基于前向 χ^2 散度的变分重要性采样

李宬睿，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 2 月 18 日

摘要

最大化边缘对数似然是学习隐变量模型中的关键,而变分推断 (variational inference, VI) 是最常用的方法. 然而, VI 在处理特别复杂的后验分布时可能无法达到很高的边缘对数似然. 为了应对这一局限性, 我们提出了一个新的变分重要性采样 (variational importance sampling, VIS) 方法, 直接估计并最大化边缘对数似然. VIS 利用由最小化前向 χ^2 散度得到的最优的建议分布来增强边缘对数似然的估计. 我们在混合模型、变分自编码器和部分可见广义线性模型这几个常见的隐变量模型上应用了 VIS. 结果表明, VIS 能够在边缘对数似然以及模型参数估计两方面均超过最先进的基线方法.代码:https://github.com/JerrySoybean/vis.

1 引言

给定隐变量 $\boldsymbol{z}$ 和观测变量 $\boldsymbol{x}$，如何寻找最大化边缘似然 $p(\boldsymbol{x};\theta)=\int p(\boldsymbol{x},\boldsymbol{z};\theta)\,\mathrm{d}\boldsymbol{z}$ 的最优参数 θ 在诸多应用问题中都十分重要．然而，当问题特别复杂时，我们只知道 $p(\boldsymbol{x},\boldsymbol{z};\theta)$ 的显式形式，而无法解析地计算边缘 $p(\boldsymbol{x};\theta)$．因此，我们通常就会使用变分推断 (variational inference, VI) [Blei et al., 2017] 或重要性采样 (importance sampling, IS) [Kloek and Van Dijk, 1978] 这种近似方法来学模型参数 θ 以及推断无法解析求解的后验 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$.

VI 用一个变分分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 来近似后验 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$.二者之间的差距用反向 KL 散度 $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta))$ 来衡量. 而最小化 KL 散度等价于最大化 $\ln p(\boldsymbol{x};\theta)$ 的证据下界 $\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$. 然后，用 ELBO 最大化 $\ln p(\boldsymbol{x};\theta)$ 不一定是最好的办法,尤其是在处理重尾、多峰等复杂后验分布时.有可能 $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta))$ 非常小，但实际上 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 和 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 都离真实的后验 $p(\boldsymbol{z}|\boldsymbol{x};\theta^{\mathrm{true}})$ 非常远，导致 ELBO 很高但是边缘对数似然却很低（如 4.1 节所示）.

尽管像基于 α 散度的下界 [Li and Turner, 2016, Hernandez-Lobato et al., 2016] 和像基于 χ^2 散度的上界等等可以用于更好的后验估计，但更直接的做法就是用 IS 直接估计 $\ln p(\boldsymbol{x};\theta)$．理想情况下，如果建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 选得好、蒙特卡洛样本足够多，IS 就可以得到一个很好的估计．但实际应用中，我们经常不清楚到底如何选择一个好的建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$，也没有明确的指标来验证 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 的质量.

Su and Chen [2021] 表明 VI 学到的变分分布可以作为 IS 的建议分布，但是已经有大量文献表明这不是最优选择 [Jerfel et al., 2021, Saraswat, 2014, Sason and Verdú, 2016, Nishiyama and Sason, 2020]. 此外，Pradier et al. [2019] 注意到了最小化前向 χ^2 散度时的数值问题和尺度问题 Finke and Thiery [2019]，需要我们细致地处理.

为了解决这些问题，我们提出了一个新的学习方法，称为变分重要性采样 (variational importance sampling, VIS). 我们会阐明用于 IS 的最优建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 可以通过在对数空间中最小化前向 χ^2 散度得到，并且这样是数值稳定的. 此外，只要有足够多的蒙特卡洛样本，估计的边缘对数似然 $\ln\hat{p}(\boldsymbol{x};\theta)$ 相较于 ELBO 是一个渐近意义上更紧的下界，从而 $\ln p(\boldsymbol{x};\theta)$ 可以被更高效地估计. 在实验部分，我们将 VIS 应用于多个模型，其中包括了没有 $p(\boldsymbol{x},\boldsymbol{z};\theta)=p(\boldsymbol{x}|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$ 这种显式分解的最一般的情形. 在合成数据集和真实数据集中，VIS 都比最常用的 VI 和其他三个最先进的基线方法 CHIVI [Dieng et al., 2017]、VBIS [Su and Chen, 2021] 以及 IWAE [Burda et al., 2015] 表现更优. 附录 A.8 用表格的形式总结了与本篇文章相关的一些工作以及相对应的我们的贡献.

2 变分推断的背景知识

我们在这里先简要介绍一下变分推断 (variational inference, VI)，以及它的估计量和偏差. VI 由反向 KL 散度导出:

$$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta))=\int q(\boldsymbol{z}|\boldsymbol{x};\phi)\ln\frac{q(\boldsymbol{z}|\boldsymbol{x};\phi)}{p(\boldsymbol{z}|\boldsymbol{x};\theta)}\,\mathrm{d}\boldsymbol{z}=-\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)+\ln p(\boldsymbol{x};\theta), \tag{1}$$

其中 $\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)\coloneqq\mathbb{E}_q[\ln p(\boldsymbol{x},\boldsymbol{z};\theta)-\ln q(\boldsymbol{z}|\boldsymbol{x};\phi)]$. 由于 ELBO 是 $\ln p(\boldsymbol{x};\theta)$ 的一个下界，最大化 $\ln p(\boldsymbol{x};\theta)$ 的问题就转换成了最大化 $\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$. VI 受欢迎的原因主要有二：1) ELBO 的形式是对数似然的数学期望，使得其数值上相较于直接处理原始的似然更稳定；2) 当模型可以进行 $p(\boldsymbol{x},\boldsymbol{z};\theta)=p(\boldsymbol{x}|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$ 分解时，ELBO 可以写成 $\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)=\mathbb{E}_q[\ln p(\boldsymbol{x}|\boldsymbol{z};\theta)]-\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z};\theta))$. 这种写法的好处在于，第二项 KL 项通常对特定的先验分布 $p(\boldsymbol{z};\theta)$ 和变分分布族 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 有解析解，比如高斯分布.

实际中，公式 1 中的目标函数 ELBO 仍然需要数值估计，也就是它的估计量

$$\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)=\frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)-\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right], \tag{2}$$

其中 $\left\{\boldsymbol{z}^{(k)}\right\}_{k=1}^{K}$ 是 K 个来自变分分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 的蒙特卡洛样本. 现在，我们就把关于 θ 最大化 $\ln p(\boldsymbol{x};\theta)$ 的问题转化成了关于 θ 和 ϕ 最大化 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ 了. 附录 A.1 介绍了 ELBO 的得分函数梯度估计 (score function gradient estimator) 和路径梯度估计 (pathwise gradient estimator).

ELBO 估计量的偏差. 注意到尽管 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ 是 ELBO 的一个无偏估计，它却是边缘对数似然 $\ln p(\boldsymbol{x};\theta)$ 的一个严格下偏估计（图 1(a)）:

$$\mathbb{E}_q\left[\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)-\ln p(\boldsymbol{x};\theta)\right]=\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)-\ln p(\boldsymbol{x};\theta)=-\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta)). \tag{3}$$

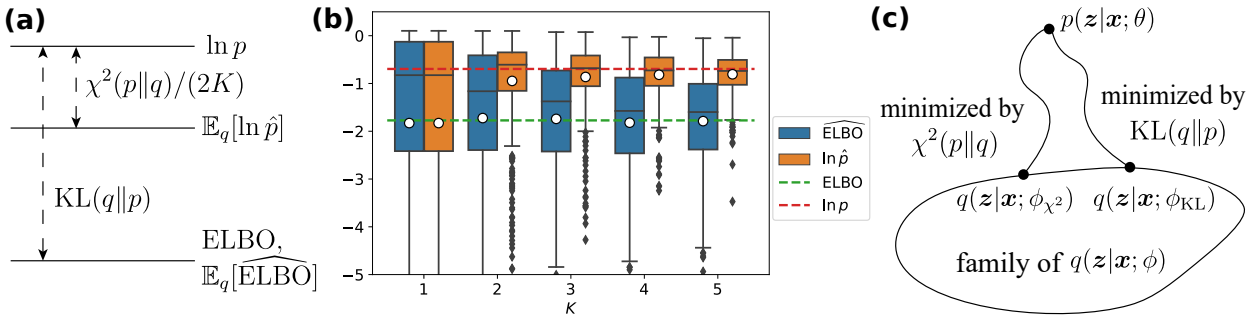

图 1: **(a)**：边缘对数似然 $\ln p(\boldsymbol{x};\theta)$ 和它的 IS 估计量的数学期望 $\mathbb{E}_q[\ln \hat{p}(\boldsymbol{x};\theta,\phi)]$，$\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$，以及 ELBO 的估计量的数学期望 $\mathbb{E}_q[\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)]$ 之间的偏差．在估计 $\ln p(\boldsymbol{x};\theta)$ 时，下偏的 IS 估计量 $\mathbb{E}_q[\ln \hat{p}(\boldsymbol{x};\theta,\phi)]$ 相较下偏的 ELBO 估计量 $\mathbb{E}_q[\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)]$ 是一个更紧的下界．**(b)**：(a) 中四个量在不同蒙特卡洛样本量 $K \in \{1,2,3,4,5\}$ 下的实际可视化．(b) 中每个箱的都是由 500 次重复得到的．其中的空心圆是 500 次的平均值．随着 K 的增加，它们差距渐近的显现出来了．**(c)**：最小化前向 χ^2 散度（做 IS 的最优方式）和最小化反向 KL 散度会得到不同的 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$．

正如之前提到的，有可能得到的 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 和 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 都离真正的后验 $p(\boldsymbol{z}|\boldsymbol{x};\theta^{\mathrm{true}})$ 很远，导致有较高的 ELBO 但却只有很低的边缘对数似然 $\ln p(\boldsymbol{x};\theta)$．

3 变分重要性采样

为了解决这一问题，我们采用重要性采样 (importance sampling, IS) 的方法来直接估计边缘对数似然 $\ln p(\boldsymbol{x};\theta)$．然而，IS 的估计质量取决于建议分布和蒙特卡洛样本数．我们首先说明，用 IS 可以得到一个比 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ 渐近更紧的 $\ln p(\boldsymbol{x};\theta)$ 的估计．之后，我们会证明这一估计量的偏差和有效性（方差）都和前向 χ^2 散度以及蒙特卡洛样本数有关．这位我们选建议分布和决定蒙特卡洛样本数提供了一个指导．最后，我们会推到其数值稳定的梯度估计量，用于求得最优的建议分布．

