一个可微分的带有前向后向信息传递的部分可见广义线性模型

李成睿，王雨乐，黎维瀚，吴安琪
计算科学与工程 School of Computational Science & Engineering
佐治亚理工学院 Georgia Institute of Technology
Atlanta, GA 30305, USA
{cnlichengrui,yulewang,weihanli,anqiwu}@gatech.edu

2024 年 6 月 2 日

摘要

当假设存在隐藏神经元时，部分可见广义线性模型 (POGLM) 是理解神经连接的有力工具．当只记录了可见神经元的放电序列时，现有的工作用变分推断来学 POGLM，但同时也意识到学这样一个隐变量模型本身是很困难的．主要问题有两个：(1) 采样得到的隐藏神经元的泊松放电数阻碍了在 VI 中使用路径梯度估计 (pathwise gradient estimator)；(2) 现有的变分模型的设计本身表示能力欠缺也不够高效，这进一步影响了模型的性能．对于 (1)，我们提出了一个新的可微分的 POGLM，其能够使用比现有模型中采用的得分函数梯度估计 (score function gradient estimator) 更好的路径梯度估计．对于 (2)，我们让变分模型采用了前向后向信息传递采样方案．全面的实验表明我们的可微分 POGLM 带上前向后向信息传递可以在一个合成数据集和两个真实世界数据集上达到更好的效果．此外，我们的方法也给出了可解释性更好的参数，这强调了在神经科学中的重要意义．代码：https://github.com/JerrySoybean/poglm.

1 引言

理解神经连接在神经科学领域是一个重要的研究问题．广义线性模型 (GLM) [Pillow et al., 2008] 及其各种变种 [Linderman et al., 2016, Roudi et al., 2015, Li et al., 2024a] 是推断神经连接的重要工具．然而，一个不可忽略的问题在于，我们记录到的神经数据可能仅仅是目标脑区神经元总体中的一小部分．针对这种不完备问题 (incomplete problem) 的 GLM 被称为部分可见 (partially observable) GLM (POGLM) [Pillow and Latham, 2007, Jimenez Rezende and Gerstner, 2014, Linderman et al., 2017]，其同时考虑可见神经元和隐藏神经元．

POGLM 的总体目标就是在仅用可见神经元的放电序列 X 去学模型参数集 θ，尤其是神经元之间的连接（包括可见神经元和隐藏神经元）．隐藏神经元的放电序列 $\boldsymbol{Z}$ 在 POGLM 中为隐变量．变分推断 (variational inference, VI) [Blei et al., 2017] 是解决这种隐变量模型最常用的方法．VI 的目标是关

于 θ 和 ϕ 最大化观测数据的证据下限 (evidence lower bound)

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{X};\theta,\phi) =& \mathbb{E}_{q(\boldsymbol{Z}|\boldsymbol{X};\phi)}\left[\ln p(\boldsymbol{X},\boldsymbol{Z};\theta) - \ln q(\boldsymbol{Z}|\boldsymbol{X};\phi)\right] \\ =& \ln p(\boldsymbol{X};\theta) - \mathrm{KL}(q(\boldsymbol{Z}|\boldsymbol{X};\phi)\|p(\boldsymbol{Z}|\boldsymbol{X};\theta)) \\ \leqslant& \ln p(\boldsymbol{X};\theta),\end{aligned} \tag{1}$$

其中 $p(\boldsymbol{X},\boldsymbol{Z};\theta)$ 是生成模型，$q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 是由 ϕ 参数化的用于估计后验 $p(\boldsymbol{Z}|\boldsymbol{X};\theta)$ 的变分模型．最大化公式 1 需要从 $q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 中采样隐藏放电序列 $\boldsymbol{Z}$，并计算 $\mathrm{ELBO}(\boldsymbol{X};\theta,\phi)$ 关于 θ 和 ϕ 的梯度估计．

然而，现有的工作已经表明了解决这样一个复杂模型本身是非常困难的．尤其有两点：
(1) $\frac{\partial\,\mathrm{ELBO}(\boldsymbol{X};\theta,\phi)}{\partial\phi}$ 的梯度估计最好是用路径梯度估计 Kingma and Welling [2013]，而离散的 $\boldsymbol{Z}$ 阻碍了路径梯度估计的使用．那就只能采用通常比路径梯度估计方差高很多的得分函数梯度估计 [Paisley et al., 2012, Bengio et al., 2013, Schulman et al., 2015].
(2) 绝大多数现有工作中，变分模型的采样方案 $\boldsymbol{Z} \sim q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 都是一个关于隐藏神经元的 GLM [Jimenez Rezende and Gerstner, 2014, Kajino, 2021, Li et al., 2024b]，也就是说隐藏神经元的放电个数只依赖于可见神经元的历史放电和采样得到的隐藏神经元的历史放电．这一设计使得采样和推断过程非常慢，并且忽略了隐藏神经元放电对未来可见神经元未来放电的影响.

处于对这些问题的考虑，我们的文章旨在解决现有工作中的这两个缺陷，并且更系统全面地研究 POGLM．在第 2 节中，我们会提出一个可微分的 POGLM，使得 VI 中能够使用路径梯度估计．我们还会介绍不同的用于采样隐藏神经元放电的变分分布族，尤其是我们新提出的前向后向信息传递．在第 3 节中，我们会做系统全面的在一个合成数据集和两个真实世界神经数据集上进行实验，比较不同推断方法（包括原始的 POGLM、我们新提出的可微分 POGLM 以及其他过渡模型）× 不同变分采样方案的组合．所有的结果会阐明我们的可微分 POGLM 和我们的前向后向信息传递采样方案的优越性.

2 模型

2.1 背景：POGLM

生成模型. 我们从研究神经元放电序列中神经元之间交互的部分可见广义线性模型 (POGLM) [Pillow and Latham, 2007, Jimenez Rezende and Gerstner, 2014, Linderman et al., 2017] 讲起．假设 N 个神经元中有 V 个是可见的，剩下的 $H = N - V$ 个是隐藏的．记 $\boldsymbol{X} \in \mathbb{N}^{T\times V}$ 为观测到的 V 个可见神经元在 T 个时间桶内的放电序列，$x_{t,v}$ 为第 v 个可见神经元在第 t 个时间桶内的放电个数．$\boldsymbol{Z} \in \mathbb{N}^{T\times H}$ 为记录自 H 个隐藏神经元在 T 个时间桶内的隐放电序列，$z_{t,h}$ 为第 h 个隐藏神经元在第 t 个时间桶内的放电个数．完备的生成模型 (complete generative model) $p(\boldsymbol{X},\boldsymbol{Z};\theta)$ 可以由图 1(a) [Pillow et al., 2008] 表示．对于一个可见神经元 v，其在 t 时的放电率为

$$f_{t,v} = \sigma\left(b_v + \sum_{v'=1}^{V} w_{v\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{h'=1}^{H} w_{v\leftarrow h'} \cdot \left(\sum_{l=1}^{L} z_{t-l,h'}\ \psi_l\right)\right), \tag{2}$$

其放电数由

$$x_{t,v} \sim \mathrm{Poisson}(f_{t,v}) \tag{3}$$

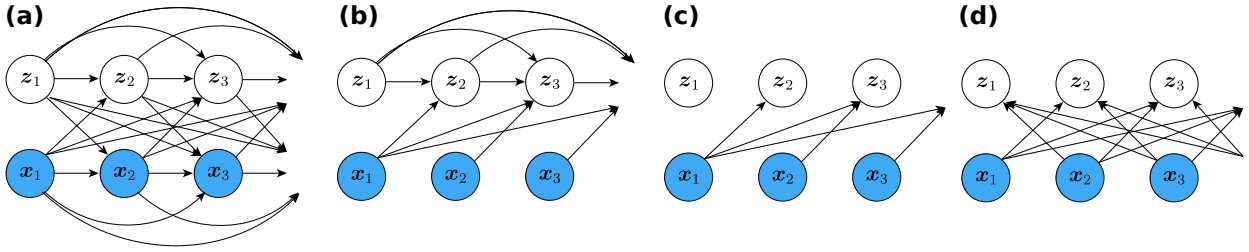

图 1: **(a)**：完备 POGLM 的生成模型 $p(\boldsymbol{X}, \boldsymbol{Z}; \theta)$. **(b)**、**(c)**、**(d)**：前向自循环、前向以及前向后向变分模型采样方案 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$.

生成. $\sigma(\cdot)$ 是一个非线性函数（比如 Softplus）；$\boldsymbol{b}_V = [b_1, b_2, \ldots, b_V]^{\mathrm{T}} \in \mathbb{R}^V$ 是 V 个可见神经元的背景强度向量；$\boldsymbol{W}_{V\leftarrow V} = [w_{v\leftarrow v'}]_{V\times V} \in \mathbb{R}^{V\times V}$ 是表示从可见神经元到可见神经元连接权重的权重矩阵；$\boldsymbol{W}_{V\leftarrow H} = [w_{v\leftarrow h'}]_{V\times H} \in \mathbb{R}^{V\times H}$ 是表示从隐藏神经元到可见神经元连接权重的权重矩阵；$\boldsymbol{\psi} = [\psi_1, \psi_2, \ldots, \psi_L]^{\mathrm{T}} \in \mathbb{R}_+^L$ 是提前给定的用于总结从 $t-L$ 到 $t-1$ 的放电历史的基函数. 类似的，对于一个隐藏神经元 h，其在 t 时的放电率为

$$f_{t,h} = \sigma\left(b_h + \sum_{v'=1}^{V} w_{h\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{h'=1}^{H} w_{h\leftarrow h'} \cdot \left(\sum_{l=1}^{L} z_{t-l,h'}\ \psi_l\right)\right), \tag{4}$$

其放电数由

$$z_{t,h} \sim \mathrm{Poisson}(f_{t,h}) \tag{5}$$

生成. 参数为 $\boldsymbol{b}_H = [b_1, b_2, \ldots, b_H]^{\mathrm{T}} \in \mathbb{R}^H$; $\boldsymbol{W}_{H\leftarrow V} = [w_{h\leftarrow v'}]_{H\times V} \in \mathbb{R}^{H\times V}$; $\boldsymbol{W}_{H\leftarrow H} = [w_{h\leftarrow h'}]_{H\times H} \in \mathbb{R}^{H\times H}$.

