2025 年 12 月 15 日至 12 月 21 日周报

何瑞杰
中山大学, 大湾区大学

目录

速 览

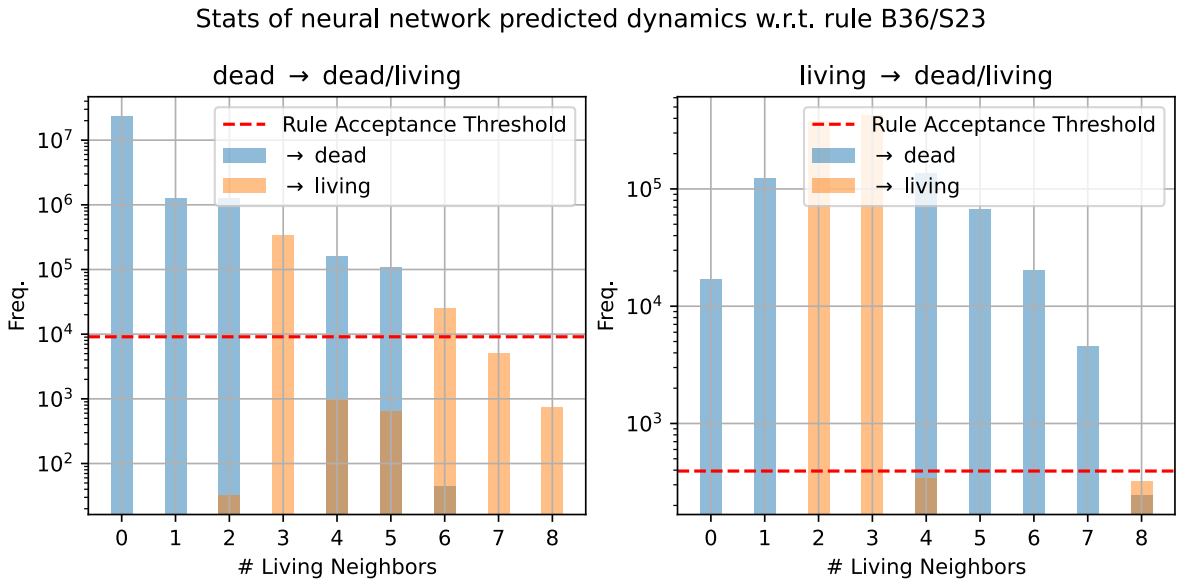
本周大部分时间花在了阅读 Schrödinger 桥论文[1], 它给出了一个以 Schrödinger 桥问题观察生成模型的视角。一般的薛定谔桥问题（或静态薛定谔桥问题）没有闭式解, 作者转而通过修改 IPF 算法使之适配分数匹配生成模型的框架, 并采用迭代的方式学习薛定谔桥。

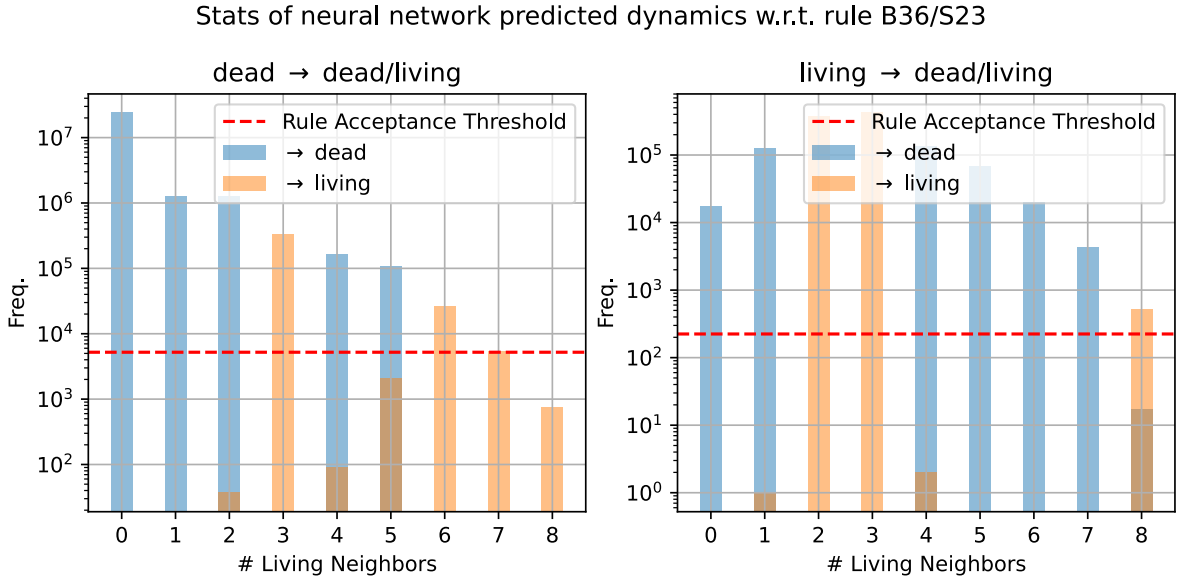
在生命游戏项目方面, 本周实现了一个极小的朴素规则提取函数, 在 B3678/S34678 数据集上可以成功提取规则。