边缘对数似然下偏的 IS 估计量． 用重要性采样 (importance sampling, IS)，边缘可以通过建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 进行估计，即，

$$p(\boldsymbol{x};\theta) = \int p(\boldsymbol{x},\boldsymbol{z};\theta)\,\mathrm{d}\boldsymbol{z} \approx \frac{1}{K}\sum_{k=1}^{K}\frac{p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)}{q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)} =: \hat{p}(\boldsymbol{x};\theta,\phi), \tag{4}$$

其中 $\left\{\boldsymbol{z}^{(k)}\right\}_{k=1}^{K}$ 是 K 个来自建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 的蒙特卡洛样本．为了数值稳定，我们需要在对数空间中计算，

$$\ln \hat{p}(\boldsymbol{x};\theta,\phi) = \mathrm{logsumexp}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right] - \ln K, \tag{5}$$

其中用到了 logsumexp 技巧．附录 A.2 给出了 $\ln p(\boldsymbol{x};\theta)$ 关于 θ 的梯度的估计

$$\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\theta} \approx \frac{\partial \ln \hat{p}(\boldsymbol{x};\theta,\phi)}{\partial\theta}. \tag{6}$$

由于

$$\mathbb{E}_q[\hat{p}(\boldsymbol{x};\theta,\phi)] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_q\left[\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\right] = \int p(\boldsymbol{x},\boldsymbol{z};\theta)\ \mathrm{d}\boldsymbol{z} = p(\boldsymbol{x};\theta), \tag{7}$$

$\hat{p}(\boldsymbol{x};\theta,\phi)$ 是 $p(\boldsymbol{x};\theta)$ 的一个无偏估计.然而,$\ln(\cdot)$ 是一个凹函数,所以由詹森不等式可以得出,$\mathbb{E}_q[\ln\hat{p}(\boldsymbol{x};\theta,\phi)] \leqslant \ln\mathbb{E}_q[\hat{p}(\boldsymbol{x};\theta,\phi)] = \ln p(\boldsymbol{x};\theta)$. 这意味着在对数空间中，估计量 $\ln\hat{p}(\boldsymbol{x};\theta)$ 是 $\ln p(\boldsymbol{x};\theta)$ 的一个下偏估计.

IS 估计量的偏差. 类似于 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$，我么也可以利用 Delta 方法 [Oehlert, 1992, Struski et al., 2022] 推导出 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 的偏差,

$$\begin{aligned}
&\mathbb{E}_q[\ln\hat{p}(\boldsymbol{x};\theta,\phi) - \ln p(\boldsymbol{x};\theta)] = \mathbb{E}_q\left[\ln\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\theta\right)}{q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)}\right)\right] \\
\approx& -\frac{1}{2K}\mathrm{Var}_q\left[\frac{p(\boldsymbol{z}|\boldsymbol{x};\theta)}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\right] = -\frac{1}{2K}\left\{\mathbb{E}_q\left[\left(\frac{p(\boldsymbol{z}|\boldsymbol{x};\theta)}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\right)^2\right] - \mathbb{E}_q^2\left[\frac{p(\boldsymbol{z}|\boldsymbol{x};\theta)}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\right]\right\} \\
=& -\frac{1}{2K}\left(\int\frac{p(\boldsymbol{z}|\boldsymbol{x};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\ \mathrm{d}\boldsymbol{z} - 1\right) = -\frac{1}{2K}\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi)),
\end{aligned} \tag{8}$$

其中 $\chi^2(p\|q)$ 是 p 和 q 之间的前向 χ^2 散度（图 1(a)）. 由于公式 8 随 $K\to\infty$ 收敛到 0，$\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 相较于 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ 是一个渐近意义上更紧的下界（图 1(a)）. 特别的，当 $K=1$ 时，$\ln\hat{p}(\boldsymbol{x};\theta,\phi) = \widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$.为了实证验证这一关系,我们基于 K 个蒙特卡洛样本重复估计 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 和 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ 500 次, 并画出了它们关于 K 的经验分布, 见图 1(b). 当有更多的蒙特卡洛样本数 K 时，$\ln\hat{p}$ 和 $\widehat{\mathrm{ELBO}}$ 都变得更稳定，但是每个箱中的空心圆表示的经验期望表明，只有 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 收敛到了边缘对数似然 $\ln p(\boldsymbol{x};\theta)$.

图 1 阐释了 IS 在较大的 K 下可以得到对 $\ln p(\boldsymbol{x};\theta)$ 更好的估计. 这就意味着用 IS 来最大化 $\ln p(\boldsymbol{x};\theta)$ 比用 ELBO 更直接. 此外, 为了更快收敛, 我们还需要选择能够最小化 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$ 的建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$，因为这一前向 χ^2 散度可以用作指示一个建议分布是否优良的指标：如果前向 χ^2 散度小，那么 IS 估计量的偏差（公式 8 的绝对值）就小.

另一方面，我们也可以写出估计量 $\hat{p}(\boldsymbol{x};\theta,\phi)$ 的有效性 [Freedman et al., 1998]，即,

$$\mathrm{Var}_q\left[\hat{p}(\boldsymbol{x};\theta,\phi)\right] = \frac{1}{K^2}K\,\mathrm{Var}_q\left[\frac{p(\boldsymbol{z}|\boldsymbol{x};\theta)p(\boldsymbol{x};\theta)}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\right] = \frac{p(\boldsymbol{x};\theta)^2}{K}\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi)), \tag{9}$$

也就是估计量的方差. 公式 8 和公式 9 巧合地共同表明想要 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 的偏差较小以及想要 $\hat{p}(\boldsymbol{x};\theta,\phi)$ 的有效性更高,都需要一个较小的 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$ 和一个较大的 K.换句话说,我们需要尽可能多的蒙特卡洛样本; 同时做 IS 的最优建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 来自最小的前向 χ^2 散度 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$ 而不是反向 KL 散度 $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta))$（图 1(c)）.

变分重要性采样 (variational importance sampling, VIS) 的算法总结在了算法 1 中. 我们首先固定当前的建议分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$，用 IS 关于 θ 最大化 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$；然后固定 θ，关于 ϕ 最小化 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$ 来得到一个更好的建议分布. 然而，关于 ϕ 最小化 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$ 没那么简单，因为我们不知道 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$. 下面我们推到一个用于最小化前向 χ^2 散度的稳定梯度估计. the following.

Algorithm 1 VIS

1: **for** i = 1:N **do**
2: 从 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 中采样 $\left\{\boldsymbol{z}^{(k)}\right\}_{k=1}^{K}$.
3: 通过公式 6 最大化 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 更新 θ.
4: 通过公式 12 或公式 24 最小化 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\phi)\|q(\boldsymbol{z}|\boldsymbol{x};\theta))$ 更新 ϕ.
5: **end for**

梯度估计. 把前向 χ^2 散度重写成

$$\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi)) = \frac{1}{p(\boldsymbol{x};\theta)^2}\int\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\,\mathrm{d}\boldsymbol{z} - 1 =: \frac{1}{p(\boldsymbol{x};\theta)^2}V(\boldsymbol{x};\theta,\phi) - 1. \tag{10}$$

所以最小化 $\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$ 等价于关于 ϕ 最小 $V(\boldsymbol{x};\theta,\phi) \coloneqq \int\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\,\mathrm{d}\boldsymbol{z}$. 这个量仍需在对数空间中最小化，以保证数值稳定性 [Pradier et al., 2019, Finke and Thiery, 2019, Geffner and Domke, 2020, Yao et al., 2018]. . 在附录 A.3 中，我们推导了 $\ln V(\boldsymbol{x};\theta,\phi)$ 可以估计如下

$$\ln V(\boldsymbol{x};\theta,\phi) \approx \operatorname{logsumexp}\left[2\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - 2\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right] - \ln K =: \ln\hat{V}(\boldsymbol{x};\theta,\phi). \tag{11}$$

$\ln V(\boldsymbol{x};\theta,\phi)$ 关于 ϕ 在 ϕ_0 处的得分函数梯度估计为

$$\frac{\partial\ln V(\boldsymbol{x};\theta,\phi)}{\partial\phi} \approx \frac{\partial}{\partial\phi}\frac{1}{2}\ln\hat{V}(\boldsymbol{x};\theta,\phi). \tag{12}$$

当重参数技巧 (reparameterization trick) 可以使用时，$\boldsymbol{z}|\boldsymbol{x};\phi = g(\boldsymbol{\epsilon}|\boldsymbol{x};\phi)$，其中 $\boldsymbol{\epsilon}\sim\boldsymbol{r}(\boldsymbol{\epsilon})$，那么我们就有变换 $q(\boldsymbol{z}|\boldsymbol{x};\phi)\,\mathrm{d}\boldsymbol{z} = r(\boldsymbol{\epsilon})\,\mathrm{d}\boldsymbol{\epsilon}$ [Schulman et al., 2015]. 现在就可以得到路径梯度估计 $\frac{\partial\ln V(\boldsymbol{x};\theta,\phi)}{\partial\phi} \approx \frac{\partial}{\partial\phi}\ln\hat{V}(\boldsymbol{x};\theta,\phi)$，其中我们如下采样 $\boldsymbol{\epsilon}\sim r(\boldsymbol{\epsilon})$ 并用 $\boldsymbol{z}^{(k)} = g\left(\boldsymbol{\epsilon}^{(k)}\middle|\boldsymbol{x};\phi\right)$ in $\ln\hat{V}(\boldsymbol{x};\theta,\phi)$. 完整的推导在附录 A.3 中.

4 实验

用于对比实验的基线方法. 我们会将 VIS 应用到三个不同的模型上，并将其与另外四个方法进行比较. 它们分别是:

- **VI**: 最广泛使用的变分推断方法，最大化 ELBO.
- **CHIVI** [Dieng et al., 2017]: 在更新 ϕ 时，同时用一个上界 CUBO（基于前向 χ^2 散度）和一个下界 ELBO（基于反向 KL 散度）来挤压后验估计 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 使其近似后验 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$.
- **VBIS** [Su and Chen, 2021]: 用 VI 学到的 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 作为 IS 的建议分布.
- **IWAE** [Burda et al., 2015]: 重要性加权自编码器 (importance-weighted autoencoder). 它用 IS 而非 VI 来学习自编码器. 这个仅作为 VAE 模型的额外比较方法.

