2026 年一月上旬工作汇报和寒假工作计划

何瑞杰
中山大学, 大湾区大学

目录

速 览

一月上半旬由于期末考试后休息、未完成的两个课程作业和其他事宜，消耗了大部分时间，因此这一期间仅进行了一些知识学习和论文阅读工作。主要内容包括对实变函数的知识做了简单的概览，并阅读了和 Neural SDE 相关的论文一篇。一月下旬的工作计划附在文后。

1. 文献阅读

1.1. Neural Stochastic Differential Equations: Deep Latent Gaussian Models in the Diffusion Limit

arxiv 2011.13456 | Belinda Tzen and Maxim Raginsky

1.1.1. 隐 Gauss 模型

机器学习常常将高维的数据 x 投射到某个维度更低的隐空间中，得到其“潜在表示” z。隐 Gauss 模型意为潜在表示服从某个 Gauss 分布。假设需要生成某个数据 y，隐 Gauss 模型尝试通过若干 Gauss 随机变量 $z_1, z_2, \cdots, z_k$ 生成之，具体操作如下：

$$\begin{aligned} x_0 &= z_0, \\ x_i &= x_{i-1} + b_i(x_{i-1}) + \sigma_i z_i, \quad i = 1, 2, \cdots, k, \\ y &\sim p_\theta(\cdot\,|x_k). \end{aligned} \tag{1}$$

其中 $z_i \sim \mathcal{N}(0, I)$，$\theta$ 是所有 b_i 和 σ_i 中的参数。和很多生成模型一样，我们要最大化训练数据 y 的对数似然 $\log p_\theta(y)$。本文中给出了一个相比于常规生成模型论文中长篇累牍的 ELBO 推导的一个更自然的视角，即 Gibbs 变分原理 (Gibbs variational principle)。

定理 1.1 (Gibbs 变分原理)：对 $\Omega = \left(\mathbb{R}^d\right)^{k+1}$ 上的任意测度 μ，和任意可测函数 $F : \Omega \to \mathbb{R}$，有

$$-\log \mathbb{E}_\mu[\exp(-F(z_1, \ldots, z_k))] = \inf_{\nu \in \mathscr{P}(\Omega)} \{D_{\mathrm{KL}}(\nu \| \mu) + \mathbb{E}_\nu[F(z_1, \ldots, z_k)]\}. \tag{2}$$

在隐 Gauss 模型的语境下，令 $F(z_1, ..., z_k) = -\log p(y|f_\theta(z_1, ..., z_k))$，就得到

$$-\log p_\theta(y) = \inf_{\nu\in\Omega}\left\{D_{\mathrm{KL}}(\nu\|\mu) - \int_\Omega \log p_\theta(y|z)\nu(\mathrm{d}(z))\right\}. \tag{3}$$

其中 μ 对应的是 $z = z_0, ..., z_k$ 的分布。等号右边取下确界的这一项称为变分自由能 F_θ。式 3 中等号右侧的积分无法计算，常用的技巧是使用一个带参后验 $\nu_\beta(\mathrm{d}z|y) = q_\beta(z|y)\mathrm{d}z$ 进行近似，最后得到带参的变分自由能 $F_{\theta,\beta}$，其形式为

$$F_{\theta,\beta} = D_{\mathrm{KL}}(\nu\|\mu) - \int_\Omega \log p(y|f_\theta(z_1, ..., z_k))\nu_\beta(\mathrm{d}z_1, ..., \mathrm{d}z_k|y). \tag{4}$$

