SO(3)最大均值差异（均值　± 标准差）使用 5,000 个样本计算生成集合与测试集之间的差异，以及生成 1K 个样本的采样时间　s.

方法	$K = 16$	$K = 32$	$K = 64$	采样时间（秒/1K）
FFF	0.0230 ± 0.0023	0.0242 ± 0.0016	0.0204 ± 0.0018	0.0040 ± 0.0003
RSGM	0.0155 ± 0.0025	0.0171 ± 0.0024	0.0190 ± 0.0043	0.4628 ± 0.0190
DSM	0.0222 ± 0.0021	0.0209 ± 0.0046	0.0178 ± 0.0031	0.2216 ± 0.0350
MAD（我们的方法）	0.0229 ± 0.0026	0.0220 ± 0.0052	0.0180 ± 0.0029	0.2305 ± 0.0360

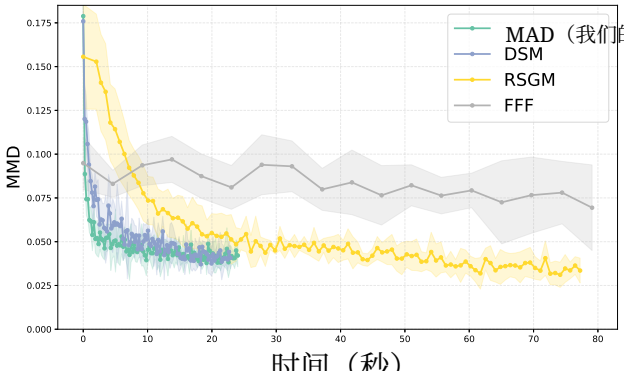


图 *4.* 最大均值差异与累积训练步时间的对比。阴影区域为 5 次运行的 ± 标准差。最大均值差异是针对 $K = 64$ 的训练集计算的，为提高效率使用了 1,000 个样本（表格使用 5,000 个）。请注意，为了更清晰地比较所有方法，所示时间最多到 80，但 FFF 在 400 之后结束。

旋转对称性的物体的三维旋转条件生成建模。具体而言，给定刚体的一张二维图像 c，目标是学习其姿态上的条件分布，$p(R|c)$其中 $R \in \mathrm{SO}(3)$，并从该后验分布中生成样本。我们的实验使用 SYMSOL I 数据集（Murphy 等人，2022年），该数据集包含简单对称实体（例如圆柱体、立方体、二十面体）的二维图像及其对应的真实旋转（见图5）。由于这些形状中许多都具有非平凡的对称群 $G \subset \mathrm{SO}(3)$，例如，圆柱体的二维图像可以由两个不同的旋转表示，因此，对称实体的条件姿态分布通常不可辨识。因此，我们采用2.3.3中所述的商空间规范化方法。条件去噪器的设计和训练详情见附录D.3。由于 SYMSOL I 提供了真实旋转，我们按建议计算均值 离散度 [2] （见（Liu 等人，2023；

[2]Spread 评估最小期望平均角偏差 $\mathbb{E}_{R\sim p(R|I)}[\min_{R'\in\{R_{gt}g|g\in G\}} d(R,R')]$,，其中 $d(R,R')$

Murphy 等人, 2022年) 作为我们的关键指标。

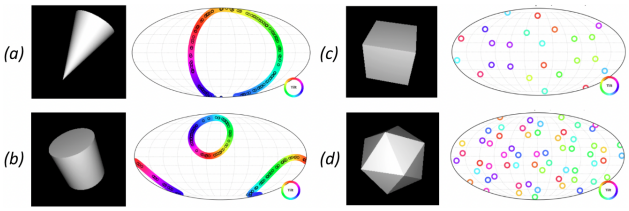


图 *5.* 随机旋转的圆锥、圆柱、立方体、二十面体（a-d）的二维图像快照，以及 SO(3) 中相应姿态分布的莫尔威德投影。（a-b）中的黑色圆圈表示来自 MAD 的样本，为视觉清晰起见，在（c-d）中予以省略。

对于所有形状，MAD 的表现均优于 DSM，再次展现出更快的收敛速度（图6）。MAD 在到达流形方面也有更可靠的保证，在各个样本上的流形偏差始终较低。值得注意的是，对于具有高阶对称性的复杂形状，例如立方体（$|G| = 24$）和二十面体（$|G| = 60$），MAD 的表现优于 DSM，这表明 MAD 将更多优化工作用于学习原本相当复杂的分布。最后，我们在表3中证明，MAD 的表现与先前方法相当，这凸显出我们的分数分解和商空间规范化能够高效而简单地解决 SO(3)中的复杂分布问题。

3.3. 离散数据

最后，一项日益重要但也颇为复杂的任务是学习从离散分布中生成样本。近期，人们对使用扩散模型完成这项任务表现出浓厚兴趣，主要是在大语言模型的背景下（例如（Shi 等人，2024））。离散分布也代表了低维结构的一种极端情况，此时支撑

表示测地距离。

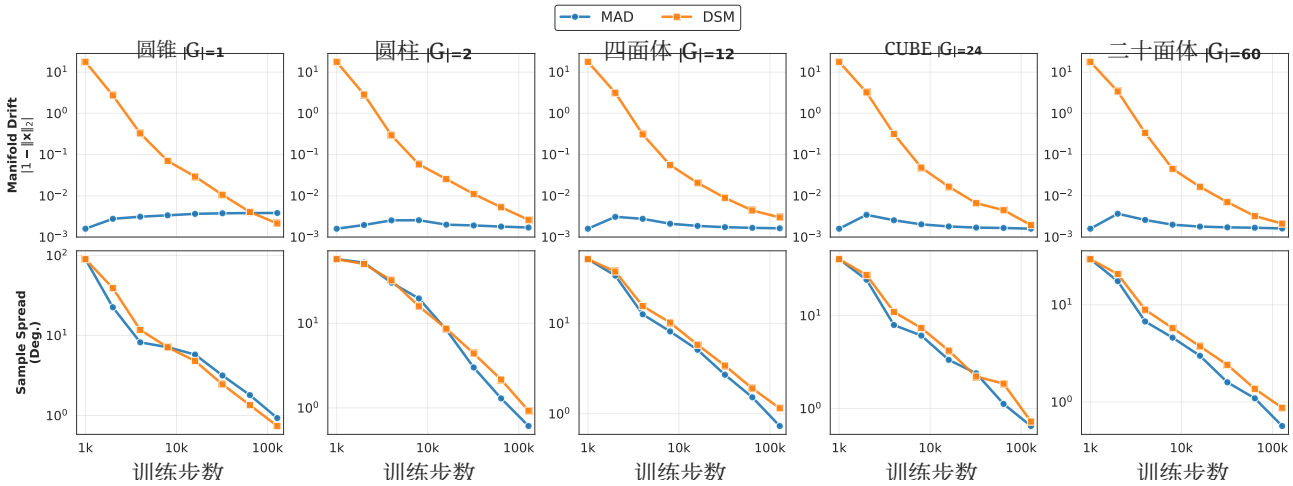
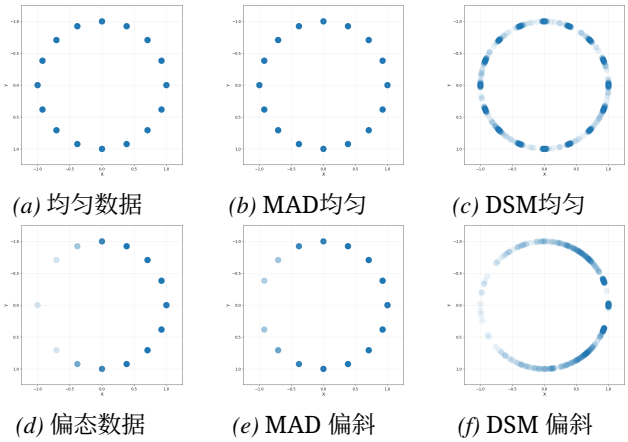


图6. 上排：流形漂移（$|1-\|x\|_2|$）。MAD（蓝色）的漂移始终较低，表明预测结果更接近有效旋转。DSM（橙色）的漂移较高，说明存在“幽灵旋转”平均现象。下排：样本离散度（越低越好）。

表3。S 在 S 上进行扩散估计YMSOL I. 数值单位为度（↓）。基线结果取自Liu 等人（202 3).