指标. 对所有的模型和数据集，我们都用不同的方法在 $\boldsymbol{x}_{\text{train}}$ 上训练模型，然后在 $\boldsymbol{x}_{\text{test}}$ 上用**边缘对数似然 (marginal log-likelihood, LL)** $p(\boldsymbol{x}_{\text{test}};\theta)$（适用于合成数据集和真实数据集）; **完备对数似然 (complete log-likelihood, CLL)** $p(\boldsymbol{x}_{\text{test}},\boldsymbol{z}_{\text{test}};\theta)$（仅适用于合成数据集，因为生成数据时有 $\boldsymbol{z}_{\text{test}}$）; 以及**隐对数似然 (hidden log-likelihood, HLL)** $q(\boldsymbol{z}_{\text{test}}|\boldsymbol{x}_{\text{test}};\phi)$（仅适用于合成数据集，原因同上）.

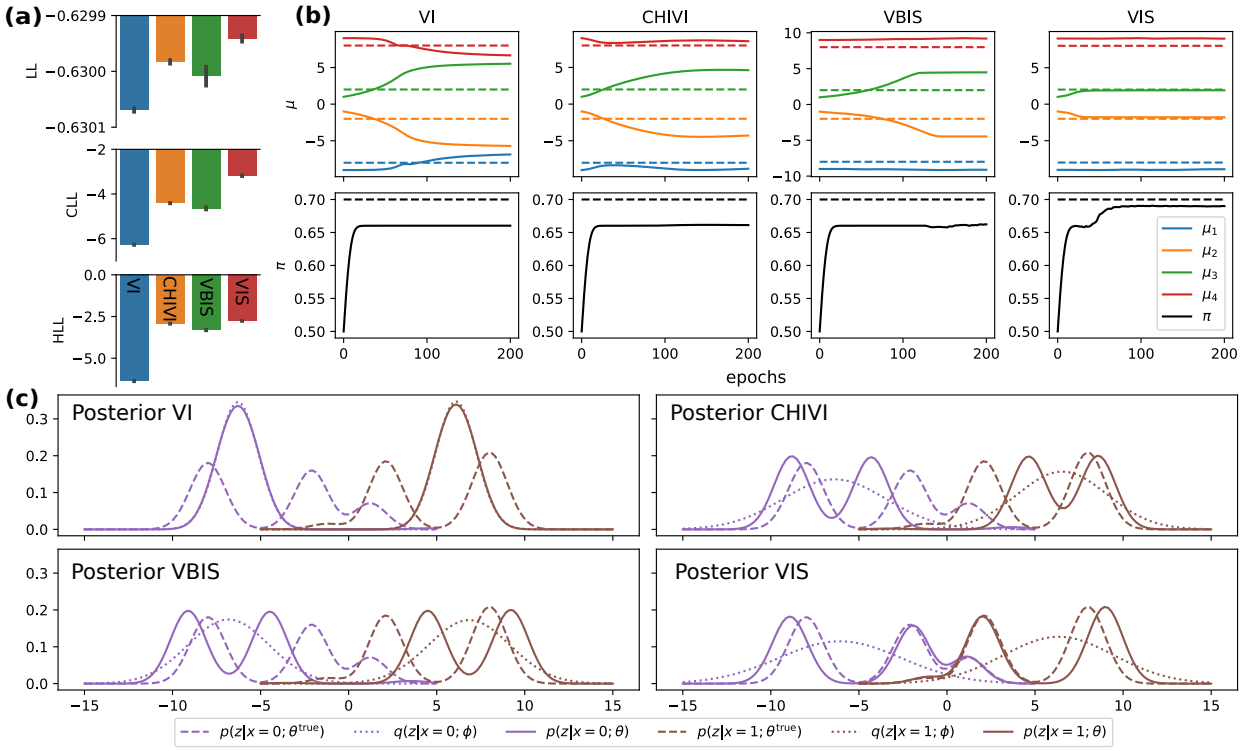

图 2: **(a)**：测试集上的 LL、CLL 以及 HLL. **(b)**：不同方法下参数集 θ 的收敛曲线. 短划线是用于生成数据的真实参数，实线是学到的参数. **(c)**：不同方法下，在 $x=0$ 和 $x=1$ 处的后验分布. 短划线是真实后验 $p(\boldsymbol{z}|x;\theta^{\text{true}})$，实线是学到的后验 $p(\boldsymbol{z}|x;\theta)$，点虚线是学到的变分/建议分布给出的近似后验 $q(z|x;\phi)$.

4.1 一个简单的混合模型

模型. 我们首先用一个简单的混合模型来阐释不同方法的一些代表性行为. 考虑生成模型 $p(z;\theta)=\sum_{i=1}^{4}\mathcal{N}(z;\mu_i,1^2)$ 其中 $\pi_1=\pi_2=\frac{1-\pi}{2}$, $\pi_3=\pi_4=\frac{\pi}{2}$; $p(x|z;\theta)=\text{Bern}(x;\text{sigmoid}(z))$. 参数集为 $\theta=\{\pi\}\cup\{\mu_i\}_{i=1}^4$, 隐变量是 $z\in\mathbb{R}$, 观测变量是 $x\in\{0,1\}$. 把变分/建议分布族选为 $q(z|x;\phi)=\mathcal{N}(z;c_x,\sigma_x^2)$, $x\in\{0,1\}$, 变分/建议参数集为 $\phi=\{c_0,c_1,\sigma_0,\sigma_1\}$. 这样一个简单的模型使我们可以画出 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 来, 帮我们理解不同方法的行为区别.

实验设置. 训练集和测试集分别包括 1000 个来自 $p(x,z;\theta^{\text{true}})$ 的样本. 我们用 Adam [Kingma and Ba, 2014] 作为优化器, 学习率设为 0.002. 每个方法我们训练 200 轮, 用批训练的方式, 每批 100 个样本点, 共 10 批. 每次采 $K=5000$ 个蒙特卡洛样本. 对每个方法, 我们以不同的随机数种子重复 10 次并报告结果.

结果. 定量的角度, VIS 在全部三个指标上表现得比其它模型一致得好（图 2(a). 图 2(b) 中, 我们画出了不同方法学习下参数集 θ 的收敛曲线. 很明显, VIS 实现了更精准的参数估计. 这进一步佐证了

好的参数估计和高的测试集边缘对数似然之间的对应关系.

为了理解不同方法估计出的近似后验 $q(z|x;\phi)$ 的效果，我们画出了真实后验 $p(z|x;\theta^{\text{true}})$（短划线）学到的后验 $p(z|x;\theta)$（实线）以及近似后验 $q(z|x;\phi)$（点虚线）分别在 $x=0$ 和 $x=1$ 上的形状在图 2(c) 中. 首先我们可以确认，真实后验 $p(z|x;\theta^{\text{true}})$ 在 $x=0$ 和 $x=1$ 的条件上都是多模态的，至少有两个峰. 比如，$p(z|x=0;\theta^{\text{true}})$ 在大约 $z=-8$ 处有一个大峰，在大约 $z=-2$ 有一个大峰，在大约 $z=1$ 处有一个小峰（见图 2(c) 中的紫色短划线）. 下面我们检查学到的后验 $p(z|x=0;\theta)$ 和近似后验 $q(z|x=0;\phi)$.

- 对于 VI，VI 中最小化反向 KL 散度带来的迫零/寻众 (zero-forcing/mode-seeking) 行为导致左边的两个大峰不得不收为一体. 并且 $q(z|x=0;\phi)$ 的有效支撑区间仅覆盖到了 $p(z|x=0;\theta^{\text{true}})$ 的左边的大峰，这导致 $p(z|x=0;\theta)$ 的形状和 $p(z|x=0;\theta^{\text{true}})$ 完全不同. 这就是之前所说的，反向 KL 散度 $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta))$ 非常小，但实际上 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 和 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 都离真实后验 $p(\boldsymbol{z}|\boldsymbol{x};\theta^{\text{true}})$ 很远的情况，导致了较高的 ELBO 但是只有较低的边缘对数似然.
- 对于 VBIS，通过重要性采样，学到的后验 $p(z|x=0;\theta)$ 保持了两个大峰，但在由于最小化反向 KL 散度带来的迫零行为，大约 $z=1$ 处的小峰仍然没有被 $q(z|x=0;\phi)$ 有效的覆盖到. 此外，由于最小化反向 KL 散度得到的 $q(z|x;\phi)$ 不是做 IS 最优的建议分布，学到的 $p(z|x;\theta)$ 没能够很好的对上 $p(z|x;\theta^{\text{true}})$.
- 对于 CHIVI，反向 KL 和前向 χ^2 散度都被考虑到了，所以 $q(z|x=0;\phi)$ 的有效支撑区间变得足够得宽，确保了两个大峰下的概率密度都能被采样得到. 然而相较于 VIS，它还是没宽到能够覆盖在大约 $z=1$ 处的小峰. 此外，由于 CHIVI 关于 θ 优化 ELBO 而不是边缘对数似然，学到的 θ 不如 VIS.
- 对于 VIS，最小化前向 χ^2 散度带来的大规模覆盖/寻均 (mass-covering/mean-seeking) 行为使得 $q(z|x=0;\phi)$ 足够宽到可以覆盖两个大峰和那个在大约 $z=1$ 处的小峰. 然而，并且由于我们在公式 8 和公式 9 说明了最小化前向 χ^2 散度使得学到的 $q(z|x;\phi)$ 是做 IS 最优的建议分布，所以学到的后验 $p(z|x;\theta)$ 的形状和真实后验 $p(z|x;\theta^{\text{true}})$ 的形状相较于其它方法匹配得最好.

4.2 变分自编码器 (Variational auto-encoder)

模型. 变分自编码器 (variational auto-encoder, VAE) [Kingma and Welling, 2013] 的生成模型可以表示为 $p(\boldsymbol{z};\theta)=\mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})$；$p(\boldsymbol{x}|\boldsymbol{z};\theta)=\mathrm{Bern}(\boldsymbol{x};\mathrm{sigmoid}(\mathrm{MLP}_{\text{dec}}(\boldsymbol{z})))$. 参数集 θ 包含了 MLP 解码器的所有参数. 变分/建议分布参数化为 $q(\boldsymbol{z}|\boldsymbol{x};\phi)=\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}(\boldsymbol{x}),\boldsymbol{\sigma}^2(\boldsymbol{x})\boldsymbol{I})$，其中 $\boldsymbol{\mu}(\boldsymbol{x})$ 和 $\boldsymbol{\sigma}(\boldsymbol{x})$ 是 MLP 解码器在给定输入 $\boldsymbol{x}$ 下的输出. 参数集 ϕ 包括了 MLP 编码器的所有参数.

