arXiv:2305.04152v2 [cs.IT] 9 May 2023

# 基于信道驱动随机梯度朗之万动力学的贝叶斯空中联邦平均

Boning Zhang*、Dongzhu Liu*、Osvaldo Simeone†和 Guangxu Zhu‡
*计算科学学院，格拉斯哥大学
† 工程系，伦敦国王学院
‡深圳大数据研究院 邮箱：b.zhang.6@research.gla.ac.uk,dongzhu.liu@glasgow.ac.uk,
osvaldo.simeone@kcl.ac.uk, gxzhu@sribd.cn

摘要——近期可扩展贝叶斯推断方法的发展，使人们重新关注将贝叶斯学习作为传统频率学派学习的替代方案；通过不确定性量化，贝叶斯学习能够改进模型校准。近期，联邦平均朗之万动力学（FALD）作为联邦平均的一种变体被提出，可在无噪声通信条件下高效实现分布式贝叶斯学习。本文提出无线联邦平均朗之万动力学（WFALD），这是一种新型协议，通过结合空中计算与信道驱动采样来执行蒙特卡罗更新，从而在无线系统中实现 FALD。与以往的无线贝叶斯学习研究不同，WFALD 支持：（i）在通信轮次之间进行多次本地更新；（ii）基于小批量计算随机梯度。本文还通过比较 WFALD 生成的样本与目标全局后验分布之间的 2-瓦瑟斯坦距离，给出了收敛分析。分析和实验表明，当信噪比足够高时，信道噪声可以完全用于蒙特卡洛采样，因此不会造成性能损失。

索引词——贝叶斯联邦学习、随机梯度朗之万动力学、功率控制、空中计算。

## 一、引言

随着机器学习工具日益广泛地应用于敏感领域，深度学习技术的可靠性受到了严格审视，推动了对不确定性量化与校准的研究（例如，见 [1], [2]）。经过良好校准的机器学习技术的黄金标准是贝叶斯学习，它将模型参数视为随机变量。可扩展贝叶斯推断领域的最新进展，使贝叶斯学习技术成为需要进行不确定性量化的应用中切实可行的竞争方案 [3]–[6]。本文聚焦于在无线系统中高效地分布式实现贝叶斯学习（见图 1）。

联邦学习（FL）是指一种分布式学习协议，使多个设备能够协同训练统计模型，而无需直接共享本地数据集 [7], [8]。训练过程需要在服务器与设备之间通信模型参数。为降低通信成本，一种广受欢迎的方法由

O. Simeone 的工作得到了欧洲研究委员会（ERC）通过欧盟地平线 2020研究与创新计划（资助协议编号 725731）、英国工程与自然科学研究委员会（EPSRC）开放奖学金（编号 EP/W024101/1），以及欧盟地平线欧洲计划项目 CENTRIC（101096379）的支持。

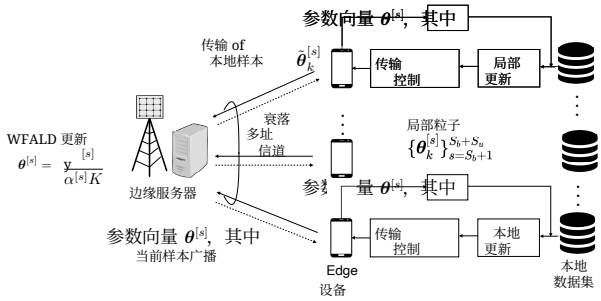


图 1. 所提出的无线联邦平均朗之万动力学（WFALD）协议，通过空中信道驱动采样以及全局聚合步骤之间的多次本地更新，实现分布式贝叶斯学习。

联邦平均（FedAvg） [7]，通过允许设备在将本地模型发送至中央服务器进行全局聚合之前执行多次本地更新，来减少通信频率。在无线信道上实现联邦学习，则提供了另一种通过空中计算（AirComp） [9]–[11]提高通信效率的机会。空中计算利用非正交多址接入（NOMA）信道传输本地更新，采用无码传输，并利用无线信道的叠加特性，实现高效的空中模型聚合。

大多数现有的联邦平均都采用频率学派学习，其目标是通过随机梯度下降（SGD）算法 [12] 估计使经验损失最小化的最优模型。实现贝叶斯学习最简单的方法之一，是采用随机梯度下降的一种变体，即随机梯度朗之万动力学（SGLD） [13]。通过在随机梯度下降更新中注入高斯噪声，随机梯度朗之万动力学生成的模型参数可以近似贝叶斯后验密度的分布。

实现贝叶斯学习的联邦协议研究较少，目前正受到广泛关注[14]。分布式 SGLD 最初于 [15] 提出，基于梯度平均，使得每轮全局聚合之间可以进行一次本地迭代。[16] 和 [17]提出的实现方案允许进行多次本地更新并对模型参数求平均。特别是， [17], 提出的最先进方法——联邦平均朗之万动力学（FALD）——已被证明具有良好的收敛性质。另一个研究方向旨在通过量化更新梯度来降低联邦 SGLD 的通信成本，

以此量化更新梯度 [18]。

上述关于分布式随机梯度朗之万动力学算法的研究均假设通信无噪声。文献 [19]–[21] 提出了信道驱动采样的概念，即将信道噪声转化为空中计算实现贝叶斯学习的资源，并通过蒙特卡洛采样完成学习。具体而言，文献 [20] 实现了一种基础版的去中心化朗之万动力学，允许在全局聚合轮次之间进行一次本地更新，并采用全批量梯度更新。文献 [22]还研究了该协议的完全去中心化版本。