因此，POGLM 是一个参数集为 $\theta = \{\boldsymbol{b}, \boldsymbol{W}\}$ 的隐变量模型，其中 $\boldsymbol{b}$ 和 $\boldsymbol{W}$ 可以表示为如下分块矩阵/向量的形式：

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_V \\ \boldsymbol{b}_H \end{bmatrix} \in \mathbb{R}^N, \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{V\leftarrow V} & \boldsymbol{W}_{V\leftarrow H} \\ \boldsymbol{W}_{H\leftarrow V} & \boldsymbol{W}_{H\leftarrow H} \end{bmatrix} \in \mathbb{R}^{N\times N}. \tag{6}$$

变分推断. 由于 POGLM 是一个隐变量模型，那么目标则为学模型参数 θ，同时推断隐变量 $\boldsymbol{Z}$. 鉴于 POGLM 本身的复杂性（图 1(a)），后验分布 $p(\boldsymbol{Z}|\boldsymbol{X}; \theta)$ 没有解析表达式. 因此，我们需要选择一个比较好的由 ϕ 参数化的变分模型 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$ 来做变分推断 (variational inference, VI) [Blei et al., 2017]. 我们会在第 2.3 节中讨论变分模型的不同选择. 现在我们就用最简单的时间齐次泊松变分模型来阐述. 隐藏神经元在 t 时的放电率为

$$f_{t,h} = \sigma(c_h), \tag{7}$$

放电数为 $z_{t,h} \sim \mathrm{Poisson}(f_{t,h})$. 变分参数集为 $\phi = \{\boldsymbol{c}_H\}$，其中 $\boldsymbol{c}_H = [c_1, \ldots, c_H]^{\mathrm{T}}$.

选定了变分模型 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$，我们就可以用 VI 了. 我们关于 θ 和 ϕ 最大化模型的证据下限 $\mathrm{ELBO}(\boldsymbol{X}; \theta, \phi)$（公式 1），从而学到的 θ 可以用于估计模型的真实参数 θ^{true}，变分分布 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$ 可以近似未知的隐变量 $\boldsymbol{Z}$ 的后验分布 $p(\boldsymbol{Z}|\boldsymbol{X}; \theta)$. 由于 POGLM 模型本身的复杂性（图 1(a)），ELBO

（公式 1）也没有解析解．因此我们需要其数值估计

$$\begin{aligned}\widehat{\mathrm{ELBO}}(\boldsymbol{X};\theta,\phi) =&\hat{\mathbb{E}}_{q(\boldsymbol{Z}|\boldsymbol{X};\phi)}[\ln p(\boldsymbol{X},\boldsymbol{Z};\theta)-\ln q(\boldsymbol{Z}|\boldsymbol{X};\phi)]\\ =&\frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{X},\boldsymbol{Z}^{(k)};\theta\right)-\ln q\left(\boldsymbol{Z}^{(k)}\middle|\boldsymbol{X};\phi\right)\right],\end{aligned} \tag{8}$$

其中 $\left\{\boldsymbol{Z}^{(k)}\right\}_{k=1}^{K}$ 是从 $q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 采样得到的 K 个蒙特卡洛样本．关于 θ 的导数很简单（附录 A.1）：

$$\frac{\partial\,\mathrm{ELBO}(\boldsymbol{X};\theta,\phi)}{\partial\theta}\approx\frac{\partial}{\partial\theta}\widehat{\mathrm{ELBO}}(\boldsymbol{X};\theta,\phi). \tag{9}$$

由于 $\boldsymbol{Z}\in\mathbb{N}^{T\times H}$ 是来自隐藏神经元的离散的放电数，其关于 ϕ 在 ϕ_0 处的导数需要得分函数梯度估计（附录 A.1）：

$$\frac{\partial\,\mathrm{ELBO}(\boldsymbol{X};\theta,\phi)}{\partial\phi}\approx\frac{1}{K}\sum_{k=1}^{K}\left\{\left[\ln p\left(\boldsymbol{X},\boldsymbol{Z}^{(k)};\theta\right)-\ln q\left(\boldsymbol{Z}^{(k)}\middle|\boldsymbol{X};\phi_0\right)\right]\frac{\partial}{\partial\phi}\ln q\left(\boldsymbol{Z}^{(k)}\middle|\boldsymbol{X};\phi\right)\right\}. \tag{10}$$

然而，之前的文献表明，关于 ϕ 优化 ELBO 的得分函数梯度估计方差很大 [Paisley et al., 2012, Bengio et al., 2013, Kingma and Welling, 2013, Schulman et al., 2015]，因此可能更好是为从 $q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 采 $\boldsymbol{Z}^{(k)}$ 找到一个重参数化技巧并使用路径梯度估计．由于泊松分布没有重参数化技巧，我们不得不将隐变量 $\boldsymbol{Z}$ 放宽成一个连续变量，然后在下面重写一个可微分的 POGLM.

2.2 一个可微分的 POGLM

为可微分而放宽． 本子节中，我们会通过 Gumbel-Softmax [Jang et al., 2016, Maddison et al., 2016] 分布重写一个可微分的 POGLM．首先先设一个足够大的上限 M 从而 $z_{t,h}\in\{0,1,\ldots,M-1\}$．实际上，$M$ 不需要特别大，因为在很短的一个时间桶内的放电数非常有限．通常，我们希望每个时间桶内的放电数很少（绝大多数都应该是 0 或 1）从而放电序列的精度能够得到保证．不失一般性的，我们在后面的实验中用 $M=5$．这样一来，我们就可以用一个分类分布来近似对应的泊松分布：

$$z_{t,h}\sim\mathrm{Cat}(\boldsymbol{\pi}(f_{t,h})), \tag{11}$$

其中

$$\boldsymbol{\pi}(f)=\left(1-\sum_{m=1}^{M-1}\frac{f^m\mathrm{e}^f}{m!},\frac{f^1\mathrm{e}^f}{1!},\ldots,\frac{f^{M-1}\mathrm{e}^f}{(M-1)!}\right) \tag{12}$$

将截断到 M 泊松分布全部展开．之后我们就可以用 Gumbel-Softmax (GS) 将离散的 $z_{t,h}$ 放宽到一个软的独热形式

$$\tilde{\boldsymbol{z}}_{t,h}=(\tilde{z}_{t,h,0},\ldots,\tilde{z}_{t,h,M-1})\sim\mathrm{GS}(\boldsymbol{\pi}(f_{t,h});\tau), \tag{13}$$

其中 $\tilde{\boldsymbol{z}}_{t,h}$ 是一个在单纯形 $\Delta^{M-1}:=\left\{\boldsymbol{z}\in[0,1]\middle|\sum_{m=0}^{M-1}z_m=1\right\}$ 上的软独热向量．具体来说

$$\tilde{z}_{t,h,m}=\frac{\exp[(\ln\pi_m(f_{t,h})+g_{t,h,m})/\tau]}{\sum_{m'=0}^{M-1}\exp[(\ln\pi_{m'}(f_{t,h})+g_{t,h,m'})/\tau]}, \tag{14}$$

其中 $g_{t,h,m} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0,1)$. 具体实现时，我们可以通过从 $\text{Uniform}(0,1)$ 中独立同分布的采 u 并计算 $g = -\ln(-\ln(u))$ 来实现对 g 的采样. $\tau > 0$ 是温度超参数，用于迫使 $\tilde{\boldsymbol{z}}_{t,h}$ 这一软独热表示接近单纯形 Δ^{M-1} 的一个角. 当 $\tau \to 0$ 时，$\tilde{\boldsymbol{z}}_{t,h}$ 成为了放电数 $z_{t,h}$ 的一个硬度热表示. 在 Gumbel-Softmax 中温度 τ 的选取范围通常是 $[0.1, 1]$. 如果 τ 太大，放宽得就太软了；如果 τ 太小，就会出现数值问题. 在我们的模型中，τ 时用于迫使软独热编码接近单纯形的一个角，所以我们尝试了 $\tau \in \{0.2, 0.5, 1\}$，发现 $\tau = 0.5$ 总能给出稳定且较好的分类分布的近似，且不会出现数值问题，为最优选择. 更多关于 Gumbel-Softmax 分布的细节及其斯然函数在 Jang et al. [2016], Maddison et al. [2016] 中.

生成和变分模型. 给定软独热的 $\tilde{z}_{t,h,m}$，我们定义其等价的软隐藏（神经元）放电数

$$z_{t,h} = \sum_{m=0}^{M-1} m \cdot \tilde{z}_{t,h,m}. \tag{15}$$

现在，我们就可以定义完整的可微分生成模型 $p\left(\boldsymbol{X}, \tilde{\boldsymbol{Z}}; \theta\right)$. 可见神经元的放点序列 $\boldsymbol{X}$ 从公式 2（$f_{t,v}$）和公式 3（泊松）生成，其中公式 2 中的 $z_{t,h}$ 现在由公式 15 定义. 隐变量 $\tilde{\boldsymbol{Z}}$ 不再从公式 4 和公式 5（泊松）生成，而是从公式 4（$f_{t,h}$）和公式 13（GS）生成. 类似的，变分模型中对 $\tilde{\boldsymbol{Z}}$ 的采样从公式 7（$f_{t,h}$）和公式 5（泊松）变为公式 7 和公式 13（GS）. 现在，这个可微分的 POGLM 的完整放电序列为 $\left\{\boldsymbol{X}, \tilde{\boldsymbol{Z}}\right\}$，其中 $\tilde{\boldsymbol{Z}} \in \left(\Delta^{M-1}\right)^{T \times H} \subsetneq [0,1]^{T \times H \times M}$.

路径梯度估计. 当生成模型和变分模型都可微分时，路径梯度估计就可用于优化 ϕ 了. 特别的，通过公式 14 的重参数化技巧（简写为 $\tilde{\boldsymbol{Z}}|\boldsymbol{X}; \phi = r(\boldsymbol{G}|\boldsymbol{X}; \phi)$ where $\boldsymbol{G} \sim \text{Gumbel}(\boldsymbol{G}; 0, 1)$），我们可以得到如下变换关系 $q\left(\tilde{\boldsymbol{Z}}\middle|\boldsymbol{X}; \phi\right)\ \mathrm{d}\tilde{\boldsymbol{Z}} = \text{Gumbel}(\boldsymbol{G}; 0, 1)\ \mathrm{d}\boldsymbol{G}$ [Schulman et al., 2015]. 那么，ELBO 关于 ϕ 的路径梯度估计为:

$$\frac{\partial \,\text{ELBO}(\boldsymbol{X}; \theta, \phi)}{\partial \phi} \approx \frac{\partial}{\partial \phi} \widehat{\text{ELBO}}(\boldsymbol{X}; \theta, \phi), \tag{16}$$

$$\widehat{\text{ELBO}}(\boldsymbol{X}; \theta, \phi) = \frac{1}{K} \sum_{k=1}^{K} \left[\ln p\left(\boldsymbol{X}, \tilde{\boldsymbol{Z}}^{(k)}; \theta\right) - \ln q\left(\tilde{\boldsymbol{Z}}^{(k)}\middle|\boldsymbol{X}; \phi\right)\right], \tag{17}$$

其中 $\tilde{\boldsymbol{Z}}^{(k)} = r\left(\boldsymbol{G}^{(k)}\middle|\boldsymbol{X}; \phi\right)$，$\left\{\boldsymbol{G}^{(k)}\right\}_{k=1}^{K}$ 为采自 $\text{Gumbel}(\boldsymbol{G}; 0, 1)$ 的 K 个蒙特卡洛样本. 具体推导细节见附录 A.1.