该形式可以通过对适当的参数化后的梯度进行 Monte-Carlo 得到对变分自由能梯度的估计。

1.1.2. Neural SDE

我们发现 式 1 就是某个 SDE 的 Euler-Maruyama 离散化形式，如果我们转而考虑形如 $\mathrm{d}X = F\mathrm{d}t + G\mathrm{d}W$ 这样的随机微分方程时，若将 F 和 G 使用神经网络替代，就得到了 Neural SDE。对比隐 Gauss 模型，SDE 语境下的隐变量在 $[0, 1]$ 上的 d 维路径组成的空间 $\mathbb{W} = \mathscr{C}\left([0, 1], \mathbb{R}^d\right)$ 中，隐变量本身是 $[0, 1]$ 上的某个连续的 Brown 运动轨道。可以利用随机积分的定义得到循序可测映射 $f : \mathbb{W} \to \mathbb{W}$，其形式为

$$[f_\theta(W)]_t = \int_0^t b\left([f_\theta(W)]_s, s; \theta\right)\ \mathrm{d}s + \int_0^t \sigma\left([f_\theta(W)]_s, s; \theta\right)\ \mathrm{d}W_s, . \tag{5}$$

上一节中，联合分布 $p_\theta(y, z)$ 可以写为

$$p_\theta(y, z) = p_\theta(y|z)p(z) = p_\theta(y|z)\phi_d(z_1)\cdots\phi_d(z_k). \tag{6}$$

其测度元可以写为

$$P_\theta(\mathrm{d}y, \mathrm{d}z) = p_\theta(y|z)\ \mathrm{d}y \prod_{i=1}^k \phi_d(z_i)\ \mathrm{d}z_i. \tag{7}$$

现在将式 7 中的隐变量替换为 W，并将 $\prod_{i=1}^k \phi_d(z_i)\ \mathrm{d}z_i$ 替换为 Wiener 测度 $\boldsymbol{\mu}(\mathrm{d}W)$，就得到了 Neural SDE 的联合测度元

$$P_\theta(\mathrm{d}y, \mathrm{d}W) = \underbrace{p_\theta(y|f_\theta(W))}_{\text{条件密度}} \cdot \underbrace{\boldsymbol{\mu}(\mathrm{d}W)}_{\text{Wiener 测度}} \cdot \underbrace{\mathrm{d}y}_{\text{Lebesgue 测度}}. \tag{8}$$

因此 y 的边缘密度为 $p_\theta(y) = \int_{\mathbb{W}} p_\theta(y|f_\theta(W))\boldsymbol{\mu}(\mathrm{d}W) = \mathbb{E}_{\boldsymbol{\mu}}[p_\theta(y|f_\theta(W))]$。对应于离散版本中用到的定理 1.1，$\mathbb{W}$ 上也有类似的结论，即

$$-\log p_\theta(y) = \inf_{\boldsymbol{\nu}\in\mathscr{P}(\mathbb{W})}\{D_{\mathrm{KL}}(\boldsymbol{\nu}\|\boldsymbol{\mu}) + \mathbb{E}_{\boldsymbol{\nu}}[F(W)]\}. \tag{9}$$

由 **Girsanov** 定理，每个 $\mathbb{W}$ 上关于 Wiener 测度绝对连续的的测度 $\boldsymbol{\nu}$ 都可以对应于一个 Brown 运动 W 加上一个漂移。可以得到对于满足 $\mathrm{d}Z = u\mathrm{d}t + \mathrm{d}W$ 的过程 $Z \in \boldsymbol{\nu}$，有

$$D_{\mathrm{KL}}(\boldsymbol{\nu}\|\boldsymbol{\mu}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}}\left[\int_0^1 \|u(t)\|^2\ \mathrm{d}t\right], \qquad \mathbb{E}_{\boldsymbol{\nu}}[F(W)] = \mathbb{E}_{\boldsymbol{\mu}}\left[F\left(W + \int_0^\bullet u(s)\ \mathrm{d}s\right)\right]. \tag{10}$$

因此 式 9 可以写成所谓变分的 Girsanov 表示:

$$-\log p_\theta(y) = \inf_u \mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{2}\int_0^1 \|u(t)\|^2 \ \mathrm{d}t + F_\theta\left(W + \int_0^\bullet u(s) \ \mathrm{d}s\right)\right]. \tag{11}$$