在此背景下，本文的主要贡献如下。

- 无线联邦平均朗之万动力学（WFALD）：我们提出了无线联邦平均朗之万动力学（WFALD），这是一种新颖的贝叶斯学习算法，将空中计算与信道驱动采样相结合，同时支持：(i) 在通信轮次之间进行多次本地更新；以及 (ii) 使用随机梯度。WFALD 只需对联邦平均的标准空中计算实现进行少量修改。

- 收敛分析：我们从 2-瓦瑟斯坦距离的角度分析 WFALD 的收敛性。分析界表明，在低信噪比（SNR）场景下，提高全局聚合频率未必总是有益，因为过多的信道噪声无法被完全重新用于 MC 采样。

- 实验：我们展示了 WFLAD 在不同全局聚合频率和信噪比设置下的性能，以验证分析结果。

## II. 系统模型

如图 1 所示，我们考虑一个由单天线边缘服务器和 $K$ 个边缘设备组成的分布式学习系统。每个设备 $k$ 都有自己的本地数据集 $\mathcal{D}_k$，其中包含 $N_k$ 个数据样本 $\mathcal{D}_k = \{\mathbf{d}_{k,\ n}\}_{n=1}^{N_k}$。相应地，我们将全局数据集记为 $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k = \{\mathbf{d}_n\}_{n=1}^{N}$，其中包含$N = \sum_{k=1}^{K} N_k$。正如本节将详细介绍的那样，单天线设备通过采用无编码调制的非正交多址接入（NOMA）信道与服务器通信，如 [20]所示。该系统的目标是通过联邦平均范式执行基于梯度的蒙特卡洛（MC）采样，该范式利用信道驱动采样 [19], [20] 和空中计算 [23]。如第 I 节所述，不同于 [20], 仅关注全局聚合步骤之间的单次梯度下降（GD）更新，我们允许进行多次本地随机梯度下降（SGD）更新。

### A. 学习问题

我们考虑由似然函数 $p(\mathbf{d}|\boldsymbol{\theta})$和先验分布 $p(\boldsymbol{\theta})$定义的学习模型。因此，设备 $k$ 的似然函数如下给出

$$p(\mathcal{D}_k|\boldsymbol{\theta}) = \prod_{n=1}^{N_k} p(\mathbf{d}_{n,k}|\boldsymbol{\theta}). \tag{1}$$

学习系统的目标是估计所关注量的平均值，例如集成预测分布（参见例如 [4, 第 12]章），相对于全局后验分布

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\boldsymbol{\theta}) \prod_{k=1}^{K} p(\mathcal{D}_k|\boldsymbol{\theta}). \tag{2}$$

全局后验分布可以分解为 $K$ 个局部子后验分布的乘积，如 $p(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{k=1}^{K} \tilde{p}(\boldsymbol{\theta}|\mathcal{D}_k)$，其中每个子后验分布为

$$\tilde{p}(\boldsymbol{\theta}|\mathcal{D}_k) \propto p(\boldsymbol{\theta})^{1/K} p(\mathcal{D}_k|\boldsymbol{\theta}). \tag{3}$$

全局后验分布 (2) 与传统的局部正则化训练损失相关

$$f_k(\boldsymbol{\theta}) = -\log p(\mathcal{D}_k|\boldsymbol{\theta}) - \frac{1}{K}\log p(\boldsymbol{\theta}) \tag{4}$$

对于所有设备 $k$。事实上，我们有比例关系 $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-f(\boldsymbol{\theta}))$，其中全局代价函数

$$f(\boldsymbol{\theta}) = \sum_{k=1}^{K} f_k(\boldsymbol{\theta}) \tag{5}$$

对应于频率学派联邦学习所研究的标准全局训练损失。

### B. 随机梯度朗之万动力学

随机梯度朗之万动力学（SGLD）是一种基于随机梯度的 MCMC 采样方案。与所有 MCMC 技术一样，其目标是生成若干粒子 $\boldsymbol{\theta}^{[s]}$with $s = 1, ..., S$，使它们近似服从所关注的后验分布，即全局后验分布（2）。随后，可利用样本 $\{\boldsymbol{\theta}^{[s]}\}_{s=1}^{S}$ 来估计关于某个函数 $g(\boldsymbol{\theta})$ 的后验 $p(\boldsymbol{\theta}|\mathcal{D})$ 平均值，即 $1/S\sum_{s=1}^{S} g(\boldsymbol{\theta}^{[s]})$。一个值得注意的例子是，当 $g(\boldsymbol{\theta}) = p(\boldsymbol{d}|\boldsymbol{\theta})$ 时，该平均值可得出关于测试数据 $\boldsymbol{d}$ 的集成贝叶斯预测器。

从任意粒子 $\boldsymbol{\theta}^{[0]}$ 开始，SGLD 使用来自全局数据集、大小为 $p_b N$ 的小批量 $\mathcal{C}^{[s]} \subseteq \mathcal{D}$，针对某个 $p_b \in (0, 1]$ as 计算梯度的随机估计

$$\hat{\nabla} f(\boldsymbol{\theta}^{[s]}) = -\frac{1}{p_b} \sum_{\mathbf{d}_n \in \mathcal{C}^{[s]}} \nabla \log p(\mathbf{d}_n|\ \boldsymbol{\theta}^{[s]}) - \frac{1}{K} \nabla \log p(\boldsymbol{\theta}^{[s]}). \tag{6}$$