放宽到一般的连续分布. 事实上，上面引入的可微分 POGLM 已经可以兼容满足下面两个条件的任意连续分布：(1) 分布由一个代表均值的参数参数化，因为 GLM 结构给出的是一个代表（等价软）放电数 $z_{t,h}$ 的均值统计量的放电率 $f_{t,h}$；(2) 这个分布在采样时有重参数化技巧. 比如，在生成模型和变分模型中，我们可以假设软隐藏放电数是来自指数分布的

$$z_{t,h} \sim \text{Exp}(1/f_{t,h}), \tag{18}$$

其中均值 $f_{t,h}$ 由公式 4 和 7 计算得出. 采样有如下重参数化技巧

$$z_{t,h} = -f_{t,h} \ln(1-u), \quad u \sim \text{Unif}(0,1). \tag{19}$$

相比于在 GS 分布中等价的软隐藏放电数靠近一个 $\{0,1,\ldots,M-1\}$ 中的整数，来自指数分布的 $z_{t,h}$ 可以是 $\mathbb{R}_{\geqslant 0}$ 的任意值．第 3 节和图 3 详细描述了一些可选的分布．

2.3 变分模型的采样方案

目前为止，我们已经提出了可微分的 POGLM 来解决第 1 节中提出的第一个问题．现在我们解决第二个问题—变分模型的选取．具体来说，我们需要设计变分模型中 $f_{t,h}$ 的公式．显然，公式 7 表示的时间齐次（时齐）模型太过简单了，以至于 $q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 的变分分布族和后验分布 $p(\boldsymbol{Z}|\boldsymbol{X};\theta)$ 相去甚远．一个与真实后验分布更接近的好的变分分布族对 VI 的成功至关重要．这里我们讨论以下五个候选模型：

• **时齐泊松**：$f_{t,h}=\sigma(c_h),\ \forall t\in\{1,\ldots,T\}$，变分参数集为 $\phi=\{\boldsymbol{c}_H\}$，其中 $\boldsymbol{c}_H=[c_1,\ldots,c_H]^{\mathrm{T}}$．然而，这个在实际情况下太简单了，不能很好的作为变分分布族．

• **非时齐泊松（平均场）**：$f_{t,h}=\sigma(c_{t,h})$，$\phi=\left\{\boldsymbol{C}_{T\times H}\in\mathbb{R}^{T\times H}\right\}$．尽管平均场理论在很多隐变量模型中有广泛的应用，但其缺少对可见放电序列 $\boldsymbol{X}$ 的依赖．对于 POGLM 来说，这样学到的 ϕ 只对应于训练放电序列 $p(\boldsymbol{Z}_{\text{train}}|\boldsymbol{X}_{\text{train}};\theta)\approx q(\boldsymbol{Z}_{\text{train}}|\boldsymbol{X}_{\text{train}};\phi)$，但无法推广到测试放电序列 $p(\boldsymbol{Z}_{\text{test}}|\boldsymbol{X}_{\text{test}};\theta)\not\approx q(\boldsymbol{Z}_{\text{test}}|\boldsymbol{X}_{\text{test}};\phi)$. 此外, 时齐和非时齐泊松都没有神经元之间的信息传递, 因此对学习神经连接矩阵 $\boldsymbol{W}$ 没有任何帮助.

• **前向自循环** [Jimenez Rezende and Gerstner, 2014, Kajino, 2021]：一个很典型又直观的想法就是，认为真实的后验 $p(\boldsymbol{Z}|\boldsymbol{X};\theta)$ 可以由一个关于 $\boldsymbol{Z}$ 的 GLM 的变分分布 $q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 近似，其中 $\boldsymbol{X}$ 是固定的（图 1(b)），即，

$$f_{t,h}=\sigma\left(c_h+\sum_{v'=1}^{V}a_{h\leftarrow v'}\cdot\left(\sum_{l=1}^{L}x_{t-l,v'}\ \psi_l\right)+\sum_{h'=1}^{H}a_{h\leftarrow h'}\cdot\left(\sum_{l=1}^{L}z_{t-l,h'}\ \psi_l\right)\right),\tag{20}$$

$\phi=\{\boldsymbol{c}_H,\boldsymbol{A}\}$．特别的，

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{O}_{V\leftarrow V} & \boldsymbol{O}_{V\leftarrow H}\\ \boldsymbol{A}_{H\leftarrow V} & \boldsymbol{A}_{H\leftarrow H}\end{bmatrix}\in\mathbb{R}^{N\times N}.\tag{21}$$

上面两个块 $\boldsymbol{O}_{V\leftarrow V},\boldsymbol{O}_{V\leftarrow H}$ 全为零因为我们不用采可见的放电序列 $\boldsymbol{X}$．$\boldsymbol{A}_{H\leftarrow V}$ 和 $\boldsymbol{A}_{H\leftarrow H}$ 分别表示可见到隐藏、隐藏到隐藏的影响．为了用公式 20，我们需要串行地从 $t=1$ 到 $t=T$ 采样，因为当前的样本 $z_{t,h}$ 依赖于之前的样本 $\boldsymbol{Z}_{t-L:t-1,1:H}$.

• **前向**：由于前向自循环采样过程很低效，一个更简单的方法是去掉隐藏到隐藏的那一块（也就是公式 20 中的第三项）．这样我们就得到了前向信息传递方案（图 1(c)）：

$$f_{t,h}=\sigma\left(c_h+\sum_{v'=1}^{V}a_{h\leftarrow v'}\cdot\left(\sum_{l=1}^{L}x_{t-l,v'}\ \psi_l\right)\right),\tag{22}$$

$\phi=\{\boldsymbol{c}_H,\boldsymbol{A}\}$．这里，

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{O}_{V\leftarrow V} & \boldsymbol{O}_{V\leftarrow H}\\ \boldsymbol{A}_{H\leftarrow V} & \boldsymbol{O}_{H\leftarrow H}\end{bmatrix}\in\mathbb{R}^{N\times N}.\tag{23}$$

前向变分分布可以并行采样，因为 $z_{t,h}$ 不再互相依赖．注意到去掉了隐藏到隐藏的块理论上来讲可能忽略了隐藏到隐藏的影响因素．但实际上，由于隐藏神经元生成过程中的长序列依赖，学到隐藏到隐藏

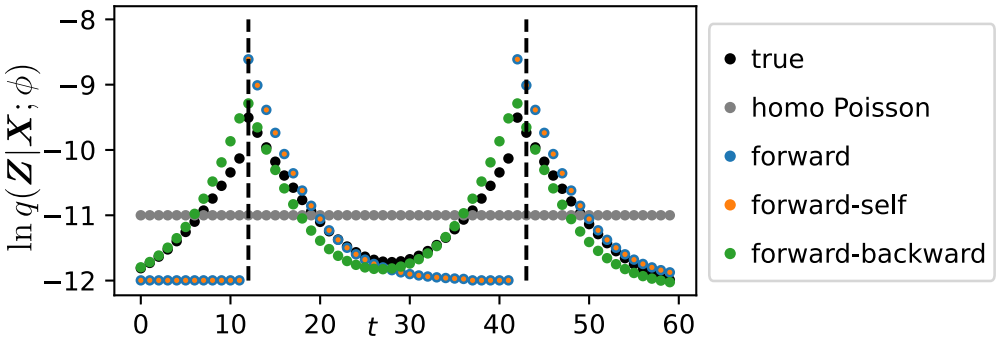

图 2: 一个比较不同变分分布 $q(\boldsymbol{Z}|\boldsymbol{X};\phi)$ 的例子. 真实后验是 $p(\boldsymbol{Z}|\boldsymbol{X};\theta)$（黑点）. 这里只有一个可见神经元和一个隐藏神经元. 两个来自可见神经元的可见放电在短划线处发生. 每条虚线表示一个隐藏放电发生在不同时间桶内的近似对数斯然. 只有前向后向准确描述了真实的后验分布. 前向和前向自循环由于缺乏反向传播的信息，都无法抓住可见放电前对数斯然的上升趋势.

的的 $\boldsymbol{W}_{H\leftarrow H}$ 是非常困难的. 此外，尽管实际情况下大多数神经元都是没有被记录到的，但人们还是通常会假设 POGLM 中的隐藏神经元之占少数（$V > H$），寄希望于这些隐藏神经元可以作为代表. 事实上，由于 POGLM 问题本身的复杂性，学很多隐藏神经元是不现实的. 如果有很多隐藏神经元，问题就会非常繁琐，目前没有任何方法能成功. 所以，在 $V > H$ 的条件下，忽略 $\boldsymbol{W}_{H\leftarrow H}$ 这一块儿可能不会很影响结果.

• **前向后向**: 前面两种采样方案都忽略了一个重要的关系—模拟生成过程 $p(\boldsymbol{X}, \boldsymbol{Z};\theta)$ 中隐藏到可见的影响 $\boldsymbol{W}_{V\leftarrow H}$. 因此，这里我们引入前向后向信息传递方案（图 1(d)），

$$f_{t,h} = \sigma\left(c_n + \sum_{v'=1}^{V} a_{h\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{v'=1}^{V} a_{v'\leftarrow h} \cdot \left(\sum_{l=1}^{L} x_{t+l,v'}\ \psi_l\right)\right), \tag{24}$$

$\phi = \{\boldsymbol{c}_H, \boldsymbol{A}\}$. 这里,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O}_{V\leftarrow V} & \boldsymbol{A}_{V\leftarrow H} \\ \boldsymbol{A}_{H\leftarrow V} & \boldsymbol{O}_{H\leftarrow H} \end{bmatrix} \in \mathbb{R}^{N\times N}, \tag{25}$$

其中 $\boldsymbol{A}_{V\leftarrow H}$ 块模拟的是生成过程 $p(\boldsymbol{X}, \boldsymbol{Z};\theta)$ 中隐藏到可见的影响 $\boldsymbol{W}_{V\leftarrow H}$. 具体来说，在采当前 $z_{t,h}$ 时，引入了来自未来可见放电序列 $\boldsymbol{X}_{t+1:t+L,1:V}$ 的贡献（即公式 24 中的第三项）.