最后本周学习了 Stein 实分析中开篇的少量内容。

1. 项目进展

1.1. 使用神经网络学习生命游戏的演化动力学





将统计用数据序列添加扰动：

$$x' = \left[\texttt{float} \circ \texttt{clamp}_{[0,1]} \circ \texttt{int}\right](x + 0.5 + 0.1 \cdot \varepsilon) \tag{1}$$

将等变网络的群改变为 $p8$ （包括以 45° 为单位的旋转变换和平移变换），只需将下列函数中参数改为 n=8:

2. 文献阅读

2.1. Scalable Diffusion Models with Transformers

https://arxiv.org/abs/2212.09748v2 | William Peebles, Saining Xie

一般的 Diffusion 系列模型所使用的主干都是 U-Net，本文提出了一个基于 Transformer 的替代方案，拥有良好的扩展性能和 SOTA (state of the art) 的 FID 生成分数。

2.1.1. 隐扩散模型 (latent diffusion model, LDM)

相比于一般的扩散模型，隐扩散模型工作在 VAE 等带有瓶颈结构的潜在空间中。其特点为相比直接在图像空间 (pixel space) 中运行扩散模型，隐空间维数远小于图像空间，这样可以显著降低计算开销和推理时间。注意 LDM 的工作空间只是从图像空间改为隐空间，因此扩散模型范畴内的方法，如 DDPM、DDIM、无类引导等方法均可挪用在 LDM 上。

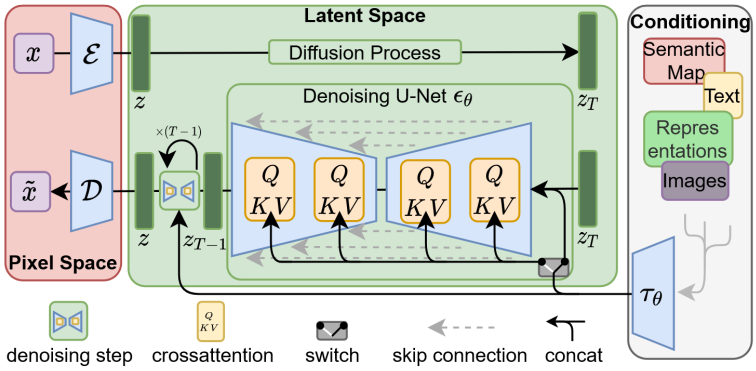


图 1 隐扩散模型的架构[2]

2.1.2. 视觉 Transformer (Vision Transformer, ViT)

视觉 Transformer 是将先前在自然语言处理 (natural language processing, NLP) 中大放异彩的 Transformer 架构引入计算机视觉 (Computer vision, CV) 领域的一次成功尝试。Transformer 架构的核心为注意力机制 (attention mechanism)，其本质为对加上位置编码的序列每项计算得到的键-值对进行匹配，其匹配方式是做内积，然后以此确定其他项相对于某一项的权重。换言之，注意力机制赋予序列中的每项不同的注意力，这是源于自然语言中词元 (token)（组）之间的依赖关系，例如代词往往指示的是一个也许在千里之外的另一个名词。

现在将视线从 NLP 转向 CV，ViT 的核心思想是在图像中构造视觉 token，其方法为将图片按照网格划分为若干小块 (patch)，然后将这些小块经过映射后作为对应于该输入图像的视觉 token，最后输入 Transformer 模块。ViT 继承了 Transformer 的有点，具有可拓展性、高并行度、可优化、全局感受野等优点。

