arXiv:2302.03314v2 [stat.ML] 7 Jul 2023

结构化潜变量模型的联邦变分推断方法

Conor Hassan
数据科学中心 昆士兰科技大学
conordaniel.hassan@hdr.qut.edu.au

Robert Salomone
数据科学中心 昆士兰科技大学
robert.salomone@qut.edu.au

Kerrie Mengersen数据科学中心 昆士兰科技大学
k.mengersen@qut.edu.au

摘要

联邦学习方法能够在分布式数据源上训练模型，同时无需将数据移出其原始位置，因此在各个领域受到越来越多的关注。然而，现有方法有所局限，无法涵盖许多结构化概率模型。我们提出了一种通用而简洁的解决方案，该方案基于贝叶斯机器学习中广泛使用的结构化变分推断，并将其调整用于联邦设置。此外，我们还提供了一种通信效率更高的变体，其设计类似于经典的 FedAvg 算法。我们展示了所提算法的有效性，并将其性能与层次贝叶斯神经网络和主题模型进行了比较。

1. 引言

联邦学习（FL）是一类算法，能够在多个数据源之间实现分布式机器学习，而无需将数据汇集到同一地点（McMahan 等人，2017；Kairouz 等人，2021）。FL 方法涉及各数据源（本文称为数据孤岛）与集中式服务器之间多轮信息通信。FL 在许多现实应用中具有巨大潜力，尤其适用于医疗保健领域（Rieke 等人，2020；Xu 等人，2021）、移动计算（Lim 等人，2020）和物联网（Khan 等人，2021）。

本研究探讨概率模型的贝叶斯联邦学习设定。针对贝叶斯联邦学习，已有多种算法，包括联邦随机变分推断方法，如 pFedBayes（Zhang 等人，2022，2023），该方法估计一个全局模型，并将其用作每个数据孤岛的局部模型的先验，从而在一定程度上使各数据孤岛的推断趋向于共享的推断。其他方法包括采用马尔可夫链蒙特卡罗（MCMC）方案，例如随机梯度朗之万动力学（SGLD）（Plassier 等人，2021；El Mekkaoui 等人，2021；Vono 等人，2022）、分布式 MCMC（Neiswanger 等人，2014；Wang 等人，2015；Scott 等人，2016）、后验矩匹配（Al-Shedivat 等人，2021），以及其他推断方法，例如期望传播（EP）（Ashman 等人，2022；Guo 等人，2023）和替代似然（Jordan 等人，2018）方法。

联邦学习设置中，限制模型性能的一个主要因素包括数据孤岛之间的数据异质性，以及数据孤岛部分参与——在算法的每次迭代中，只有部分数据孤岛参与。贝叶斯联邦学习方法，例如 pFedBayes（Zhang 等人，2022，2023）和*FedPop*（Kotelevskii 等人，2022），在应对这两个问题方面展现出良好的效果。据作者所知，FedPop 是首篇将

预印本。正在审阅。

混合效应建模的思想（Gelman and Hill，2006）融入贝叶斯联邦学习领域的论文。每个数据孤岛独立于其他数据孤岛，为贝叶斯神经网络的最终层估计其参数。这些方法在异质数据上的成功，源于所选概率模型对数据层次结构的自然整合。

然而，现有算法的集合限制了适用模型的类型。据作者所知，目前不存在易于使用且适用范围广泛的方法来处理不具备以下结构的模型：所有模型参数和潜变量都依赖于所有数据孤岛中的全部数据观测值。后一种结构（如图1中间和最右侧所示）适用于广义线性模型、神经网络和高斯过程回归，但排除了相当多值得关注的潜在模型。尽管 FedPop 朝这个方向迈出了一步，但它只能优化共享参数，并要求每个数据孤岛具有相同的局部潜变量结构。

本研究旨在通过引入一种基于变分推断技术的通用近似方法，扩展贝叶斯联邦学习所适用的模型类别和推断类型。我们考虑的设置是如下形式的概率图模型

$$\boldsymbol{Z}_G \sim p_{\boldsymbol{\theta}}(\boldsymbol{Z}_G), \tag{1}$$

$$\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G \overset{\text{ind}}{\sim} p_{\boldsymbol{\theta}}(\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G), \quad j = 1, \ldots, J \tag{2}$$

$$\boldsymbol{y}_j|\boldsymbol{Z}_G, \boldsymbol{Z}_{L_j} \overset{\text{ind}}{\sim} p_{\boldsymbol{\theta}}(\boldsymbol{y}_j|\boldsymbol{Z}_{L_j}, \boldsymbol{Z}_G), \quad j = 1, \ldots, J. \tag{3}$$

上述内容中，$\boldsymbol{Z}_G$ 以及 $\boldsymbol{Z}_{L_j}$ 的集合是潜变量，分别称为全局和局部潜变量。向量 $\boldsymbol{\theta}$ 是一个可训练参数，用于参数化概率模型。我们以最一般的形式给出上述内容，但需要指出的是，在许多情况下，模型的不同部分会使用 $\boldsymbol{\theta}$ 的不同子集，或者可能完全不依赖于它。例如，我们可能有$\boldsymbol{\theta} = (\boldsymbol{\theta}_G^\top, \boldsymbol{\theta}_L^\top)^\top$，其中 $\boldsymbol{\theta}_G$ 是全局变量的先验超参数，而 $\boldsymbol{\theta}_L$ 是各个局部变量不同的先验超参数。下文中，$\boldsymbol{Z}_G$ 的维度以及每个 $\boldsymbol{Z}_{L_j}$的维度分别记为 n_G和n_{L_j}。每个 j都索引单个数据孤岛级数据$\boldsymbol{y}_j$（其还可以进一步分解为单个观测值$\boldsymbol{y}_{j,1}$，……，$\boldsymbol{y}_{j,N_j}$）及其对应的数据孤岛级潜变量 $\boldsymbol{Z}_{L_j}$（同样可能分解为$\boldsymbol{Z}_{L_j,1}$，……，$\boldsymbol{Z}_{L_j,N_j}$）。我们关注的目标是同时优化边际似然

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{z}_G) \prod_{j=1}^{J} p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{z}_{L_j}|\boldsymbol{z}_G) d\boldsymbol{z}_{L_j} d\boldsymbol{z}_G, \tag{4}$$

关于模型参数 $\boldsymbol{\theta}$，同时还要对 $\boldsymbol{Z}_G, \boldsymbol{Z}_{L_1}, \ldots, \boldsymbol{Z}_{L_J}|\boldsymbol{y}$ 的分布进行后验推断。请注意，当 $\boldsymbol{\theta} = \emptyset$ 时，任务就简化为对所有潜变量集合的后验分布进行完全贝叶斯推断。尽管每个 $\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G$ 都必须条件独立，但我们不要求它们服从相同的分布。每个 $\boldsymbol{y}_j|\boldsymbol{Z}_G, \boldsymbol{Z}_{L_j}$ 的分布也是如此。上述设定涵盖了许多类模型，例如层次混合模型（Gelman and Hill，2006）、深度潜高斯模型和变分自编码器（Kingma 和 Welling，2014；Rezende 等人，2014）、潜在狄利克雷分配系列中的主题模型（Blei 等，2003；Srivastava 和 Sutton，2017），以及 FedPop（Kotelevskii 等人，2022）等。

需要进一步说明的是，鉴于本文所研究模型的特殊类型，我们关注的是一种广义的联邦学习设置，其中数据以及任何局部潜变量后验近似的相关信息都不会离开各自的数据孤岛。据作者所知，文献中尚未探讨过这一目标：协作估计全局参数 $\boldsymbol{\theta}$ 和全局潜变量$\boldsymbol{Z}_G$，同时保护有关局部潜变量 $\boldsymbol{Z}_{L_j}$ 和数据$\boldsymbol{y}_j$的信息。我们施加这一额外限制，因为在估计 $\boldsymbol{Z}_{L_j}$时，保密至关重要；潜变量可能包含少部分观测数据点的敏感信息。图1 展示了本设置与以往设置之间的区别。

本文具体研究变分推断方法，即用下界目标函数替代边际似然（4）的方法。这类方法优雅之处在于，只需进行最少修改便可广泛应用于不同设置，而且比传统推断方法（例如涉及 MCMC 的方法）更具可扩展性。