然后，它应用带噪声扰动的随机梯度下降更新

$$\text{(SGLD)} \quad \boldsymbol{\theta}^{[s+1]} = \boldsymbol{\theta}^{[s]} - \eta \hat{\nabla} f(\boldsymbol{\theta}^{[s]}) + \sqrt{2\eta}\boldsymbol{\xi}^{[s]}, \tag{7}$$

其中 $\boldsymbol{\xi}^{[s]}$对 $s = 1$、2、.....，这是标准高斯变量 $\mathcal{N}(0, \mathrm{I})$ 构成的独立同分布序列。SGLD更新 (7) 需要访问整个数据集 $\mathcal{D}$，因此无法直接在去中心化环境中实现。

### C. 联邦平均朗之万动力学（FALD）

文献 [16] 近期提出了 SGLD 的一种联邦版本，称为 FALD；该方法在全局联邦平均式聚合步骤之间融入了多次本地更新。设备按照公共时间步长 $s = 1$、2、.....运行，并假设它们拥有公共随机性的来源，即一个公共种子，用于生成共享随机变量。

在每次迭代 $s$，设备 $k$ 随机抽取 $p_b N_k$ 个数据样本，其中 $p_b \in (0,1]$, 作为一个批次 $\mathcal{C}_k^{[s]} \subseteq \mathcal{D}_k$，取自其本地数据集），并按照式（6）计算随机梯度。然后，它执行类似 SGLD 的更新

$$\tilde{\boldsymbol{\theta}}_k^{[s+1]} = \boldsymbol{\theta}_k^{[s]} - \eta \hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]}) + \sqrt{2\eta}\boldsymbol{\xi}_k^{[s]}, \tag{8}$$

其中，i.i.d. 加性噪声 $\boldsymbol{\xi}_k^{[s]}$ with $s = 1$、2、... 通常在设备之间 $k$ 具有相关性。具体而言，加性噪声项 $\{\boldsymbol{\xi}_k^{[s]}\}_{k=1}^K$ 之间的相关性由参数 $\tau^{[s]} \in [0,1]$ 控制，具体为

$$\boldsymbol{\xi}_k^{[s]} = \sqrt{\frac{\tau^{[s]}}{K}}\tilde{\boldsymbol{\xi}}_c^{[s]} + \sqrt{1-\tau^{[s]}}\tilde{\boldsymbol{\xi}}_k^{[s]}, \tag{9}$$

其中，噪声项 $\tilde{\boldsymbol{\xi}}_c^{[s]} \sim \mathcal{N}(0, \mathrm{I})$ 在所有设备之间共享，而 $\tilde{\boldsymbol{\xi}}_k^{[s]} \sim \mathcal{N}(0, \mathrm{I})$ 在不同设备 $k$ 之间相互独立。

在每次本地更新之后的每次迭代 $s$ 中，所有设备利用公共随机性共同决定是否与服务器进行通信。因此，所有设备均以概率 $p_c$ 抽取一个公共伯努利随机变量 $B^{[s]}$。当 $B^{[s]} = 1$ 时，每个设备 $k$ 将其在式 (8) 中当前更新后的粒子 $\tilde{\boldsymbol{\theta}}_k^{[s]}$ 发送给服务器；而当 $B^{[s]} = 0$ 时，不进行任何通信，所有设备继续进入下一次迭代 $s+1$。服务器接收到所有设备发送的粒子后，通过应用标准的联邦平均更新获得一个全局样本。

$$\text{(Global Aggregation)} \quad \boldsymbol{\theta}^{[s+1]} = \frac{1}{K}\sum_{k=1}^K \tilde{\boldsymbol{\theta}}_k^{[s+1]} \tag{10}$$

并将其广播给所有设备，用作下一次迭代的初始化值。

具体而言，将用作下一次迭代初始化值的本地粒子 $\boldsymbol{\theta}_k^{[s+1]}$ 设置为

$$\text{(Update)}\ \boldsymbol{\theta}_k^{[s+1]} = (1-B^{[s]})\tilde{\boldsymbol{\theta}}_k^{[s+1]} + \frac{B^{[s]}}{K}\sum_{k=1}^K \tilde{\boldsymbol{\theta}}_k^{[s+1]}. \tag{11}$$

注意，根据（9），在完成全局聚合（10）后，系统以有效学习率$\eta/K$执行SGLD更新（7）。

### D. 通信模型

在上行链路中，设备通过共享的非正交多址接入（NOMA）信道与边缘服务器进行通信。每台设备均采用无编码模拟传输，通过空中计算 [9]–[11], 实现高效的粒子聚合步骤（10），同时利用信道噪声作为（9）中的相关噪声项。正如我们将在下一节中详细说明的，我们强调，后一个思想是本文的一项关键创新贡献。与现有技术 [10], [11], 一样，我们假设所有节点均具有完美的信道状态信息（CSI），从而能够针对空中计算进行功率控制。

我们假设信道系数在 $d$ 次信道使用期间保持不变的块平坦衰落信道。该假设使得可以在一个数据块内传输 $d$ 维样本向量 $\{\boldsymbol{\theta}_k^{[s]}\}_{k=1}^K$。所有设备之间的符号级同步可通过长期演进和 5G新空口中的定时提前过程等标准协议实现 [24]。

