复现论文：Federated Learning over Wireless Networks: Optimization Model Design and Analysis

摘要

新机器学习技术联邦学习越来越受欢迎，在这种技术中，模型训练分布在移动用户设备 (UE) 上，每个 UE 通过基于其本地训练数据独立计算梯度来贡献学习模型。虽然现有的大部分工作都集中在设计具有可证明收敛时间的学习算法上，但其他问题，如无线信道的不确定性和具有异构功率约束和本地数据大小的 UE，尚未得到充分探讨。这些问题特别影响到各种权衡：(1) 由学习精度级别决定的计算和通信延迟之间的权衡，以及 (2) 联邦学习时间和 UE 能耗之间的权衡，即通过将无线网络上的联邦学习作为捕获两者权衡的优化问题 FEDL。实际表现为一个非凸的优化问题，因此，复现工作主要基于文章中的数学模型、算法描述以及理论分析结果进行，即包括对文章中的算法、计算模型、通信模型、问题分解和求解等。同时在复现过程中，对其进行了深入的分析和扩展。例如，探讨了不同参数设置对系统性能的影响，并提出了可能的改进方向。

关键词：无线网络上的分布式机器学习；联邦学习；优化分解

1 引言

维护消费者数据隐私的兴趣日益浓厚，这导致了一类新的机器学习技术的出现，这种技术利用了大量移动电话用户的参与。其中一种流行的技术被称为联邦学习 [1]。这种学习技术允许用户协作构建共享学习模型，同时将所有训练数据保存在他们自己的用户设备（UE）上。通过这种方式，用户数据隐私得到了很好的保护，因为本地训练数据是不共享的，因此它将机器学习与传统方法在数据中心获取、存储和训练数据分离开来。联邦学习基于数据隐私前提，并且拥有现代强大的处理器和低延迟的移动边缘网络支持，然而无线信道的不确定性和具有异构功率约束的移动设备尚未在联邦学习中得到充分探索。这些问题尤其影响到如下方面：计算和通信延迟由学习准确度水平决定；联邦学习时间和移动设备能耗之间的关系。

2 相关工作

由于大数据应用和深度学习等复杂的模型，训练机器学习模型需要分布在多台机器上，从而产生了分散机器学习的研究 [2,3]。然而，这些工作中的大多数算法都是为具有平衡和 id 数据并连接到高吞吐量网络（如数据中心）的机器设计的。

出于不同的动机，联邦学习（以及相关的设备上智能方法）最近引起了许多关注，利用移动设备之间的协作，这些移动设备可能数量众多，互联网连接缓慢和/或不稳定，并且具有非id。局部数据不平衡。然而，这些工作大多集中在设计算法以提高学习时间的收敛性，而不关心其他限制因素，如无线通信和移动 UE 的能量限制性质，这些因素会影响联邦学习的性能。

联邦学习与其他机器学习方案类似，联邦学习最关键的性能指标之一是收敛到预定义准确度水平所需的学习时间。然而，与传统的机器学习方法不同，联邦学习时间不仅包括移动设备 (UE) 的计算时间还包括通信时间 (取决于移动设备的信道增益和更新数据大小)。

因此，提出两个问题，第一个问题是：UE 是否应该在计算上花费更多时间以实现高学习精度和更少通信更新，反之亦然？另一方面，由于参赛者资源有限，如何分配 UE 的计算和传输功率等资源以最大限度地降低能耗是主要关注的问题。第二个问题是：如何在最小化联邦学习时间和 UE 能源消耗这两个相互冲突的目标之间取得平衡。

为了解决这些问题，提出了基于无线网络的联合学习问题设计和分析，以此研究 UE 的计算和通信特性如何影响其能耗、学习时间收敛和联邦学习的准确性水平，并考虑了异构 UE 在数据大小、信道增益、计算和传输功率方面的能力。

3 本文方法

3.1 本文方法概述

本文的研究方法具体可概括为：

（1）提出了无线网络上的联邦学习问题（FEDL）[4]，该问题捕获了两个权衡：(i) 使用帕累托效率模型的学习时间与 UE 能量消耗，以及（ii）通过寻找最佳学习精度参数的计算与通信学习时间，尽管 FEDL 具有非凸性质，但可以利用其特殊结构并使用变量分解方法将 FEDL 拆分并转换为三个凸子问题。