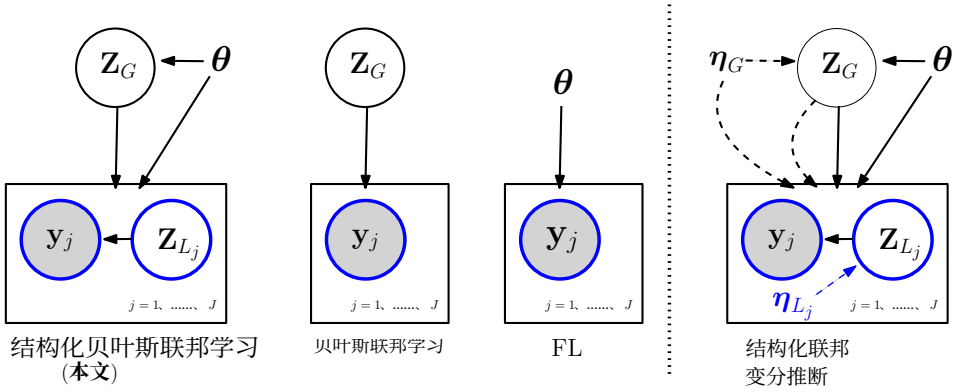


图1：不同联邦学习场景的板图。训练期间不会离开数据孤岛的变量以蓝色圈出或标注。虚线表示变分近似中的关系。

贡献。本文的主要贡献如下：（i）结构化联邦变分推断（SFVI），一种分布式推断算法，是首个适用于结构化贝叶斯联邦学习设置的方法，并且不受数据孤岛间数据分区方式的影响；（ii）*SFVI-Avg*：SFVI 的通信高效版本，类似于经典联邦学习算法*FedAvg*（McMahan 等人，2017）；（iii）一项数值研究，对上述两种方法进行比较，包括在 MNIST 上进行高维分类，以及使用乘积潜在狄利克雷分配（ProdLDA）模型的主题建模示例。

论文结构。第2节介绍随机梯度变分推断方法的必要背景。第3 节介绍 SFVI 算法、条件结构化变分族以及一种通信高效的 SFVI-Avg 算法。第4 节包含数值实验，第5节总结全文。

记号。两个向量之间的 Hadamard（逐元素）乘积记为 $\odot$。向量以粗体表示（例如 $\boldsymbol{y}$），矩阵以罗马字体表示（例如 X）。记号$\mathbf{1}$ 和$\mathbf{0}$ 分别表示适当维度的全 1 列向量或全 0 列向量。类似地，在需要时，记号 0 表示全 0 矩阵。对于矩阵 M，vec(M) 表示将 M 的各列依次堆叠得到的向量。

2. 变分推断

变分推断是一类方法，其核心是将式（4）中难以处理的边际似然（模型证据）项替换为一个下界目标函数，该函数涉及对所有潜变量的联合后验分布进行近似，即 $\boldsymbol{Z}|\boldsymbol{y}$。请注意，对于本文所考虑的模型，$\boldsymbol{Z} = (\boldsymbol{Z}_G^\top, \boldsymbol{Z}_{L_1}^\top, \ldots, \boldsymbol{Z}_{L_J}^\top)^\top$。该任务的标准目标函数是证据下界（ELBO），

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{Z}\sim q_{\boldsymbol{\eta}}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{y}) - \log q_{\boldsymbol{\eta}}(\boldsymbol{Z})]. \tag{5}$$

注意，如果 $q_{\boldsymbol{\eta}}(\boldsymbol{Z}) \equiv p_{\boldsymbol{\theta}}(\boldsymbol{Z}|\boldsymbol{y})$，那么 $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(\boldsymbol{y})$。在 $\boldsymbol{\theta}$ 固定或 $\boldsymbol{\theta} = \emptyset$ 的情况下，上述所谓的变分目标等价于最小化 $\mathrm{KL}\big(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{y})\big)$。分布 $q_{\boldsymbol{\eta}}$ 被称为变分近似或变分后验，因为它近似于潜在参数的真实模型后验。标准方法是通过随机优化，分别对 $\boldsymbol{\theta}$ 和 $\boldsymbol{\eta}$ 最大化 ELBO，具体做法是估计式（5）中 $\widehat{\ }$ 期望值的梯度。对于 $\boldsymbol{\theta}$，梯度估计量的计算很直接，即计算 $\nabla_{\boldsymbol{\theta}}\mathcal{L}$，其中 $\widehat{\nabla}\mathcal{L} := \log p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{Z}) - \log q_{\boldsymbol{\eta}}(\boldsymbol{Z})$，对于 $\boldsymbol{Z} \sim q_{\boldsymbol{\eta}}$ 而言，它是 $\mathcal{L}$ 的一个估计量。然而，为了使优化在实践中可行，需要低方差的 $\nabla_{\boldsymbol{\eta}}\mathcal{L}$ 估计量。现代机器学习中的标准策略是使用式（5）的梯度进行优化。为了以计算上可行的方式使用此类目标，对于连续分布，通常采用重参数化梯度估计量（也称为重参数化技巧）；下面我们将对此进行简要推导。假设对于 $\boldsymbol{Z} \sim q_{\boldsymbol{\eta}}$，存在一个随机表示 $\boldsymbol{Z} = f_{\boldsymbol{\eta}}(\boldsymbol{\epsilon})$ $\boldsymbol{\epsilon} \sim q_{\boldsymbol{\epsilon}}$，其

中某个分布 $q_{\boldsymbol{\epsilon}}$ 满足 $q_{\boldsymbol{\epsilon}}$ 不依赖于 $\boldsymbol{\eta}$。然后，对 $p_{\boldsymbol{\epsilon}}$ 求期望，并在允许交换梯度算子与期望算子的温和正则条件下，所需梯度向量的一个（单样本）无偏蒙特卡洛估计量为

$$\widehat{\nabla}_{\boldsymbol{\eta}}\mathcal{L} := \frac{\partial f_{\boldsymbol{\eta}}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\eta}}^{\top} \nabla_{\boldsymbol{Z}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{y}) - \log q_{\boldsymbol{\eta}}(\boldsymbol{Z})], \tag{6}$$

其中 $\boldsymbol{\epsilon} \sim q_{\boldsymbol{\epsilon}}$。式（6）中的估计量是 *sticking the landing*（STL）梯度估计量（Roeder 等人，2017）。这类估计量通常具有足够低的方差，因此每次迭代仅使用一个样本即可实现稳定优化。对变分分布族具有可重参数化形式的要求并不特别苛刻。许多分布都存在可重参数化形式，包括精度矩阵具有稀疏性、可适应条件独立结构的高斯分布（Tan 等人，2020）以及因子协方差结构（Ong 等人，2018）。其他变分族还包括归一化流（Papamakarios 等人，2021）和隐式 copula 构造（Tran 等人，2015；Smith and Loaiza-Maya，2021）。

3. 结构化联邦随机变分推断

我们考虑以下生成形式的变分近似：

$$\boldsymbol{Z}_G \sim q_{\boldsymbol{\eta}_G}(\boldsymbol{Z}_G), \tag{7}$$

$$\boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G \overset{\text{ind}}{\sim} q_{\boldsymbol{\eta}_{L_j}}(\boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G), \quad j = 1, \ldots, J, \tag{8}$$

其中，$\boldsymbol{\eta} = (\boldsymbol{\eta}_G^\top, \boldsymbol{\eta}_{L_1}^\top, \ldots, \boldsymbol{\eta}_{L_J}^\top)^\top$ 是用于参数化联合分布 $q_{\boldsymbol{\eta}}$ 的变分参数。采用上述形式，是因为变分近似的条件独立结构与目标后验分布的条件独立结构相匹配。具有这一性质的近似称为结构化近似。此类近似值得采用，因为它们能够提高参数效率并改善模型性能（Hoffman 和 Blei，2015；Ambrogioni 等人，2021），而且通常是获得高质量后验近似所必需的（Tan 和 Nott，2018；Tan 等人，2020；Quiroz 等人，2022）。我们很快将展示，这种近似对于本文研究的模型类别还具有其他理想性质，同时满足引言中讨论的联邦学习设置要求。如第 1 节所述，对模型（1）–（3）结构的一项关键洞见是，上述变分近似所需的计算能够以适当方式进行因子分解，从而实现联邦推断算法。用于说明这一点的相关计算见补充材料。

算法1给出了结构化联邦变分推断（SFVI）算法。值得强调的是，使用自动微分框架（如 JAX（Bradbury 等人，2018）计算向量-雅可比积项十分直接。

备注（***SFVI*** 不受数据分区影响）。根据构造，SFVI 得到的结果与将所有数据放在单个数据孤岛中运行 SFVI（或以任何其他方式将数据分配到各数据孤岛中）所得的相同。

3.1 结构化高斯变分族

研究表明，结构化高斯近似能够为各种模型提供良好的近似（例如，Quiroz 等人（2022）；Tan 和 Nott（2018））并且易于用于对单个 $\boldsymbol{Z}_{L_j}$ 中的更多条件独立结构进行建模，这可通过稀疏精度矩阵实现。有时更倾向于使用高斯分布族，因为与更复杂的分布族相比，它们更容易优化（Dhaka 等人，2021）。借鉴Tan 等人（2020）的类似思路——他们指出，全高斯近似可以具有条件结构——我们采用如下联合高斯变分族，

$$\boldsymbol{Z}_G \sim \mathcal{N}(\boldsymbol{\mu}_G, \Sigma_{GG}), \text{ and } \boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G \sim \mathcal{N}\left(\bar{\boldsymbol{\mu}}_j + \mathrm{C}_j(\boldsymbol{Z}_G - \boldsymbol{\mu}_G), \Sigma_{L_j L_j}\right),$$