实验设置. 我们将 VAE 模型应用在了 MMIST 数据集 [LeCun et al., 1998] 上. 训练集有 60000 个样本，测试集有 10000 个样本. 每个样本是一个 28×28 的灰度手写体数字图像，所以 $\boldsymbol{x}\in[0,1]^{784}$. 为了可视化，我们令 $\boldsymbol{z}\in\mathbb{R}^2$. 类似于 [Kingma and Welling, 2013]，我们将编码器和解码器的结构设为

$$\begin{aligned}
&\mathrm{MLP}_{\text{dec}}(\boldsymbol{z})=\boldsymbol{W}_{\text{dec},2}\boldsymbol{h}_{\text{dec}}+\boldsymbol{b}_{\text{dec},2},\quad \boldsymbol{h}_{\text{dec}}=\tanh\left(\boldsymbol{W}_{\text{dec},1}\boldsymbol{z}+\boldsymbol{b}_{\text{dec},1}\right),\quad \boldsymbol{h}_{\text{dec}}\in\mathbb{R}^{128},\\
&\begin{cases}\boldsymbol{\mu}(\boldsymbol{x})=\boldsymbol{W}_{\boldsymbol{\mu}}\boldsymbol{h}_{\text{enc}}+\boldsymbol{b}_{\boldsymbol{\mu}}\\ \ln\boldsymbol{\sigma}(\boldsymbol{x})=\boldsymbol{W}_{\boldsymbol{\sigma}}\boldsymbol{h}_{\text{enc}}+\boldsymbol{b}_{\boldsymbol{\sigma}}\end{cases},\quad \boldsymbol{h}_{\text{enc}}=\tanh\left(\boldsymbol{W}_{\text{enc}}\boldsymbol{x}+\boldsymbol{b}_{\text{enc}}\right),\qquad \boldsymbol{h}_{\text{enc}}\in\mathbb{R}^{128}.
\end{aligned}\tag{13}$$

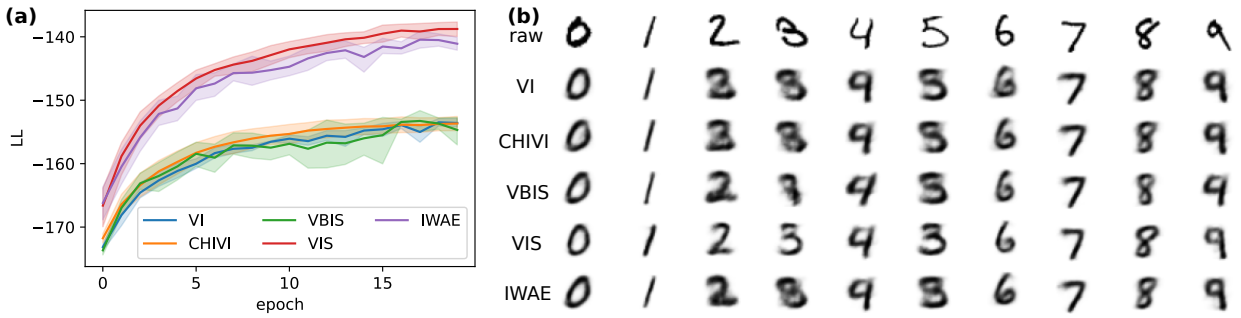

图 3: **(a)**：测试集上的边缘对数似然关于训练轮数的曲线．**(b)**：一个原始图像和不同方法重构的图像.

我们用 Adam [Kingma and Ba, 2014] 作为优化器，学习率设为 0.005．每个方法我们跑 20 轮．批大小为 64．采隐变量的蒙特卡洛样本数设置为 $K = 500$．每个方法我们用不同的随机数种子重复 5 次，并报告测试集上的对数似然.

结果. 图 3(a) 画出了学习过程中边缘对数似然的变化．VI 作为经典的求解方法表现地和 CHIVI 以及 VBIS 差不多，不过 VI 的收敛曲线更稳定．然而如果和 IWAE 以及 VIS，作比较 IWAE 更好，VIS 最好．图 3(b) 所示的重构图像也表明，VIS 求解得到的 VAE 能够重构出和原始图像最相似的重构图像．不同方法学到的隐变量流形贴在了附录 A.4 中.

4.3 部分可见广义线性模型

模型. 我们首先给出经典的广义线性模型 (generalized linear model, GLM) [Pillow et al., 2008]，其用于研究神经放电序列中神经元之间的相互影响．我们把来自 N 个神经元、持续 T 个时间桶的神经放电序列了数据记为 $\boldsymbol{Y} \in \mathbb{N}^{T\times N}$．$y_{t,n}$ 是第 n 个神经元在第 t 个时间桶内的放电次数．当给出 $\boldsymbol{Y}$ 时，经典的 GLM 用下面的公式预测第 n 个神经元在第 t 个时间桶内的放电率 $f_{t,n}$，

$$f_{t,n} = \sigma\left(b_n + \sum_{n'=1}^{N} w_{n\leftarrow n'} \cdot \left(\sum_{l=1}^{L} y_{t-l,n'}\psi_l\right)\right), \quad \text{with spike } y_{t,n} \sim \text{Poisson}(f_{t,n}), \tag{14}$$

其中 $\sigma(\cdot)$ 是一个非线性函数（比如 Softplus）；b_n 是第 n 个神经元的基准放电强度（偏置），其向量形式为 $\boldsymbol{b} \in \mathbb{R}^N$；$w_{n\leftarrow n'}$ 是第 n' 个神经元给第 n 个神经元的影响，其矩阵形式为 $\boldsymbol{W} \in \mathbb{R}^{N\times N}$；$\boldsymbol{\psi} \in \mathbb{R}_+^L$ 是提前给定的基函数，用于合计神经元从 $t-L$ 到 $t-1$ 的放电历史.

经典的 GLM 不是给隐变量模型．然而我们可以把 GLM 扩展成一个部分可见的 GLM (partially observable GLM, POGLM) [Pillow and Latham, 2007]，这就成了一个隐变量模型．特别的，POGLM 在神经元放电序列数据是部分可见的情况下研究神经元之间的相互影响．这恰恰是神经科学中经常遇到的情况，因为在大脑目标区域收集所有神经元的数据通常是不现实的．考虑 N 个神经元中 V 个是可见的神经元，剩余 H 个是隐藏的神经元（$N = V + H$）．给定一个放电序列 $\boldsymbol{Y}$，它左边 V 列为 $\boldsymbol{X} = \boldsymbol{Y}_{1:T,1:V} \in \mathbb{N}^{T\times V}$ 包含了可见神经元的放电序列，右边 H 列为 $\boldsymbol{Z} = \boldsymbol{Y}_{1:T,V+1:N} \in \mathbb{N}^{T\times H}$ 包含了隐

藏神经元的放电数据．对可见神经元和隐藏神经元，放电率都是

$$f_{t,n} = \sigma\left(b_n + \sum_{n'=1}^{V} w_{n\leftarrow n'} \cdot \left(\sum_{l=1}^{L} x_{t-l,n'}\psi_l\right) + \sum_{n'=1+V}^{N} w_{n\leftarrow n'} \cdot \left(\sum_{l=1}^{L} z_{t-l,n'-V}\psi_l\right)\right). \tag{15}$$

由于隐藏神经元是观察不到的，POGLM 就成了一个隐变量模型，观测变量为 $x_{t,n}$，隐变量为 $z_{t,n}$．模型参数 θ 就是 $\{\boldsymbol{b}, \boldsymbol{W}\}$．POGLM 的图模型画在了图 4(a) 上一.

为了在 POGLM 上做 VI、VIS 或其它方法，一个常用的变分/建议分布 [Rezende and Gerstner, 2014, Kajino, 2021] 是 $q(z_{t,n}|x_{1:t-1,1:V}, z_{1:t-1,1:H}) = \text{Poisson}(f_{t,n})$，其中 $f_{t,n}$ 也是由公式 15 定义．注意到当用公式 15 定义变分/建议分布时，$\{\boldsymbol{b}, \boldsymbol{W}\}$ 就构成了变分/建议分布参数集 ϕ．变分/建议分布的图模型画在了图 4(a) 下一.

4.3.1 合成数据集

实验设置. 我们随机生成了 10 组 GLM 模型的参数集 θ 用于生成数据，对应了 10 个试次．共有 $N = 5$ 个神经元，其中前 $V = 3$ 个神经元是可见的，剩下 $H = 2$ 个神经元是隐藏的．对每个试次，我们都模拟 40 个训练序列和 20 个测试序列．每个序列的长度为 $T = 100$ 个时间箱．线性．训练模型的线性权重和偏置都初始化为 0．我们用 Adam [Kingma and Ba, 2014] 作为优化器，学习率设为 0.01．每个方法跑 20 轮，每轮有 4 个批，批大小为 10．采样隐藏神经元放电的蒙特卡洛数为 $K = 2000$．每个方法我们以不同的随机数种子重复 10 次并报告结果.

结果. 从图 4(b) 的条形图中可以看出，VIS 从所有三个指标（LL、CLL 和 HLL）来看，都表现得比其他三个方法显著的好．和之前简单的混合模型类似，我们可以检查参数的估计情况，并和用于生成数据的真实参数作比较．权重平均误差和偏置平均误差是图 4(b) 中最右边两个条形图．VIS 的权重误差是最小的．对于偏置误差，VBIS 和 VIS 都是最小的，它们两个显著小于 VI 和 CHIVI.

在图 4(c) 中，我们还可视化了不同方法的参数恢复情况．对于偏置向量，我们可以看出 VI 和 CHIVI 是不如 VBIS 和 VIS 的．比如，神经元 2 的偏置本身是正的，但只有 VIS 恢复出了这一正值．对于可见到可见的权重（权重部分的左上块），四种方法都可以匹配上真实值．对于隐藏到可见的权重（权重部分的右上块），VI 和 CHIVI 因为最大化 ELBO 而没有得到足够的梯度，所以那些权重仍然保持在 0 附近．对于可见到隐藏的权重（权重部分的左下块），VI、CHIVI 和 VBIS 给出的全是看上去随机的、没什么信息量的估计值，但 VIS 却能够和真实值很好地匹配上．对于隐藏到隐藏的权重（权重部分的右下块），四种方法都没有给出好结果．隐藏到可见以及隐藏到隐藏这两块的较差结果也反映出了变分/建议分布族本身的局限性.