方法	CONE	CYL.	TET.	CUBE	ICO.	AVG.
DENG 等人. (2020年)	10.1	15.2	16.7	40.7	29.5	22.4
IPDF（MURPHYETAL。，2022年）	1.4	1.4	4.6	4.0	8.4	4.0
DNF（LIU ETAL。，2023）	**0.5**	**0.5**	**0.6**	**0.6**	1.1	**0.7**
DSM + 商	0.7	0.9	1.2	0.7	0.9	0.9
MAD + 商（我们的方法）	0.9	0.6	0.7	**0.6**	**0.6**	**0.7**

集由孤立点构成。这一设置可直接检验我们的方法融入支撑信息的能力。我们使用单位圆上离散点集上的一个分布，如图 7 所示。请注意，我们不会将样本投影到支撑集上，以便观察生成样本与支撑集之间的距离。首先，我们考虑一个简单任务，假设分布为均匀分布（图 7a）。不出所料，我们的方法能够轻松学习该分布，因为基准分数与真实得分之差为零，因而只有基准分数会起作用。相比之下，DSM 会在各支撑点之间生成分布外样本。接下来，我们将目标分布设为偏态分布，以增加任务难度（图 7d和附录C.3）。此时，两者的差异更加明显：MAD 生成的样本与目标分布高度相似，而 DSM 则会产生许多分布外样本。根据定理 2.1，MAD 能够生成接近支撑集的样本，这一点并不意外；定理表明，随着 σ_t 减小，分数中剩余的分量将近似为 0。这也与（Li 等人，2025）的研究发现有关：要恢复真实分布，学习得到的得分函数与真实得分函数之间的误差必须小于 $o(1)$ 。



(a) 均匀数据 *(b)* MAD均匀 *(c)* DSM均匀

(d) 偏态数据 *(e)* MAD 偏斜 *(f)* DSM 偏斜

图 7. 生成离散分布——均匀（上）与偏斜（下）。数据（左）以及各方法的生成样本（中、右）。样本未投影到支撑集上。

4. 相关工作

4.1. 流形上方法

流形上方法会直接在数据流形上对生成过程进行显式建模，通常通过在流形几何结构中定义常微分方程或随机微分方程（Falorsi&Forr´e，2020年；Lou 等人，2020年；Mathieu & Nickel，2020年；

Rozen 等人，2021；Huang 等人，2022年；Ben-Hamu 等人，2022年；De Bortoli 等人，2022年；Chen & Lipman，2023）。这些方法通常假设流形是黎曼流形，因为它们需要使用切空间。近期具有代表性的方法包括黎曼分数生成模型（RSGMs（De Bortoli 等人，2022年）），这类方法在黎曼流形上定义随机微分方程，并通过测地线随机游走对其进行近似。尽管这类方法能够生成高质量样本，但通常会在训练和采样期间带来额外计算开销。此外，这些方法要求流形满足黎曼流形假设。尽管该假设适用范围广，但会限制它们在离散空间中的应用。

4.2. 低维欧几里得方法

与流形上类别中的方法不同，这些方法并不显式定义流形上的生成过程，例如基于流形的随机微分方程或常微分方程；而是将欧几里得生成过程应用于低维欧几里得空间，并利用图册或其他手段（Gemici 等人，2016；Brehmer & Cranmer，2020年；Kalatzis 等人，2021）将各空间中生成的样本映射回流形。这通常需要进行多次映射，可能会增加优化的复杂性。此外，这些方法通常假设流形是光滑的，但实际情况未必如此。使用此类映射还可能引入失真，或者因映射的选择不同而产生不同的生成分布。最后，由于低维流形将由单个点构成，这些方法无法直接应用于离散情形。

4.3. 环境空间方法

我们将那些先在环境空间中生成样本、再将其从环境空间投影到流形上的方法归为这一类别。将 DSM 应用于低维流形时，可以把它视为此类方法。具体而言，如（Song &Ermon，2019）所述，将 DSM 应用于低维流形时，假设目标分布经过轻微加噪，从而得到其全维近似。学习这一近似后，还需将样本从环境空间投影到流形上，才能获得流形上的样本。这种方法计算速度快且易于部署，不过，环境空间中的近似必须足够准确且接近流形，投影才能取得良好效果。由于 DSM 只能通过目标分布提供的样本来感知流形，因此必须先花费一定精力学习流形本身，正如（Li 等人，2025）所示。

流形自由形式流 (FFF) (Sorrenson 等人，2024) 也可视为环境空间方法，因为它们依赖于编码器–解码器架构，在环境空间中生成样本后再进行投影到流形上。尽管 FFF 能够快速采样，但它同样假设存在黎曼流形结构，并引入多个辅助损失项，其中一项用于鼓励环境空间样本与投影后的样本彼此接近。这些额外目标增加了训练的复杂性。

MAD 属于环境空间类别，但不同之处在于，它通过已知的基准分数融入了对流形的先验知识。这些信息会在采样过程中推动样本朝流形移动；与其他环境空间方法相比，这可能减轻学习流形的负担，同时不会引入额外的几何假设或辅助训练目标。

5. 讨论与局限性

本工作通过使用基准分数，使 DSM 能够显式考虑流形结构。我们的实验表明，MAD 不仅提高了收敛速度，还能捕捉更精细的分布细节。这可能归因于我们的分数分解有效地将流形支撑集的学习与该支撑集上的密度学习解耦，并表明 MAD 对（Li 等人，2025）所观察到的现象提供了部分回应。具体而言，通过基准分数编码支撑集的几何结构，MAD 绕过了标准去噪分数匹配所需的“支撑集恢复”第一阶段，这一优势在复杂分布（例如偏斜离散集和对称实体）中最为明显。

我们将 MAD 与环境空间 DSM 进行比较的结果表明，MAD 能使生成样本更接近流形。基准分数起到全局吸引子的作用，确保生成过程遵循流形结构。这表明，在数据稀疏性使隐式流形恢复变得困难或数据稀缺的场景中，MAD 可能尤其有价值。

我们方法的一个局限是依赖解析基准分数。未来，我们计划探究近似是否可作为可行的替代方案。尽管我们侧重于低维基准，但我们的理论结果适用于内禀维数任意的流形（$n > 0$）。在高维场景以及更复杂的真实世界数据集上评估 MAD——在这些场景中，流形感知或许能带来更大的益处——是未来研究的重要方向。

致谢

本研究部分由工程与自然科学研究委员会（EPSRC）通过生成模型人工智能中心 [资助，项目编号为 EP/Y028805/1], ；另获 Isambard-AI 资助。该项目由布里斯托大学运营，并由英国科学、创新与技术部（DSIT）通过英国研究与创新署提供资金；此外还获科学技术设施委员会资助（ST/AIRR/I-A-I/1023）。ALJ 衷心感谢 Google DeepMind 和 Schmidt Sciences 的支持。AP 获得 EPSRC [2928912] 和阿斯利康的支持，资助形式为机器学习哲学博士 iCASE奖项。JC 通过现代统计与统计机器学习（StatML）博士培训中心项目获得 EPSRC 资助，项目编号为 EP/S023151/1。

影响声明

本文旨在改进流形上的基于分数的生成模型的学习。我们的工作或许有助于提升药物设计、地球与气候科学以及文本生成等重要领域中的生成能力，但同时也可能加剧生成能力的滥用所带来的影响。在这方面，我们工作的伦理与社会影响与其他生成模型以及机器学习方法总体上的影响相似。

参考文献

Amin, A. N., Gruver, N., 和 Wilson, A. G. 为什么掩码扩散有效：以跳跃调度为条件，改进离散扩散。*arXiv* 预印本 *arXiv:2506.08316*，2025。