在第 $s$ 个通信轮次中，服务器接收到的信号为

$$\mathbf{y}^{[s]} = \sum_{k=1}^K h_k^{[s]}\mathbf{x}_k^{[s]} + \mathbf{z}^{[s]} \tag{12}$$

其中，$h_k^{[s]}$ 是第 $s$ 轮中的 [channel gain] for device $k$，$\mathbf{x}_k^{[s]} \in \mathbb{R}^d$ 是包含局部粒子信息的发射信号，$\mathbf{z}^{[s]}$ 是信道噪声，服从独立同分布的 $\mathcal{N}(0, N_0\mathrm{I})$。每个设备都有发射功率约束。

$$\|\mathbf{x}_k^{[s]}\|^2 \le P, \text{ for all } k \tag{13}$$

对每个通信块进行计量。我们将信号与噪声比定义为 $\text{SNR} = P/(dN_0)$。与大多数其他相关论文一样，我们假设下行通信无噪声。当边缘服务器通过功率和带宽约束不如设备严格的基站进行通信时，这一假设在实践中是合理的。

### E. 关于局部代价函数的假设

最后，我们列出若干关于式 (4) 中局部代价函数 $f_k(\boldsymbol{\theta})$ 及其梯度的标准假设（参见例如 [20]）。

假设 1（光滑性）。局部代价函数 $f_k(\boldsymbol{\theta})$、$k = 1$、...、$K$ 具有常数 $L > 0$ 的光滑性，即它是连续可微的，且梯度 $\nabla f_k(\boldsymbol{\theta})$ 是具有常数 $L$ 的利普希茨连续，即，

$$\|\nabla f_k(\boldsymbol{\theta}) - \nabla f_k(\boldsymbol{\theta}')\| \le L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d.$$

假设 2（强凸性）。对于任意局部代价函数 $f_k(\boldsymbol{\theta})$，$k = 1, \cdots, K$，存在常数 $\mu > 0$，使得下列不等式成立

$$[\nabla f_k(\boldsymbol{\theta}) - \nabla f_k(\boldsymbol{\theta}')]^{\mathsf{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}') \ge \mu\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2, \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d. \tag{14}$$

假设 3（无偏且方差有界的随机梯度）。随机梯度 $\nabla f_k(\boldsymbol{\theta})$ 是局部梯度 $\nabla f_k(\boldsymbol{\theta})$的无偏估计，且其方差有界，满足 $\sigma_k^2$ as

$$\mathbb{E}\left[\left\|\hat{\nabla} f_k(\boldsymbol{\theta}) - \nabla f_k(\boldsymbol{\theta})\right\|^2\right] \le \sigma_k^2, \forall \boldsymbol{\theta} \in \mathbb{R}^d, k = 1, \cdots, K. \tag{15}$$

假设 4（有界局部梯度）。局部梯度有界，满足

$$\|\nabla f_k(\boldsymbol{\theta})\| \le G, \forall \boldsymbol{\theta} \in \mathbb{R}^d,\ k = 1, \cdots, K. \tag{16}$$

## III. 无线联邦平均朗之万动力学

本节介绍所提出的 WFALD 协议，该协议将上一节介绍的 FALD [16], 与信道驱动采样 [19], [20] 和空中计算 [23]相结合。我们的目标是定义一种高效的分布式贝叶斯学习协议，只需对基于空中计算的传统频率学派联邦学习做少量改动。因此，与 [20], 不同，我们允许根据第 II-C 节所述的随机调度机制进行多次本地更新。这项工作的核心思想是在全局聚合步骤中利用信道噪声，作为 FALD 更新式 (9) 中的公共噪声项。

A. WFALD概述

正如 [17, 定理 1], 所证明的，在式 (9) 中加入公共噪声项，对于缩小式 (10) 中聚合粒子的分布与目标全局后验分布 (2) 之间的差异至关重要。在[16], 提出的实现方案中，这要求节点共享足够的公共随机性，以便在每次迭代 $s$ 时生成独立同分布的高斯随机变量，如式 (9) 所示。WFALD 基于以下观察。

在任意一次迭代 $s$ 中进行全局聚合时，即 $B^{[s]}$ =1时，每个设备 $k$ 都会传输更新后的本地粒子

$$\mathbf{x}_k^{[s]} = \alpha_k^{[s]}\left[\boldsymbol{\theta}_k^{[s]} - \eta\hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]})\right] \tag{17}$$

使用无码传输，通过NOMA信道（12）并采用某个功率控制参数 $\alpha_k^{[s]}$ 。如下文小节所详述，功率控制的设计确保接收信号（12）可以乘以因子 $c^{[s]}$ 进行缩放，从而近似表示式（10）中的更新后粒子 $\boldsymbol{\theta}^{[s+1]}$ ，其中 $\tau^{[s]}=1$, i.e.,

$$\frac{\mathbf{y}^{[s]}}{c^{[s]}} \approx \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k^{[s]} - \frac{\eta}{K}\sum_{k=1}^{K}\hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]}) + \sqrt{\frac{2\eta}{K}}\tilde{\boldsymbol{\xi}}_c. \tag{18}$$

对于不进行全局聚合的迭代 $s$ ，即$B^{[s]} = 0$, WFALD规定每个设备 $k$ 使用 $\tau^{[s]} = 0$执行SGLD更新（8），也就是说，仅添加独立局部噪声项 $\tilde{\boldsymbol{\xi}}_k^{[s]}$ 。这样选择是为了简化实现；如果设备拥有足够的公共随机性来生成共享高斯噪声，则可以省去这一步。