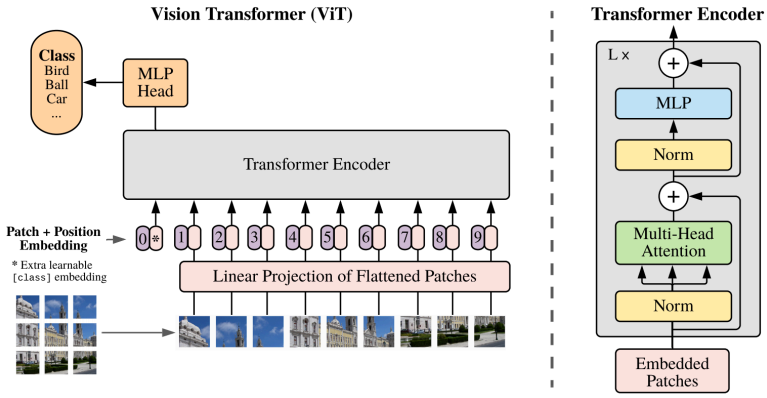


图 2 视觉 Transformer 的架构[3]

2.1.3. 扩散 Transformer (Diffusion Transformer, DiT)

DiT 可以说是融合了 ViT 和 LDM 这两个框架，或者说 DiT 是将 LDM 中主干换成 ViT 后的产物。为适配条件生成等任务，需要在模型中引入一个时间标志 t 和一个类别标志 c。其中前者是将标量时间映射后的时间向量，后者可以是离散的标签，或文生图中对应生成提示词 (prompt) 的嵌入向量 (embedding vector)。作者提出了三种 DiT 模块

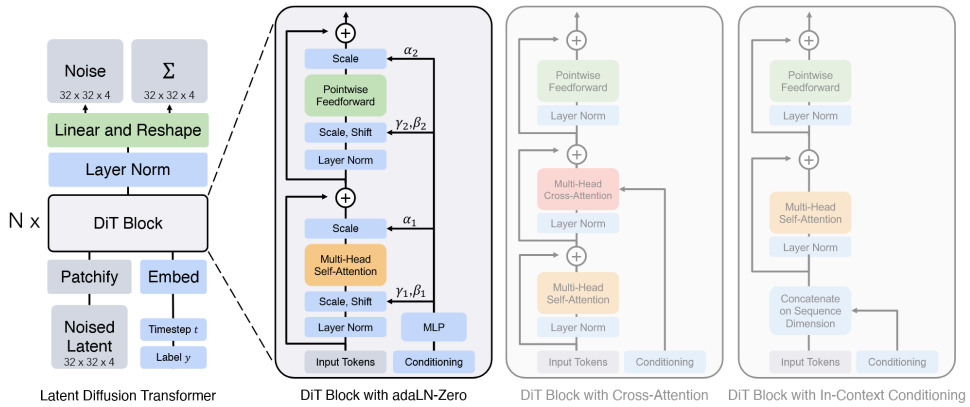


图 3 隐扩散模型的架构[2]

1. 上下文条件（**In-context conditioning**）： 将时间标志和类别标志作为两个额外的 token 附加到输入序列中，与图像 token 平等。
2. 交叉注意力块（**Cross-attention**）： 将时间标志和类别标志连接成一个短的序列，与图像 token 序列分开，并在中间的交叉注意力模块中进入。
3. 自适应层归一化（**adaLN**）： 将层归一化 (layer norm, LN) 的参数改为由时间标志和类别标志而非从图像 token 中学习得到，并同等地施加在全体图像 token 上。
   - **adaLN-Zero 块：** 对 adaLN 的修改，引入一个自时间标志和类别标志学习得到的缩放系数，作用于多头注意力后的缩放模块中。MLP 对应于输出 α 的部分为零初始化，这样在训练初期整个 DiT 模块近似为恒等映射。因为在扩散模型中，若加噪系数 $\sigma \to 0$，那么就有 $x_k \approx x_{k+1}$，这样有利于模型的快速训练。

2.1.4. 实验

Transformer 发力了。

2.1.4.1. DiT 模块比较

作者训练了四个最高计算量的 DiT-XL/2 模型，每个使用不同的 DiT 块。adaLN-Zero 块产生的 FID 低于其他两种，但计算效率最高（上下文条件 119.4 GFlops，交叉注意力模块 119.4 Gflops，adaLN 和 adaLN-Zero 118.6 Gflops）。adaLN-Zero 显著优于普通 adaLN。