图 4(d) 中，我们可视化了学到的不同方法给出的预测放电率 $f_{t,n}$．图 4(d) 上一、上二展示出，四种方法中，VIS 学到的 $p(\boldsymbol{X}, \boldsymbol{Z}; \theta)$ 给出的预测放电率最精确到真实的放电率．特别的，由于只有 VIS 学到了比较好的可见到隐藏的权重，VI、CHIVI 和 VBIS 预测的放电率显著差于 VIS（图 4(d) 上一、上二）．这些对应于图 4(b) 中的 CLL 条形图．图 4(d) 下一表明，VIS 学到的建议分布可以采样到和真实隐藏神经元放电序列更接近的隐藏神经元放电序列，这可以提高学习效果，并得到更好的参数恢复．此外，图 4(d) 下二、下一中除了 VIS 以外的方法揭示了所谓 $q(\boldsymbol{Z}|\boldsymbol{X}; \theta)$ 和 $p(\boldsymbol{Z}|\boldsymbol{X}; \theta)$ 在反向 KL 散度上很接近但二者都离真实后验很远的情况，导致了和 VIS 相比较高的 ELBO 但是较低的边缘对数似然.

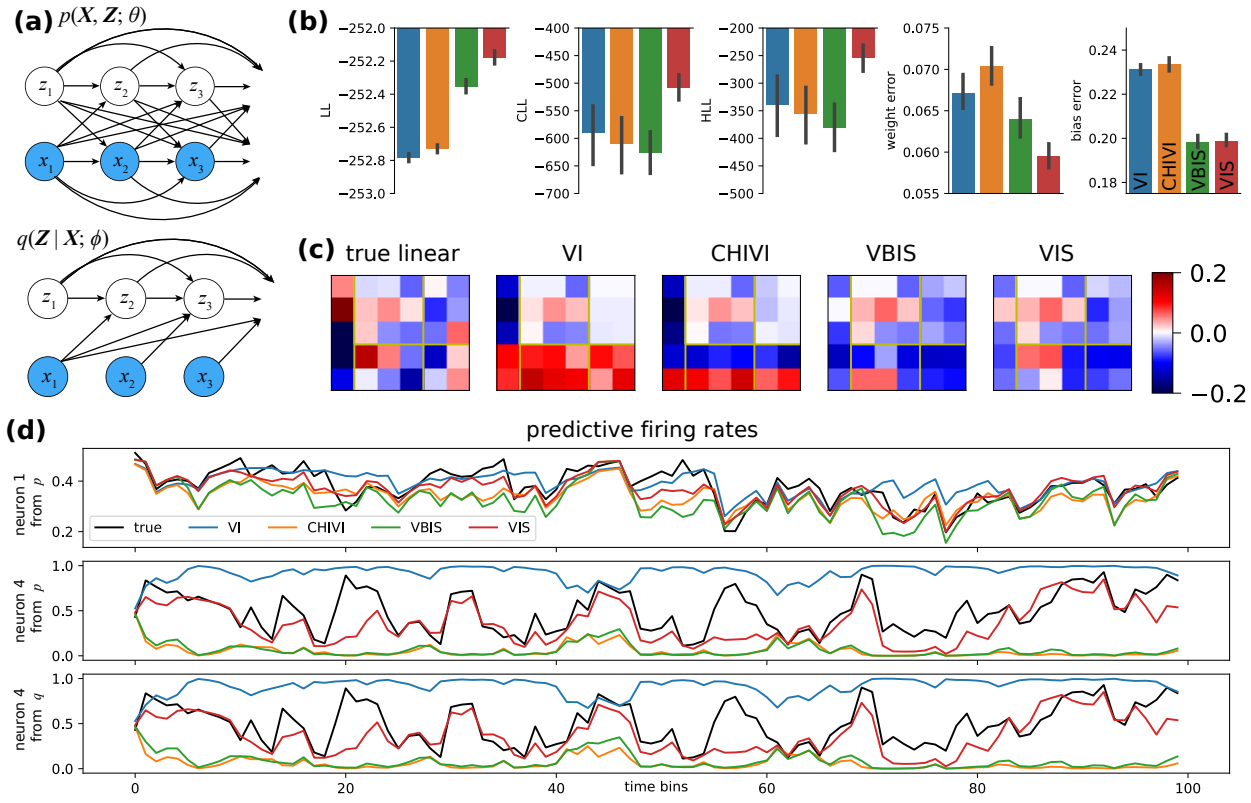

图 4: **(a)**: $p(\boldsymbol{X}, \boldsymbol{Z}; \theta)$ 和 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$ 的图模型. **(b)**: 测试集上的 LL, CLL, HLL，以及线性映射中的权重平均误差和偏置平均误差. **(c)**: 第一个试次里，真实的参数以及不同方法估计的参数. 对于每个矩阵，最左侧一列是偏置 $\boldsymbol{b}$，剩下的右边一大块是权重矩阵 $\boldsymbol{W}$. 权重的左上块是可见到可见的部分，右上块是隐藏到可见的部分，左下块是可见到隐藏的部分，右下块是隐藏到隐藏的部分. **(d)**: 不同方法得到的预测放电率. 具体来说，给定一个完整的测试放电序列 $\boldsymbol{Y} = [\boldsymbol{X}, \boldsymbol{Z}]$，我们可以由完整的生成模型 $p(\boldsymbol{X}, \boldsymbol{Z}; \theta)$ 并通过公式 14 为可见神经元（如，神经元 1）和隐藏神经元（如，神经元 4）计算预测放电率. 对于隐藏神经元（如，神经元 4），我们还可以用 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$ 来预测放电率.

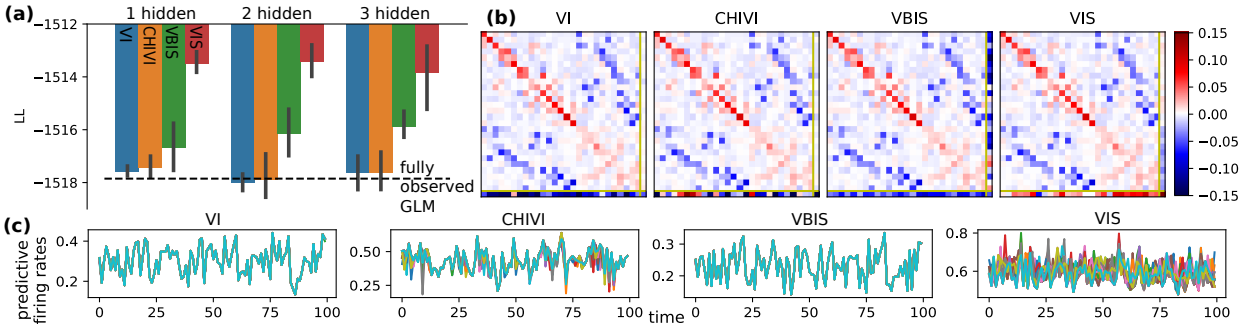

图 5: **(a)**：不同隐藏神经元个数下，测试段上的边缘对数似然. **(b)**：不同方法估计出的权重矩阵. **(c)**：由不同方法的变分/建议分布随机采样 20 次得到的预测放电率.

4.3.2 视网膜神经节细胞 (Retinal ganglion cell, RGC) 数据集

数据集. 数据集是一只老鼠在进行一项大约 20 min 的视觉实验时，$V = 27$ 个神经元的放电序列记录 [Pillow and Scott, 2012]. 神经元 1-16 是 OFF 细胞，神经元 17-27 是 ON 细胞.

实验设置. 我们用前 $\frac{2}{3}$ 段作为训练集，后 $\frac{1}{3}$ 段作为测试集. 原始的放电序列被转换成每 50 ms 时间箱内的放电数序列. 为了采用随机梯度下降算法，我们把完整的序列切成许多小段. 每一小段的长度是 100 个时间箱. 首先我们先学了一个完全可见的 GLM 作为基线. 之后我们假设有 $H \in \{1, 2, 3\}$ 个隐变量神经元代表，然后用不同的方法学习 POGLM. 我们用 Adam [Kingma and Ba, 2014] 作为优化器，学习率设为 0.01. 每个方法训练 10 轮. 批大小为 32. 蒙特卡洛样本数在 $H = 1, 2, 3$ 下分别是 1000、2000 和 3000. 我们对每个方法以不同的随机数种子重复 10 次并报告结果.

结果. 和完全可见 GLM（图 5(a) 中的短划线）相比，添加隐藏神经元并用 VBIS 或 VIS 学习显著提升了模型在测试集上预测放电事件的能力. 这反映在了图 5(a) 中 VBIS 和 VIS 的高测试边缘对数似然上. 特别的，VIS 总能比其它三个方法打到更高的测试边缘对数似然.

我们还在图 5(b) 中可视化了一个隐藏神经元时四种方法学到的权重矩阵. VIS 学到的那一个隐藏神经元，隐藏神经元到几乎所有 OFF 细胞的权重都是正的，到所有 ON 细胞的权重都是负的. 这意味着这一隐藏神经元代表表现得如果一个 OFF 细胞. 从这一隐藏神经元到其它可见神经元的权重符号清楚的表达出了这些突出后神经元的细胞类型. 其它方法给出的权重矩阵的最后一列都没有这一显著的区分性.

由于真实数据集中，我们没有真实的隐藏神经元放电序列，我们就从变分/建议分布 $q(\boldsymbol{Z}|\boldsymbol{X}; \phi)$ 中采样一些隐藏神经元放电序列，并计算用于采样隐藏神经元序列所对应的放电率. 图 5(c) 中，我们画出了 20 次随机采样得到的一个隐藏神经元情况下的隐藏神经元预测放电率. 很显然，由于最小化前向 χ^2 散度带来的大规模覆盖/寻均行为，VIS 给出的预测放电率提供了更宽有效支撑区间用于采样. 这一变化性提高了学 $\ln p(\boldsymbol{X}; \theta)$ 的有效性. 相比于 VIS，VI 和 VBIS 学到的变分/建议分布是非常受限、集中的，使得采隐藏神经元放电序列时变化性较小. 由于 CHIVI 同时最小化前向 χ^2 和反向 KL 散度，其变分/建议分布的变化性适中.