对于 $j = 1, \ldots, J$。很容易证明，上述内容可得 $\boldsymbol{Z} = (\boldsymbol{Z}_G^\top, \boldsymbol{Z}_{L_j}^\top)^\top$ 服从联合高斯分布，其协方差为 $\mathbb{C}\mathrm{ov}(\boldsymbol{Z}_G, \boldsymbol{Z}_{L_j}) = \Sigma_{GG} C_j$。实验中使用的上述重参数化生成形式为 $(\boldsymbol{\epsilon}_G^\top, \boldsymbol{\epsilon}_{L_1}^\top, \ldots, \boldsymbol{\epsilon}_{L_J}^\top) \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$，然后设定 $\boldsymbol{Z}_G = \boldsymbol{\mu}_G + \boldsymbol{\sigma}_G \odot \mathrm{L}_G \boldsymbol{\epsilon}_G$，以及 $\boldsymbol{Z}_{L_j} = \bar{\boldsymbol{\mu}}_j + \mathrm{C}j(\boldsymbol{Z}_G - \boldsymbol{\mu}_G) + \boldsymbol{\sigma}_j \odot \mathrm{L}j \boldsymbol{\epsilon}_{L_j}$，其中 $j = 1, \ldots, J$，$\mathrm{L}_G \in \mathbb{R}^{n_G \times n_G}$ 和每个 $\mathrm{L}_j \in \mathbb{R}^{n_{L_j} \times n_{L_j}}$ 均为下单位三角矩阵。

算法 1： SFVI

输入： 服务器：初始参数 $\boldsymbol{\theta}$，初始变分参数 $\boldsymbol{\eta}_G$，迭代次数 N。
数据孤岛：初始变分参数 $\boldsymbol{\eta}_{L_j}$.

输出： 参数 $\boldsymbol{\theta}$，全局变分参数 $\boldsymbol{\eta}_G$。每个数据孤岛 j 都可以访问其各自的局部变分参数 $\boldsymbol{\eta}_{L_j}$.

对于 $i = 1, \ldots, N$ **do**

  服务器 计算

    绘制 $\boldsymbol{\epsilon}_G \sim \mathcal{N}(\mathbf{0},\ \mathrm{I}_{n_G})$

    发送 $\boldsymbol{\theta}$，$\boldsymbol{\eta}_G$，$\boldsymbol{\epsilon}_G$ 至每个数据孤岛

  对 每个数据孤岛 $j = 1, \ldots, J$ 并行**执行**

    从服务器接收 $\boldsymbol{\theta}$，$\boldsymbol{\eta}_G$，$\boldsymbol{\epsilon}_G$

    绘制 $\boldsymbol{\epsilon}_{L_j} \sim \mathcal{N}(\mathbf{0},\ \mathrm{I}_{n_{L_j}})$

    $\boldsymbol{Z}_G,\ \boldsymbol{Z}_{L_j} \leftarrow f_{\boldsymbol{\eta}_G}(\boldsymbol{\epsilon}_G)\ f_{\boldsymbol{\eta}'_{L_j}}(\boldsymbol{\epsilon}_{L_j})$

    $\boldsymbol{\eta}_{L_j} \leftarrow$ 优化器。步骤 $\left(\frac{\partial f_{\boldsymbol{\eta}'_{L_j}}(\boldsymbol{\epsilon}_G,\ \boldsymbol{\epsilon}_{L_j})}{\partial \boldsymbol{\eta}_{L_j}}^{\top} \nabla_{\boldsymbol{Z}_{L_j}} \text{对数}\ \frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{y}_j | \boldsymbol{Z}_G, \boldsymbol{Z}_{L_j}\right)}{q_{\boldsymbol{\eta}_{L_j}}} \right)$

    $\boldsymbol{g}_j^{\boldsymbol{\eta}} \leftarrow \frac{\partial f_{\boldsymbol{\eta}_G}(\boldsymbol{\epsilon}_G)}{\partial \boldsymbol{\eta}_G}^{\top} \nabla_{\boldsymbol{Z}_G} \text{对数}\ \frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{y}_j | \boldsymbol{Z}_G, \boldsymbol{Z}_{L_j}\right)}{q_{\boldsymbol{\eta}_{L_j}}(\boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G)} + \frac{\partial f_{\boldsymbol{\eta}'_{L_j}}(\boldsymbol{\epsilon}_G,\ \boldsymbol{\epsilon}_{L_j})}{\partial \boldsymbol{\eta}_G}^{\top} \nabla_{\boldsymbol{Z}_{L_j}} \text{对数}\ \frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{y}_j | \boldsymbol{Z}_G, \boldsymbol{Z}_{L_j}\right)}{q_{\boldsymbol{\eta}_{L_j}}(\boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G)}$

    $\boldsymbol{g}_j^{\boldsymbol{\theta}} \leftarrow \nabla_{\boldsymbol{\theta}} \text{对数}\ p_{\boldsymbol{\theta}}(\boldsymbol{y}_j,\ \boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G)$

    发送 $\boldsymbol{g}_j^{\boldsymbol{\theta}}, \boldsymbol{g}_j^{\boldsymbol{\eta}}$ 到服务器

  服务器 计算

    接收 $\{\boldsymbol{g}_j^{\boldsymbol{\theta}}, \boldsymbol{g}_j^{\boldsymbol{\eta}}\}_{j=1}^{J}$

    $\boldsymbol{g}^{\boldsymbol{\theta}}, \boldsymbol{g}^{\eta} \leftarrow \sum_{j=1}^{J} \boldsymbol{g}_j^{\boldsymbol{\theta}}, \sum_{j=1}^{J} \boldsymbol{g}_j^{\boldsymbol{\eta}}$

    $\boldsymbol{\theta}, \boldsymbol{\eta}_G \leftarrow$ 优化器.步骤$\left(\boldsymbol{g}^{\boldsymbol{\theta}}, \boldsymbol{g}^{\boldsymbol{\eta}}\right)$

3.2 SFVI-Avg

SFVI 主要致力于通过扩展可进行联邦分析的模型类别，推动概率联邦学习的发展。在此，我们针对结构化模型设置，考虑一种通信高效的SFVI 变体。如前所述，可以将 SFVI 视为基线*FedSGD* 联邦学习算法McMahan 等人(2017)所研究方法的类比。他们随后提出了一种名为联邦平均(FedAvg) 的替代算法。该算法在各个数据孤岛上，使用各自可用的数据训练模型，同时定期将模型参数发送给中央服务器；中央服务器对参数进行聚合和平均，再将更新后的模型参数发回各数据孤岛，以重复这一过程。McMahan 等人(2017)表明，他们的方法适用于神经网络训练，但仅适用于标准联邦学习模型设置（图1，从右数第二个）。

本文介绍一种算法 *SFVI-Avg*，该算法在结构化贝叶斯联邦学习设置中考虑了这些思路。我们首先指出，对于全局参数 $\boldsymbol{\theta}$，可以像通常在 *FedAvg* 中那样进行平均。不过，我们还考虑了三个方面，以确保该算法适用于变分参数：

1. **只需对与 $\boldsymbol{Z}_G$ 的后验变分近似相关的量进行平均**。局部潜变量只与各自的数据孤岛有关；其后验分布仅取决于该数据孤岛内的数据和全局值 $\boldsymbol{Z}_G$。因此，如果模型是针对单个数据孤岛的数据拟合的，只需对全局潜变量值进行平均。

2. **要使平均有意义，模型需要额外的结构**。尽管 SFVI 允许非常一般的模型结构，但若要通过对分布取平均来近似全局潜变量 $\boldsymbol{Z}_G$，并获得准确结果，那么所有模型都必须对其中 $\boldsymbol{y}$包含的所有 N 单个观测值使用相同的似然。为了捕捉这一点，我们牺牲一些一般性，假设采用一个简单的设定：不同数据孤岛中的观测集合具有相似的潜变量结构和相同的似然。

具体而言，模型具有以下形式：

$$\begin{aligned}\boldsymbol{Z}_G &\sim p_{\boldsymbol{\theta}}(\boldsymbol{Z}_G),\\ \boldsymbol{Z}_{L_{j,k}}|\boldsymbol{Z}_G &\overset{\text{iid}}{\sim} p_{\boldsymbol{\theta}}(\boldsymbol{Z}_{L_{j,k}}|\boldsymbol{Z}_G),\quad j=1,\ldots,J,\quad k=1,\ldots,N_j\\ \boldsymbol{y}_{j,k}|\boldsymbol{Z}_G,\boldsymbol{Z}_{L_{j,k}} &\overset{\text{iid}}{\sim} p_{\boldsymbol{\theta}}(\boldsymbol{y}_{j,k}|\boldsymbol{Z}_{L_{j,k}},\boldsymbol{Z}_G),\quad j=1,\ldots,J,\quad k=1,\ldots,N_j.\end{aligned}$$