其中$W + \int_0^\bullet u(s) \ \mathrm{d}s$是由 Brown 运动轨道生成的新的随机过程$Z$; $F_\theta(w) = -\log p_\theta\big(y|[f_\theta(w)]_1\big)$。注意此时期望下标从 $\boldsymbol{\nu}$ 变成了 $\boldsymbol{\mu}$。

在使用 SDE 的无限层隐 Gauss 模型中，考虑对上面的下确界使用平均场近似。令 $u_s = \tilde{b}(y, s; \beta), s \in [0, 1]$，其中 $\tilde{b}$ 是确定的神经网络，并满足 $\int_0^1 \|\tilde{b}(y, t; \beta)\| < \infty$，平均场的思想体现在我们采用确定的偏移项而不是随机的，这样使式 11 的第一项不再需要计算期望，同时变为对所有 β 求下确界，得到负对数似然的上界:

$$-\log p_\theta(y) \leqslant \inf_\beta \left\{\frac{1}{2}\int_0^1 \|\tilde{b}(y, s; \beta)\|^2 \ \mathrm{d}t + \mathbb{E}_{\boldsymbol{\mu}}\left[F_\theta\left(W + \int_0^\bullet \tilde{b}(y, s; \beta) \ \mathrm{d}s\right)\right]\right\} \tag{12}$$

在这里 $\boldsymbol{\beta}$ 的作用是作为 $\tilde{\boldsymbol{b}}$ 的参数构成所谓变分网络，用于估计 $\boldsymbol{p_\theta(W|y)}$，因此 $\boldsymbol{\theta}$ 可以看做解码器参数，$\boldsymbol{\beta}$ 可以看做编码器参数。

1.1.3. 对 Neural SDE 求梯度

研究 Neural SDE 的变分自由能

$$F_{\theta,\beta} = \frac{1}{2}\int_0^1 \|\tilde{b}(y, s; \beta)\|^2 \ \mathrm{d}t + \mathbb{E}_{\boldsymbol{\mu}}\left[F_\theta\left(W + \int_0^\bullet \tilde{b}(y, s; \beta) \ \mathrm{d}s\right)\right]. \tag{13}$$

等式右侧的第一项是确定函数的积分，可以使用通常的梯度求解方法；我们关注第二项。第二项 F_θ 括号里面是下面的 Itô 过程

$$X_t^{\theta,\beta} = W_t + \int_0^t b\big(X_t^{\theta,\beta}, s; \theta\big) \ \mathrm{d}s + \int_0^t \tilde{b}(y, s; \beta) \ \mathrm{d}s + \int_0^t \sigma\big(X_s^{\theta,\beta}, s; \theta\big) \ \mathrm{d}W_s, t \in [0, 1]. \tag{14}$$

多出来的蓝色的一项是因为先前使用了 Girsanov 定理。右边这一项就对应着 $-\mathbb{E}\big[\log p\big(y|X_1^{\theta,\beta}\big)\big]$。而要对其求导，我们希望内部的函数 $\log p\big(y|X_1^{\theta,\beta}\big)$ 的性质足够好，是的偏导数算子和期望算子可以交换:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\mathbb{E}\big[\log p\big(y|X_1^{\theta,\beta}\big)\big] &= \mathbb{E}\left[\frac{\partial}{\partial x}\log p\big(y|X_1^{\theta,\beta}\big)\frac{\partial}{\partial\beta}X_1^{\theta,\beta}\right], \\
\frac{\partial}{\partial\theta}\mathbb{E}\big[\log p\big(y|X_1^{\theta,\beta}\big)\big] &= \mathbb{E}\left[\frac{\partial}{\partial x}\log p\big(y|X_1^{\theta,\beta}\big)\frac{\partial}{\partial\theta}X_1^{\theta,\beta}\right].
\end{aligned} \tag{15}$$