WFALD的上述设计有双重优势：既减少了对共享全局随机性的需求，更重要的是，还减轻了信道噪声的影响，因为信道噪声被重新用作学习的有用资源[19], [20]。任何额外噪声都被视为干扰，因此需要修改[17], 中给出的收敛分析，接下来我们将对此进行讨论。

B. 信号设计与功率控制

如上一小节所述，对于满足$B^{[s]}=1$的迭代，所有设备均按照 (17) 传输其更新后的局部粒子。与 [10], 中一样，我们采用信道反演，并选择功率控制参数 $\alpha_k^{[s]} = \alpha^{[s]}/h_k^{[s]}$ 以实现信号对齐，其中 $\alpha^{[s]} > 0$ 是待设计的公共增益参数。为了估计全局更新 (10)，随后按照 (18) 对接收信号 (12) 进行缩放，得到

$$c^{[s]} = K\alpha^{[s]} \tag{19}$$

为解释设计 (19) 背后的原理，将 (12)、(17) 和 (19) 代入 (18)，得到

(无线全局聚合)

$$\boldsymbol{\theta}^{[s+1]} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k^{[s]} - \frac{\eta}{K}\sum_{k=1}^{K}\hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]}) + \frac{\mathbf{z}^{[s]}}{\alpha^{[s]}K} \tag{20}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k^{[s]} - \frac{\eta}{K}\sum_{k=1}^{K}\hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]}) + \sqrt{\frac{2\eta}{K}}\tilde{\boldsymbol{\xi}}_c^{[s]} + \sqrt{\beta^{[s]}}\boldsymbol{\Delta}^{[s]} \tag{21}$$

其中， $\boldsymbol{\Delta}^{[s]} \sim \mathcal{N}(0, \mathrm{I}_d)$ 在迭代索引 $s$ 上独立同分布，且我们定义了

$$\beta^{[s]} = \max\left\{0, \frac{N_0}{(\alpha^{[s]}K)^2} - \frac{2\eta}{K}\right\}. \tag{22}$$

比较 (21) 和 (18) 可知，估计值 (19) 与全局更新 (18) 相对应，但多出了噪声项$\sqrt{\beta^{[s]}}\boldsymbol{\Delta}^{[s]}$。

通过设置$\alpha^{[s]} = \sqrt{N_0}/(2\eta K)$，可以使附加噪声项为零。然而，由于功率约束 (13)，这一选择并非总是可行，因此需满足条件 $\alpha^{[s]} \le \min_k \sqrt{P}|h_k^{[s]}|/\|\boldsymbol{\theta}_k^{[s]} - \eta\hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]})\|$。考虑到这些因素，WFALD 将功率控制增益选为

$$\alpha^{[s]} = \min\left\{\sqrt{\frac{N_0}{2\eta K}}, \min_k \frac{\sqrt{P}|h_k^{[s]}|}{\|\boldsymbol{\theta}_k^{[s]} - \eta\hat{\nabla} f_k(\boldsymbol{\theta}_k^{[s]})\|}\right\}. \tag{23}$$

C. 收敛分析

为分析 WFALD 经过 $s$ 次迭代后产生的样本 $\{\boldsymbol{\theta}_k^{[s]}\}_{k=1}^K$ 的质量，与 [17], 中类似，我们研究平均局部粒子的分布 $p(\tilde{\boldsymbol{\theta}}^{[s]})$

$$\tilde{\boldsymbol{\theta}}^{[s]} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k^{[s]}. \tag{24}$$

具体而言，我们分析分布 $p(\tilde{\boldsymbol{\theta}}^{[s]})$ 与式（2）中的目标全局后验分布 $p(\boldsymbol{\theta}|\mathcal{D})$ 之间的标准 2-瓦瑟斯坦距离$W_2(p(\tilde{\boldsymbol{\theta}}^{[s]})\ p(\boldsymbol{\theta}|\mathcal{D}))$，并在下述定理中给出其上界。

定理1. 对于学习率 $0 < \eta \le 2/L$，在假设1–4成立的条件下，经过 $s$ 次迭代，WFALD 产生的平均本地样本的分布与全局后验分布之间的 2-瓦瑟斯坦距离有如下上界：

$$\mathbf{W}_2^2(p(\tilde{\boldsymbol{\theta}}^{[s]}), p(\boldsymbol{\theta}|\mathcal{D})) \le \left(\frac{1+\gamma}{2}\right)^{2s}\mathbf{W}_2^2(p(\tilde{\boldsymbol{\theta}}^{[0]}), p(\boldsymbol{\theta}|\mathcal{D})) + \sum_{j=0}^{s-1}\left(\frac{1+\gamma}{2}\right)^{2(s-j)} p_c\beta^{[s]}d + \frac{8(1+\gamma)}{3(1-\gamma)^2}\left[\frac{\eta^4L^3d}{3K} + \eta^3L^2d + \left(\frac{\eta^2}{K} + \frac{4\eta^4L^2}{Kp_c}\right)\sum_{k=1}^{K}\sigma_k^2 + \frac{6\eta^4L^2G^2}{p_c^2} + \frac{4\eta^3L^2(K-1)d}{Kp_c}\right], \tag{25}$$

其中，对于 $0 < \eta \le 2/(\mu+L)$， $\gamma = 1-\eta\mu$ ；对于 $2/(\mu+L)$, $\gamma = \eta L - 1 \le \eta \le 2/L$； $\sigma_k^2$ 是随机局部梯度的有界方差。