Andrews, G. E., Askey, R. 和 Roy, R. 贝塞尔函数与合流超几何函数，第 187–239 页。数学及其应用百科全书。剑桥大学出版社，1999年。

Ben-Hamu, H., Cohen, S., Bose, J., Amos, B., Grover, A., Nickel, M., Chen, R. T. 和 Lipman, Y. 流形上的归一化流与概率路径匹配。*arXiv* 预印本 *arXiv:2207.04711*，2022年。

Brakenridge, G. 全球大型洪水事件活跃档案库。http://floodobservatory.colorado.edu/Archives/index.html, 2017年。

Brehmer, J. 和 Cranmer, K. 用于同时进行流形学习与密度估计的流。神经信息处理系统进展，33:442–453，2020年。

Buttenschoen, M.、Morris, G. M. 和 Deane, C. M. Posebusters：基于人工智能的对接方法未能生成物理上有效的姿态，或泛化到新的序列。化学科学，15:3130–3139，2024。doi: 10. 1039/D3SC04185A。网址 http://dx.doi.org/10.1039/D3SC04185A。

Chen, R. T. 和 Lipman, Y. 一般几何上的流匹配。*arXiv* 预印本 *arXiv:2302.03660*，2023。De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y. W. 和 Doucet, A. 黎曼基于分数的生成建模。神经信息处理系统进展，35:2406–2422，2022年。

Deng, H., Bui, M., Navab, N., Guibas, L., Ilic, S., 和 Birdal, T. 深度Bingham网络：应对姿态估计中的不确定性和歧义，2020年。网址：https://arxiv.org/abs/2012.11002。

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., 和 Fei-Fei, L. ImageNet：一个大型分层图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页，2009年。doi：10.1109/CVPR.2009.5206848。

DLMF。*NIST*数学函数数字图书馆。https://dlmf.nist.gov/，2025年03月15日发布的1.2.4版。网址：https://dlmf.nist.gov/10.29。由 F. W. J. Olver、A. B. Olde Daalhuis、D. W. Lozier、B. I. Schneider、R. F. Boisvert、C. W. Clark、B. R. Miller、B. V. Saunders、H. S. Cohl 和 M. A. McClain 编。

EOSDIS。由美国国家航空航天局地球科学数据与信息系统（ESDIS）运营的陆地、大气近实时 EOS（LANCE）系统。https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data，2020年。

Falorsi, L. 和 Forr´e, P. 流形上的神经常微分方程。*arXiv* 预印本 *arXiv:2006.06663*，2020年。

Feng, G.、Geng, Y.、Guan, J.、Wu, W.、Wang, L. 和 He, D. 扩散语言模型的理论优势与局限。*arXiv* 预印本 *arXiv:2502.09622*，2025。

Gemici, M. C.、Rezende, D. 和 Mohamed, S. 黎曼流形上的归一化流。*arXiv* 预印本 *arXiv:1611.02304*，2016。

何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习，2015年。网址 https://arxiv.org/abs/1512.03385。

Huang, C.-W., Aghajohari, M., Bose, J., Panangaden, P., and Courville, A. C. 黎曼扩散模型。神经信息处理系统进展，35: 2750–2761，2022年。

Huang, K., Yang, Y., Fu, K., Chu, Y., Cong, L., and Wang, M. 用于可控 RNA 序列生成的潜在扩散模型。*arXiv* 预印本 *arXiv:2409.09828*，2024。

Kalatzis, D., Ye, J. Z., Pouplin, A., Wohlert, J. 和 Hauberg, S. 使用归一化流在光滑流形上进行密度估计。*arXiv* 预印本 *arXiv:2106.03500*，2021。

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A. 和 Kumar, V. 面向地球科学的机器学习：挑战与机遇。*IEEE*知识与数据工程汇刊，31(8):1544–1554，2018。

Karras, T., Aittala, M., Aila, T. 和 Laine, S. 阐明基于扩散的生成模型的设计空间，2022。网址 https://arxiv.org/abs/2206.00364。

Lee, J. M. 流形上的积分，第 400–439 页。Springer New York，纽约，NY，2012。

Li, X., Shen, Z., Hsieh, Y.-P., 和 He, N. 当分数学会几何：流形假设下的速率差异。*arXiv* 预印本 *arXiv:2509.24912*，2025。

Liu, Y., Liu, H., Yin, Y., Wang, Y., Chen, B., 和 Wang, H. 深入探究 SO(3)流形上的离散归一化流及其在概率旋转建模中的应用，2023。网址 https://arxiv.org/abs/2304.03937。

Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., 和 Cresswell, J. C. 从流形假设的视角看深度生成模型：综述与新联系。*arXiv* 预印本 *arXiv:2404.02954*，2024。

Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化，2019。网址https://arxiv.org/abs/1711.05101。

Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S. N. 和 De Sa, C. M. 神经流形常微分方程。神经信息处理系统进展，33:17548–17558，2020年。

Lou, A., Xu, M. 和 Ermon, S. 黎曼扩散模型的扩展，2023。网址 https://arxiv.org/abs/2310.20030。

Lynch, K. M. 和 Park, F. C. 现代机器人学。剑桥大学出版社，2017年。

Mardia, K. V. 和 Jupp, P. E. 方向统计学。约翰威立，2009年。

Mathieu, E. 和 Nickel, M. 黎曼连续正规化流。神经信息处理系统进展，33:2503–2515，2020年。

Mehraban, Z. 和 Pichler, A. 最大¨均值¨ 差异距离下概率测度的量化。*arXiv* 预印本 *arXiv:2503.11868*，2025。Murphy, K.、Esteves, C.、Jampani, V.、Ramalingam, S. 和 Makadia, A. Implicit-PDF：旋转流形上概率分布的非参数表示，2022年。网址 https://arxiv.org/abs/2106.05965。（NGDC/WDS），N. G. D. C. . W. D. S. NCEI/WDS全球重大地震数据库。
h
t
t
ps://
www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153，2022a。

（NGDC/WDS），N. G.D. C.. W.D. S. NCEI/WDS全球重大火山喷发数据库。
https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G10147，2022b。

Prat, A., Zhang, L., Deane, C. M., Teh, Y. W., and Morris, G. M. SigmaDock：通过基于片段的SE(3)扩散解开分子对接中的纠缠，2025。
URLhttps://arxiv.org/abs/2511.04854。

Rozen, N., Grover, A., Nickel, M., and Lipman, Y. Moser流：流形上的基于散度的生成建模。神经信息处理系统进展，34:17669–17680，2021。

Senanayake, R. and Ramos, F. 方向网格图：动态环境中的多模态角不确定性建模。载于*2018 IEEE/RSJ* 智能机器人与系统国际会议（*IROS*），第3241–3248页。IEEE，2018年。

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. 面向离散数据的简化通用掩码扩散。神经信息处理系统进展，37:103131–103167，2024年。

Song, Y. 和 Ermon, S. 通过估计数据分布的梯度进行生成建模。神经信息处理系统进展，32，2019。
Song, Y.、Sohl-Dickstein, J.、Kingma, D. P.、Kumar, A.、Er- mon, S. 和 Poole, B. 通过随机微分方程进行基于分数的生成建模。*arXiv* 预印本 *arXiv:2011.13456*，2020年。

Sorrenson, P.、Draxler, F.、Rousselot, A.、Hummerich, S. 和 K¨othe, U. 使用自由形式流学习流形上的分布。神经信息处理系统进展，37：66961–66994，2024。