图 2 可视化地比较了不同变分分布，帮我们理解前向后向采样在近似真实后验分布上的优越性.

3 实验

为了全面的对比分析，我们考虑不同**推断方法** × 不同**变分采样方案**的**方法组合**.

推断方法. 我们考虑七种推断方法. 每一种推断方法由隐藏放电关于放电率的分布 $\mathbb{P}[z; f]$（见表 1）和 VI 中使用的梯度估计来确定

• **泊松 (Pois)**: 这是原始的 POGLM（公式 2、3、4, 5、7）. 这是命名自原始 POGLM 的放电序列 $\boldsymbol{Z}$

表 1: 在生成模型和变分模型中使用的不同的隐藏放电关于放电率的分布 $\mathbb{P}[z; f]$. 出于简洁的目的，我们忽略了 z、$\tilde{\boldsymbol{z}} = (\tilde{z}_0, \ldots, \tilde{z}_{M-1})$ 和 f 中用于指示隐藏神经元和时间桶的下标.

分布	样本	似然	是否可用路径梯度估计
泊松 (Pois)	$z \sim \text{Poisson}(f)$	$\mathbb{P}[z; f] = \frac{f^z \mathrm{e}^{-z}}{z!}$	✗
分类 (Cat)	$z \sim \text{Cat}(\boldsymbol{\pi}(f))$	$\mathbb{P}[z; f] = \boldsymbol{\pi}(f)_z$	✗
Gumbel-Softmax (GS)	$\tilde{\boldsymbol{z}}_{t,h} \sim \text{GS}(\boldsymbol{\pi}(f_{t,h}); \tau)$	Eq. 36	✓
指数 (Exp)	$z \sim \text{Exp}\left(\frac{1}{f}\right)$	$\mathbb{P}[z; f] = \frac{1}{f} \exp\left(-fz\right)$	✓
Rayleigh (Ray)	$z \sim \text{Ray}\left(\sqrt{\frac{2}{\pi}} f\right)$	$\mathbb{P}[z; f] = \frac{\pi z}{2f^2} \exp\left(-\frac{\pi z^2}{4f^2}\right)$	✓
Half-normal (HN)	$z \sim \text{HN}\left(\sqrt{\frac{\pi}{2}} f\right)$	$\mathbb{P}[z; f] = \frac{2}{\pi f} \exp\left(-\frac{z^2}{\pi f^2}\right)$	✓

所属的泊松分布. 由于这是一个离散分布，VI 中之能使用得分函数梯度估计.

• **分类 (Cat)**: 这是第一个原始 POGLM 和可微分 POGLM 之间的过渡模型, 其中我们不使用公式 13 中的 Gumbel-Softmax 来近似，但保留了分类分布（公式 11）. 和泊松分布一样，VI 中之能使用得分函数梯度估计.

• **Gumbel-Softmax-score (GS-s)**: 这个是以 GS（公式 13）作为软隐藏放电数分布的可微分 POGLM. 只不过我们在更新 ϕ 时仍然使用得分函数梯度估计（公式 10），尽管这个模型已经可微分了.

• **Gumbel-Softmax-pathwise (GS-p)**: 这个是以 GS（公式 13）作为软隐藏放电数分布的可微分 POGLM. 在更新 ϕ 时使用路径梯度估计（公式 16）. 我们期望这个推断方法比前面三个要好. 为了实验从 GS 推广到其它单参数连续分布，我们还试了下面三种分布，并使用路径梯度估计.

• **指数 (Exp)**.

• **Rayleigh (Ray)**.

• **Half-normal (HN)**.

引入两个过渡模型（Cat 和 GS-s）的目的是使模型从原始的以泊松分布作为隐藏放电数分布且使用得分函数梯度估计的 POGLM 一步步地变到以 GS 作为隐藏放电数分布且使用路径梯度估计的可微分 POGLM. 通过这一控制变量法，我们可以更好的理解最终的带有连续软隐藏放电数分布的可微分 POGLM. 由于我们也不知道什么样的单参数分布更好, 我们就试了三个常用的: Exp、Ray 和 HN. 图 3 画出了这些分布.

变分采样方案. 我们考虑三种采样方案.

• **前向 (F)**: 公式 20 和图 1(c) 展示的采样方案.

• **前向自循环 (FS)**: 公式 22 和图 1(b) 展示的采样方案.

• **前向后向 (FB)**: 公式 24 和图 1(d) 展示的采样方案.

时齐和非时齐泊松在之后的实验中不再考虑了，因为它们过于简单或无法兼容测试集（在 2.3 节中阐述了）.

原始的解决 POGLM 的方法可以视为 Poisson × FS [Pillow and Latham, 2007, Jimenez Rezende and Gerstner, 2014, Linderman et al., 2017]. 我们新提出的基于 GS 和其它连续分布的且带有 FB 信

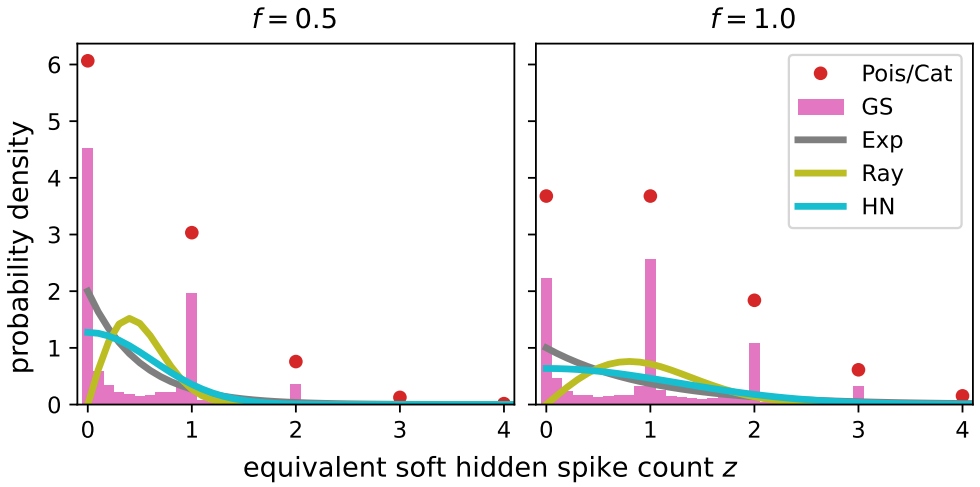

图 3: 在放电率分别为 $f = 0.5$ 和 $f = 1.0$ 时，不同（软）隐藏放电数分布的选取的可视化．绝大多数用于近似原始泊松分布的来自 GS 的 z 都靠近整数点，但那三个连续分布（Exp、Ray 和 HN）则没有这样的性质．

息传递方案的可微分 POGLM 应当是我们期望的最优组合．为了更加清楚地理解这些组合，附录 A.2 给出了一份这些方法组合的完整总结．

评价． 尽管我们有不同的推断方法，在测试集上评价对数似然 (log-likelihood, LL) 时，所有推断方法都重新设置成原始的 POGLM 形式．也就是说为了公平比较，我们全部使用泊松对数似然（公式 3 和公式 5）．LL 指标可以在合成数据集和真实世界数据集上使用．此外，出于我们在第 1 节中解释了，直接应用 VI 求解本身就很困难的 POGLM 无法得到理想的参数恢复，我们很关心参数的恢复情况．所以我们还会在合成数据集上比较参数估计的平均绝对值误差．

3.1 合成数据集

合成数据集旨在全面比较不同方法组合．在知道参数真实值的情况下，我们也可以验证更好的表现对应更小的参数误差．我们还可以检查应用路径梯度估计带来的好处．

数据集． 我们随机创建了 10 组参数用于生成合成数据集，其中参数来自 $w_{n \leftarrow n'} \overset{\text{i.i.d.}}{\sim} \text{Unif}(-2, 2)$ 以及 $b_n \overset{\text{i.i.d}}{\sim} \text{Unif}(-0.5, 0.5)$．每组参数对应一个试次，所以共有 10 个试次．模型中包括 5 个神经元，其中前 3 个是可见的，剩下 2 个是隐藏的．每一个试次我们都生成 40 个放电序列用于训练，再生成 20 个放电序列用于测试．每个放电序列有 100 个时间桶．

实验设置． 模型线性权重和偏置的初值也像上面一样初始化．我们选用 Adam 优化器 [Kingma and Ba, 2014]，学习率为 0.05．优化跑 20 轮，每轮包括 4 个批，批大小为 10．每个试次我们都用不同的随机数种子重复优化十次，并基于此画出结果和误差线．

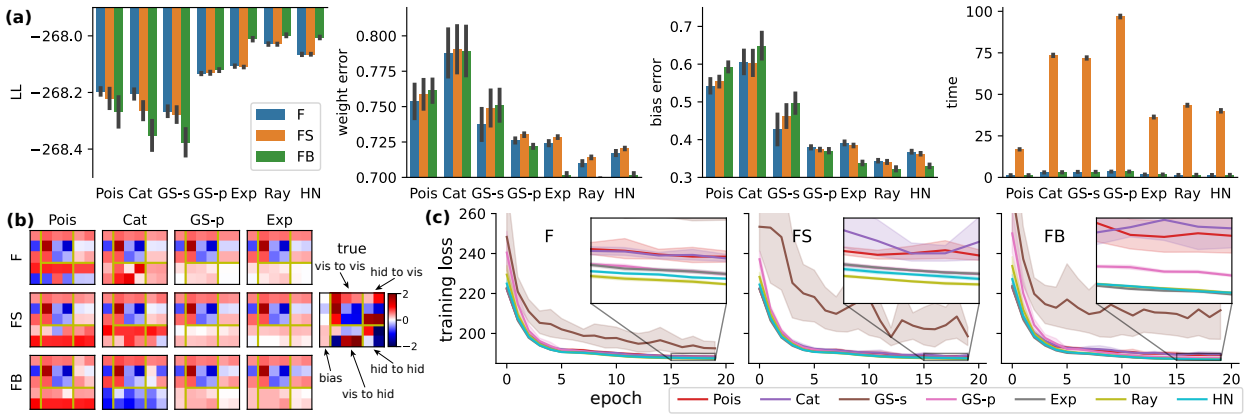

图 4: **(a)**：不同方法组合的测试集对数似然 (LL)、权重误差、偏置误差以及运行时间. **(b)**：一些方法组合在某一试次上学到的权重矩阵和偏置向量与真实值的比较. 所有方法组合的可视化结果在附录 A.3 中的图 8 中. **(c)**：不同方法组合的学习曲线.