其依据是，在这类情形下，将 $\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G$ 的对数先验项和对数似然乘以因子 N/N_j（其中 N 是所有数据孤岛中的观测值总数， N_j 是当前用于拟合模型的数据孤岛中的观测值），即可近似所有数据孤岛上联合对数密度函数的整体尺度，

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{Z}_L,\boldsymbol{y}|\boldsymbol{Z}_G)\approx\frac{N}{N_j}\sum_{k=1}^{N_j}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(\boldsymbol{Z}_{L_{j,k}},\boldsymbol{y}_{j,k}|\boldsymbol{Z}_G).$$

3. 平均方式很重要。直接对变分参数取平均，会因变分族的参数化方式不同而产生截然不同的结果。因此，我们考虑一种更合理的方法：在分布空间而非参数空间中进行平均。Wasserstein重心已用于分布式MCMC算法中的后验平均（Srivastava et al., 2018; Ou et al., 2021）。我们使用Wasserstein重心，对各数据孤岛中的全局潜变量的变分近似进行平均。我们的算法首次在联邦学习场景中考虑了这种平均方案。对于一组测度 $\{\pi_j\}_{j=1}^J$ 和 $\mathcal{W}$（ν， π 表示Wasserstein度量（Panaretos and Zemel, 2019）），Wasserstein重心定义为 $\nu^\star=\arg\min_\nu\sum_k\mathcal{W}(\nu,\pi_k)^2$。当 $\{\pi_j\}_{j=1}^J$ 中的元素均为多元高斯分布时，重心是唯一的，且本身也是高斯分布（Mallasto and Feragen, 2017, 定理4），其均值向量为 $\boldsymbol{\mu}_\star=J^{-1}\sum_j\boldsymbol{\mu}_j$，协方差矩阵则是求根问题 $\Sigma_\star=J^{-1}\sum_j(\Sigma_\star^{1/2}\Sigma_j\Sigma_\star^{1/2})^{1/2}$ 的唯一解。后者可通过不动点迭代进行数值计算（Álvarez-Esteban et al., 2016）；ott软件包中提供了相关实现（Cuturi et al., 2022）。此外，当每个 Σ_j 都是对角矩阵时，可得解析解，即对角矩阵

$$\Sigma_\star=\left(J^{-1}\sum_j\Sigma_j^{1/2}\right)^2.$$

考虑到上述几点，算法 2 给出了带平均的结构化联邦变分推断（SFVI-Avg），其索引为一个严格正整数 m，该整数决定了每一步平均之前要进行的局部 迭代次数。

备注（摊销推断）。SFVI 和 SFVI-Avg 都可以轻松扩展到摊销推断设置（Kingma 和 Welling，2014；Rezende 和 Mohamed，2015）。在这种设置下，局部潜变量的变分参数并非直接训练，而是通过某个神经网络f_ϕ（由 ϕ 参数化）定义，该网络称为推断网络。训练的是推断网络参数 ϕ，而不是直接训练变分参数 $\boldsymbol{\eta}_{L_j}$（用我们的符号表示为 $\phi\in\boldsymbol{\theta}$）。对算法所需的修改很直接；只需设置 $\boldsymbol{\eta}_{L_{j,\,k}}=f_\phi(\boldsymbol{y}_{j,\,k},\ \boldsymbol{Z}_G)$，因此 $\log\ q(\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G)$ $=\sum_{k=1}^{N_j}\log\ q(\boldsymbol{Z}_{L_{j,\ k}};\ f_\phi(\boldsymbol{y}_{j,\ k}))$。

4. 数值实验

本节展示一系列数值示例，主要旨在说明所提出的算法能够以多么简单、灵活的方式拟合复杂模型，而这些模型迄今为止尚无法应用于联邦学习设置中。另一个目标是评估所提出的*SFVI-Avg*算法相较于*SFVI*的性能。

值得强调的是，我们的算法有望助力复杂层次模型的开发，这些模型可用于异构联邦学习设置，而迄今为止尚无可用的推断算法。第一个示例将考察此类模型和推断方法。然而，尽管所考察的模型表现尚可，并且在某些情况下优于两个

算法 2： SFVI-Avg(m)

输入：服务器：初始参数 $\boldsymbol{\theta}$、初始全局变分参数 $\boldsymbol{\eta}_G$。数据孤岛：初始局部变分参数 $\boldsymbol{\eta}_{L_j}$、观测值总数 N、每个数据孤岛的观测值数量 N_j，以及轮次数量 R。输出：参数 $\boldsymbol{\theta}$、变分参数 $\boldsymbol{\eta}_G$，以及 $\boldsymbol{\eta}_{L_j}$（后者由每个数据孤岛 j 持有）。对于 $i = 1$，……，R，执行 服务器计算 将 $\boldsymbol{\theta}$、$\boldsymbol{\eta}_G$ 发送至每个数据孤岛 对于每个数据孤岛 $j = 1$，……，J，并行执行 从服务器接收 $\boldsymbol{\theta}$、$\boldsymbol{\eta}_G$ $\boldsymbol{\theta}^{(j)}$、$\boldsymbol{\eta}_G^{(j)} \leftarrow \boldsymbol{\theta}$、$\boldsymbol{\eta}_G$通过使用（6）的 $\nabla_{\boldsymbol{\theta}} \log\, p_{\boldsymbol{\theta}}(\boldsymbol{Z})$：$= \nabla_{\boldsymbol{\theta}} \log\, p_{\boldsymbol{\theta}}(\boldsymbol{Z}_G) + \frac{N}{N_j} \nabla_{\boldsymbol{\theta}} \log\, p_{\boldsymbol{\theta}}(\boldsymbol{Z}_{L_j},\ \boldsymbol{y}_j | \boldsymbol{Z}_G)$，经过 m 个局部随机梯度变分推断优化步骤，更新 $\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}_G^{(j)}, \boldsymbol{\eta}_{L_j}\right)$ 将 $\boldsymbol{\theta}^{(j)}$、$\boldsymbol{\eta}_G^{(j)}$ 发送至服务器 服务器计算 接收 $\left\{\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta}_G^{(j)}\right\}_{j=1}^{J}$ $\boldsymbol{\theta} \leftarrow J^{-1} \sum_{j=1}^{J} \boldsymbol{\theta}^{(j)}$ $\boldsymbol{\eta}_G \leftarrow$与由参数 $\boldsymbol{\eta}_G^{(j)}$ 所诱导的分布的 Wasserstein重心相对应的变分参数，对于 $j = 1$，……，J。

最先进的方法。值得强调的是，我们认为本文的主要贡献在于提供了可用于拟合一大类概率模型的算法，而这些模型此前无法在联邦学习环境中拟合。为联邦学习任务确定能够取得最佳结果的最优层次深度学习模型，超出了本文的研究范围，仍有待进一步探究。

所有实验均使用 Adam 优化器（Kingma and Ba，2015），其实现采用 optax软件包（Babuschkinetal.，2020）。请注意，实验实际上隐式地与在单个数据孤岛中使用全部数据进行变分近似的方式进行了比较，因为 SFVI 算法正是采用这种方式（参见前文说明）。

Add有关实验的更多细节以及其他实验，可参见补充材料。

4.1 异质数据的层次贝叶斯神经网络

本例使用手写数字 MNIST 数据集（Deng，2012），并采用异质数据划分设置。我们将训练集和测试集分别划分为 10 个或 50 个数据孤岛，每个数据孤岛包含相同数量的观测值。然而，每个数据孤岛中 90% 的观测值对应同一个数字（标签），其余九个数字则在剩余 10% 的观测值中近似均匀分布。对于 $i = 1$、…、784，$k = 1$、…、64，以及 $j = 1$、…、10，该模型为$\mu_{ik} \sim \mathcal{N}(0, 1)$、$\sigma \sim \mathcal{N}_+(0, 1)$、$\epsilon_{ik}^{(j)} \sim_{\text{ind}} \mathcal{N}(0, 1)$、$W_{ik}^{(1,j)} = \mu_{ik} + \sigma\epsilon_{ik}^{(j)}$ 和 $W_{ik}^{(2,j)} \sim_{\text{iid}} \mathcal{N}(0, 1)$。然后，数据孤岛 j 的个性化神经网络模型为

$$f_j(\boldsymbol{x}) = \text{softmax}\left(\mathrm{W}^{(2,j)} \text{ReLU}\left(\mathrm{W}^{(1,j)} \boldsymbol{x}\right)\right).$$

在上述设定中，$\boldsymbol{Z}_G = ((\mu_{ik}), (\sigma_{ik}))$、$\boldsymbol{Z}_{L_j} = \left(\mathrm{W}^{(1,j)},\ \left(\epsilon_{ik}^{(j)}\right)\right)$，对于 $j = 1$，……，J，以及 $\boldsymbol{\theta} = \emptyset$。