Zhang, Z.、Wang, M. 和Liu, Q. Flexsbdd：采用柔性蛋白质建模的基于结构的药物设计。进展

发表于神经信息处理系统，37:53918– 53944，2024。

A. SO(3) 的四元数表示

表示旋转的一种有效方法是使用单位四元数（也称为 versor（单位四元数））（关于此主题的实用入门介绍，参见 Lynch & Park, 2017）。三维旋转 $R \in \mathrm{SO}(3)$ 可由一个单位向量 $u = (u_1, u_2, u_3)$ 表示的旋转轴和角度 $\theta \in [0, \pi]$ 指定。随后，它可以恰好表示为两个 versor（单位四元数）： $x = \cos(\theta/2) + \sin(\theta/2)(u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k})$，以及它的对跖点 $-x$。所有 versor（单位四元数）构成拓扑群 $\mathcal{S}^3(1)$，其 Haar 测度是球面 σ^3 上的均匀分布。二对一映射 $f\colon \mathcal{S}^3(1) \to \mathrm{SO}(3)$ 将 $x, -x$ 映射回对应的旋转，同时保留其原始概率分布。

A.1. 四元数坐标中的$\mathrm{SO}(3)/G$规范化

推论 A.1 (SO(3) 的四元数规范化 $\mathrm{SO}(3)/G$)。 设 $G \subset \mathrm{SO}(3)$是一个通过右乘作用的有限对称群。用单位四元数 $q \in \mathcal{S}^3$ 表示旋转，并用单位四元数表示每个 $g \in G$ （为每个群元素固定选择一个符号），因此 q 的对称轨道为 $\{qg\colon \ g \in G\} \subset \mathcal{S}^3$。

定义规范化映射 $\psi\colon \ \mathcal{S}^3 \to \mathcal{F}$ 如下:

$$q_{\text{canon}} \ = \ \psi(q) \ = \text{取最大值}_{g \in G} \ |\mathrm{Re}(qg)|, \tag{18}$$

其中，$\mathrm{Re}(q)$ 表示 $q =$ 的标量分量（$w, \ x, \ y, \ z$），即 $\mathrm{Re}(q) = w$。最后，通过施加以下条件来消除四元数坐标的符号歧义:

$$\mathrm{Re}(q_{\text{canon}}) \geq 0. \tag{19}$$

因此， q_{canon} 是等价类 $[R] \in \mathrm{SO}(3)/G$ 的确定性代表，而在规范化目标上训练扩散模型，可以消除 $p(R \mid c)$ 中由对称性导致的多模态性；其中 c 是具有隐式对称性的观测，例如图像、分子群、点云等。

在推理时，可以通过抽取 $g \sim \mathrm{U}(G)$ 将样本映射回 SO(3)，并返回 $q_{\text{canon}} g$。

注释。 最大化 $|\mathrm{Re}(qg)|$，会选出与恒等旋转距离最近的对称等价代表，因为对于单位四元数 q，SO(3) 上到恒等旋转的测地距离为 $2 \arccos(|\mathrm{Re}(q)|)$。请注意，我们仅对单位四元数执行此操作，以将 q_{canon} 作为特定条件 c 下的参考目标。前向扩散过程采用标准的方差爆炸型随机微分方程。

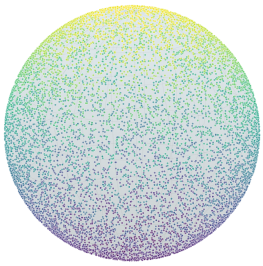


图 *8.* 展示 S^2 的基准分数——仅使用基准分数生成的样本（均匀分布在 $\mathcal{S}^2$ 上的分数函数），其中 $\sigma_{\min} = 1e^{-4}$。样本未投影到流形上。颜色根据 z 轴取值编码，仅为便于可视化。

B. 证明

B.1. 定理的证明2.1

定理 B.1. 设 $N \in \mathbb{N}$、 $N > 1$和 $\mathcal{M} = \{u_i\}_{i=1}^{N} \subset \mathbb{R}^n$，其中某个 $n \in \mathbb{N}_{>0}$。设 μ 是 $\mathcal{M}$ 上的均匀归一化计数测度， p 是关于 μ 的密度函数，其支撑集为 $\mathcal{M}$。设 $x \in \mathbb{R}^n$，使得 x 到 $\mathcal{M}$ 中的所有点的距离并不相等。那么，当 $t \to 0$($\sigma_t \to 0$) 时:

$$\|s(x,t) - s^{base}(x,t)\| \to 0 \tag{20}$$

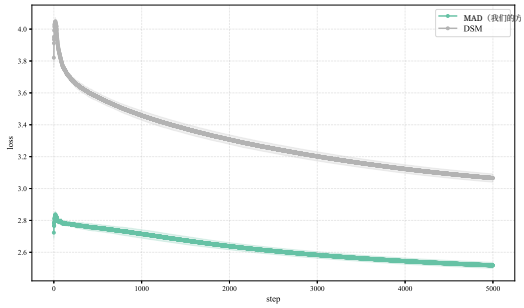


图*9.* 与 DSM 相比，MAD 的训练损失在整个训练过程中始终更低。损失函数针对 SO(3) 及 $K = 64$。阴影区域表示 5 个随机种子下 ±标准差。

证明。 我们使用以下对所有 $x \in \mathbb{R}^n$均成立的恒等式：

$$\begin{aligned} \|s(x,t) - s^{base}(x,t)\| &= \left\| \frac{\mathbb{E}[x_0|x] - x}{\sigma_t^2} - \frac{\mathbb{E}^{base}[x_0|x] - x}{\sigma_t^2} \right\| \\ &= \left\| \frac{\mathbb{E}[x_0|x] - \mathbb{E}^{base}[x_0|x]}{\sigma_t^2} \right\| \end{aligned} \tag{21}$$

因此，我们希望证明：

$$\lim_{\sigma_t \to 0} \frac{1}{\sigma_t^2} \left\| \mathbb{E}[x_0|x] - \mathbb{E}^{base}[x_0|x] \right\| = 0 \tag{22}$$

在我们的情况下：

$$\mathbb{E}[x_0|x] = \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x - u_i) p(u_i)}{\sum_{i=1}^N N_{\sigma_t}(x - u_i) p(u_i)} \tag{23}$$

$$\mathbb{E}^{base}[x_0|x] = \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x - u_i)}{\sum_{i=1}^N N_{\sigma_t}(x - u_i)} \tag{24}$$

因此，我们的目标变为：

$$\lim_{\sigma_t \to 0} \frac{1}{\sigma_t^2} \left\| \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x - u_i) p(u_i)}{\sum_{i=1}^N N_{\sigma_t}(x - u_i) p(u_i)} - \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x - u_i)}{\sum_{i=1}^N N_{\sigma_t}(x - u_i)} \right\| = 0 \tag{25}$$

范数内的表达式可写为：

$$\sum_{i=1}^N u_i \left(\frac{N_{\sigma_t}(x - u_i) p(u_i)}{\sum_{i=1}^N N_{\sigma_t}(x - u_i) p(u_i)} - \frac{N_{\sigma_t}(x - u_i)}{\sum_{i=1}^N N_{\sigma_t}(x - u_i)} \right) \tag{26}$$

利用三角不等式，再结合集合 $\mathcal{M} = \{u_i\}_{i=1}^N$ 由 $C := \max_i \|u_i\|$ 有界这一事实，我们得到：

$$\left\|\sum_{i=1}^{N} u_i \left(\frac{N_{\sigma_t}(x-u_i)p(u_i)}{\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)p(u_i)} - \frac{N_{\sigma_t}(x-u_i)}{\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)} \right) \right\| \tag{27}$$

$$\leq C \sum_{i=1}^{N} \left| \frac{N_{\sigma_t}(x-u_i)p(u_i)}{\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)p(u_i)} - \frac{N_{\sigma_t}(x-u_i)}{\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)} \right| \tag{28}$$

因此，只需证明:

$$\lim_{\sigma_t \to 0} \frac{1}{\sigma_t^2} \sum_{i=1}^{N} \left| \frac{N_{\sigma_t}(x-u_i)p(u_i)}{\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)p(u_i)} - \frac{N_{\sigma_t}(x-u_i)}{\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)} \right| = 0 \tag{29}$$