5 讨论

本篇文章中，我们介绍了变分重要性采样 (variational importance sampling, VIS) 这样一个新的基于前向 χ^2 散度的隐变量模型参数学习方法．和变分推断 (variational inference, VI) 这种最大化证据下界的方法不同，VIS 直接估计并最大化边缘对数似然来学习模型参数．我们的分析表明，估计的边缘对数似然的质量在足够多的蒙特卡洛样本和一个最优的由最小化前向 χ^2 散度量化的建议分布下就可以得到保证．这强调了选择建议分布时的统计意义．三个不同模型上的实验结果证明了 VIS 在得到更高边缘对数似然和更好模型参数估计上的能力．这强调了 VIS 是解决复杂隐变量模型的有力工具．然而需要注意的是，这种重要性采样的建议分布的选择只是统计意义上的最优，其在特定应用场景下的实际意义可能需要进一步的研究和验证．

参考文献

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.

Teun Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

Yingzhen Li and Richard E Turner. Variational inference with rényi divergence. *Statistics*, 1050, 2016.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International conference on machine learning*, pages 1511–1520. PMLR, 2016.

Xiao Su and Yuguo Chen. Variational approximation for importance sampling. *Computational Statistics*, 36(3):1901–1930, 2021.

Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A Heller, Yian Ma, and Michael I Jordan. Variational refinement for importance sampling using the forward kullback-leibler divergence. In *Uncertainty in Artificial Intelligence*, pages 1819–1829. PMLR, 2021.

Ram Naresh Saraswat. Chi square divergence measure and their bounds. In *3rd International Conference on "Innovative Approach in Applied Physical, Mathematical/Statistical", Chemical Sciences and Emerging Energy Technology for Sustainable Development*, page 55, 2014.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62 (11):5973–6006, 2016.

Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and χ 2-divergence, generalizations and applications. *Entropy*, 22(5):563, 2020.

Melanie F Pradier, Michael C Hughes, and Finale Doshi-Velez. Challenges in computing and optimizing upper bounds of marginal likelihood based on chi-square divergences. In *Second Symposium on Advances in Approximate Bayesian Inference*, 2019.

Axel Finke and Alexandre H Thiery. On importance-weighted autoencoders. *arXiv preprint arXiv:1907.10477*, 2019.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. *Advances in Neural Information Processing Systems*, 30, 2017.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Gary W Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

Łukasz Struski, Marcin Mazur, Paweł Batorski, Przemysław Spurek, and Jacek Tabor. Bounding evidence and estimating log-likelihood in vae. *arXiv preprint arXiv:2206.09453*, 2022.

David Freedman, Robert Pisani, and Roger Purves. Statistics. w. w, 1998.

Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. *arXiv preprint arXiv:2010.09541*, 2020.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

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

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

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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.

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(ARTICLE):38, 2014.

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

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

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. *Advances in neural information processing systems*, 31, 2018.

Ömer Deniz Akyildiz and Joaquín Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31:1–17, 2021.

附录 A

A.1 变分推断中的梯度估计

$\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$ 关于 θ 的导数估计为

$$\begin{aligned}\frac{\partial \mathrm{ELBO}(\boldsymbol{x};\theta,\phi)}{\partial\theta} =& \int \frac{\partial \ln p(\boldsymbol{x},\boldsymbol{z};\theta)}{\partial\theta} q(\boldsymbol{z}|\boldsymbol{x};\phi)\ \mathrm{d}\boldsymbol{z} \\ \approx& \frac{1}{K}\sum_{k=1}^{K}\frac{\partial \ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)}{\partial\theta} \\ =& \frac{\partial}{\partial\theta}\frac{1}{K}\sum_{k=1}^{K}\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right).\end{aligned} \tag{16}$$

对于 $\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$ 关于 ϕ 在 ϕ_0 处的导数，得分函数梯度估计为

$$\begin{aligned}\frac{\partial \mathrm{ELBO}(\boldsymbol{x};\theta,\phi)}{\partial\phi} =& \int \left[\ln p(\boldsymbol{x},\boldsymbol{z};\theta) - \ln q(\boldsymbol{z}|\boldsymbol{x};\phi_0)\right]\frac{\partial q(\boldsymbol{z}|\boldsymbol{x};\phi)}{\partial\phi} - q(\boldsymbol{z}|\boldsymbol{x};\phi_0)\frac{\partial \ln q(\boldsymbol{z}|\boldsymbol{x};\phi)}{\partial\phi}\ \mathrm{d}\boldsymbol{z} \\ =& \int \left[\ln p(\boldsymbol{x},\boldsymbol{z};\theta) - \ln q(\boldsymbol{z}|\boldsymbol{x};\phi_0)\right] q(\boldsymbol{z}|\boldsymbol{x};\phi_0)\frac{\partial \ln q(\boldsymbol{z}|\boldsymbol{x};\phi)}{\partial\phi}\ \mathrm{d}\boldsymbol{z} \\ & - \frac{\partial}{\partial\phi}\int q(\boldsymbol{z}|\boldsymbol{x};\phi)\ \mathrm{d}\boldsymbol{z} \\ \approx& \frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi_0\right)\right]\frac{\partial \ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)}{\partial\phi} - 0 \\ =& \frac{\partial}{\partial\phi}\frac{-1}{2K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right]^2.\end{aligned} \tag{17}$$

当重参数化技巧可以使用时，$\boldsymbol{z}|\boldsymbol{x};\phi = g(\boldsymbol{\epsilon}|\boldsymbol{x};\phi)$，其中 $\boldsymbol{\epsilon}\sim\boldsymbol{r}(\boldsymbol{\epsilon})$，那么

$$q(\boldsymbol{z}|\boldsymbol{x};\phi)\ \mathrm{d}\boldsymbol{z} = r(\boldsymbol{\epsilon})\ \mathrm{d}\boldsymbol{\epsilon}. \tag{18}$$

现在，就可以得出路径梯度估计，

$$\begin{aligned}\frac{\partial \mathrm{ELBO}(\boldsymbol{x};\theta,\phi)}{\partial\phi} =& \frac{\partial}{\partial\phi}\int q(\boldsymbol{z}|\boldsymbol{x};\phi)\left[\ln p(\boldsymbol{x},\boldsymbol{z};\theta) - \ln q(\boldsymbol{z}|\boldsymbol{x};\phi)\right]\ \mathrm{d}\boldsymbol{z} \\ =& \frac{\partial}{\partial\phi}\int r(\boldsymbol{\epsilon})\left[\ln p(\boldsymbol{x},g(\boldsymbol{\epsilon}|\boldsymbol{x};\phi)) - \ln q(g(\boldsymbol{z}|\boldsymbol{x};\phi)|\boldsymbol{x};\phi)\right]\ \mathrm{d}\boldsymbol{\epsilon} \\ \approx& \frac{\partial}{\partial\phi}\frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},g\left(\boldsymbol{\epsilon}^{(k)}\middle|\boldsymbol{x};\phi\right);\theta\right) - \ln q\left(g\left(\boldsymbol{\epsilon}^{(k)}\middle|\boldsymbol{x};\phi\right)\middle|\boldsymbol{x};\phi\right)\right].\end{aligned} \tag{19}$$

A.2 重要性采样的梯度估计

$\ln p(\boldsymbol{x};\theta)$ 关于 θ 在 θ_0 处的导数估计为

$$\begin{aligned}\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\theta} =& \frac{1}{p(\boldsymbol{x};\theta_0)}\int \frac{\partial p(\boldsymbol{x},\boldsymbol{z};\theta)}{\partial\theta}\,\mathrm{d}\boldsymbol{z}\\ =& \frac{1}{p(\boldsymbol{x};\theta_0)}\int p(\boldsymbol{x},\boldsymbol{z};\theta_0)\frac{\partial \ln p(\boldsymbol{x},\boldsymbol{z};\theta)}{\partial\theta}\,\mathrm{d}\boldsymbol{z}\\ \approx& \frac{1}{\hat{p}(\boldsymbol{x};\theta_0)}\frac{1}{K}\sum_{k=1}^{K}\frac{p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta_0\right)}{q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi_0\right)}\frac{\partial \ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)}{\partial\theta}\\ =& \frac{1}{\hat{p}(\boldsymbol{x};\theta_0)}\frac{\partial}{\partial\theta}\frac{1}{K}\sum_{k=1}^{K}\exp\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)-\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right]\\ =& \frac{1}{\hat{p}(\boldsymbol{x};\theta_0)}\frac{\partial \hat{p}(\boldsymbol{x};\theta)}{\partial\theta} = \frac{\partial \ln \hat{p}(\boldsymbol{x};\theta)}{\partial\theta}.\end{aligned} \tag{20}$$

由于 $\hat{p}(\boldsymbol{x};\theta_0)$ 出现在分母，$\frac{\partial \ln \hat{p}(\boldsymbol{x};\theta,\phi)}{\partial\phi}$ 是 $\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\phi}$ 的一个模长偏大的估计．但是 $\frac{\partial \ln \hat{p}(\boldsymbol{x};\theta,\phi)}{\partial\phi}$ 的方向是无偏的：

$$\begin{aligned}\mathbb{E}_q\left[\frac{\partial \hat{p}(\boldsymbol{x};\theta,\phi)}{\partial\theta}\right] =& \mathbb{E}_q\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)}\frac{\partial p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)}{\partial\theta}\right]\\ =& \mathbb{E}_q\left[\frac{1}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\frac{\partial p(\boldsymbol{x},\boldsymbol{z};\theta)}{\partial\theta}\right] = \int\frac{\partial p(\boldsymbol{x},\boldsymbol{z};\theta)}{\partial\theta}\,\mathrm{d}\boldsymbol{z}\\ =& \frac{\partial}{\partial\theta}\int p(\boldsymbol{x},\boldsymbol{z};\theta)\,\mathrm{d}\boldsymbol{z} = \frac{\partial p(\boldsymbol{x};\theta)}{\partial\theta}.\end{aligned} \tag{21}$$

A.3 VIS 中更新建议分布的梯度估计

本节中，我们推导用于最小化前向 χ^2 散度的得分函数梯度估计和路径梯度估计，这也等价于最小化公式 11 中的 $\ln V(\boldsymbol{x};\theta,\phi)$.