需要解决的核心问题为如何计算 $\dfrac{\partial}{\partial\bullet}X_1^{\theta,\beta}$。论文中给出了两种方法：第一种为求解后微分，第二种为先微分后求解。具体而言，因为 $X_1^{\theta,\beta}$ 是一个随机过程在 $t = 1$ 处的取值，它是按照式 14 生成的，对其求导必须要经过这个随机积分。于是第一种方法考虑先将 $X_1^{\theta,\beta}$ 离散化，选取一个划分 $0 = t_0 < t_1 < \cdots < t_N = 1$，然后采样独立同分布的 $Z_1, Z_2, ..., Z_N$，做 Euler-Maruyama 算法:

$$\hat{X}_{t_{i+1}}^{\theta,\beta} = \hat{X}_{t_i}^{\theta,\beta} + (t_{i+1} - t_i)\Big[b\Big(\hat{X}_{t_i}^{\theta,\beta}, t_i; \theta\Big) - \tilde{b}(y, t_i, \beta)\Big] + \sqrt{t_{i+1} - t_i}\,\sigma\Big(\hat{X}_{t_i}^{\theta,\beta}, t_i; \theta\Big)Z_{i+1}, \tag{16}$$

然后求得 $X_1^{\theta,\beta}$ 的数值解后调用自动求导工具进行求导，其时间复杂度大约为 $O\big[N\big(\mathsf{T}(b) + \mathsf{T}(\tilde{b}) + \mathsf{T}(\sigma)\big)\big]$。

另一种方法是尝试直接得出 $X_1^{\theta,\beta}$ 对 θ 或 β 的 SDE 解析形式，这个 SDE 中仅包含通常意义上的导数计算，然后根据这些求导结果，进行 SDE 的数值计算。具体而言，有下面的定理

定理 1.2 (Neural SDE 求导方程)：设 $X_i^{\theta,\beta}$ 满足式 14，且满足

1. $b(x,t;\theta)$ 和 $\sigma(x,t;\theta)$ 在 $[0,1]$ 上一致 Lipschitz 连续，且它们对 x 和 θ 的 Jacobian 矩阵在 $[0,1]$ 上也一致 Lipschiz 连续；
2. $\tilde{b}(y,t;\beta)$ 在 $[0,1]$ 上一致 Lipschitz 连续，且对 β 的 Jacobian 矩阵在 $[0,1]$ 上也一致 Lipschiz 连续；

则随机过程 $X = X^{\theta,\beta}$ 对 θ 和 β 的路径导数为

$$
\begin{aligned}
\frac{\partial X_t}{\partial \beta^{(i)}} &= \int_0^t \left(\frac{\partial b_s}{\partial x} \frac{\partial X_s}{\partial \beta^{(i)}} + \frac{\partial \tilde{b}_s}{\partial \beta^{(i)}} \right) \mathrm{d}s + \sum_{l=1}^{d} \int_0^t \frac{\partial \sigma_{s,l}}{\partial x} \frac{\partial X_s}{\partial \beta^{(i)}} \mathrm{d}W_s^{(l)} \qquad \text{(a)} \\
\frac{\partial X_t}{\partial \theta^{(j)}} &= \int_0^t \left(\frac{\partial b_s}{\partial \theta^{(j)}} + \frac{\partial b_s}{\partial x} \frac{\partial X_s}{\partial \theta^{(j)}} \right) \mathrm{d}s + \sum_{l=1}^{d} \int_0^t \left(\frac{\partial \sigma_{s,l}}{\partial \theta^{(j)}} + \frac{\partial \sigma_{s,l}}{\partial x} \frac{\partial X_s}{\partial \theta^{(j)}} \right) \mathrm{d}W_s^{(l)} \text{(b)}
\end{aligned}
\tag{17}
$$