我们将其称为分层贝叶斯神经网络，因为第一层通过使用非中心参数化，在各数据孤岛之间共享第一层权重的高斯先验，从而具有层次结构。此类

表 1：在五次运行中，MNIST 示例在数据孤岛间存在严重数据异质性时的测试准确率结果。FedSOUL 和 pFedBayes 使用标准参数运行。

模型	推断	$J = 10$ 准确率。%（标准差）	$J = 50$ 准确率。%（标准差）	轮次
分层贝叶斯神经网络	SFVI-平均值	96.6 (0.68)	**93.6** (1.68)	20
	SFVI	**97.5** (0.50)	96.0 (1.03)	10^4
全贝叶斯 FedPop	SFVI-Avg	97.0 (0.62)	94.2 (1.33)	20
	SFVI	**97.5** (0.46)	96.2 (1.05)	10^4
FedPop (Kotelevskii 等人,2022)	FedSOUL	97.3 (1.41)	95.4 (8.57)	**20**
pFedBayes (Zhang 等人，2022)	pFedBayes	**94.1** (2.20)	**97.1** (1.18)	10

建模方法已被广泛证明能够有效处理数据异质性。更多详情请参阅例如 Gelman和Hill （2006）。

在联邦学习场景下，使用具有数据孤岛个性化潜变量的贝叶斯神经网络处理异质数据，这一方法首次出现在所谓的 FedPop 模型中（Kotelevskii 等人，2022）。然而，相关推断算法 FedSOUL（同样见（Kotelevskii 等人，2022））存在一些局限：该算法为全局潜变量提供最大后验估计 $\boldsymbol{Z}_G$ （而非完全贝叶斯推断），并且所有数据孤岛必须采用相同的神经网络架构。

所提出的推断算法 SFVI 和 SFVI-Avg 不存在这些局限，并且能够对所提出的分层贝叶斯神经网络以及 FedPop 模型的完全贝叶斯版本进行推断。在此示例中，我们还将其与 pFedBayes（Zhang 等人，2022）进行比较。

表 4.1 展示了各方法在五次独立运行中的比较结果；每次运行都采用不同的数据孤岛划分，划分方式如本小节开头所述。值得注意的是，SFVI o 的表现优于所有联邦学习算法。FedSOULalgorithm sings the FedPop 在联邦学习算法中取得了最佳测试预测准确率。不过，FedSOUL 算法在各数据孤岛上的测试预测准确率的标准误，远高于使用 SFVI-Avg 拟合的任一模型。

4.2 乘积潜在狄利克雷分配

第二个实验探讨了一个主题建模示例。ProdLDA（Srivastava 和 Sutton，2017）是对 Blei 等（2003）提出的原始潜在狄利克雷分配模型的轻微改进。针对该模型，Srivastava 和 Sutton（2017）证明，即使对经典 LDA 使用塌缩吉布斯 MCMC 抽样方法，ProdLDA 上的变分推断方法仍优于经典 LDA。该模型

$$\boldsymbol{T}_j \sim_{\text{iid}} \text{Dirichlet}(\beta \mathbf{1}_{n_{\text{vocab}}}), \quad j = 1, \ldots, n_{\text{topics}},$$

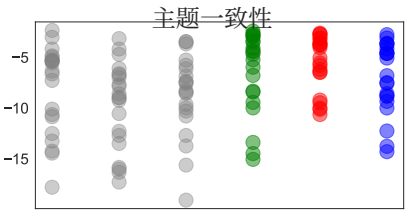
$$\boldsymbol{W}_k \sim_{\text{iid}} \mathcal{N}(\alpha \mathbf{1}_{n_{\text{topics}}}, 1), \quad k = 1, \ldots, n_{\text{docs}},$$

$$\boldsymbol{c}_k | \mathrm{T}, \boldsymbol{W}_k \sim_{\text{ind}} \text{Multinom}(l_k, \text{softmax}(\mathrm{T}\boldsymbol{W}_k)), \quad k = 1, \ldots, n_{\text{docs}}$$

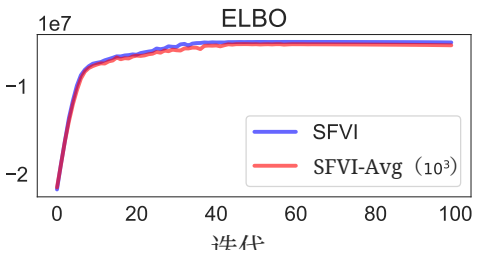
其中，$T := (\boldsymbol{T}_1, \ldots, \boldsymbol{T}_{n_{\text{topics}}}) \in \mathbb{R}^{n_{\text{vocab}} \times n_{\text{topics}}}$，且 l_k 表示文档 k 的长度（词元数量）。上文中，$\boldsymbol{\theta} = (\alpha, \beta)$、$\boldsymbol{Z}_G = (\boldsymbol{T}_1^\top, \ldots, \boldsymbol{T}_{n_{\text{topics}}}^\top)^\top$以及 $\boldsymbol{Z}_L = (\boldsymbol{W}_1^\top, \ldots, \boldsymbol{W}_{n_{\text{docs}}}^\top)^\top$。注意，总维度为 $n_{\text{vocab}} \times n_{\text{topics}}$，而 $\boldsymbol{Z}_L$是 $n_{\text{topics}} \times n_{\text{docs}}$维的。

此示例使用 20Newsgroups 数据集，其中包含 $n_{\text{docs}} = 18884$ 条来自 20 个不同在线新闻组的消息。数据经过预处理（移除停用词，并使用 gensim 软件包进行词形还原 Rehurekand Sojka (2010)），然后随机、均等地拆分到三个（3）个虚拟数据孤岛中。词汇表由语料中最常见的2000 个词构成，即 $n_{\text{vocab}} = 2000$，主题数为 $n_{\text{topics}} = 21$。鉴于后验分布的高维特性，整体近似族选为具有对角 协方差矩阵的多元高斯分布。

为比较不同方法的模型质量，我们绘制了各主题的 *(UMass)* 一致性得分 (Mimno et al., 2011)。图 4.2 (a) 展示了不同训练方法下各个主题的一致性得分。结果表明，我们提出的方法优于在每个数据孤岛中分别拟合模型。尽管 SFVI-Avg 算法在证据下界方面更高（见图 10^2 个通信轮次和 10^3 个本地步数时，在主题一致性方面优于 SFVI（图 4.2(b)）。



（a）使用以下方法训练的 ProdLDA 模型的各主题一致性值（数值越高越好）。



（b）ProdLDA 训练过程中的证据下界值在 20NewsGroups 数据集上的示例。迭代次数是图中 SFVI 所报告迭代次数的 10^3 倍，该次数标注在 x轴上。

图 2：ProdLDA 的主题一致性和证据下界值。

5. 讨论

本文通过提出一种简单而通用的解决方案，并灵活且审慎地选择近似族，扩展了可进行联邦推断的模型范围。本文还探索了在上述设置中实现通信高效推断的初步方法，并取得了令人鼓舞的结果。据作者所知，本文讨论的方法是针对结构化概率模型最通用且易于应用的方法。

结构化概率模型在 FL 语境中的兴起是一项重大进展，并具有广泛的潜在应用。现实世界数据固有的复杂性和异质性，使结构化模型能够纳入变量之间多样而复杂的关系，从而更准确地捕捉底层数据结构。这些模型中的层次表示与 FL 设置十分契合；在该设置中，数据天然分布于不同的数据孤岛之中，并为解决 FL 中显著的数据异质性问题提供了稳健的方法。

本文的工作在多个方向上都有许多可能的拓展。首先，尽管本文聚焦于证据下界这一规范的变分目标，但本文的思路很容易推广到其他目标，例如重要性加权目标（Burda 等人，2016）及其双重重参数化梯度估计器（Tan 等人，2020），或局部增强界（Geffner 和 Domke，2022）。一个值得探索的方向是，利用差分隐私框架，对包含数据或局部潜变量的信息隐私建立正式的数学保证。这方面的途径可能包括应用差分隐私随机梯度下降（Abadi 等人，2016），并拓展以往差分隐私变分推断方法中的思路（Jälkö 等人，2016）。例如，从收敛性等角度对 SFVI-Avg 的性质开展理论研究也很有意义，其他改进同样如此。例如，基于密度比估计理念的改进型平均方案（Sugiyama 等人，2012）或许会展现出良好的前景。

参考文献

Martin Abadi、Andy Chu、Ian Goodfellow、H Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。差分隐私深度学习。载于*2016 年 ACM SIGSAC 计算机与通信安全会议论文集*，2016 年。

Maruan Al-Shedivat、Jennifer Gillenwater、Eric Xing 和 Afshin Rostamizadeh。通过后验平均实现联邦学习：一种新视角与实用算法。载于国际学习表征会议（*ICLR*），2021 年。

Pedro C Álvarez-Esteban、E Del Barrio、JA Cuesta-Albertos 和 C Matrán。Wasserstein空间中重心的一种不动点方法。数学分析与应用杂志，2016。