记为 $\Delta_{u_i} = {}^{N_{\sigma_t}(x-u_i)p(u_i)\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)p(u_i)}\!\!-\!\!{}_{N_{\sigma_t}(x-u_i)\sum_{i=1}^{N} N_{\sigma_t}(x-u_i)}$，可进一步将其拆分为正项和负项，从而得到等价表达式:

$$\lim_{\sigma_t \to 0} \frac{1}{\sigma_t^2} \sum_{u_i \in U^+} \Delta_{u_i} + \sum_{u_i \in U^-} -\Delta_{u_i} = 0 \tag{30}$$

其中， U^+和 U^-分别表示满足 $\Delta_{u_i} > 0$和 $\Delta_{u_i} < 0$的点集 $u_i \in \mathcal{M}$。现在，我们将分别对这些和进行上界估计。首先注意到，由于分子非零，我们可以写为:

$$\Delta_{u_i} = \frac{1}{1 + \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)p(u_j)}{N_{\sigma_t}(u_i - x)p(u_i)}} - \frac{1}{1 + \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)}{N_{\sigma_t}(u_i - x)}} \tag{31}$$

For $u_i \in U^+$ 我们可以通过将所有 $p(u_j)$ 减小至 m 来对此给出上界: $= \min_{i \in [N]} p(u_i)$，从而得到 :

$$\Delta_{u_i} \leq \frac{1}{1 + \frac{m}{p(u_i)} \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)}{N_{\sigma_t}(u_i - x)}} - \frac{1}{1 + \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)}{N_{\sigma_t}(u_i - x)}} \tag{32}$$

对于 $u_i \in U^-$，则可通过将所有 $p(u_j)$ 增大至 M: $= \max_{i \in [N]} p(u_i)$ 来得到:

$$-\Delta_{u_i} \leq \frac{1}{1 + \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)}{N_{\sigma_t}(u_i - x)}} - \frac{1}{1 + \frac{M}{p(u_i)} \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)}{N_{\sigma_t}(u_i - x)}} \tag{33}$$

现在我们仅关注 $u_i \in U^+$，因为对 U^-的分析是对称的。

为简 洁起见，我们使用以下记号表示出现在两个分母中的比值 :

$$r_{-i,i} := \frac{\sum_{j \neq i} N_{\sigma_t}(u_j - x)}{N_{\sigma_t}(u_i - x)}$$

采用这一记法并通分后，式32 中的界变为:

$$\frac{r_{-i,i}\left(1 - \frac{m}{p(u_i)}\right)}{1 + r_{-i,i} + \frac{m}{p(u_i)} r_{-i,i} + \frac{m}{p(u_i)} r_{-i,i}^2} = \frac{1 - \frac{m}{p(u_i)}}{\frac{1}{r_{-i,i}} + 1 + \frac{m}{p(u_i)} + r_{-i,i} \frac{m}{p(u_i)}} \tag{34}$$

其中，最后一个表达式是将$r_{-i\text{除以}i} \neq 0$后得到的。将该表达式以及用同样方法得到的$u_i \in U^-$对应表达式代入式32，并与我们的目标结合，可知只需证明所得上界收敛于0，即：

$$\lim_{\sigma_t \to 0} \frac{1}{\sigma_t^2} \left(\sum_{u_i \in U^+} \frac{1 - \frac{m}{p(u_i)}}{\frac{1}{r_{-i,i}} + 1 + \frac{m}{p(u_i)} + r_{-i,i}\frac{m}{p(u_i)}} + \sum_{u_i \in U^-} \frac{\frac{M}{p(u_i)} - 1}{\frac{1}{r_{-i,i}} + \frac{M}{p(u_i)} + 1 + r_{-i,i}\frac{M}{p(u_i)}} \right) = 0 \tag{35}$$

同样，我们只关注第一个求和项，因为第二项的分析是对称的。由于分子关于 σ_t 是常数，只需关注分母，并证明对所有 $i \in [N]$:

$$\lim_{\sigma_t \to 0} \sigma_t^2 \left(\frac{1}{r_{-i,i}} + r_{-i,i}\frac{m}{p(u_i)} \right) = \infty \tag{36}$$

令 $i \in [N]$。注意，由于指数收敛快于多项式收敛，因此 $r_{-i,\ i}$ 的收敛占主导地位；又由于 $m \neq 0$（p 的支撑集是 $\mathcal{M}$），上述结论在 $\lim_{\sigma_t \to 0} r_{-i,\ i} \in \{\infty, 0\}$ 时成立（关键观察是，$r_{-i,\ i}$ 同时出现在分子和分母中，因此如果 $\lim_{\sigma_t \to 0} r_{-i,\ i} \in \{\infty, 0\}$，那么至少有一个涉及 $r_{-i,\ i}$ 的项收敛到 ∞）。现在我们将证明确实如此。

我们将 $r_{-i,\ i}$ 改写为:

$$r_{-i,i} = \sum_{j \neq i} \exp\left(\frac{-\|u_j - x\|^2 + \|u_i - x\|^2}{2\sigma_t^2} \right) \tag{37}$$

考虑求和中的单个项。它不收敛到 ∞,0 之一的唯一情况是分子为 0，这当且仅当 $\|u_j - x\|^2 = \|u_i - x\|^2$，也就是说，$x$ 与 u_i, u_j 等距；但根据假设，这种情况并不成立。 □

注意，条件 x 与 $\mathcal{M}$ 中的所有点并非等距，是一个必要条件。对于这种情况，考虑式22（即得到的等价目标 25），并令 x 与 $\mathcal{M}$ 中的所有点等距（例如，可以设想$S^0 \subset \mathbb{R}$，使得 $x = 0$）。结果是一个常数除以 σ_t 的极限；当 $\sigma_t \to 0$时，该极限趋于 ∞。这表明，为了提高数值稳定性，最好避免这些情况（尽管对于所有 t，相对于 p_t，这种情况发生的概率应为 0）。

B.2. 定理证明2.2

定理 B.2.（离散分布的基准分数）对于支撑在有限点集$\mathcal{M} = \{u_i\}_{i=1}^N$上的离散分布，

$$s^{base}(x_t, t) = \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x_t - u_i)}{\sum_{i=1}^N N_{\sigma_t}(x_t - u_i)} \tag{38}$$

证明。 对于支撑在有限个点 $\mathcal{M} = \{u_i\}_{i=1}^N$ 上的离散分布，我们有，

$$s^{base}(x_t, t) = \nabla_{x_t} \log \int N_{\sigma_t}(x_t - x_0)\mu(dx_0) \tag{39}$$

$$= \nabla_{x_t} \log \frac{1}{N} \sum_{i=1}^N N_{\sigma_t}(x_t - u_i) \tag{40}$$

$$= \frac{\mathbb{E}^{base}[x_0|x_t] - x_t}{\sigma_t^2} \tag{41}$$

$$\mathbb{E}^{base}[x_0|x_t] = \frac{\sum_{i=1}^N u_i N_{\sigma_t}(x_t - u_i)}{\sum_{i=1}^N N_{\sigma_t}(x_t - u_i)} \tag{42}$$

□

这是在均值为 u_i、方差为 σ_t^2 的高斯混合模型下，高斯均值的后验均值。

B.3. 定理2.3的证明

定理 B.3。（n-球面的基准分数）假设 $n \in \mathbb{N}_{>0}$。令 $\mathcal{M} = \mathcal{S}^n(1) := \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$，且令 μ 为 $\mathcal{M}$ 上的归一化球面测度（通常记作 σ^n）。那么对于 $x_t \neq 0$,

$$s^{base}(x_t, t) = -\frac{x_t}{\sigma_t^2} + \frac{1-n}{2}\frac{x_t}{\|x_t\|^2} + \frac{1}{2I_{\frac{n-1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)}\left(I_{\frac{n-3}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right) + I_{\frac{n+1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)\right)\frac{x_t}{\sigma_t^2\|x_t\|} \tag{43}$$