首先我们推导公式 11.

$$\begin{aligned}\ln V(\boldsymbol{x};\theta,\phi) \approx& \ln\frac{1}{K}\sum_{k=1}^{K}\frac{p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)^2}{q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)^2}\\ =& \operatorname{logsumexp}\left[2\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)-2\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right]-\ln K\\ =:& \ln\hat{V}(\boldsymbol{x};\theta,\phi).\end{aligned} \tag{22}$$

公式 11 中 $\ln V(\boldsymbol{x};\theta,\phi)$ 的得分函数梯度估计为

$$\begin{aligned}\frac{\partial \ln V(\boldsymbol{x};\theta,\phi)}{\partial\phi} =& \frac{1}{V(\boldsymbol{x};\theta,\phi_0)}\int p(\boldsymbol{x},\boldsymbol{z};\theta)^2\frac{\partial}{\partial\phi}\frac{1}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\,\mathrm{d}\boldsymbol{z}\\ =& \frac{1}{V(\boldsymbol{x};\theta,\phi_0)}\int -\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi_0)}\frac{\partial}{\partial\phi}\ln q(\boldsymbol{z}|\boldsymbol{x};\phi)\,\mathrm{d}\boldsymbol{z}\\ \approx& \frac{1}{\hat{V}(\boldsymbol{x};\theta,\phi_0)}\frac{1}{K}\sum_{k=1}^{K}-\frac{p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)^2}{q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi_0\right)^2}\frac{\partial \ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)}{\partial\phi}\\ =& \frac{1}{\hat{V}(\boldsymbol{x};\theta,\phi_0)}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\frac{\partial}{\partial\phi}\exp\left[2\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)-2\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right]\\ =& \frac{\partial}{\partial\phi}\frac{1}{2}\ln\hat{V}(\boldsymbol{x};\theta,\phi).\end{aligned} \tag{23}$$

当重参数化技巧可以使用时，$\boldsymbol{z}|\boldsymbol{x};\phi = g(\boldsymbol{\epsilon}|\boldsymbol{x};\phi)$，其中 $\boldsymbol{\epsilon}\sim\boldsymbol{r}(\boldsymbol{\epsilon})$，那么就有变换 $q(\boldsymbol{z}|\boldsymbol{x};\phi)\,\mathrm{d}\boldsymbol{z} = r(\boldsymbol{\epsilon})\,\mathrm{d}\boldsymbol{\epsilon}$ [Schulman et al., 2015]. 那么，

$$\begin{aligned}\frac{\partial \ln V(\boldsymbol{x};\theta,\phi)}{\partial\phi} =& \frac{1}{V(\boldsymbol{x};\theta,\phi_0)}\frac{\partial}{\partial\phi}\int q(\boldsymbol{z}|\boldsymbol{x};\phi)\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi)^2}\,\mathrm{d}\boldsymbol{z}\\ =& \frac{1}{V(\boldsymbol{x};\theta,\phi_0)}\frac{\partial}{\partial\phi}\int r(\boldsymbol{\epsilon})\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi)^2}\,\mathrm{d}\boldsymbol{\epsilon}\\ \approx& \frac{1}{V(\boldsymbol{x};\theta,\phi)}\frac{\partial}{\partial\phi}\frac{1}{K}\sum_{k=1}^{K}\exp\left[2\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right)-2\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right]\\ =& \frac{\partial}{\partial\phi}\ln\hat{V}(\boldsymbol{x};\theta,\phi).\end{aligned} \tag{24}$$

A.4 MNIST 数据集的隐变量流形

下图是不同方法学到的 MNIST 数据集的隐变量流形.

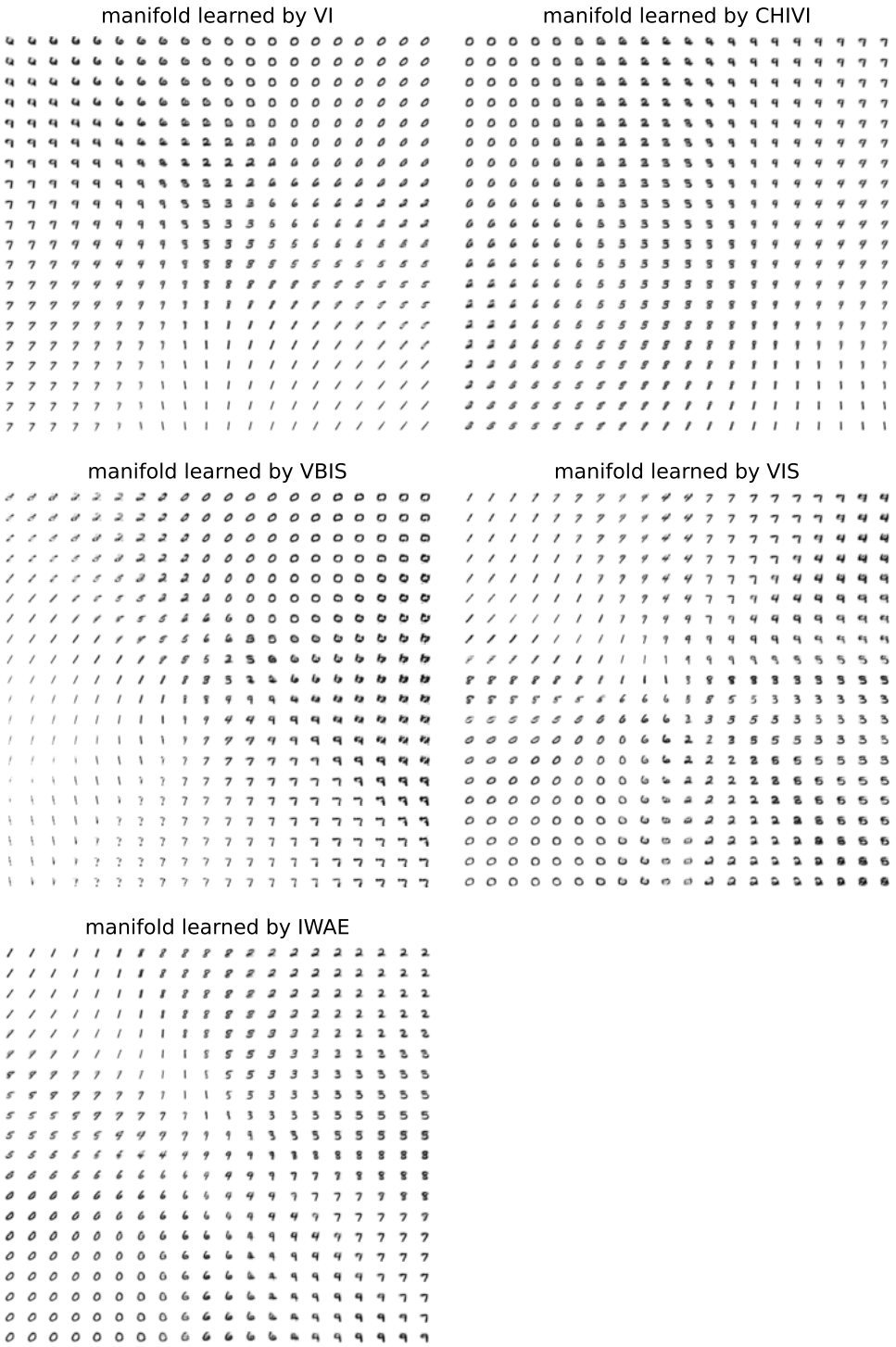

图 6: 不同方法学到的 MNIST 数据集的隐变量流形.

A.5 公式 11 的不同梯度估计

考虑到很多之前的很多工作 [Pradier et al., 2019, Finke and Thiery, 2019, Geffner and Domke, 2020, Yao et al., 2018] 已经发现最小化 χ^2 散度存在数值问题，我们再在简单的混合模型（4.1 节）上跑 VIS 方法，并用 {得分函数、路径} 梯度估计在 {对数、原} 空间中最小化前向 χ^2 散度．图 7 的结果表明，得分函数梯度估计在最小化前向 χ^2 散度上比路径梯度估计更好．此外，在对数空间中估计是非常重要的，以保证得分函数梯度估计的数值稳定性．

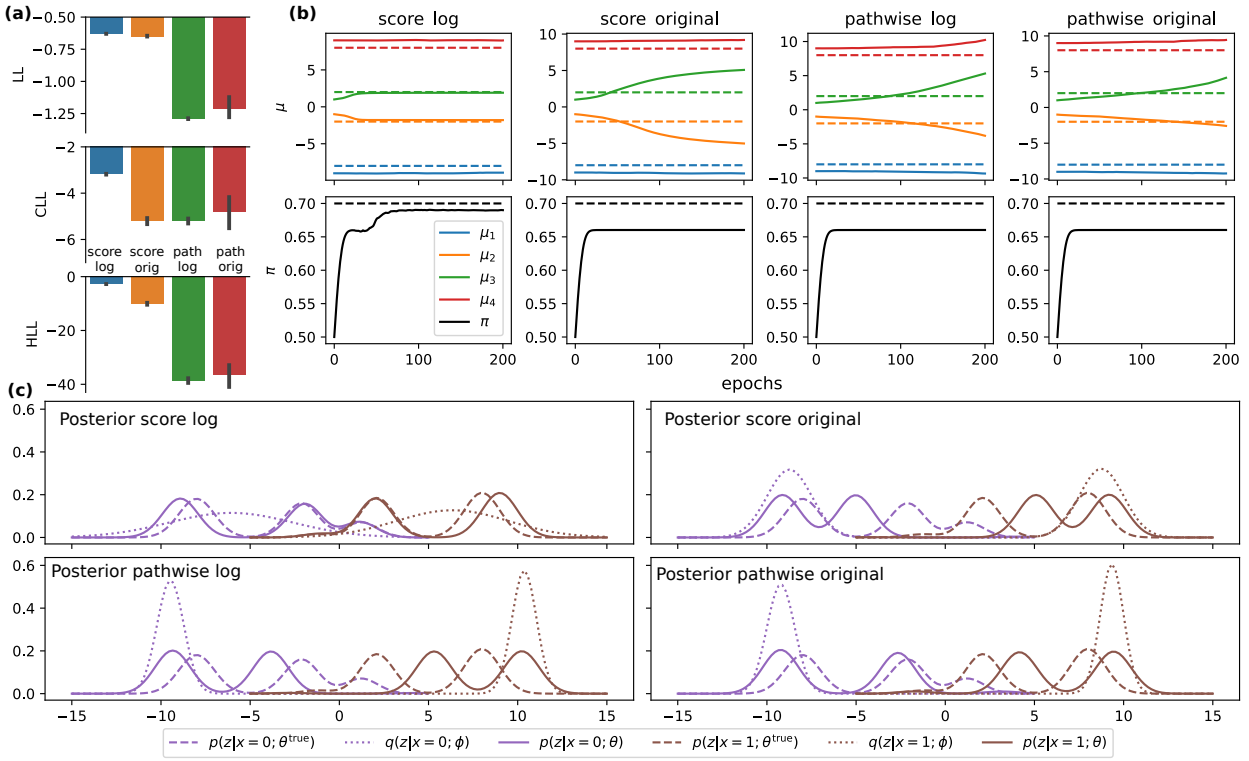

图 7: **(a)**：测试集上的 LL、CLL 以及 HLL．**(b)**：不同梯度估计学到的参数集 θ 的收敛曲线．短划线是用于生成数据的真实参数，实线是学到的参数．**(c)**：不同方法下，在 $x=0$ 和 $x=1$ 处的后验分布．短划线是真实后验 $p(\boldsymbol{z}|x;\theta^{\text{true}})$，实线是学到的后验 $p(\boldsymbol{z}|x;\theta)$，点虚线是学到的变分/建议分布给出的近似后验 $q(z|x;\phi)$．