结果. 在图 4(a) 中，我们可以看到每个采样方案下使用分类分布或 GS 去近似离散泊松分布时，测试集 LL 都会下降. 然而当我们将梯度估计从得分函数换成路径时，GS 的 LL 显著超过了原始的泊松. 当我们把分布从 GS 换成 Exp、Ray 或 HN 时，LL 进一步提升. 这意味着尽管隐藏放电时来自泊松分布，后验可能不是泊松的而是一个在形状上可能更接近比如指数分布的某一离散分布. 从不同采样方案看，当推断方法选得好时，FB 就比 F 和 FS 要强了，比如推断方法是 GS-p 或 Exp 时. 没有好的推断方法时，前向采样方案（F）比较好，因为它简单. 总结来看，一个可微分的 POGLM 使用路径梯度估计 × FB 采样方案能得到较好的结果.

图 4(a) 所示的权重误差和偏置误差量化的验证了好的 LL 对应更小的参数误差. 和 LL 柱状图一致，权重误差和偏置误差的柱状图也表明 Exp、Ray 和 HN 优于 GS-p 优于其它. 当使用连续分布以及路径梯度估计作为推断方法时，FB 是最优的采样方案. 图 4(b) 展示了一些方法组合恢复出的权重矩阵和偏置向量以及用于生成数据集的真实值. 它们之间的主要区别在于和隐藏神经元相关的权重块以及隐藏神经元的偏置.

除性能外，图 4(a) 还比较了不同方法组合的运行时间. 由于 FS 的串行依赖，其运行时间显著长于 F 和 FB. 这解释了除去变分模型中“自循环”信息传递的好处. 此外，将离散的泊松分布转化为其硬（Cat）或软（GS）近似分布需要额外花些时间.

3.2 视网膜神经节细胞 (Retinal Ganglion Cell, RGC) 数据集

这一数据集旨在理解不同方法组合在真实世界数据集上的表现，以及估计出的模型参数的可解释性.

数据集. 接下来我们就把不同方法用在分析真实神经放电序列上. 放电序列来自 27 个视网膜神经节神经元. 记录时小鼠在进行一项时长约 20 分钟的视觉任务 [Pillow and Scott, 2012]. 具体的，神经元 1–16 位 OFF 细胞，而 17–20 为 ON 细胞.

实验设置. 我们将放电序列分为训练集和测试集，用前 2/3 段作为训练集，后 1/3 段作为测试集．原始的放电序列用宽度为 50 ms 的时间桶来离散化．为了使用随机梯度下降算法，我们把完整序列分成了 14400 段，每段包含 100 个时间桶．由于我们不知道设置多少个隐藏神经元合适，我们先训练了一个完全可见的 GLM 作为基线．然后，我们假设有 $H \in \{1, 2, 3\}$ 个隐藏代表神经元，基于此用不同的方法组合训练模型．我们用 Adam 优化器、0.02 学习率来进行优化．整个训练过程有 20 轮，批大小为 32．我们用不同的随机数种子重复每个训练、测试过程十次，并基于此报告结果和误差线．

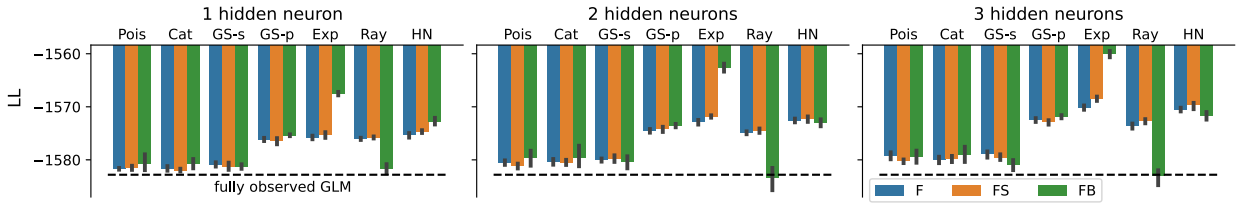

图 5: 在假设 $H \in \{1, 2, 3\}$ 个隐藏神经元的情况下，不同方法组合的测试集对数似然 (LL)．黑色短划线表示完全可见 GLM 的测试集 LL，也就是基线.

结果. 和合成数据集类似，我们在图 5 中画出了不同方法组合在 1、2、3 个隐藏神经元下的测试集 LL．此外，图 5 中还包括了作为基线的完全可见 GLM．首先，当假设存在隐藏神经元时，所有方法组合都都变得比基线好了，除了 Ray × FB．第二，可微分推断方法普遍比不可微分推断方法要好．尤其是 Exp × FB 显著优于其它方法组合．第三，对于绝大多数方法组合来说，增加隐藏神经元的数量能够提高 LL，尤其是 GS-p 和 Exp.

除了知道可微分推断方法 × FB 比其它好以外，我们还关心学到的模型参数的可解释性．以一个隐藏神经元为例，图 6 展示了，Exp × FB 学到的隐藏神经元可以视为一个负反馈单元．具体来讲，这个隐藏代表到所有 OFF 细胞的权重都是负的，到所有 ON 细胞的权重都是正的；几乎所有 OFF 细胞到这个隐藏代表的权重都是正的，所有 ON 细胞到这一隐藏代表的权重都是负的．也就是说，这一隐藏代表到可见神经元的权重的正负情况清楚的表明了可见神经元的类型．类似但较弱的结果也能在 GS-p 中看到，但在泊松中就没有.

附录 A.3 中的图 9 也能够在更多隐藏神经元的情况下支持可解释性．比如，当有从 RGC 数据集中学三个隐藏代表神经元时，Exp × FB 学到的隐藏到可见的权重具有特定且非随机的样式．这个样式代表来一个隐藏代表神经元和可见神经元之间更复杂多变的交互方式.

3.3 PVC-5 数据集

这一数据集旨在研究不同方法组合的性能关于隐藏神经元数量的变化.

数据集. 最后，我们把不同方法组合应用到记录自初级视觉皮层 (primary visual cortex) 的数据集（PVC-5）Chu et al. [2014][1]．这一数据集记录了 15 分钟猕猴（恒河猴）在没有视觉刺激的情况下，初

[1] https://crcns.org/data-sets/pvc/pvc-5

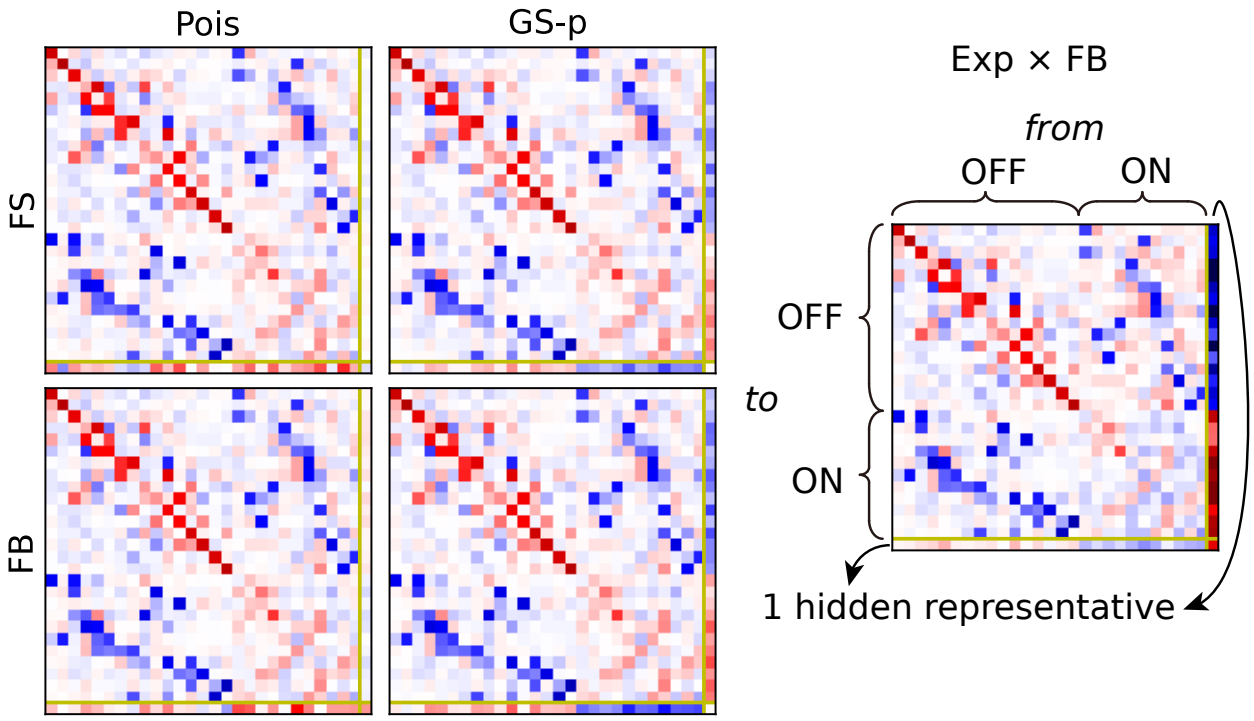

图 6: The learned weight matrix of selected method combinations. Visualization of all method combinations is in Fig. 9 in Appendix. A.3.

级视觉皮层 (VI) 的放电序列.

实验设置. 和 RGC 数据集类似，我们用开始的 7.5 min 作为训练集，用后面的 7.5 min 作为测试集. 原始放电序列用宽度为 20 ms 的时间桶离散化成放电数. 训练集被均等地分为 225 段. 用于训练的批大小为 25. 由于仅有三个可见神经元，我们可以尝试更多数量的隐藏神经元 $H \in \{1, \ldots, 9\}$ 以理解性能关于隐藏神经元个数的变化，尤其是当 $H \gg V$ 时. 我们用 Adam 优化器优化 20 轮，学习率设为 0.1. 我们用不同的随机数种子重复训练和测试过程 10 次，并基于此报告性能和误差线.

结果. 图 7 展示了不同方法组合的性能随隐藏神经元个数的变化曲线. 无论我们选哪种方法组合，最优的隐藏神经元个数总是不超过 3. 这意味着假设存在更多隐藏神经元不一定总能保证性能的提升，还有可能引入冗余，导致效果下降. 当有更多隐藏神经元时，那些不可微分的推断方法甚至变得比完全可见 GLM 还差.