如附录所述，通过修改 [20, 命题 1] 的证明，将多次随机本地更新所导致的散度（亦称客户端漂移） [25]纳入考量，即可得到这一结果。

界（25）主要揭示了全局聚合率 $p_c$ 的作用。具体而言，在高信噪比区间，当 $\beta^{[s]}=0$ 时，设置 $p_c=1$ 可降低收敛界，这表明通信总是有益的。相反，当信噪比低于某个阈值时，增大 $p_c$ 可能会使界增大，这表明过于频繁的通信可能不利于

收敛。这是由于更新（21）中存在功率为 $\beta^{[s]}$ 的非零残余信道噪声。因此，在低信噪比区间，存在一个最优通信频率 $p_c$，能够在客户端漂移误差与残余信道噪声之间实现最佳权衡。

最后需要指出，[20, 命题 1] 中的结果针对的是一种更简单的协议，该协议在每一步都进行全局聚合，即 $p_c = 1$，并使用全批次梯度，即 $p_b = 1$，这意味着 $\sigma_k = 0$。因此，界（25）并不能简化为 [20, 命题 1], 中的界，因为如第三节 A 部分所述，这两种方案采用了不同的传输策略。

## IV. 数值结果

为了定量了解 WFALD 的性能，如 [20], 所示，我们考虑一个标准的高斯线性回归问题，其似然函数为

$$p(v_n \mid \boldsymbol{\theta}, \mathbf{u}_n) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(v_n - \boldsymbol{\theta}^\top \mathbf{u}_n\right)^2}, \tag{26}$$

先验 $p(\boldsymbol{\theta})$假设服从高斯分布$\mathcal{N}(0, \mathrm{I})$。因此，后验 $p(\boldsymbol{\theta}|\mathcal{D})$服从高斯分布 $\mathcal{N}((\mathrm{UU}^\mathsf{T} + \mathrm{I})^{-1}\mathrm{Uv}$ ， $= [\ _1 \cdots \ _N]$ $= [v_1, \cdots, v_N]^\mathsf{T}$ $(\mathrm{UU}^\mathsf{T} + \mathrm{I})^{-1})$，其中 U u , , u 是数据矩阵，v 是目标变量向量。我们使用合成数据集$\{\mathrm{d}_n = (\mathrm{u}_n, v_n)\}_{n=1}^N$，其中 $N = 1200$ 遵循模型（26），协变量 $\mathrm{u}_n \in \mathbb{R}^d$ 独立同分布地取自高斯分布 $\mathcal{N}(0, \mathrm{I})$，其中 $d = 5$。通过似然函数（26）生成目标变量 $v$ 的真实模型参数为 $\boldsymbol{\theta}^\star = [-0.0615, -1.6057, 1.7629, 1.0240, -1.5902]^\mathsf{T}$。除非另有说明，数据集均匀分配给 $K = 30$ 个设备；所有通信轮次和设备的信道$h_k^{[s]}$ 均设为 1；学习率为 $\eta = 3 \times 10^{-3}$；且 $p_b$ 为 0.4。

我们运行 $S = 200$ 次迭代。为了评估粒子上的感兴趣均值，我们假设预热期为$S_b = 100$，因此仅保留剩余的 $S_u = S - S_b$ =100个样本用于评估。作为测试函数，我们考虑两种选择。首先，考虑恒等函数 $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$，因此将预热期后各设备 $k$ 上由 WFALD 生成的粒子均值 $1/S_u \sum_{s=S_b+1}^{S_b+S_u} \boldsymbol{\theta}_k^{[s]}$ 与全局后验分布 (2) 的真实均值 $\boldsymbol{\mu}_p = (\mathrm{UU}^\mathsf{T} + \mathrm{I})^{-1}\mathrm{Uv}$进行比较。为此，我们测量所有设备上经验均值与真实均值之间的均方误差:

$$\mathrm{MSE} = \frac{1}{K} \sum_{k=1}^{K} \left\| \frac{1}{S_u} \sum_{s=S_b+1}^{S_b+S_u} \boldsymbol{\theta}_k^{[s]} - \boldsymbol{\mu}_p \right\|^2. \tag{27}$$

接下来，为了能够与传统的频率学派无线联邦平均（WFedAvg）进行比较，我们考虑测试函数 $g(\boldsymbol{\theta}) = p(v|\boldsymbol{\theta}, \mathrm{u})$，并评估在每台设备使用包含 500 个样本的测试集时的平均二次误差。对于贝叶斯学习，使用集成平均预测器 $\frac{1}{S_u} \sum_{s=S_b+1}^{S_b+S_u} p(v|\boldsymbol{\theta}_k^{[s]}, \mathrm{u})$ 得出的期望；对于 WFedAvg，则使用通过最后一次迭代得到的预测器 $p(v|\boldsymbol{\theta}_k^{[S]}, \mathrm{u})$。WFedAvg 使用与 WFALD 相同的小批量大小和全局聚合轮次频率来实现，模型传输的功率控制