Luca Ambrogioni、Kate Lin、Emily Fertig、Sharad Vikram、Max Hinne、Dave Moore 和 Marcel van Gerven。自动化结构化变分推断。载于人工智能与统计国际会议，2021。

Matthew Ashman、Thang D Bui、Cuong V Nguyen、Efstratios Markou、Adrian Weller、Siddharth Swaroop 和 Richard E Turner。分区变分推断：概率联邦学习框架。*arXiv*预印本 *arXiv:2202.12275*，2022。

Igor Babuschkin、Kate Baumli、Alison Bell、Surya Bhupatiraju、Jake Bruce、Peter Buchlovsky、David Budden、Trevor Cai、Aidan Clark、Ivo Danihelka、Claudio Fantacci、Jonathan Godwin、Chris Jones、Ross Hemsley、Tom Hennigan、Matteo Hessel、Shaobo Hou、Steven Kapturowski、Thomas Keck、Iurii Kemaev、Michael King、Markus Kunesch、Lena Martens、Hamza Merzic、Vladimir Mikulik、Tamara Norman、John Quan、George Papamakarios、Roman Ring、Francisco Ruiz、Alvaro Sanchez、Rosalia Schneider、Eren Sezener、Stephen Spencer、Srivatsan Srinivasan、Luyu Wang、Wojciech Stokowiec 和 Fabio Viola。DeepMind JAX生态系统，2020。网址http://github.com/deepmind。

David M Blei、Andrew Y Ng 和 Michael I Jordan。潜在狄利克雷分配。《机器学习研究杂志》，2003。

James Bradbury、Roy Frostig、Peter Hawkins、Matthew James Johnson、Chris Leary、Dougal Maclaurin、George Necula、Adam Paszke、Jake VanderPlas、Skye Wanderman-Milne 和 Qiao Zhang。JAX：Python+NumPy 程序的可组合变换，2018。网址http://github.com/google/jax。

Yuri Burda、Roger B. Grosse 和 Ruslan Salakhutdinov。重要性加权自编码器。国际学习表征会议（*ICLR*），2016。

Marco Cuturi、Laetitia Meng-Papaxanthos、Yingtao Tian、Charlotte Bunne、Geoff Davis 和 Olivier Teboul。最优传输工具（ott）：用于 Wasserstein 各类应用的 JAX 工具箱。*arXiv* 预印本 *arXiv:2201.12324*，2022。

Li Deng。用于机器学习研究的 MNIST 手写数字图像数据库。*IEEESignal ProcessingMagazine*，2012。

Akash Kumar Dhaka、Alejandro Catalina、Manushi Welandawe、Michael R Andersen、Jonathan Huggins 和 Aki Vehtari。高维变分推断中的挑战与机遇。神经信息处理系统进展，2021。

Khaoula El Mekkaoui、Diego Mesquita、Paul Blomstedt 和 Samuel Kaski。联邦随机梯度朗之万动力学。载于人工智能中的不确定性，第 1703–1712 页。PMLR，2021。

Garrett M Fitzmaurice 和 Nan M Laird。一种分析纵向二元响应的基于似然的方法。*Biometrika*，1993。

Tomas Geffner 和 Justin Domke。利用局部增强界对层次模型进行变分推断。载于国际机器学习会议 *(ICML)*，2022。

Andrew Gelmanand Jennifer Hill。使用回归和多层次/层次模型进行数据分析。剑桥大学出版社，2006。

Han Guo、Philip Greengard、Hongyi Wang、Andrew Gelman、Yoon Kim 和 Eric Xing。《将联邦学习视为变分推断：一种可扩展的期望传播方法》。载于第十一届国际学习表征会议。PMLR，2023。

Matthew D Hoffman 和 David M Blei。结构化随机变分推断。载于人工智能与统计国际会议，2015。Matthew D Hoffman、Andrew Gelman 等。无反转采样器：在哈密顿蒙特卡洛中自适应设置路径长度。《机器学习研究杂志》，2014。Joonas Jälkö、Onur Dikmen 和 Antti Honkela。非共轭模型的差分隐私变分推断。arXiv 预印本 arXiv:1610.08749，2016。Michael I Jordan、Jason D Lee 和 Yun Yang。通信高效的分布式统计推断。《美国统计学会杂志》，2018。Peter Kairouz、H Brendan McMahan、Brendan Avent、Aurélien Bellet、Mehdi Bennis、Arjun Nitin Bhagoji、Kallista Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。联邦学习的进展与开放问题。《机器学习基础与趋势》，2021。Latif U Khan、Walid Saad、Zhu Han、Ekram Hossain 和 Choong Seon Hong。物联网联邦学习：最新进展、分类体系与开放挑战。IEEE通信调查与教程，2021。Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于国际学习表征会议（ICLR），2015。Diederik P. Kingma 和 Max Welling。自编码变分贝叶斯。载于国际学习表征会议（ICLR），2014。Nikita Kotelevskii、Maxime Vono、Eric Moulines 和 Alain Durmus。FedPop：一种个性化联邦学习的贝叶斯方法。神经信息处理系统进展，2022。Wei Yang Bryan Lim、Nguyen Cong Luong、Dinh Thai Hoang、Yutao Jiao、Ying-Chang Liang、Qiang Yang、Dusit Niyato 和 Chunyan Miao。移动边缘网络中的联邦学习：全面综述。IEEE通信调查与教程，2020。Anton Mallasto 和 Aasa Feragen。从不确定曲线中学习：用于高斯过程的2-瓦瑟斯坦度量。神经信息处理系统进展，2017。Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。利用去中心化数据进行深度网络的通信高效学习。载于人工智能与统计国际会议，2017。David Mimno、Hanna Wallach、Edmund Talley、Miriam Leenders 和 Andrew McCallum。优化主题模型中的语义连贯性。载于2011年自然语言处理实证方法会议论文集，2011。Kevin P. Murphy。《概率机器学习：高级主题》。麻省理工学院出版社，2023。网址：http://probml.github.io/book2。Willie Neiswanger、Chong Wang 和 Eric P. Xing。渐近精确且易于并行化的MCMC。载于第三十届人工智能不确定性会议（UAI），2014。Victor M-H Ong、David J Nott 和 Michael S Smith。具有因子协方差结构的高斯变分近似。《计算与图形统计学杂志》，2018。Rihui Ou、Deborshee Sen 和 David Dunson。通过分而治之实现时间序列的可扩展贝叶斯推断。arXiv 预印本 arXiv:2106.11043，2021。Victor M Panaretos 和 Yoav Zemel。瓦瑟斯坦距离的统计学方面。《统计学及其应用年度评论》，2019。George Papamakarios、Eric T Nalisnick、Danilo Jimenez Rezende、Shakir Mohamed 和 Balaji Lakshminarayanan。用于概率建模与推断的归一化流。《机器学习研究杂志》，2021。

Du Phan、Neeraj Pradhan 和 Martin Jankowiak。《NumPyro 中用于灵活且加速概率编程的可组合效应》。*arXiv* 预印本 *arXiv:1912.11554*，2019。Vincent Plassier、Maxime Vono、Alain Durmus 和 Eric Moulines。《Dg-lmc：通过 Gibbs 算法中的朗之万蒙特卡罗实现的一种即用型、可扩展的同步分布式 MCMC 算法》。载于国际机器学习会议，第 8577–8587 页。PMLR，2021。Matias Quiroz、David J Nott 和 Robert Kohn。《高维状态空间模型的高斯变分近似》。*Bayesian Analysis*，2022。Radim Rehurek 和 Petr Sojka。《面向大规模语料库主题建模的软件框架》。载于*LREC 2010* 新自然语言处理框架挑战研讨会论文集，2010。Danilo Rezende 和 Shakir Mohamed。《利用归一化流进行变分推断》。载于国际机器学习会议（*ICML*），2015。Danilo Jimenez Rezende、Shakir Mohamed 和 Daan Wierstra。《深度生成模型中的随机反向传播与近似推断》。载于国际机器学习会议（*ICML*），2014。Nicola Rieke、Jonny Hancox、Wenqi Li、Fausto Milletari、Holger R Roth、Shadi Albarqouni、Spyridon Bakas、Mathieu N Galtier、Bennett A Landman、Klaus Maier-Hein 等。《联邦学习与数字健康的未来》。*NPJ Digital Medicine*，2020。Geoffrey Roeder、Yuhuai Wu 和 David K Duvenaud。《sticking the landing：用于变分推断的简单低方差梯度估计量》。神经信息处理系统进展，2017。Steven L Scott、Alexander W Blocker、Fernando V Bonassi、Hugh A Chipman、Edward I George 和 Robert E McCulloch。《贝叶斯与大数据：共识蒙特卡罗算法》。国际管理科学与工程管理杂志，2016。Michael Stanley Smith 和 Rubén Loaiza-Maya。《隐式Copula变分推断》。*arXiv* 预印本 *arXiv:2111.09511*，2021。Akash Srivastava 和 Charles Sutton。《面向主题模型的自编码变分推断》。载于国际学习表征会议（*ICLR*），2017。Sanvesh Srivastava、Cheng Li 和 David B Dunson。《通过Wasserstein空间中的重心实现可扩展贝叶斯方法》。《机器学习研究杂志》，2018。Masashi Sugiyama、Taiji Suzuki 和 Takafumi Kanamori。《style id='25'>机器学习中的密度比估计。剑桥大学出版社，2012。Linda SL Tan 和 David J Nott。《具有稀疏精度矩阵的高斯变分近似》。统计与计算，2018。Linda SL Tan、Aishwarya Bhaskaran 和 David J Nott。《带重要性权重的条件结构化变分高斯近似》。统计与计算，2020。Dustin Tran、David Blei 和 Edo M Airoldi。《Copula变分推断》。神经信息处理系统进展，2015。Maxime Vono、Vincent Plassier、Alain Durmus、Aymeric Dieuleveut 和 Eric Moulines。《QLSD：用于贝叶斯联邦学习的量化朗之万随机动力学》。载于国际人工智能与统计学会议，2022。Xiangyu Wang、Fangjian Guo、Katherine A Heller 和 David B Dunson。《利用随机划分树并行化MCMC》。神经信息处理系统进展，2015。Jie Xu、Benjamin S Glicksberg、Chang Su、Peter Walker、Jiang Bian 和 Fei Wang。《面向医疗信息学的联邦学习》。医疗信息学研究杂志，2021。