在所有方法组合中，Exp × FB 同时隐藏神经元个数小于 3 看上去是最好的. GS-p（所有三种采样方案）相较于其它推断方法，对不同的隐藏神经元个数更稳健. 此外，由于使用了路径梯度估计，可微分的推断方法比不可微分的推断方法，测试集对数似然更小.

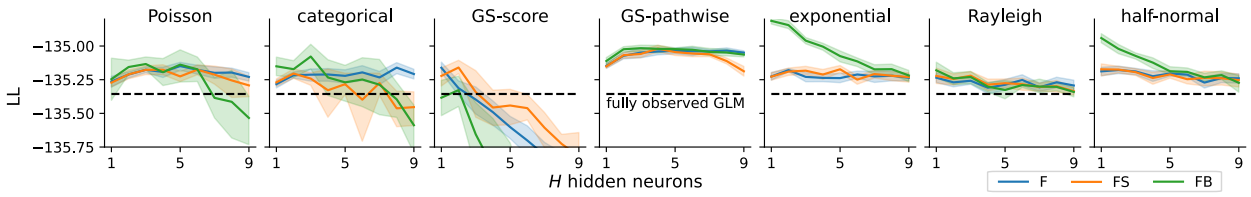

图 7: 不同方法组合下，测试集 LL 随隐藏神经元数量 H 的变化曲线.

4 相关工作

一些之前的工作考虑了 POGLM 的点过程形式，也就是广义多元部分可见 Hawkes 过程 (generalized multivariate partially observable Hawkes process)，其中放电序列不用时间桶离散化成放电数，而是保留其原始的放电时间戳形式. 比如，Zhou and Sun [2021], Shelton et al. [2018], Mei et al. [2019] 把这一问题视为数据丢失问题（所有隐藏神经元的数据全部丢失）；Kajino [2021] 提出了一个可微分的点过程模型，使点过程也能使用路径梯度估计.

然而使用点过程的话，存储放电时间戳的数据结构通常不理想 [Xu, 2018]. 具体的原因在附录 A.4 中. 本文我们只关注 POGLM，也就是提前用时间桶离散化的放电数序列. 更多关于（离散）GLM 和（连续）广义 Hawkes 过程的关系的详细讨论也在附录 A.4 中.

5 讨论

本文我们提出了一个可微分版本的部分可见广义线性模型（POGLM），使得在变分推断 (VI) 时能够使用路径梯度估计. 由于现有的前向自循环采样方案表达性很差、采样效率很低剋哦们提出了新的前向后向信息传递采样方案，引入了从隐藏神经元到可见神经元的信息传递，从而使变分模型更好的服务于 VI. 不同方法组合在一个合成数据集和两个真实世界数据集上的全面比较表明，带有前向后向采样方案的可微分推断方法可以在测试集上达到更高的似然，也可以取得更好的参数恢复.

注意到从 Gumbel-Softmax 分布放宽到一般连续分布丧失了 $\boldsymbol{Z}$ 作为放电数的意义，但能学得更好. 研究是否一个一般的连续分布会比离散的泊松分布更接近真实后验分布，既有趣又富有挑战性. 这一局限性可能是一个很大的话题，值得未来进一步研究.

参考文献

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Scott Linderman, Ryan P Adams, and Jonathan W Pillow. Bayesian latent structure discovery from multi-neuron recordings. *Advances in neural information processing systems*, 29, 2016.

Yasser Roudi, Benjamin Dunn, and John Hertz. Multi-neuronal activity and functional connectivity in cell assemblies. *Current opinion in neurobiology*, 32:38–44, 2015.

Chengrui Li, Soon Ho Kim, Chris Rodgers, Hannah Choi, and Anqi Wu. One-hot generalized linear model for switching brain state discovery. In *International Conference on Learning Representations*, 2024a.

Jonathan Pillow and Peter Latham. Neural characterization in partially observed populations of spiking neurons. *Advances in Neural Information Processing Systems*, 20, 2007.

Danilo Jimenez Rezende and Wulfram Gerstner. Stochastic variational learning in recurrent spiking networks. *Frontiers in computational neuroscience*, 8:38, 2014.

Scott W Linderman, Yixin Wang, and David M Blei. Bayesian inference for latent hawkes processes. *Advances in Neural Information Processing Systems*, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. *Advances in neural information processing systems*, 28, 2015.

Hiroshi Kajino. A differentiable point process with its application to spiking neural networks. In *International Conference on Machine Learning*, pages 5226–5235. PMLR, 2021.

Chengrui Li, Yule Wang, Weihan Li, and Anqi Wu. Forward χ^2 divergence based variational importance sampling. In *International Conference on Learning Representations*, 2024b.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jonathan Pillow and James Scott. Fully bayesian inference for neural models with negative-binomial spiking. *Advances in neural information processing systems*, 25, 2012.

CCJ Chu, Ping F Chien, and CP Hung. Multi-electrode recordings of ongoing activity and responses to parametric stimuli in macaque v1. *CRCNS.org*, 10:K0J1012K, 2014.

Zihan Zhou and Mingxuan Sun. Multivariate hawkes processes for incomplete biased data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 968–977. IEEE, 2021.

Christian Shelton, Zhen Qin, and Chandini Shetty. Hawkes process inference with missing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Hongyuan Mei, Guanghui Qin, and Jason Eisner. Imputing missing events in continuous-time event streams. In *International Conference on Machine Learning*, pages 4475–4485. PMLR, 2019.

Hongteng Xu. Poppy: a point process toolbox based on pytorch. *arXiv preprint arXiv:1810.10122*, 2018.

附录 A

A.1 ELBO 的梯度估计

这里我们给出 $\mathrm{ELBO}(\boldsymbol{X};\theta,\phi)$ 关于 θ 和 ϕ 的梯度估计的详细推导过程. 关于 θ 的导数为:

$$\begin{aligned}\frac{\partial \mathrm{ELBO}(\boldsymbol{X};\theta,\phi)}{\partial\theta} =& \frac{1}{|\mathbb{N}^{T\times H}|}\sum_{\boldsymbol{Z}\in\mathbb{N}^{T\times H}} q(\boldsymbol{Z}|\boldsymbol{X};\phi)\frac{\partial}{\partial\theta}\left[\ln p(\boldsymbol{X},\boldsymbol{Z};\theta)-\ln q(\boldsymbol{Z}|\boldsymbol{X};\phi)\right]\\ \approx& \frac{1}{K}\sum_{k=1}^{K}\frac{\partial}{\partial\theta}\left[\ln p\left(\boldsymbol{X},\boldsymbol{Z}^{(k)};\theta\right)-\ln q\left(\boldsymbol{Z}^{(k)}\middle|\boldsymbol{X};\phi\right)\right]\\ =& \frac{\partial}{\partial\theta}\widehat{\mathrm{ELBO}}(\boldsymbol{X};\theta,\phi).\end{aligned} \tag{26}$$

ELBO 关于 ϕ 在 ϕ_0 处的导数的得分函数梯度估计（公式 10）为:

$$\begin{aligned}\frac{\partial \mathrm{ELBO}(\boldsymbol{X};\theta,\phi)}{\partial\phi} =& \frac{1}{|\mathbb{N}^{T\times H}|}\sum_{\boldsymbol{Z}\in\mathbb{N}^{T\times H}} \frac{\partial}{\partial\phi}q(\boldsymbol{Z}|\boldsymbol{X};\phi)\left[\ln p(\boldsymbol{X},\boldsymbol{Z};\theta)-\ln q(\boldsymbol{Z}|\boldsymbol{X};\phi_0)\right]\\ &+ q(\boldsymbol{Z}|\boldsymbol{X};\phi_0)\frac{\partial}{\partial\phi}\left[\ln p(\boldsymbol{X},\boldsymbol{Z};\theta)-\ln q(\boldsymbol{Z}|\boldsymbol{X};\phi)\right]\\ =& \frac{1}{|\mathbb{N}^{T\times H}|}\sum_{\boldsymbol{Z}\in\mathbb{N}^{T\times H}}\left[\ln p(\boldsymbol{X},\boldsymbol{Z};\theta)-\ln q(\boldsymbol{Z}|\boldsymbol{X};\phi_0)\right]q(\boldsymbol{Z}|\boldsymbol{X};\phi)\frac{\partial}{\partial\phi}\ln q(\boldsymbol{Z}|\boldsymbol{X};\phi)\\ &- \frac{1}{|\mathbb{N}^{T\times H}|}\sum_{\boldsymbol{Z}\in\mathbb{N}^{T\times H}}\frac{\partial}{\partial\phi}q(\boldsymbol{Z}|\boldsymbol{X};\phi)\\ \approx& \frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{X},\boldsymbol{Z}^{(k)};\theta\right)-\ln q\left(\boldsymbol{Z}^{(k)}\middle|\boldsymbol{X};\phi_0\right)\right]\frac{\partial}{\partial\phi}\ln q\left(\boldsymbol{Z}^{(k)}\middle|\boldsymbol{X};\phi\right)-0.\end{aligned} \tag{27}$$

ELBO 关于 ϕ 的导数的路径梯度估计（公式 16）为:

$$\begin{aligned}\frac{\partial \mathrm{ELBO}(\boldsymbol{X};\theta,\phi)}{\partial\phi} =& \frac{\partial}{\partial\phi}\int_{\tilde{\boldsymbol{Z}}} q\left(\tilde{\boldsymbol{Z}}\middle|\boldsymbol{X};\phi\right)\left[\ln p\left(\boldsymbol{X},\tilde{\boldsymbol{Z}};\theta\right)-\ln q\left(\tilde{\boldsymbol{Z}}\middle|\boldsymbol{X};\phi_0\right)\right]\,\mathrm{d}\tilde{\boldsymbol{Z}}\\ =& \frac{\partial}{\partial\phi}\int_{\boldsymbol{G}}\mathrm{Gumbel}(\boldsymbol{G};0,1)\left[\ln p\left(\boldsymbol{X},r(\boldsymbol{G}|\boldsymbol{X};\phi);\theta\right)-\ln q(r(\boldsymbol{G}|\boldsymbol{X};\phi)|\boldsymbol{X};\phi)\right]\,\mathrm{d}\boldsymbol{G}\\ \approx& \frac{\partial}{\partial\phi}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{X},r\left(\boldsymbol{G}^{(k)}\middle|\boldsymbol{X};\phi\right);\theta\right)-\ln q\left(r\left(\boldsymbol{G}^{(k)}\middle|\boldsymbol{X};\phi\right)\middle|\boldsymbol{X};\phi\right)\right]\\ =& \frac{\partial}{\partial\phi}\widehat{\mathrm{ELBO}}(\boldsymbol{X};\theta,\phi).\end{aligned} \tag{28}$$

A.2 方法组合

我们把不同的方法组合在这里总结出来.