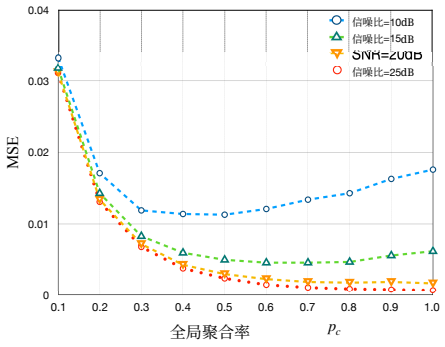


图 2. 不同信噪比下，MSE(27) 随全局聚合率 $p_c$ 的变化。

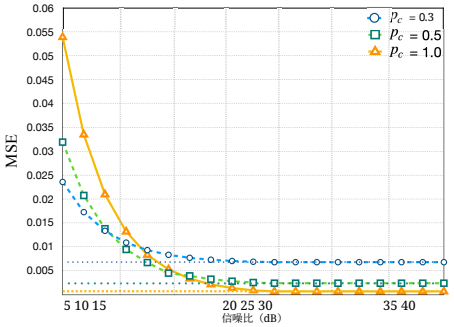


图 3. 不同全局聚合率下，均方误差 (27) 随信噪比变化的曲线。

基于缩放信道反演，如 [9] 中所述。所有结果均取 100 次实验的平均值。

我们首先通过图 2 展示均方误差 (27) 随全局聚合率 $p_c$ 变化的曲线，考察全局聚合率 $p_c$ 对 WFALD 性能的影响；其中信噪比取不同值，定义为 SNR $= P/(dN_0)$。与第 III 节的分析一致，当信噪比较大时，提高全局聚合步骤的频率 $p_c$ 并无不利影响；在 [17]研究的无噪声情形下也是如此。相反，正如上一节的分析所预期的那样，低信噪比时信道引入的额外噪声会使 $p_c$ 的最优值通常小于 1。

我们在图 3 中研究信噪比对 WFALD 性能的影响。对于所考虑的全局聚合率，信噪比降低至约 25 dB 之前不会损害性能；继续降低则会因 WFALD 更新中的额外噪声而导致性能退化。此外，与上一节的分析一致，我们观察到，在低信噪比区域，应谨慎选择聚合率 $p_c$，以获得最佳性能。

最后，图 4 绘出了使用 WFALD 提供的集成预测器和 WFedAvg 生成的传统预测器 [9] 对变量 $v$ 的预测均方误差。与所考虑的信噪比范围内的 WFedAvg 相比，WFLAD 表现更优；由于能够通过信道驱动采样利用信道噪声，因此在低信噪比范围内取得了显著提升。

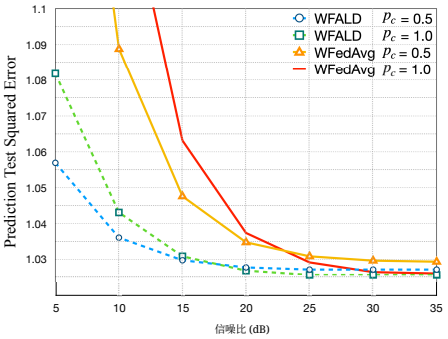


图4. WFALD 生成的集成预测器与传统 WFedAvg [9] 的预测测试平方误差（对所有用户取平均）。

## 五、结论

本文提出了一种新颖的协议，用于在无线系统中高效实现联邦平均朗之万动力学（FALD）。作为未来的研究方向，值得探讨数据异质性所带来的影响（如 [17] 中所述），并研究规模更大的应用。

## 附录

为证明定理1，首先通过朗之万扩散过程（LDP）生成的样本与从WFALD获得的样本之间的2-范数期望，对2-瓦瑟斯坦距离进行上界估计，遵循 [26, 定理 1], 。随后，按照证明 [20, 命题1], 中的类似步骤，我们分别跟踪量化误差、随机梯度误差和客户端漂移的贡献。与 [20, 命题 1], 相比，对后者的分析需要一个新的结果，以下引理对此进行了概括，其证明与 [17, 引理 10]类似。

引理1（客户端漂移的上界）。对于任意迭代 $s$，定义为 $V_c^{[s]} _= \|\nabla f(\tilde{\boldsymbol{\theta}}^{[s]}) \sum_{k=1}^{K} \nabla f_k\ \boldsymbol{\theta}_k^{[s]}\ \|^2$ ˆ ( ) 的客户端漂移，其上界为

$$\mathbb{E}[V_c^{[s]}] \leq K^2 L^2 \mathbb{E}[V_{\boldsymbol{\theta}}^{[s]}] + K \sum_{k=1}^{K} \sigma_k^2, \tag{28}$$

其中，我们将局部模型偏差定义为 $V_{\boldsymbol{\theta}}^{[s]}$ =。后者的上界为 $\frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{\theta}_k^{[s]} - \tilde{\boldsymbol{\theta}}^{[s]}\|^2$

$$\mathbb{E}[V_{\boldsymbol{\theta}}^{[s]}] \leq \frac{2(1-p_c)}{p_c}\Big[\frac{(2+p_c)\eta^2}{p_c} G^2 + \frac{\eta^2}{K}\sum_{k=1}^{K}\sigma_k^2 + \frac{2(K-1)\eta d}{K}\Big]. \tag{29}$$