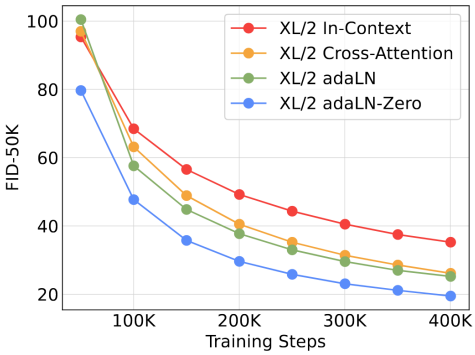


图 4 DiT 模块比较

2.1.4.2. 模型规模与块大小的规模法则

作者训练了 12 个 DiT 模型，覆盖模型配置（S, B, L, XL）和块大小 (8, 4, 2)。实验结果显式增加模型规模和减小块大小都能显著改善扩散模型。当保持块大小不变增加模型大小时，FID 显著降低。保持模型规模不变减小块大小时，FID 显著降低。模型参数量并不能唯一确定模型的质量，当保持模型规模不变而减小块大小时，Transformer 的总参数基本不变，只有计算量增加。这些结果表明，模型的计算量是其性能的关键。

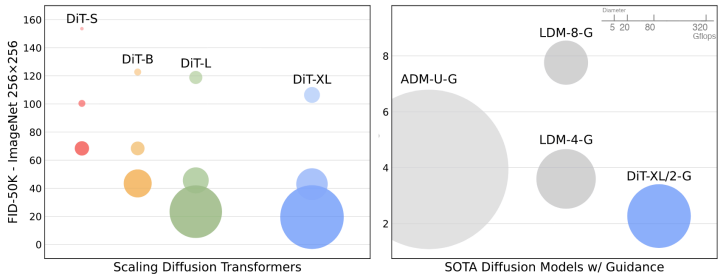


图 5 不同规模 DiT 模型的性能（左）及与 SOTA 模型的比较（右）
圆圈大小代表模型的计算量

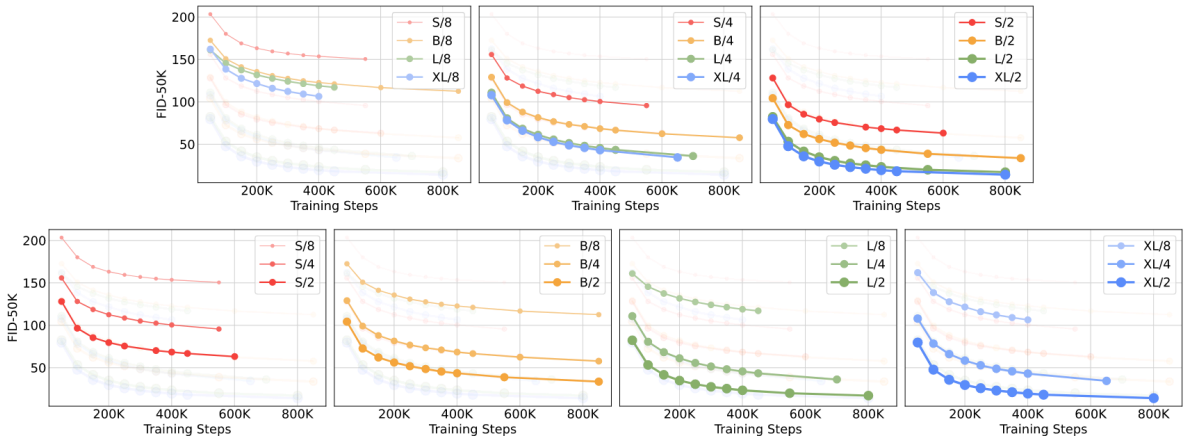


图 6 增加模型大小和减小块大小均能显著提升模型性能

2.1.4.3. 与 SOTA 模型的比较

在 ImageNet 256×256 的条件生成基准上，使用无分类器引导的 DiT-XL/2 优于所有先前的扩散模型。DiT-XL/2 在各种评估指标上均优于所有先前生成模型，包括之前的 SOTA StyleGAN-XL。在 512×512 分辨率上，DiT-XL/2 再次优于所有先前的扩散模型。即使 token 数量增加，DiT-XL/2 仍保持计算效率。

2.1.4.4. 模型计算 vs 采样计算

较小模型每图像使用的采样计算比较大模型多 5 倍，较大模型仍保持更好的 FID。一般来说，增加采样计算无法弥补模型的不足。我的推测是小模型无法对噪声和协方差矩阵做很好的估计，这样添加再多采样步数也无济于事。