Xu Zhang、Yinchuan Li、Wenpeng Li、Kaiyang Guo 和 Yunfeng Shao。《通过变分贝叶斯推断实现个性化联邦学习》。载于国际机器学习会议，第 26293–26310 页。PMLR，2022。

Xu Zhang、Wenpeng Li、Yunfeng Shao 和 Yinchuan Li。《通过变分贝叶斯推断实现联邦学习：个性化、稀疏性与聚类》，2023。

补充材料

S1. SFVI 更新的推导

首先，请注意，作为该结构的一个结果

$$\boldsymbol{Z}_G \sim q_{\boldsymbol{\eta}_G}(\boldsymbol{Z}_G), \tag{S1}$$

$$\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G \overset{\text{ind}}{\sim} q_{\boldsymbol{\eta}_{L_j}}(\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G), \quad j=1,\ldots,J, \tag{S2}$$

重参数化变量必然分解为 $\boldsymbol{\epsilon} = \left(\boldsymbol{\epsilon}_G^\top,\ \boldsymbol{\epsilon}_{L_1}^\top,\ \ldots\ldots,\ \boldsymbol{\epsilon}_{L_J}^\top\right)^\top$ 和

$$\boldsymbol{Z}_{\boldsymbol{\eta}_G} \coloneqq f_G(\boldsymbol{\epsilon}_G), \quad \boldsymbol{Z}_{L_j} \coloneqq f_{\boldsymbol{\eta}'_{L_j}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_j}),\ j=1,\ldots,J,$$

其中 $\boldsymbol{\eta}'_{L_j} \coloneqq (\boldsymbol{\eta}_G^\top,\ \boldsymbol{\eta}_{L_j}^\top)$，因此 $f_{\boldsymbol{\eta}}(\boldsymbol{\epsilon})$ 等于

$$\left(f_{\boldsymbol{\eta}_G}(\boldsymbol{\epsilon}_G)^\top, f_{\boldsymbol{\eta}'_{L_1}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_1})^\top, \ldots, f_{\boldsymbol{\eta}'_{L_J}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_J})^\top\right)^\top,$$

其转置雅可比矩阵具有如下形式的分块上三角结构

$$\begin{pmatrix} \frac{\partial f_{\boldsymbol{\eta}_G}(\boldsymbol{\epsilon}_G)}{\partial \boldsymbol{\eta}_G}^\top & \frac{\partial f_{\boldsymbol{\eta}'_{L_1}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_1})}{\partial \boldsymbol{\eta}_G}^\top & \cdots & \frac{\partial f_{\boldsymbol{\eta}'_{L_J}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_J})}{\partial \boldsymbol{\eta}_G}^\top \\ 0 & \frac{\partial f_{\boldsymbol{\eta}'_{L_1}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_1})}{\partial \boldsymbol{\eta}_{L_1}}^\top & \vdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{\partial f_{\boldsymbol{\eta}'_{L_J}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_J})}{\partial \boldsymbol{\eta}_{L_J}}^\top \end{pmatrix}.$$

由于上述稀疏结构，STL 梯度估计器（Roeder 等人，2017）

$$\widehat{\nabla}_{\boldsymbol{\eta}}\mathcal{L} \coloneqq \frac{\partial f_{\boldsymbol{\eta}}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\eta}}^\top \nabla_{\boldsymbol{Z}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{y}) - \log q_{\boldsymbol{\eta}}(\boldsymbol{Z})], \tag{S3}$$

可分解为

$$\widehat{\nabla}_{\boldsymbol{\eta}}\mathcal{L} = \left((\widehat{\nabla}_{\boldsymbol{\eta}_G}\mathcal{L})^\top, (\widehat{\nabla}_{\boldsymbol{\eta}_{L_1}}\mathcal{L})^\top, \ldots, (\widehat{\nabla}_{\boldsymbol{\eta}_{L_J}}\mathcal{L})^\top\right)^\top,$$

其中，上述左侧各项的定义与（S3）类似。接下来，令

$$\widehat{\mathcal{L}}_0 \coloneqq \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{Z}_G)}{q_{\boldsymbol{\eta}_G}(\boldsymbol{Z}_G)}, \text{and } \widehat{\mathcal{L}}_j \coloneqq \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G)}{q_{\boldsymbol{\eta}_{L_j}}(\boldsymbol{Z}_{L_j}|\boldsymbol{Z}_G)},$$

对于 $j=1,\ \ldots\ldots,\ J$，注意到 $\widehat{\mathcal{L}} = \sum_{j=0}^J \widehat{\mathcal{L}}_j$。经过一些代数运算，并去掉在某些梯度下为零的项，可得

$$\widehat{\nabla}_{\boldsymbol{\eta}_G}\mathcal{L} = \frac{\partial f_{\boldsymbol{\eta}_G}(\boldsymbol{\epsilon}_G)}{\partial \boldsymbol{\eta}_G}^\top \nabla_{\boldsymbol{Z}_G}\widehat{\mathcal{L}}_0 + \sum_{j=1}^J \boldsymbol{g}_j^{\boldsymbol{\eta}}, \tag{S4}$$

其中

$$\boldsymbol{g}_j^{\boldsymbol{\eta}} := \frac{\partial f_{\boldsymbol{\eta}_G}(\boldsymbol{\epsilon}_G)}{\partial \boldsymbol{\eta}_G}^\top \nabla_{\boldsymbol{Z}_G} \widehat{\mathcal{L}}_j + \frac{\partial f_{\boldsymbol{\eta}'_{L_j}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_J})}{\partial \boldsymbol{\eta}_G}^\top \nabla_{\boldsymbol{Z}_{L_j}} \widehat{\mathcal{L}}_j. \tag{S5}$$

最后，经过直接计算可得 $j = 1$、.....、 J，

$$\widehat{\nabla}^{\text{STL}}_{\boldsymbol{\eta}_{L_j}} \mathcal{L} = \frac{\partial f_{\boldsymbol{\eta}'_{L_j}}(\boldsymbol{\epsilon}_G, \boldsymbol{\epsilon}_{L_j})}{\partial \boldsymbol{\eta}_{L_j}}^\top \nabla_{\boldsymbol{Z}_{L_j}} \widehat{\mathcal{L}}_j. \tag{S6}$$

由上可知，更新 $\boldsymbol{\eta}_{L_j}$ 所需的唯一项是 $\boldsymbol{\epsilon}_G$ 和 $\boldsymbol{\eta}_G$。此外，更新 $\boldsymbol{\eta}_G$ 所需的唯一项是根据（S6）定义的 $\boldsymbol{g}_j$，它们不需要了解 $\boldsymbol{Z}_{L_j}$（由变分近似模拟得到），也不需要知道生成这些样本的分布的参数。最后，相对于 $\boldsymbol{\theta}$ 的梯度更新可简化为

$$\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}} = \log p_{\boldsymbol{\theta}}(\boldsymbol{Z}_G) + \sum_{j=1}^{J} \boldsymbol{g}_j^{\boldsymbol{\theta}}, \tag{S7}$$

其中

$$\boldsymbol{g}_j^{\boldsymbol{\theta}} := \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{Z}_{L_j} | \boldsymbol{Z}_G). \tag{S8}$$

因此，数据孤岛只需与服务器共享 $\boldsymbol{g}_j^{\boldsymbol{\theta}}$，即可执行 $\boldsymbol{\theta}$ 所需的更新。该算法可直接由上述简化结果得到。