## 参考文献

[1] C. Guo、G. Pleiss、Y. Sun 和 K. Q. Weinberger，“现代神经网络的校准问题”，载于国际机器学习会议（ICML）论文集（澳大利亚悉尼），第1321–1330页，2017年8月。[2] M. Abdar、F. Pourpanah、S. Hussain、D. Rezazadegan、L. Liu、M. Ghavamzadeh、P. Fieguth、X. Cao、A. Khosravi、U. R. Acharya 等，“深度学习中的不确定性量化综述：技术、应用与挑战”，Information Fusion，第76卷，297，2021年243页。–。[3] E. Angelino、M. J. Johnson 和 R. P. Adams，“可扩展贝叶斯推断的模式”，*[在线]*。可查阅：https://arxiv.org/pdf/1602.05221.pdf，2016年。[4] O. Simeone，《面向工程师的机器学习》。剑桥大学出版社，2022年。[5] A. G. Wilson 和 P. Izmailov，“贝叶斯深度学习与泛化的概率视角”，载于神经信息处理系统大会（NeurIPS）论文集，第33卷（线上），第4697–4708页，2020年12月。[6] M. E. Khan 和 H. Rue，“贝叶斯学习规则”，*[在线]*。可查阅：https://arxiv.org/pdf/2107.04562.pdfs，2021年。[7] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，“基于去中心化数据的深度网络通信高效学习”，载于人工智能与统计学国际会议（AISTATS）论文集（美国劳德代尔堡），第1273–1282页，PMLR，2017年4月。[8] J. Kone˘cn`y、H. B. McMahan、F. X. Yu、P. Richt´arik、A. T. Suresh 和 D. Bacon，“联邦学习：提升通信效率的策略”，*[在线]*。可查阅：https://arxiv.org/pdf/1610.05492.pdf，2016年。[9] G. Zhu、Y. Wang 和 K. Huang，“面向低时延联邦边缘学习的宽带模拟聚合”，IEEE无线通信汇刊，第19卷，第491–506页，2019年10月。[10] D. Liu 和 O. Simeone，“免费隐私：通过无码传输和自适应功率控制实现无线联邦学习”，IEEE精选通信领域期刊，第39卷，第170–185页，2020年11月。[11] X. Cao、G. Zhu、J. Xu 和 S. Cui，“面向网络边缘空中联邦平均的传输功率控制”，IEEE精选通信领域期刊，第40卷，第5期，第1571–1586页，2022年。[12] T. Sery、N. Shlezinger、K. Cohen 和 Y. C. Eldar，“基于异构数据的空中联邦学习”，IEEE信号处理汇刊，第69卷，第3796–3811页，2021年6月。[13] M. Welling 和 Y. W. Teh，“通过随机梯度朗之万动力学进行贝叶斯学习”，载于国际机器学习会议（ICML）论文集（华盛顿州贝尔维尤），第681–688页，2011年6月。[14] L. Cao、H. Chen、X. Fan、J. Gama、Y.-S. Ong 和 V. Kumar，“贝叶斯联邦学习：综述”，2023年。[15] S. Ahn、B. Shahbaba 和 M. Welling，“分布式随机梯度MCMC”，载于国际机器学习会议（ICML）论文集（中国北京），第1044–1052页，2014年6月。[16] W. Deng、Q. Zhang、Y.-A. Ma、Z. Song 和 G. Lin，“关于联邦平均朗之万动力学的收敛性”，*[在线]*。可查阅：https://arxiv.org/pdf/2112.05120.pdf，2021年。[17] V. Plassier、E. Moulines 和 A. Durmus，“联邦平均朗之万动力学：迈向统一理论与新算法”，载于人工智能与统计学国际会议（AISTATS）论文集（西班牙瓦伦西亚），第5299–5356页，PMLR，2023年4月。[18] M. Vono、V. Plassier、A. Durmus、A. Dieuleveut 和 E. Moulines，“QLSD：用于贝叶斯联邦学习的量化朗之万随机动力学”，载于人工智能与统计学国际会议（AISTATS）论文集（西班牙瓦伦西亚），第6459–6500页，PMLR，2022年3月。[19] D. Liu 和 O. Simeone，“面向无线数据中心贝叶斯分布式学习的信道驱动蒙特卡罗采样”，IEEE精选通信领域期刊，第40卷，第2期，第562–577页，2021年。[20] D. Liu 和 O. Simeone，“无线联邦朗之万蒙特卡罗：利用信道噪声实现贝叶斯采样与隐私保护”，IEEE无线通信汇刊，第1–1页，2022年。[21] Y. Zhang、D. Liu 和 O. Simeone，“利用信道噪声，通过量化联邦朗之万蒙特卡罗实现采样与隐私保护”，载于IEEE无线通信信号处理进展研讨会（SPAWC）论文集（芬兰奥卢），第1–5页，IEEE，2022年7月。[22] L. Barbieri、O. Simeone 和 M. Nicoli，“面向D2D网络可信决策的信道驱动去中心化贝叶斯联邦学习”，*[在线]*。可查阅：https://arxiv.org/pdf/2210.10502.pdf，2022年。[23] G. Zhu、J. Xu 和 K. Huang，“面向大规模物联网无线数据聚合的空中计算”，*[在线]*。可查阅：https://arxiv.org/pdf/2009.02181.pdf，2020年。[24] A. Mahmood、M. I. Ashraf、M. Gidlund、J. Torsner 和 J. Sachs，“5G无线边缘的时间同步：关键MTC的要求与解决方案”，IEEE通信杂志，第57卷，第45–51页，2019年12月。[25] S. P. Karimireddy、S. Kale、M. Mohri、S. Reddi、S. Stich 和 A. T. Suresh，“Scaffold：用于联邦学习的随机控制平均”，载于国际机器学习会议（ICML）论文集（奥地利维也纳），第5132–5143页，PMLR，2020年7月。[26] A. Dalalyan，“采样与优化之间更深入、更强的类比：朗之万蒙特卡罗与梯度下降”，载于学习理论会议（COLT）论文集（荷兰阿姆斯特丹），2017年7月。