其中 $\theta \in \mathbb{R}^n$，$\beta \in \mathbb{R}^k$，$b_s = b\left(X_s^{\theta,\beta}, s; \theta\right)$，$\sigma_{s,l}$ 是 $\sigma\left(X_s^{\theta,\beta}, s; \theta\right)$ 的第 l 列，其他记号以此类推。

对于 (a)，我们要求 b 对 x 的导数、$\tilde{b}$ 对 β 的导数和 σ 对 x 的导数。解 SDE 时每次迭代求解这些导数的时间复杂度为 $O\left[k\left(d\mathsf{T}(b) + \min(d,k)\mathsf{T}\left(\tilde{b}\right) + d\mathsf{T}(\sigma)\right)\right]$，对 (b)，需要求的是 b 和 σ 分别对 θ 和 x 的导数，其时间复杂度为 $O[n(\min(d,n)+d)(\mathsf{T}(b)+\mathsf{T}(\sigma))]$。

2. 学习工作心得

本学期为我在中大就读的第一学期，也是尝试使用 typst 撰写周报和课程报告的第一个学期。我的体验是比较顺畅，和 markdown 和 LaTeX 的知识兼容性高，上手迅速，编译速度快。唯一缺陷也许是功能目前可能受限（在我暂时用不到的地方）以及 typst 的 slides 功能（如 `touying` 模块）的文档不够详细完善，暂时无法满足较为精细的调整，不过整体观感问题不大。

论文阅读方面，本学期共读论文 11 篇，主要涉及 SDE 和 ODE 建模的生成模型，低于课题组每周一篇论文的基本要求。这和我时长分心，难以集中精力、未能科学规划时间有关（包括本次报告迟交 1 天）。这个缺点还反映在各个方面，例如课程作业、考试复习、科研项目的时间安排。不过仔细阅读这些论文中的例如 Yang Song 关于 Score Matching 和 SDE 的论文、DDPM、群等变卷积等相对入学前读过论文的理论深一些的论文，我都期望可以完整的把握其中的核心思想，再不济也需要掌握大概，这让我不能容忍在撰写周报的论文阅读部分时遇到的任何一个不理解或是理解模糊的点。虽然本学期我的论文阅读效率不够，但我能感到自己主动走出舒适区解决困难问题的能力正在提升。

项目方面，我主要参与了生命游戏规则学习的项目，这是个人项目，除去指导老师外没有其他组员协作，这同样考验个人的时间管理和项目组织能力。我常常若干周没有实质性的进展，面临从神经网络中提取显式规则的问题，我由于缺乏相关经验和检索不到相关的成熟方案而因此对其望而却步，一拖再拖。另外该项目也暴露了我在主线方面不确定的问题，例如在某次会议前我临时将学习的目标改为少量轨道上利用主动学习逐步抽取规则，但项目仍未交付完整的基础版本。

理论学习方面，本学期我选了一门随机过程课，了解了 Poisson 过程、离散事件有限状态 Markov 过程和更新过程的基本概念，另外自学了 Evans 所著的 SDE 小册子的基础知识，目前这本小册子的前五章已经学习完毕，这些关于随机积分和随机微分方程的知识对我理解用到它们的论文很有帮助。

3. 一月上旬工作计划

3.1. 论文阅读

寒假我将阅读和 SDE 与生成模型相关的论文，同时逐步将重心转移至 Neural SDE、薛定谔桥、随机控制和与神经动力学相关的论文，并尝试寻找它们和脑机接口之间的联系。

3.2. 项目进度

寒假我将在年前完成最基本的生命游戏规则学习的完整可运行项目，并抽时间与指导老师约定线上会议时间，同时修改线上论文草稿。

3.3. 理论学习

寒假期间我将结束对 Evans 的 SDE 小册子的学习；并开始补足必要的严格数学基础知识，即实分析和泛函分析；除此之外，我将使用富裕时间开始学习量子力学和动力系统的基础知识。