S2. 论文中的实验详情

S2.1 层次化贝叶斯神经网络

对于 SFVI 和 SFVI-Avg 的结果，所有实验均使用对角高斯变分近似。pFedBayes 和 FedPop 的结果采用各自论文中用于 MNIST 数据的默认超参数。

S3. 其他实验

S3.1 贝叶斯逻辑混合模型

本例考察逻辑混合模型的完全贝叶斯分析。这里，我们考虑六城市数据集（Fitzmaurice 和 Laird，1993），该数据集取自一项针对537 名儿童的纵向研究，旨在评估空气污染对健康的影响。每年记录一次二元响应 y_{ij}，用于表示是否出现喘息，记录时间为儿童 7 至 10 岁期间，其中 i 表示儿童的索引，j 表示每名儿童的四次观测值。模型包含两个协变量：母亲吸烟状况的二元指标 x_i^{smoke}，以及以九岁为中心化基准的儿童当前年龄x_{ij}^{age}。模型如下：

$$\begin{aligned}
y_{ij} | \boldsymbol{\beta}, Z_i &\sim_{\text{ind}} \text{Bern}(p_{ij}), \quad i = 1, \ldots 537,\ j = 1, \ldots, 4, \\
p_{ij} &= \text{logit}^{-1}\big(\beta_0 + \beta_1 x_i^{\text{smoke}} + \beta_2 x_{ij}^{\text{age}} + \beta_3 x_i^{\text{smoke}} \cdot x_{ij}^{\text{age}} + b_i\big), \\
\beta_k &\sim_{\text{iid}} \mathcal{N}(0, 10^2), \quad k = 0, \ldots, 3, \\
\omega &\sim \mathcal{N}(0, 10^2), \\
b_i | \omega &\sim_{\text{iid}} \mathcal{N}(0, \exp(-2\omega)), \quad i = 1, \ldots, 537.
\end{aligned}$$

上文中，每个 b_i 都是一个随机效应项，形式为随机截距。在此例中，我们有 $\boldsymbol{Z}_G = (\boldsymbol{\beta}^\top, \omega)^\top$、$\boldsymbol{Z}_L = \boldsymbol{b}$和 $\boldsymbol{\theta} = \emptyset$。为了进一步捕捉目标后验分布的结构，我们令 $\text{L}_j \equiv \text{I}$，因为给定 $\boldsymbol{Z}_G$ 和观测数据后，每个 $\boldsymbol{b}_i$ 在后验分布中条件独立。我们将数据随机划分为两个数据孤岛，分别包含 300 名和 237 名儿童。尽管该模型看似简单，但众所周知，即使对 Tan 等人（2020）来说，它也颇具挑战性；一般而言，在完全贝叶斯框架下，混合模型需要采用结构化近似，才能得到合理的后验边缘近似（例如，Tan 和 Nott（2018））。此外，数据划分不均往往会导致各数据孤岛独立拟合数据时，全局变量的推断结果大相径庭。上述因素再加上小样本数据这一设定，使得此例对 SFVI-FedAvg 构成了挑战。

我们将 SFVI 与在全部数据上拟合的 MCMC（使用 Hoffman 等人，2014提出并在 NumPyro (Phan 等人，2019）中实现的无反转采样器）以及独立数据孤岛进行比较。图 S1 显示，SFVI

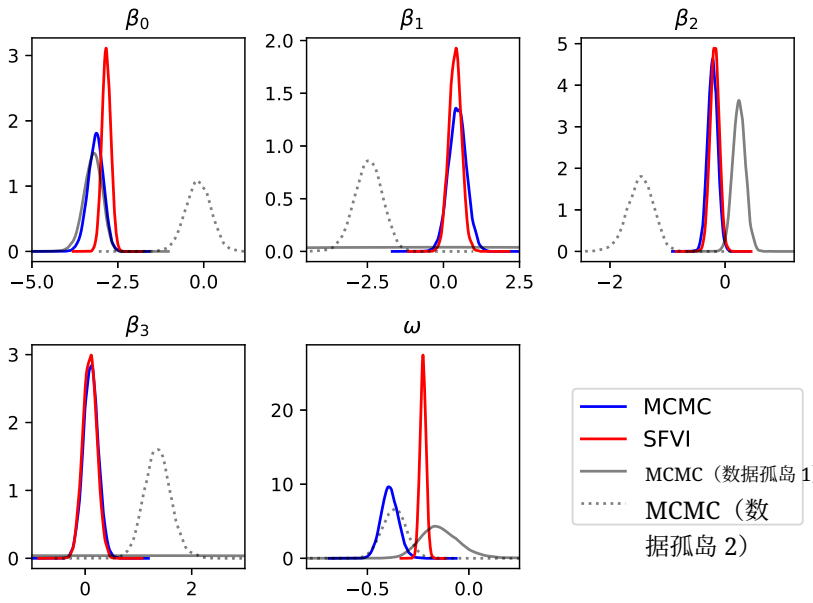


图 S1：贝叶斯广义线性混合模型示例的边缘后验参数。

能够准确捕获回归参数 $\boldsymbol{\beta}$ 的边缘后验分布，即使各独立数据孤岛的边缘密度几乎没有重叠。使用不同随机种子的多次运行均得到了相符的结果。遗憾的是，SFVI-Avg 未能为此示例给出合理的结果。这表明，尽管它在其他示例中表现十分出色，但并非适用于所有场景的通信高效型联邦推断万能方案。

S3.2 经验贝叶斯多项式回归

第一个示例使用 MNIST 手写数字数据集（Deng，2012），其中每个样本都是一幅 28×28 像素图像（共 784 个像素），并附有表示图像中手写数字（0–9）的标签。模型为

$$
\begin{aligned}
W_{jk} &\sim_{\text{iid}} \mathcal{N}(0, \sigma_{\text{W}}^2), \quad j = 1, \ldots, 10, \; k = 1, \ldots, 784 \\
b_j &\sim_{\text{iid}} \mathcal{N}(\mathbf{0}, \sigma_b^2), \quad j = 1, \ldots, 10, \\
\boldsymbol{c}_k | \text{W}, \boldsymbol{b} &\sim_{\text{iid}} \text{Multinomial}(10, \text{logit}^{-1}(\text{W}\boldsymbol{x}_k + \boldsymbol{b})),
\end{aligned}
$$

对于 $k = 1, \ldots, n$，其中 n 是数据点总数，σ_{W}^2 和 σ_b^2 是训练过程中学习的正值超参数。因此，对于此示例，$\boldsymbol{Z}_G = (\text{vec(W)}\ \boldsymbol{b}^\top)^\top$，

$\boldsymbol{Z}_L = \emptyset$　　$\boldsymbol{\theta} = \ \sigma_W^2\ \sigma_b^2$　　2023,，且 (,)。

这是一个经验贝叶斯推断过程（参见 Murphy（第 3.9 章）），因为我们正在优化潜变量先验分布的超参数。数据均匀地分布在各数据孤岛中。由于 $\boldsymbol{Z}_G$（$n_G = 7850$）的维度很高，因此采用独立性近似（即 $\text{L}_G \equiv \text{I}$），以便通过解析解高效地计算 SFVI-Avg 的重心。表 S1 报告了训练集和测试集上的平均后验预测准确率。值得注意的是，在训练过程中多次使用 SFVI-Avg，似乎会对泛化性能产生不可忽视的影响（与仅平均一次相比）。此外，我们在本文的实验中发现，仅使用 SFVI-Avg 进行五个通信轮次、每轮执行 10^3 个本地步数来初始化默认 SFVI，就可以使达到收敛所需的步数减少一半，如图 S2 所示。

表 S1：MNIST 数据集上的多项式回归结果。

J	N_j	方法	准确率 (%) 训练	Test	轮次
25	200	独立	73.0	**51.5**	0
		SFVI-Avg($5 \cdot 10^4$)	**65.7**	63.9	1
		SFVI-Avg(10^3)	71.3	69.3	50
		SFVI	83.8	**84.0**	$5 \cdot 10^4$
5	10^4	独立	88.7	86.8	0
		SFVI-Avg($5 \cdot 10^4$)	90.4	90.4	1
		SFVI-Avg(10^3)	90.1	90.2	50
		SFVI	**90.7**	90.8	$5 \cdot 10^4$

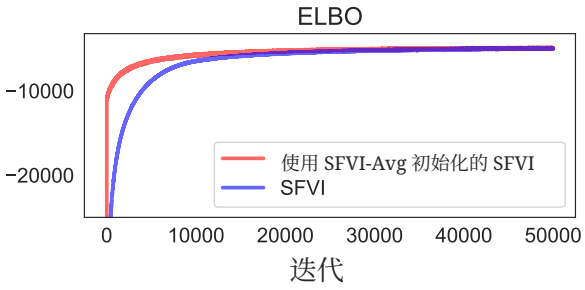


图 S2：SFVI-Avg 初始化 SFVI，而非从头开始。