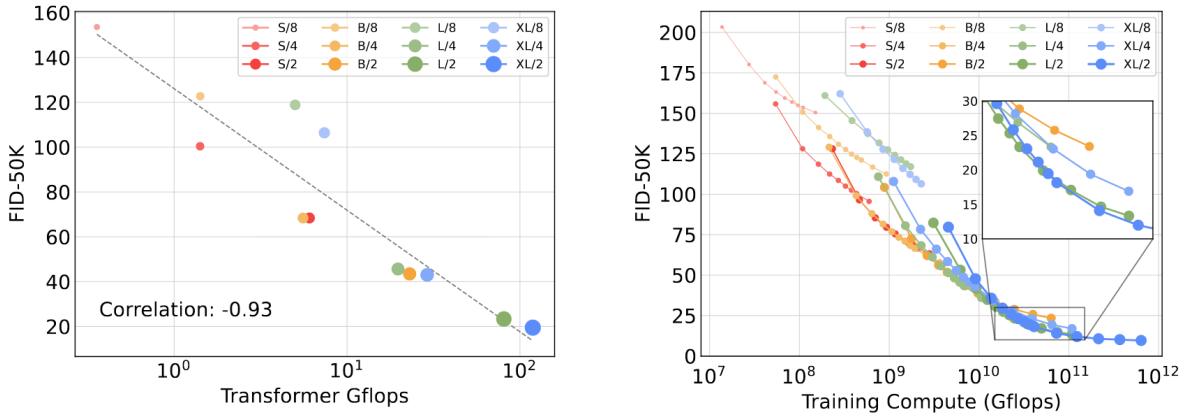


图 7 左：DiT 模块的计算量与 FID 分数显著相关；右：模型越大，对计算性能的利用越高效。

3. 学习进度

3.1. 生成模型理论

本周参阅了 MIT 的生成模型课程笔记，这是一本五十页的小册子，本周读完了大半部分内容。由于先前阅读了不少相关方面的论文，阅读起来没什么障碍，不过该讲义依然给予我了一些比较优雅的视角。

3.2. 随机微分方程

3.2.1. 一维解的构造

设 $b:\mathbb{R}\to\mathbb{R}$ 为 $\mathscr{C}^1$ 函数且 $|b'|\le L$。考虑 SDE:

$$\mathrm{d}X = b(X)\mathrm{d}t + \mathrm{d}W, X(0) = x_0 \in \mathbb{R} \tag{2}$$

通过 Picard 迭代构造解 $X^{(0)}(t)\equiv x_0$递推式为

$$X^{(k+1)}(t) := x_0 + \int_0^t b\big(X^{(k)}\big)\mathrm{d}s + W(t) \tag{3}$$

并定义

$$D^{(k)}(t) := \max_{0\le s\le t}|X^{(k+1)}(s) - X^{(k)}(s)| \tag{4}$$

可以用归纳法证明

$$D^{(k)}(t) \le C\frac{L^k}{k!}t^k \tag{5}$$

由此可得$X^{(k)}$是 Cauchy 列，几乎必然一致收敛到 SDE 的解X。

3.2.2. 变量替换法

对一般 SDE $\mathrm{d}X = b(X)\mathrm{d}t + \sigma(X)\mathrm{d}W,\quad X(0)=x_0$，设 $X=u(Y)$，其中 Y 满足 $\mathrm{d}Y = f(Y)\mathrm{d}t + \mathrm{d}W,\quad Y(0)=y_0$。由 Itô 公式可得

$$\mathrm{d}X = \left[u'(Y)f(Y) + \frac{1}{2}u''(Y)\right]\mathrm{d}t + u'(Y)\mathrm{d}W \tag{6}$$

因此需要满足:

$$u'(Y) = \sigma(u(Y)),\quad u'(Y)f(Y) + \frac{1}{2}u''(Y) = b(u(Y)),\quad u(y_0) = x_0 \tag{7}$$

可以先解 ODE $u'(z)=\sigma(u(z)), u(y_0)=x_0$，然后定义:

$$f(z) := \frac{\left[b(u(z)) - \frac{1}{2}u''(z)\right]}{\sigma(u(z))} \tag{8}$$