A.2.1 生成过程

t 时一个可见神经元 v 的放电率 $f_{t,v}$ 以及一个隐藏神经元 h 的放电率 $f_{t,h}$ 分别为:

$$\begin{cases} f_{t,v} = \sigma\left(b_v + \sum_{v'=1}^{V} w_{v\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{h'=1}^{H} w_{v\leftarrow h'} \cdot \left(\sum_{l=1}^{L} z_{t-1,h'}\ \psi_l\right)\right), \\ f_{t,h} = \sigma\left(b_h + \sum_{v'=1}^{V} w_{h\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{h'=1}^{H} w_{h\leftarrow h'} \cdot \left(\sum_{l=1}^{L} z_{t-1,h'}\ \psi_l\right)\right). \end{cases} \tag{29}$$

参数集为 $\theta = \{\boldsymbol{b}, \boldsymbol{W}\}$, 其中 $\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_V \\ \boldsymbol{b}_H \end{bmatrix} \in \mathbb{R}^N$, $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{V\leftarrow V} & \boldsymbol{W}_{V\leftarrow H} \\ \boldsymbol{W}_{H\leftarrow V} & \boldsymbol{W}_{H\leftarrow H} \end{bmatrix} \in \mathbb{R}^{N\times N}$. 对于可见神经元, 总是 $x_{t,n} \sim \text{Poisson}(f_{t,n})$.

A.2.2 变分采样方案

- **时齐泊松**

$$f_{t,h} = \sigma(c_h). \tag{30}$$

变分参数集为 $\phi = \{\boldsymbol{c}_H\}$.

- **非时齐泊松**

$$f_{t,h} = \sigma(c_{t,h}). \tag{31}$$

变分参数集为 $\phi = \{\boldsymbol{C}_{T\times H}\}$.

- **前向（F）**

$$f_{t,h} = \sigma\left(c_h + \sum_{v'=1}^{V} a_{h\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right)\right). \tag{32}$$

变分参数集为 $\phi = \{\boldsymbol{c}_H, \boldsymbol{A}\}$, 其中 $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O}_{V\leftarrow V} & \boldsymbol{O}_{V\leftarrow H} \\ \boldsymbol{A}_{H\leftarrow V} & \boldsymbol{O}_{H\leftarrow H} \end{bmatrix} \in \mathbb{R}^{N\times N}$.

- **前向自循环（FS）**

$$f_{t,h} = \sigma\left(c_h + \sum_{v'=1}^{V} a_{h\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{h'=1}^{H} a_{h\leftarrow h'} \cdot \left(\sum_{l=1}^{L} z_{t-l,h'}\ \psi_l\right)\right). \tag{33}$$

变分参数集为 $\phi = \{\boldsymbol{c}_H, \boldsymbol{A}\}$, 其中 $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O}_{V\leftarrow V} & \boldsymbol{O}_{V\leftarrow H} \\ \boldsymbol{A}_{H\leftarrow V} & \boldsymbol{A}_{H\leftarrow H} \end{bmatrix} \in \mathbb{R}^{N\times N}$.

- **前向后向（FB）**

$$f_{t,h} = \sigma\left(c_n + \sum_{v'=1}^{V} a_{h\leftarrow v'} \cdot \left(\sum_{l=1}^{L} x_{t-l,v'}\ \psi_l\right) + \sum_{v'=1}^{V} a_{v'\leftarrow h} \cdot \left(\sum_{l=1}^{L} x_{t+l,v'}\ \psi_l\right)\right) \tag{34}$$

变分参数集为 $\phi = \{\boldsymbol{c}_H, \boldsymbol{A}\}$, 其中 $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O}_{V\leftarrow V} & \boldsymbol{A}_{V\leftarrow H} \\ \boldsymbol{A}_{H\leftarrow V} & \boldsymbol{O}_{H\leftarrow H} \end{bmatrix} \in \mathbb{R}^{N\times N}$.

A.2.3 隐藏放电数的分布

表 2: 在生成模型和变分模型中使用的不同的隐藏放电关于放电率的分布 $\mathbb{P}[z; f]$. 出于简洁的目的，我们忽略了 z、$\tilde{\boldsymbol{z}} = (\tilde{z}_0, \ldots, \tilde{z}_{M-1})$ 和 f 中用于指示隐藏神经元和时间桶的下标.

分布	样本	似然	是否可用路径梯度估计
泊松 (Pois)	$z \sim \text{Poisson}(f)$	$\mathbb{P}[z; f] = \frac{f^z \mathrm{e}^{-z}}{z!}$	✗
分类 (Cat)	$z \sim \text{Cat}(\boldsymbol{\pi}(f))$	$\mathbb{P}[z; f] = \boldsymbol{\pi}(f)_z$	✗
Gumbel-Softmax (GS)	$\tilde{\boldsymbol{z}}_{t,h} \sim \text{GS}(\boldsymbol{\pi}(f_{t,h}); \tau)$	Eq. 36 bellow	✓
指数 (Exp)	$z \sim \text{Exp}\left(\frac{1}{f}\right)$	$\mathbb{P}[z; f] = \frac{1}{f} \exp\left(-fz\right)$	✓
Rayleigh (Ray)	$z \sim \text{Ray}\left(\sqrt{\frac{2}{\pi}} f\right)$	$\mathbb{P}[z; f] = \frac{\pi z}{2f^2} \exp\left(-\frac{\pi z^2}{4f^2}\right)$	✓
Half-normal (HN)	$z \sim \text{HN}\left(\sqrt{\frac{\pi}{2}} f\right)$	$\mathbb{P}[z; f] = \frac{2}{\pi f} \exp\left(-\frac{z^2}{\pi f^2}\right)$	✓

在上表中，我们使用函数 $\boldsymbol{\pi}$ 来将泊松分布截断成一个分类分布，In the above table, we used the function $\boldsymbol{\pi}$ to truncate a Poisson distribution to a categorical distribution,

$$\boldsymbol{\pi}(f) = \left(1 - \sum_{m=1}^{M-1} \frac{f^m \mathrm{e}^f}{m!}, \frac{f^1 \mathrm{e}^f}{1!}, \ldots, \frac{f^{M-1} \mathrm{e}^f}{(M-1)!}\right). \tag{35}$$

GS 的似然函数为

$$\mathbb{P}[\tilde{\boldsymbol{z}}; f] = \Gamma(M) \tau^{M-1} \left(\sum_{m=0}^{m-1} \frac{\boldsymbol{\pi}(f)_m}{\tilde{z}_m^\tau}\right) \prod_{m=0}^{M-1} \frac{\boldsymbol{\pi}(f)_m}{\tilde{z}_m^{\tau+1}} \tag{36}$$

A.3 补充图

A.3.1 合成数据集

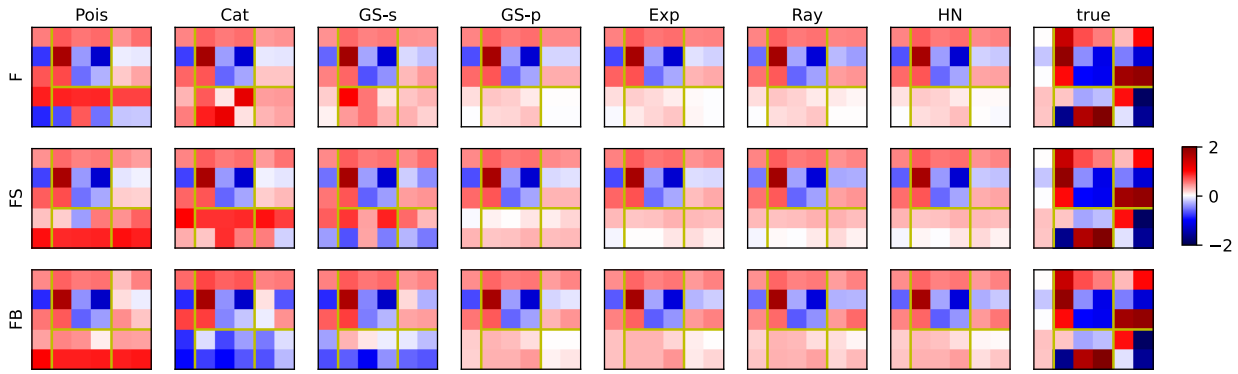

图 8: 所有方法组合在合成数据集的某一个试次上学到的权重矩阵和偏置向量，以及和真实值的比较.