A.6 不同方法的运行时间

图 8 展示了 POGLM 的合成数据集（4.3 节）上，不同方法测试集上的 LL 以及对应的运行时间关于蒙特卡洛样本数 K 的关系曲线. 一般来说，运行时间和蒙特卡洛数之间呈线性关系. 所有的方法都会随着蒙特卡洛样本数的增加而变好，而 VIS 总是一致得比其它方法好，尤其是在 K 很大的时候. 当 K 比较小的时候，所有的方法都不行，因为 POGLM 问题本身很复杂、困难. 这意味着对于复杂的图模型和较高维度的隐变量空间，我们的确需要足够多的样本以确保这些基于采样的方法是有效的. 因此，蒙特卡洛数与方法的选择无关，而是需要契合模型/问题本身的复杂度.

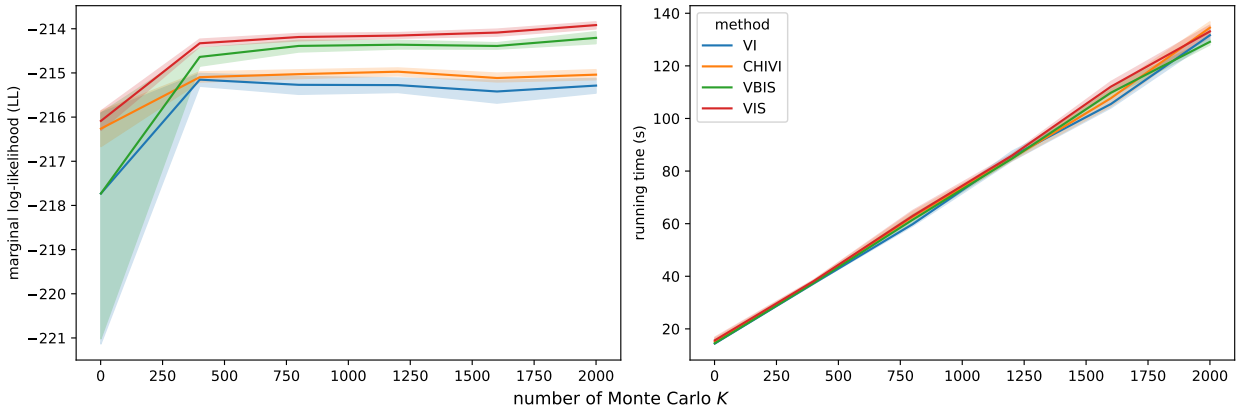

图 8: 在 POGLM 的合成数据集（4.3 节）上，不同方法测试集上的 LL 以及对应的运行时间关于蒙特卡洛样本数 K 的关系曲线.

A.7 前向 KL 散度

[Jerfel et al., 2021] 注意到了反向 KL 散度的缺点，所以他们考虑用前向 KL 散度作为更新建议分布的目标函数．然而，根据 [Sason and Verdú, 2016] 和 [Nishiyama and Sason, 2020]，$\mathrm{KL}(p\|q)$ 可以被 $\chi^2(p\|q)$ 定界：

$$\mathrm{KL}(p\|q) \leqslant \ln\left(1+\chi^2(p\|q)\right) \leqslant \chi^2(p\|q), \tag{25}$$

但反之则不行．因此，最小化前向 KL 散度可能无法得到最优的建议分布，因为最优的建议分布是最小化前向 χ^2 散度得到的．为了实证这一点，我们再次在简单的混合模型（4.1 节）上比较最小化前向 χ^2 散度 (VIS) 和最小化前向 KL 散度 (forward KL)．结果在图 9 中．

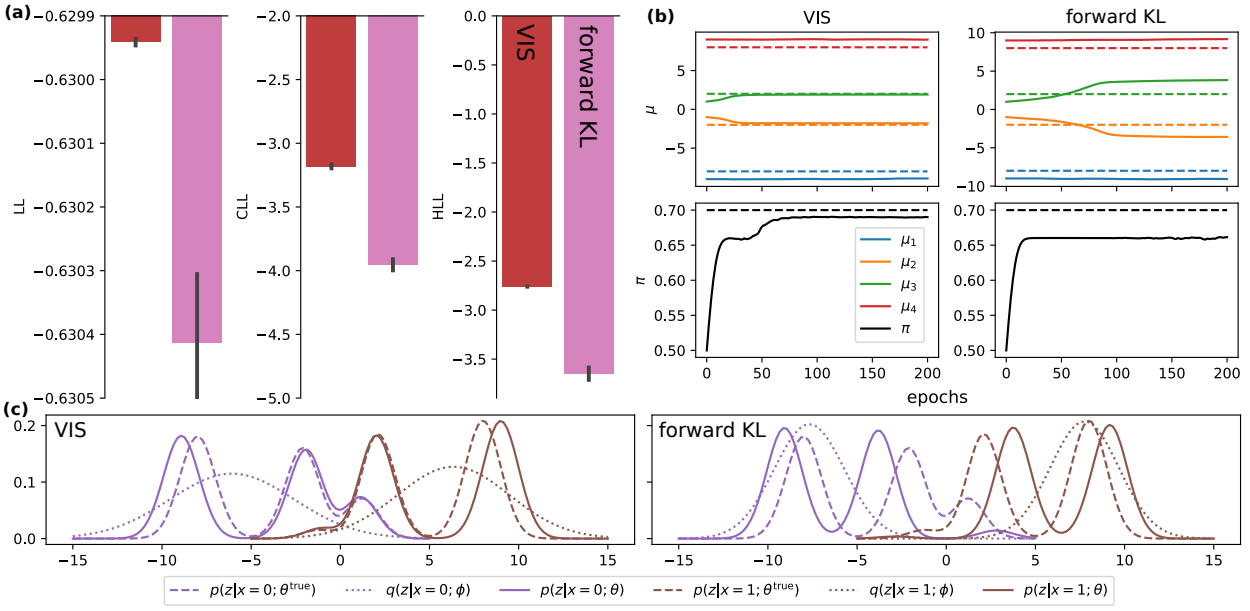

图 9: **(a)**：测试集上的 LL、CLL 以及 HLL．**(b)**：VIS 和用 forward KL 学到的参数集 θ 的收敛曲线．短划线是用于生成数据的真实参数，实线是学到的参数．**(c)**：不同方法下，在 $x=0$ 和 $x=1$ 处的后验分布．短划线是真实后验 $p(\boldsymbol{z}|x;\theta^{\text{true}})$，实线是学到的后验 $p(\boldsymbol{z}|x;\theta)$，点虚线是学到的变分/建议分布给出的近似后验 $q(z|x;\phi)$．

A.8 相关工作和贡献表

这里，我们旨在提供一个简练的相关工作和贡献总结.

表 1: 贡献.

贡献	先前文献
启发自 IS 的有效性	[3] [5] [6] [7]
旨在学 θ	[1] [3] [4] [5] [6] [7]
对 q 分布族没有限制	[1] [2] [3]
不用替代散度而是直接优化前向 χ^2 散度	[2] [3] [5] [7]
启发自 IS 在对数空间中的偏差	
对数空间中数值稳定的梯度估计	
广泛的实验于没有分解 $p(\boldsymbol{x},\boldsymbol{z};\theta)=p(\boldsymbol{x}\|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$ 的情况	
可视化推断的隐变量以及参数集 θ 的恢复	

- **启发自 IS 在对数空间中的偏差**：我们从比较 $\ln\hat{p}(\boldsymbol{x};\theta,\phi)$ 和 $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ 的偏差开始，分析为什么要做 IS 以及做 IS 的最优方式. 而且最小化前向 χ^2 散度的结论刚好与提高 IS 估计有效性是重合的（图 1）.
- **对数空间中数值稳定的梯度估计**：之前的工作已经推导了最小化 χ^2 散度的梯度估计，但是是在原空间中而不是在对数空间中. 这就导致了数值不稳定问题和无法适配高维空间的问题. 我们论证了在对数空间中得到其数值稳定、形式简洁的梯度估计的关键性，尤其是对得分函数梯度估计（图 7）.
- **广泛的实验于没有分解 $p(\boldsymbol{x},\boldsymbol{z};\theta)=p(\boldsymbol{x}|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$ 的情况**：绝大多数之前的工作都只在有显示分解 $p(\boldsymbol{x},\boldsymbol{z};\theta)=p(\boldsymbol{x}|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$ 的生成模型上进行实验，不像 POGLM. 然而，当这种显式分解不存在时，亦或是生成模型的后验分布 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 和近似后验分布 $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ 不是高斯时，ELBO 表示成 $\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)=\mathbb{E}_q[\ln p(\boldsymbol{x}|\boldsymbol{z};\theta)]-\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z};\theta))$ 就没有意义了，从而 ELBO 失去了其很多优势. 因此，我们确实需要很多不同形式的图模型来理解不同方法的表现.
- **可视化推断的隐变量以及参数集 θ 的恢复**：尽管本文中的理论内容已经展示了 VIS 的优势，我们仍然需要借助可视化来帮助我们直观的理解 VIS 比其它方法好的方式和缘由.

[1] Burda et al. [2015]
[2] Dieng et al. [2017]
[3] Finke and Thiery [2019]
[4] Jerfel et al. [2021]
[5] Domke and Sheldon [2018]
[6] Su and Chen [2021]
[7] Akyildiz and Míguez [2021]