即为原 SDE 的解。

3.2.3. 存在唯一性定理

定理 **3.1** (Gronwall 不等式): 设f,φ是 $[0,T]$ 上的非负连续函数，若

$$\varphi(t) \le C_0 + \int_0^t f\varphi\mathrm{d}s \tag{9}$$

则

$$\varphi(t) \le C_0\exp\left(\int_0^t f\mathrm{d}s\right) \tag{10}$$

证明：令 $\Phi = C_0 + \int_0^t f\varphi \mathrm{d}s$ ，则 $\Phi' = f\varphi \le f\Phi$。计算 $\left[\exp\left(-\int_0^t f\mathrm{d}s\right)\Phi\right]' \le 0$，故 $\exp\left(-\int_0^t f\mathrm{d}s\right)\Phi(t) \le C_0$，得证。 □

定理 3.2 (存在唯一性定理)：设 $b : \mathbb{R}^n x[0,T] \to \mathbb{R}^n$ 和 $B : \mathbb{R}^n x[0,T] \to \mathbb{R}^{mxn}$ 满足一致 Lipschitz 条件：

$$|b(x,t) - b(\hat{x},t)| \le L|x - \hat{x}| \tag{11}$$

$$|B(x,t) - B(\hat{x},t)| \le L|x - \hat{x}| \tag{12}$$

及线性增长条件：

$$|b(x,t)| \le L(1 + |x|) \tag{13}$$

$$|B(x,t)| \le L(1 + |x|) \tag{14}$$

若 $E|X_0|^2 < \infty$，则 SDE:

$$\mathrm{d}X = b(X,t)\mathrm{d}t + B(X,t)\mathrm{d}W, \quad X(0) = X_0 \tag{15}$$

存在唯一解 $X \in L_n^2(0,T)$，且在概率1意义下唯一。

3.3. 实分析

我们希望扩展长度、面积或是体积的概念，使之适用于一般的集合。形式上，我们希望存在一个这样的映射 $\mu : \mathcal{A} \to [0, \infty]$，其中 $\mathcal{A} \subset 2^{\Omega}$，并满足下面的性质：

1. 非负性 (non-negativity)：对任意 $A \in \mathcal{A}$，有 $\mu(A) \geq 0$；
2. 空集的测度为零 (null empty set)：$\mu(\emptyset) = 0$；
3. 可数可加性 (countable additivity)：对任意可数个两两不交的集合 $\{A_i\}_{i \in \mathbb{N}} \subset \mathcal{A}$，有

$$\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i) \tag{16}$$

3.4. 量子力学

在 1921 年和 1922 年，O.Stern 和 W.Gerlach 进行了下面的实验。他们将银放在一个留有一个小孔的加热炉中加热，然后在银原子逃逸的路径上设置一个非均匀磁场。根据经典力学的预测，银原子束在通过磁场后应该会发生扩散，因为每个原子的磁矩方向是随机的。然而实验结果显示，银原子束在通过磁场后分裂成了两个离散的部分。如果这个磁矩是由旋转角动量产生的，那么我们应该观察到一个连续的分布。因此这说明存在一个未知的内秉角动量，它在某一方向上只能取两个值。

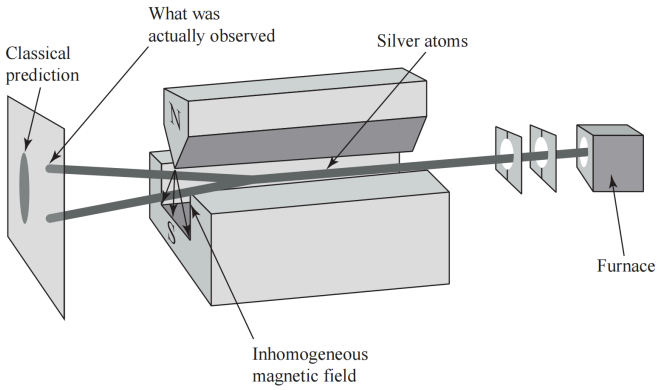


图 8 Stern 和 Gerlach 的实验[1]

故事没有结束，接下来再看下面一个级联的 Stern-Gerlach 实验。上述实验对应下图中 (a) 子图的左边。当在 $\hat{\mathbf{z}}$ 方向施加非均匀磁场时，银原子束分裂成两束。现在遮挡其中一束，对另一束考虑下面三种处理