其中，I_k 表示阶数为 k 的第一类修正贝塞尔函数。

证明。 我们推导出一个闭式解（只需使用第一类修正贝塞尔函数）：

$$s^{base}(x_t, t) = \nabla_{x_t} \log \int_{\mathcal{S}^n(1)} N_{\sigma_t}(x_t - x_0)\mu(dx_0) \tag{44}$$

$$\begin{aligned}\int_{\mathcal{S}^n(1)} N_{\sigma_t}(x_t - x_0)\mu(dx_0) &= (2\pi\sigma_t^2)^{-\frac{n+1}{2}} \int_{\mathcal{S}^n(1)} \exp\left(-\frac{\|x_t - x_0\|^2}{2\sigma_t^2}\right)\mu(dx_0) \\ &= (2\pi\sigma_t^2)^{-\frac{n+1}{2}} \int_{\mathcal{S}^n(1)} \exp\left(-\frac{\|x_t\|^2 - 2x_t \cdot x_0 + 1}{2\sigma_t^2}\right)\mu(dx_0) \\ &= (2\pi\sigma_t^2)^{-\frac{n+1}{2}} \exp\left(-\frac{\|x_t\|^2 + 1}{2\sigma_t^2}\right) \int_{\mathcal{S}^n(1)} \exp\left(\frac{x_t \cdot x_0}{\sigma_t^2}\right)\mu(dx_0)\end{aligned} \tag{45}$$

其中，在第二个等式中，我们用到了这一事实：$\|x_0\| = 1$。

记作：

$$J := \int_{\mathcal{S}^n(1)} \exp\left(\frac{x_t \cdot x_0}{\sigma_t^2}\right)\mu(dx_0) =$$

我们采用（Mardia 和 Jupp，2009）中的以下命题（为便于阅读，我们在此陈述）：

命题 B.4. 令 $k > 0$，$\|x\| = 1$。则：

$$\int_{\mathcal{S}^n(1)} e^{kx \cdot y} d\sigma(y) = \Gamma\left(\frac{n+1}{2}\right)\left(\frac{k}{2}\right)^{\frac{1-n}{2}} I_{\frac{n-1}{2}}(k)$$

其中 I 是第一类修正贝塞尔函数。

由于 $x_t \neq 0$，该结果适用于 J，其中 $k = \frac{\|x_t\|}{\sigma_t^2}$，$x = \frac{x_t}{\|x_t\|}$ 和 $y = x_0$，因此：

$$J = \Gamma\left(\frac{n+1}{2}\right)\left(\frac{\|x_t\|}{2\sigma_t^2}\right)^{\frac{1-n}{2}} I_{\frac{n-1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right) \tag{46}$$

代入 J 的表达式，我们得到：

$$\int_{\mathcal{S}^n(1)} N_{\sigma_t}(x_t - x_0)\mu(dx_0) = (2\pi\sigma_t^2)^{-\frac{n+1}{2}} \exp\left(-\frac{\|x_t\|^2+1}{2\sigma_t^2}\right)\Gamma\left(\frac{n+1}{2}\right)\left(\frac{\|x_t\|}{2\sigma_t^2}\right)^{\frac{1-n}{2}} I_{\frac{n-1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right) \tag{47}$$

因此：

$$\nabla_{x_t} \log \int_{\mathcal{S}^n(1)} N_{\sigma_t}(x_t - x_0)\mu(dx_0) = \underbrace{\nabla_{x_t}\left(-\frac{\|x_t\|^2+1}{2\sigma_t^2}\right)}_{\text{(a)}} + \underbrace{\nabla_{x_t} \log \left(\frac{\|x_t\|}{2\sigma_t^2}\right)^{\frac{1-n}{2}}}_{\text{(b)}} + \underbrace{\nabla_{x_t} \log I_{\frac{n-1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)}_{\text{(c)}} \tag{48}$$

and:

$$(a) = -\frac{x_t}{\sigma_t^2}$$

$$(b) = \frac{1-n}{2}\left(\frac{1}{\|x_t\|/2\sigma_t^2}\right)\left(\frac{x_t}{2\sigma_t^2\|x_t\|}\right) = \frac{1-n}{2}\frac{x_t}{\|x_t\|^2}$$

$$(c) = \frac{1}{I_{\frac{n-1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)}\frac{1}{2}\left(I_{\frac{n-3}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right) + I_{\frac{n+1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)\right)\frac{x_t}{\sigma_t^2\|x_t\|}$$

□

B.4. 定理的证明2.4

推论 B.5. （2-球面的基准分数）对于 $\mathcal{M} = \mathcal{S}^2(1) = \{x \in \mathbb{R}^3 : \ ||x|| = 1\}$ 以及 μ $\mathcal{S}^2(1)$上通常的归一化球面测度，

$$s^{base}(x_t, t) = x_t\left(-\frac{1}{\sigma_t^2} - \frac{1}{\|x_t\|^2} + \frac{1}{\sigma_t^2\|x_t\|\tanh\frac{\|x_t\|}{\sigma_t^2}}\right) \tag{49}$$

证明。 在这种情况下，我们可以得到 $s^{base}(x_t, t)$ 的闭式表达式，而无需第一类修正贝塞尔函数，因为 (c) 仅包含 $I_{\frac{1}{2}}, I_{-\frac{1}{2}}$和 $I_{\frac{3}{2}}$。前两者已知具有闭式表达式（Andrews 等人，1999年）：

$$I_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\sinh x \tag{50}$$

$$I_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\cosh x \tag{51}$$

并且 $I_{\frac{3}{2}}$ 可利用已知的递推关系由它们以闭式得到（DLMF ）：

$$2nI_n(x) = x\left(I_{n-1}(x) - I_{n+1}(x)\right) \tag{52}$$

具体而言，我们将递推公式应用于 $n = \frac{1}{2}$，从而得到对所有 $x \neq 0$：

$$\begin{aligned} I_{\frac{3}{2}}(x) &= I_{-\frac{1}{2}}(x) - \frac{1}{x}I_{\frac{1}{2}}(x) \\ &= \sqrt{\frac{2}{\pi x}}\cosh x - \frac{1}{x}\sqrt{\frac{2}{\pi x}}\sinh x \end{aligned} \tag{53}$$

综合这些结果，(c) 变为以下闭式表达式：

$$\begin{aligned}(c) &= \frac{1}{I_{\frac{1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)}\frac{1}{2}\left(I_{-\frac{1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)+I_{\frac{3}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)\right)\frac{x_t}{\sigma_t^2\|x_t\|}\\ &= \frac{1}{I_{\frac{1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)}\frac{1}{2}\left(2I_{-\frac{1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)-\frac{\sigma_t^2}{\|x_t\|}I_{\frac{1}{2}}\left(\frac{\|x_t\|}{\sigma_t^2}\right)\right)\frac{x_t}{\sigma_t^2\|x_t\|}\\ &= \left(\frac{\cosh\frac{\|x_t\|}{\sigma_t^2}}{\sinh\frac{\|x_t\|}{\sigma_t^2}}-\frac{\sigma_t^2}{2\|x_t\|}\right)\frac{x_t}{\sigma_t^2\|x_t\|}\\ &= \left(\frac{1}{\tanh\frac{\|x_t\|}{\sigma_t^2}}-\frac{\sigma_t^2}{2\|x_t\|}\right)\frac{x_t}{\sigma_t^2\|x_t\|}\end{aligned} \tag{54}$$

将其代入 (a) 和 (b)，并结合 $n=2$，可得对所有 $x_t\neq 0$ 成立的以下闭式解：

$$\begin{aligned}s^{base}(x_t,t) &= -\frac{x_t}{\sigma_t^2}-\frac{1}{2}\frac{x_t}{\|x_t\|^2}+\left(\frac{1}{\tanh\frac{\|x_t\|}{\sigma_t^2}}-\frac{\sigma_t^2}{2\|x_t\|}\right)\frac{x_t}{\sigma_t^2\|x_t\|}\\ &= -\frac{x_t}{\sigma_t^2}-\frac{x_t}{\|x_t\|^2}+\frac{1}{\tanh\frac{\|x_t\|}{\sigma_t^2}}\frac{x_t}{\sigma_t^2\|x_t\|}\\ &= x_t\left(-\frac{1}{\sigma_t^2}-\frac{1}{\|x_t\|^2}+\frac{1}{\sigma_t^2\|x_t\|\tanh\frac{\|x_t\|}{\sigma_t^2}}\right)\end{aligned} \tag{55}$$