A.3.2 RGC 数据集

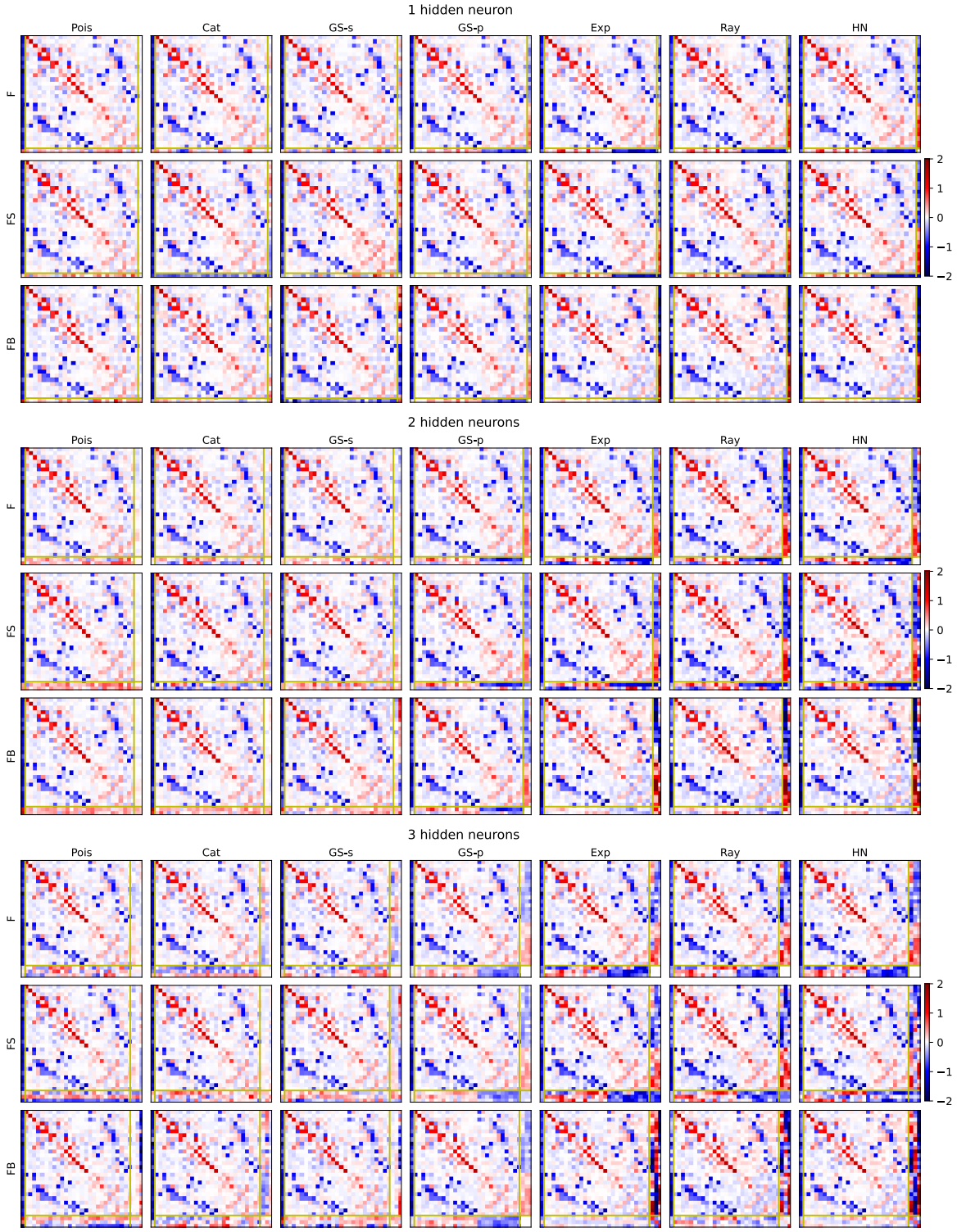

图 9: 假设不同数量的隐藏神经元，所有方法组合在 RGC 数据集上学到的权重矩阵和偏置向量.

A.4 点过程

A.4.1 广义 Hawkes 过程是点过程版本的 GLM

一个广义多元 Hawkes 过程 (GMHP) 是一个的时间点过程（右连续），其由条件强度函数描述

$$\lambda_n^*(t) = \sigma\left(b_n + \sum_{t_n<t} w_{n\leftarrow n_i}\cdot\psi(t-t_i)\right), \tag{37}$$

其中 n 为 N 个神经元的下标. (t_i, n_i) 为历史（放电）序列中第 i 个事件第到达时间和所属神经元. 历史放电序列是按到达时间 t_i 排好序的，即 $t_1 < t_2 < \cdots < t$. $\lambda_n^*(t)$ 是条件依赖于 t 之前的放电序列的 $\{(t_n, d_n)_{t_n<t}\}$. $b_n \in \mathbb{R}$ 是第 n 个神经元的背景强度，$w_{n\leftarrow n'} \in \mathbb{R}$ 是从第 n' 个神经元到第 n 个神经元的链接权重，$\psi(\cdot)$ 是一个基函数，通常积分到 1. σ 是一个非线性函数. 此外，注意到 $\lambda_n^*(t)$ 是一个**左连续**函数. 为了保证因果关系，我们还需要 $\psi(t) = 0,\ \forall t \leqslant 0$. 令 $\theta = \{\boldsymbol{b}, \boldsymbol{W}\} = \left\{[b_1, \ldots, b_N]^{\mathrm{T}}, [w_{n\leftarrow n'}]_{N\times N}\right\}$ 为我们要估计到参数集. 一条连续实验戳序列格式的放电序列 $\mathcal{X} = \{(t_i, n_i)\}_{i=1}^I$ 在一段观察时间 $[0, T]$ 内的数据似然（概率密度函数）为

$$\mathbb{P}(\mathcal{X};\theta) = \prod_{i=1}^{I}\lambda_{n_i}^*(t_i)\cdot\exp\left[-\sum_{n=1}^{N}\int_0^T \lambda_n^*(t)\ \mathrm{d}t\right]. \tag{38}$$

条件强度与到达时间间隔的关系. 对任意时间点过程（非离散），在 t 到 $t+\Delta t$ 之间发生的事件个数服从泊松分布

$$X \sim \mathcal{P}\left(\int_t^{t+\Delta t}\lambda^*(s)\ \mathrm{d}s\right) = \mathcal{P}(\Lambda(t+\tau) - \Lambda(t)), \tag{39}$$

其中 $\Lambda(t) \coloneqq \int_0^t \lambda^*(s)\ \mathrm{d}s$ 为补偿器 (compensator). 如果 $\Delta t \to 0$,

$$\int_t^{t+\Delta t}\lambda^*(s)\ \mathrm{d}s \approx \lambda^*(t)\Delta t. \tag{40}$$

强度函数 $\lambda^*(t)$ 和（从当前时间 t 到）下一到达时间间隔 τ 的分布 (PDF) 的关系为

$$\begin{aligned} f(\tau) =& \underbrace{\lambda^*(t+\tau)}_{\text{an event happens at } t+\tau\text{, density}} \underbrace{\frac{\left(\int_t^{t+\tau}\lambda^*(s)\ \mathrm{d}s\right)^0 \mathrm{e}^{-\int_t^{t+\tau}\lambda^*(s)\ \mathrm{d}s}}{0!}}_{\text{no event happens in the interval } (t,t+\tau)\text{, probability}} \\ =& \lambda^*(t+\tau)\mathrm{e}^{-\int_t^{t+\tau}\lambda^*(s)\ \mathrm{d}s}. \end{aligned} \tag{41}$$

累计分布函数 (CDF) 为 is

$$F(\tau) = 1 - \mathrm{e}^{-\int_t^{t+\tau}\lambda^*(s)\ \mathrm{d}s}. \tag{42}$$

因此，对强度函数建模和对到达时间间隔建模是等价的.

采样/模拟. 我们介绍两种等价的采样方法: Ogata 细化 (Ogata's thinning) 和先到先服务 (first-come-first-serve, FCFS). 下一个事件在时间 $t+\tau$ 时发生在第 n 个神经元上的概率密度为

$$f(\tau, n) = \prod_{n'=1}^{N}\mathrm{e}^{-\int_t^{t+\tau}\lambda_{n'}^*(s)\ \mathrm{d}s}\ \lambda_n^*(t+\tau), \tag{43}$$

这可以视为 FCFS. 而用 Ogata 细化方法,

$$f(\tau,n)=\mathrm{e}^{-\int_t^{t+\tau}\sum_{n'=1}^N\lambda_{n'}^*(s)\ \mathrm{d}s}\ \sum_{n'=1}^N\lambda_{n'}^*(t+\tau)\ \frac{\lambda_n^*(t+\tau)}{\sum_{n'=1}^N\lambda_{n'}^*(t+\tau)}. \tag{44}$$

A.4.2 连续点过程的离散化

一般来讲, 在完整数据似然的积分没有解析解. 通常的解决办法时蒙特卡洛积分, 尽管这不是最优的方案. 另一个更好的选择是数值积分法, 比如 Simpson 法. 然而, 不论使用什么数值积分方法, 都会损失过程本身的连续性. 此外, 存储放电序列时间戳的数据结构也不理想. 其一, 计算点过程的对数似然需要串行搜索一个时间戳列表, 这比时间桶化的放电序列中直接使用矩阵乘法要复杂得多. 其二, 采隐藏放电序列的时间戳需要排序, 这个也十分浪费时间. 此外, 对数似然函数中的积分项一般也没有解析解, 还是需要把时间离散化. 因此, 在一开始就把点过程数据用时间桶离散成放电数是比处理连续时间戳序列更方便的.

这里我们讲解离散化的过程. 令 $\boldsymbol{X}\in\mathbb{N}^{S\times N}$ 为离散放电序列, 其中 $S=\frac{T}{\Delta t}$ 为时间桶的总数. 那么 $x_{s,n}$ 就表示第 n 个神经元在时间区间 $((s-1)\Delta t,s\Delta t)$ 内的放电数. 现在, GMHP 就被离散化成了一个广义线性模型 (GLM): $x_{s,n}\sim\mathcal{P}(\lambda_{s,n}^*\Delta t)$,

$$\lambda_{s,n}^*=b_n+\sum_{x_{s',n'}>0,s'<s}x_{s',n'}w_{n\leftarrow n'}\psi((s-s')\Delta t)=b_n+\sum_{n'=1}^N w_{n\leftarrow n'}\sum_{l=1}^L x_{s-l,n'}\psi_l, \tag{45}$$

其中 $\lambda_{s,n}^*$ 还是由 θ 参数化, $\boldsymbol{\psi}^{\mathrm{T}}=[\psi(\Delta t),\dots,\psi(L\Delta t)]$. 完整数据似然为

$$\mathbb{P}(\boldsymbol{X};\theta)=\prod_{n=1}^N\prod_{s=1}^S\frac{(\lambda_{s,n}^*\Delta t)^{x_{s,n}}\mathrm{e}^{-\lambda_{s,n}^*\Delta t}}{x_{s,n}!}. \tag{46}$$

现在我们要证明这个似然随时间桶宽度 $\Delta t\to 0$ 收敛到连续的形式. 当取极限 $\Delta t\to 0$ 时, $x_{s,n}$ 要么是 0 要么是 1. 因此,

$$\begin{aligned}\lim_{\Delta t\to 0}\mathbb{P}(\boldsymbol{X};\theta)=&\prod_{x_{s,n}=1}\lambda_{s,n}^*\Delta t\ \mathrm{e}^{-\lambda_{s,n}^*\Delta t}\prod_{x_{s,n}=0}\mathrm{e}^{-\lambda_{s,n}^*\Delta t}\\ =&(\Delta t)^I\prod_{i=1}^I\lambda_{n_i}^*(t_i;\theta)\ \exp\left[-\sum_{n=1}^N\int_0^T\lambda_n^*(t;\theta)\ \mathrm{d}t\right].\end{aligned} \tag{47}$$

由于 $(\Delta t)^I$ 为一个常数, 我们可以把上式除以 $(\Delta t)^I$ 然后只取概率密度, 这样就和连续情况下的公式 38 一样了. 因此, 只要 Δt 足够小, 求解离散问题就和求解原始连续问题等价. 问题的解也会随 $\Delta t\to 0$ 收敛到连续问题的解. 事实上, 公式 47 所示的过程适用于任何点过程及其离散版本.