1. 再次通过一个 $\hat{\mathbf{z}}$ 方向的非均匀磁场，银原子束不再分裂；
2. 通过一个 $\hat{\mathbf{x}}$ 方向的非均匀磁场，银原子束分裂成两束；
3. 通过一个 $\hat{\mathbf{x}}$ 方向的非均匀磁场后，遮挡其中一束；再通过一个 $\hat{\mathbf{z}}$ 方向的非均匀磁场，银原子束再次分裂成两束。

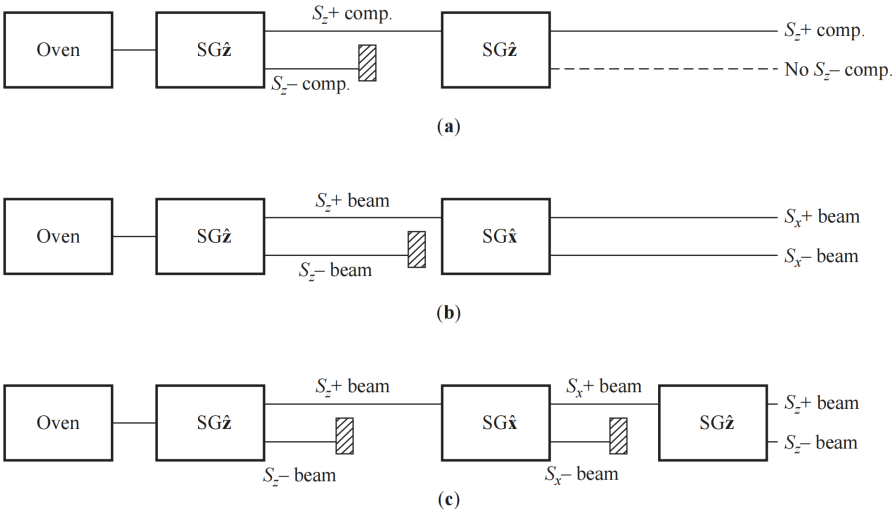


图 9 级联 Stern-Gerlach 的实验

该实验表明，每次经过某一方向的磁场，相当于对银原子的内秉角动量进行一次测量，测量结果只能取两个值的其中一个。如果已在某个方向上测量过，不经过其他操作再次测量时得到的结果不会变化。然而若在进行了某一方向的一次测量后再经过另一方向的测量，最初的测量结果会被“抹去”，换言之，不能同时确定两个方向上的测量结果。

[1] 本节内容参考自[4]

为研究如上的量子现象，需要引入对应的数学工具。量子力学研究复 Hilbert 空间 $\mathscr{H}$、其上的线性算子以及其对偶对象（复对偶 Hilbert 空间 $\mathscr{H}^*$ 和其上的线性算子）。在量子力学中，称复 Hilbert 空间中的元素为态矢量 (state vector)，写作 $|a\rangle$，因此也称右矢量 (ket)，其对偶对象 $\langle a|$ 为左矢量 (bra)。复 Hilbert 空间中两个右矢 $|a\rangle$ 和 $|b\rangle$ 的内积为 $\langle a|b\rangle$，英文就叫 bra(c)ket。注意它可以写成 $\langle a||b\rangle$，而左边一项是右矢 $|a\rangle$ 的对偶，这是由 Riesz 表示定理保证的。由内积公理，对任意右矢 $|a\rangle$，有 $\langle a|a\rangle \geqslant 0$；对任意两个右矢 $|a\rangle$ 和 $|b\rangle$，内积有共轭对称性 $\langle a|b\rangle = \langle b|a\rangle^*$。若 $\mathscr{H}$ 为至多可数维，那么存在一个正交归一基 (orthonormal basis) $\{e_i\}_{i\in\mathbb{N}}$，可以表出任意右矢 $|x\rangle$ 为

$$|x\rangle = \sum_{i\in\mathbb{N}} x_i\ |e_i\rangle \tag{17}$$

若将基矢 $|e_j\rangle$ 之对偶作用于 $|x\rangle$，得到

$$\langle e_j|x\rangle = \sum_{i\in\mathbb{N}} x_i\langle e_j|e_i\rangle = \sum_{i\in\mathbb{N}} x_i\delta_{i,j} = x_j \tag{18}$$

这样我们就能“形式上”地将右矢 $|x\rangle$ 写作一个“列矢量” x 的形式。需要注意的是，列矢量 x 是 $\mathscr{H}$ 中的抽象元素 $|x\rangle$ 在基（或称表象）$\{e_i\}_{i\in\mathbb{N}}$ 下的表示，类似地不难证明其对偶 $\langle x|$ 可以写成 $x^{\top *} = x^\dagger$。