□

B.5. 定理2.5 的证明

推论 B.6. （3-球面的基准分数）对于 $\mathcal{M}=\mathcal{S}^3(1)=\{x\in\mathbb{R}^4:\ ||x||=1\}$ 以及 μ 上通常的归一化球面测度(1)，

$$\begin{aligned}s^{base}(x_t,t) = -\frac{x_t}{\sigma_t^2}+\frac{1-n}{2}\frac{x_t}{\|x_t\|^2}+\\ \left(\frac{I_0\left(\frac{\|x_t\|}{\sigma_t^2}\right)}{I_1\left(\frac{\|x_t\|}{\sigma_t^2}\right)}-\frac{\sigma_t^2}{\|x_t\|}\right)\frac{x_t}{\sigma_t^2\|x_t\|}\end{aligned} \tag{56}$$

证明。请注意，差异在于求和中的第三项，该项在n中被记为表达式 (c) 包含了第一类修正贝塞尔函数 I_1，I_0，I_2。我们使用方程 52 中的递推关系，以 I_1，I_0表示解。具体而言，将 $n=1$ 代入递推关系后，对于所有 $x\neq 0$：

$$I_2(x)=I_0(x)-2\frac{I_1(x)}{x} \tag{57}$$

因此，(c) 变为：

$$
\begin{aligned}
(c) &= \frac{1}{I_1\left(\frac{\|x_t\|}{\sigma_t^2}\right)} \frac{1}{2}\left(I_0\left(\frac{\|x_t\|}{\sigma_t^2}\right) + I_2\left(\frac{\|x_t\|}{\sigma_t^2}\right)\right) \frac{x_t}{\sigma_t^2\|x_t\|} \\
&= \frac{1}{I_1\left(\frac{\|x_t\|}{\sigma_t^2}\right)} \left(I_0\left(\frac{\|x_t\|}{\sigma_t^2}\right) - \frac{\sigma_t^2 I_1\left(\frac{\|x_t\|}{\sigma_t^2}\right)}{\|x_t\|}\right) \frac{x_t}{\sigma_t^2\|x_t\|} \\
&= \left(\frac{I_0\left(\frac{\|x_t\|}{\sigma_t^2}\right)}{I_1\left(\frac{\|x_t\|}{\sigma_t^2}\right)} - \frac{\sigma_t^2}{\|x_t\|}\right) \frac{x_t}{\sigma_t^2\|x_t\|}
\end{aligned}
\tag{58}
$$

然后将其添加到 (a) 和 (b) 中（其中 $n = 3$ 位于 (b) 中），以获得结果。 □

C. 数据集

对于 S^2 地球数据和SO(3)合成数据，我们首先运行了RSGM，并保存了其在所有随机种子下运行过程中生成的精确训练集、验证集和测试集。这确保了所有方法看到的是同一数据集。DSM、MAD和FFF很容易适配以接受这些数据。

C.1. $\mathcal{S}^2$

我们使用地球与气候事件，并假设地球为一个2-球面。如上文一般说明所述，这些数据由RSGM使用，并由其训练代码划分为训练集、验证集和测试集。随后在训练期间进行重采样。我们保存了在所有训练步数中由重采样产生的所得训练集。因此，我们纳入了与RSGM相同的事件：火山喷发（，NGDC/WDS），地震（，NGDC/WDS），洪水（Brakenridge，2017年）和野火（EOSDIS，2020年）。请注意，洪水数据集的链接已不再有效，但这曾是该数据集最初报告的来源。

C.2. SO(3)

我们使用由 RSGM 生成的数据（这些数据以在线方式采样，每个批次均由包含 K 个分量的卷绕正态分布混合生成）。有关这些数据的生成方式，请参阅原论文。

C.3. 离散

我们生成两种类型：一种是在单位圆上均匀采样的 8 个等间距离散点，另一种（偏斜型）使用以下方法（我们也在代码库中发布了该方法）：

$$
\begin{aligned}
n_{\text{coords}} &= |\text{coordinates}| \\
i_{\text{peak}} &= \left\lfloor \frac{n_{\text{coords}}}{4} \right\rfloor \\
d_i &= \min(|i - i_{\text{peak}}|,\ n_{\text{coords}} - |i - i_{\text{peak}}|)\,, \quad i = 0, \ldots, n_{\text{coords}} - 1 \\
p_i &= \exp(-0.8\, d_i) \\
p_i &= \frac{p_i}{\sum_{j=0}^{n_{\text{coords}}-1} p_j} \\
\text{indices} &\sim \text{Multinomial}(p,\ n_{\text{samples}},\ \text{with replacement}) \\
\text{data} &= \text{coordinates}[\text{indices}]
\end{aligned}
$$

D. 训练规格

一般说明:

这些说明适用于 SO(3) 合成（而非物体）、地球和离散实验。

在所有实验中，DSM 和 MAD 使用的设置和超参数调优完全相同。唯一的差异来自基准分数的使用，具体体现在损失函数和采样阶段。为确保设置尽可能具有可比性，我们采用了与 RSGM 用于 SO(3) 的相同网络架构——由五个隐藏层组成的多层感知机，每层有 512 个神经元。我们使用了与 RSGM 相同的 Adam优化器。

所有关于 SO(3) 合成数据、离散数据和地球数据的实验均在标准 PC 上运行。

D.1. 地球实验

所有方法均运行了 2,000 步，批次大小为 512。

DSM 和 MAD ——正如本章节开头所述，除必要的差异外，所有设置均相同。这包括使用 $\sigma_{\min} \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ 进行超参数搜索，所用尺度数为 100，学习率为 $\{10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 7\times10^{-4}, 9\times10^{-4}, 10^{-3}\}$。以下列出了每种方法表现最佳的设置，并在结果章节中采用了这些设置。

MAD 的最佳结果如下：

- 火灾：学习率 $= 7\times10^{-4}, \sigma_{\text{最小值}} = 10^{-6}$
- 地震：学习率 $= 5\times10^{-4}, \sigma_{\text{最小值}} = 10^{-5}$
- 火山：学习率 $= 7\times10^{-4}, \sigma_{\text{最小值}} = 10^{-6}$
- 洪水：学习率 $= 9\times10^{-4}, \sigma_{\text{最小值}} = 10^{-6}$

DSM 的最佳结果如下：

- 火灾：学习率 $= 9\times10^{-4}$， $\sigma_{\text{分钟}} = 10^{-6}$
- 地震：学习率 $= 10^{-3}$， $\sigma_{\text{分钟}} = 10^{-6}$
- 火山：学习率 $= 7\times10^{-4}$， $\sigma_{\text{分钟}} = 10^{-6}$
- 洪水：学习率 $= 10^{-3}$， $\sigma_{\text{分钟}} = 10^{-6}$

RSGM – 我们使用了发表于（De Bortoli 等人，2022年）的公开代码库。我们采用了如下命令中的设置（对于 ’实验’ 的其他取值（例如 ’火灾’），以及 ’随机种子’ 处于 0-4（含）的情况，也同样如此）。对于所有其他设置，我们使用从相应配置文件加载的默认值。我们没有针对该基线进行广泛的超参数搜索，因为原始工作并未说明哪些超参数设置是关键的，而穷举搜索在计算上不可行。不过，我们仍对生成器测试了 ’无散’ 和 ’环境空间’ 两者，因为它们都被指定为可选项，并且我们认为生成器的质量可能是一个重要因素，而对其进行测试在计算上是可行的。

命令：
python main.py -m 实验=地震 模型=RSGM 生成器=divfree 损失函数=ism 步数=2000 种子=随机种子 验证频率=3000
请注意，由于验证频率并未用于早停，而仅用于训练期间的日志记录，我们将其设置为高于步数的值，以避免额外计算。

D.2.$\mathrm{SO}(3)$ 实验

所有方法均运行了 5,000 步，批次大小为 512。所有批次均使用 RSGM 的代码重新采样，并保存下来，以便其他方法使用完全相同的数据运行。这相当于 $5000 * 512 =\sim 2.5M$ 个独特样本。所有方法使用的样本完全相同。

DSM 和 MAD ——正如本章节开头所述，除必要的差异外，所有设置都相同。这包括使用 $\sigma_{\min} \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ 进行超参数搜索，以及 lr $\in \{3\times10^{-4}, 5\times10^{-4}, 7\times10^{-4}, 9\times10^{-4}, 10^{-3}\}$。使用的尺度数量为 100。以下是各方法表现最佳的设置，并且结果章节中使用的正是这些设置。

MAD 的最佳结果如下：

- K=16：学习率 $= 7 \times 10^{-4}$，$\sigma_{\text{最小值}} = 10^{-4}$
- K=32：学习率 $= 9 \times 10^{-4}$，$\sigma_{\text{最小值}} = 10^{-4}$
- K=64：学习率 $= 7 \times 10^{-4}$，$\sigma_{\text{最小值}} = 10^{-4}$

DSM 的最佳结果如下：

- K=16：学习率 $= 7 \times 10^{-4}$，$\sigma_{\text{分钟}} = 10^{-4}$
- K=32：学习率 $= 9 \times 10^{-4}$，$\sigma_{\text{分钟}} = 10^{-4}$
- K=64：学习率 $= 9 \times 10^{-4}$，$\sigma_{\text{分钟}} = 10^{-4}$

RSGM ——我们使用了地球数据相关上下文中提到的同一个代码库。我们从建议的两个较小学习率（5e-4、2e-4）中选择了第一个，并为另一个超参数 flow.beta 采用默认设置。我们使用的命令如下（针对三个 K 值和五个随机种子运行（下方的 SEED 为 0–4，含端点）：

python main.py -m 实验=so3 模型=rsgm 数据集.K=K 步骤=5000 损失函数=dsmv 生成器=李代数 op- tim.学习率=5 e-4随机种子=SEEDval 频率=50 -

其中，val freq 仅设为较低的 50，以便每 50 步记录一次训练步耗时。为提高效率，在其他情况下则将其设为大于 5000 的值。另请参阅附录 F 中的说明，了解我们如何计算训练步耗时以获得最佳受控设置。

FFF – 在修改 special orthogonal.py 中的 make so data 函数以加载数据集后，我们只需针对不同的 K 值和随机种子运行以下命令：

python -mlightning trainable.launcher.fit configs/m-fff/so3.yaml 数据 set.K=64 数据 set.stored 随机种子=随机种子 –name 'special-orthogonal'

D.3. SymSol：对称 SO(3) 实验

D.3.1. 架构

我们采用由两个联合训练的分量组成的条件分数生成模型：视觉编码器 E_ϕ和条件去噪网络 D_θ。为简洁起见，我们将二者统称为 s_θ。

视觉编码器。 为使生成过程以输入图像为条件，我们采用 ResNet-50 主干网络（He 等人，2015年），并使用在 ImageNet 数据集（Deng 等人，2009）上预训练的权重进行初始化。我们移除最后的全连接分类层，并从全局平均池化层提取 2048 维特征向量。我们将 ResNet 编码器与去噪器一起进行端到端联合微调。这使模型能够针对灰度图像上的三维姿态估计任务专门调整视觉特征；该任务还需要捕捉细微的几何线索（例如，区分圆柱的正面与背面），而这些线索在标准分类任务中可能具有不变性。投影头将视觉特征映射到隐藏维度 d_{model}。

分数网络的参数化。 我们使用神经网络对分数函数进行参数化 $s_\theta(q_t,\ \sigma_t,\ c)$ F_θ (残差多层感知机)，它会预测经过预条件处理的输出。遵循Karras 等人 (2022年的方法，我们通过 $q_{\text{in}} = q_t/\sqrt{\sigma_t^2 + 1}$对网络输入进行缩放，以保持单位方差。为了遵循 SO(3)的全局几何结构，我们必须考虑双覆盖性质：四元数 q 和 $-q$ 表示相同的旋转。这意味着分数场必须满足奇偶等变性，即$s_\theta(-q_t) = -s_\theta(q_t)$。我们通过对网络输出进行反对称化，在结构上强制满足这一约束：

$$F_\theta^{\text{odd}}(q_{\text{in}}, \sigma_t, c) = \frac{1}{2}\left(F_\theta(q_{\text{in}}, \sigma_t, c) - F_\theta(-q_{\text{in}}, \sigma_t, c)\right).$$

最终分数通过 $s_\theta(q_t,\ \sigma_t,\ c) = F_\theta^{\text{odd}}(q_{\text{in}},\ \sigma_t,\ c)/\sigma_t$ 得到。我们使用去噪分数匹配损失（$\mathcal{L}_{\text{dsm}}$）训练该网络。采用退火加权调度 $\lambda(\sigma_t) = \sigma_t^2$ 后，这种参数化在数学上等价于训练 F_θ 预测负噪声（$-\epsilon$），即最小化 $\|F_\theta + \epsilon\|^2$。

训练 我们使用 `PyTorch Lightning` 实现我们的框架，以确保可复现性和可扩展的训练。我们对联合架构（ResNet编码器和分数网络）进行端到端训练，共训练 128,000 步，批次大小为 64。我们采用 AdamW优化器（Loshchilov与Hutter，2019），峰值学习率为 4×10^{-4}，权重衰减为1×10^{-3}，并使用默认的动量参数（β_1, β_2）。为确保稳定收敛，我们采用线性预热与余弦退火调度器，并对模型权重使用指数移动平均（EMA）。

架构和训练超参数的完整汇总见表 4 .

表*4*。 SYMSOL I 实验的超参数配置。

类别	数值
优化	
优化器	AdamW（默认 β 值）
批次大小	64
训练步数	128,000
峰值学习率	4×10^{-4}
最小学习率	1×10^{-4}
权重衰减	1×10^{-3}
预热初始学习率	1×10^{-6}
预热步数	12,800 (10%)
梯度裁剪	10.0
精度	16位混合精度
EMA衰减	0.999
评分网络架构	
时间嵌入维度	64
时间编码	傅里叶
隐藏维度 (d_{MODEL})	128
图像嵌入维度	128
层	4
扩散过程（随机微分方程）	
$\sigma_{\min}$	0.002
$\sigma_{\max}$	20.0
σ_{DATA}	0.5
训练噪声分布	对数均匀分布 $[\sigma_{\text{MIN}}, \sigma_{\text{MAX}}]$

采样 在推理时，我们采用 EDM 提出的多项式噪声调度（Karras 等人，2022年），其中 $\rho = 4$。噪声水平序列定义如下:

$$\sigma_i = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{i}{N-1}(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}})\right)^{\rho}. \tag{59}$$

我们对所有评估均使用 $N = 40$ 个步骤。采样在环境空间 $\mathbb{R}^4$ 中进行，采用改编自（Karras 等人，2022年）的标准 Euler–Maruyama 随机微分方程求解器。为计算使用测地距离的指标，我们将结果投影到单位球面 $\mathcal{S}^3$ 上，以确保满足四元数约束。

提升。 扩散模型 $\hat{}q \sim p(q \mid c)$ 生成的样本位于基本域 $\mathcal{F}$。为恢复 SO(3) 上的完整分布，我们采用提升过程。我们均匀采样一个对称元素 $g \sim G$，并应用变换 $q_{\text{final}} = \hat{}q \cdot g$。这样便恢复了分布的多模态特性，这是正确评估对称物体所必需的。

E. 离散

我们运行 2,000 步（fre sh 从分布中批量采样 512 个样本），并为 DSM 进行超参数调优
以及 MAD，学习率为 率：3e-4、5e-4、7e-4、9e-4、1e-3，以及最小西格玛值：0.01、0.001、0.0001

F. 测量步进训练时间

在每种方法中，我们都将损失函数和训练步骤封装起来，以测量所用时间。也就是说，计时从批次加载完成后、计算损失函数和训练步骤之前开始，到计算完成后结束。