量子力学中称 $\mathscr{H}$ 上的线性算子 $\hat{A}:\mathscr{H}\to\mathscr{H}$ 为算符。算符 $\hat{A}$ 作用在右矢 $|x\rangle$ 上写作 $\hat{A}|x\rangle$，得到另一个右矢，其对偶的情形为 $\langle x|\hat{A}^\dagger$，其中 $\square^\dagger$ 表示算子的 Hermite 共轭。若将一个右矢和一个左矢形式上地乘在一起，即 $|a\rangle\langle b|$，它定义了一个算符。事实上上述结果也可以写成张量积的形式，即 $\boldsymbol{|a\rangle\langle b| = |a\rangle\otimes\langle b|}$。现在还是考虑一个基 $\{e_i\}_{i\in\mathbb{N}}$。一方面，右矢 $|x\rangle$ 可以写成基矢的线性组合，也可以被基矢作用得到在基矢“方向”上的“分量”或投影，有

$$|x\rangle = \sum_{i\in\mathbb{N}}|e_i\rangle\cdot x_i = \sum_{i\in\mathbb{N}}|e_i\rangle\langle e_j|x\rangle = \left[\sum_{i\in\mathbb{N}}|e_i\rangle\langle e_j|\right]\ |x\rangle = 1\ |x\rangle. \tag{19}$$

于是我们成功构造了恒等算符 1。另一方面，算符 $\hat{A}$ 作用在右矢 $|x\rangle$ 得到新的右矢 $|y\rangle$，用一个基矢 $|e_j\rangle$ 作用在 $|y\rangle$ 上得到

$$\langle e_j|y\rangle = \langle e_j|\hat{A}|x\rangle = \langle e_j|\hat{A}\cdot 1\cdot|x\rangle = \sum_{i\in\mathbb{N}}\langle e_j|\hat{A}\ |e_i\rangle\langle e_j|x\rangle \tag{20}$$

若令 $A_{j,i} = \langle e_j|A|e_i\rangle$，则有 $y_j = \langle e_j|y\rangle = \sum_{i\in\mathbb{N}} A_{j,i}x_i$。这表明算符 $\boldsymbol{\hat{A}}$ 在基 $\boldsymbol{\{e_i\}_{i\in\mathbb{N}}}$ 下的矩阵表示为 $\boldsymbol{A = (A_{j,i})}$，相应地，其对偶算符 $\boldsymbol{A^\dagger}$ 的矩阵表示为 $\boldsymbol{A^\dagger = (A^*_{i,j})}$，即为原算符矩阵的共轭转置。考虑算符 $\hat{A}$ 同时作用在右矢 $|x\rangle$ 和左矢 $\langle y|$ 上，有

$$\langle x|\hat{A}|y\rangle = \langle x|\left[\hat{A}|y\rangle\right] = \left[\langle y|\hat{A}^\dagger|x\rangle\right]^* \tag{21}$$

如果 $\hat{A} = \hat{A}^\dagger$，我们称其为 **Hermite 算符**，Hermite 算符的所有本征值（特征值）为实数，每个本征值对应的本征态（特征向量）都是正交的。

4. 下周计划

论文阅读

1. 生成模型
   - 薛定谔桥
   - DDIM

项目进度

1. 使用神经网络学习生命游戏的演化动力学
   - 考虑另外两种方法的实现
   - 更新在线 Overleaf 文档
2. 耦合约瑟夫森结
   - 将 MATLAB 模拟代码全部迁移至 Python
   - 考虑简单的 Neural SDE 方法解带参 OU 过程的参数

理论学习

1. 随机过程课程
   - 复习 Poisson 过程和 Markov 过程
2. 随机微分方程
   - 第五章完成

5. 附录

参考文献

[1] V. D. Bortoli, J. Thornton, J. Heng, and A. Doucet, "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling." [Online]. Available: https://arxiv.org/abs/2106.01357

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models." 2021.

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, 2020, [Online]. Available: https://arxiv.org/abs/2010.11929

[4] J. J. Sakurai, *Modern Quantum Mechanics (Revised Edition)*, 1st ed. Addison Wesley, 1993. [Online]. Available: http://www.worldcat.org/isbn/0201539292