（2）证明了前两个子问题可以单独求解，然后用它们的解来得到第三个子问题的解。通过分析每个子问题的封闭解，获得了帕雷托有效控制旋钮对最优的影响的定性见解：(i) 计算和通信学习时间，(ii) UE 资源分配，以及（iii）学习精度。最后，所有子问题的组合解可以提供 FEDL 的全局最优解。

（3）进一步提供了广泛的数值结果来检验：(i) UE 异质性的影响，(ii) UE 能量成本与系统学习时间之间的帕累托曲线 [5]，以及（iii）计算时间与通信时间的比例对最佳精度水平的影响。

3.2 系统模型

3.2.1 FEDL 模型

考虑由一个基站 (BS) 和 N 个 UE 的集合组成的无线多用户系统，其中每个参与方 UE 存储本地数据集 D_n ，然后我们可以通过以下方式来定义总数据大小 $D = D_1 + ... + D_n$。在监督学习设置的示例中，D_n 被定义作为一组输入输出对 (x_i, y_i)，其中 x_i 是具有 d 个特征的输入样本向量，y_i 则是对应标签。

数据可以通过使用 UE 来生成 (例如通过与移动应用程序的交互)，基于 UE 数据机器学习应用程序可以用于无线网络。在典型的学习问题中，对于输入样本数据 (x_i, y_i) 其任务是基

于损失函数 $f_i(W)$ 并表征输出 y_i 的模型参数 w。关于 n 个 UE 数据集上损失函数被定义为:

$$J_n(w) := \frac{1}{D_n} \sum_{i \in D_n} f_i(w)$$

然后，学习模型是最小化全局损失函数问题:

$$\min_{w \in R^d} J(w) := \sum_{i \in D_n}^{N} f_i(w)$$

基于无线网络的联邦学习：在 UE 参与方，更新分为两个阶段 (计算和通信)：首先，每个 UE 在其本地计算问题:

$$w_n^{(t)} = \arg\min_{w_n \in R^d} F_n(w_n | w^{(t-1)}, \nabla J^{(t-1)})$$

接着，所有 UE 共享无线环境并向 BS 发送 $W_n^{(t)}$ 和梯度 $\nabla J_n^{(t)}$。在 BS 处则汇总了以下所有来自 UE 信息，并反馈给所有参与 UE:

$$w^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} w_n^{(t)}$$
$$\nabla J^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \nabla J_n^{(t)}$$

进一步而言，FEDL 通过多次全局迭代以达到全局精度水平 ϵ，并基于 UE 和 BS 交互完成每次全局迭代。参与方 UE 在每个计算阶段将使用本地训练数据 D_n 最小化本地损失函数 $F_n(W_n)$，通过多次本地迭代达到精度阈值 θ，然后使用无线媒体共享方案 (例如分时 TDMA) 将更新发送到 BS。随后，BS 分别聚合接收的局部模型参数和梯度，在下一次全局迭代中更新并然后广播给所有参与方 UE 以最小化全局损失函数。

FEDL 同时受全局精度 ϵ 和局部精度 θ 的影响，当 ϵ 和 θ 较小 (更准确) 时，FedL 需要运行更多的全局迭代次数。此外，每次全局迭代都包含计算时间和上行通信时间，其中计算时间取决于局部迭代的次数，经过证明可知其收敛时间上界为 $O(log(\frac{1}{\theta}))$，如果用 T_{cmp} 表示一次局部迭代的时间，那么对于某个条件数 v [6] 其在一个全局迭代中的计算时间是 $vlog(\frac{1}{\theta})T_{cmp}$，接着我们定义 T_{com} 表示一次全局迭代的通信时间，因此 FEDL 一次全局迭代的总时间定义为:

$$T_{glob}(T_{cmp}, T_{com}, \theta) := T_{com} + vlog(\frac{1}{\theta})T_{cmp}$$

在本文中，我们考虑一个固定的全局精度　所需要的时间开销，因此在这里将 $O(log(\frac{1}{\theta}))$ 归一化为 1 以便于表示 $K(\theta) = \frac{1}{1-\theta}$，此外我们还将 v 归一化为 1 从而可以将 v 吸收到 T_{cmp} 中作为一次局部计算迭代的上界。综上来看，可以得出 FEDL 学习时间上界为 $K(\theta)T_{glob}(\theta)$ 。

3.2.2 计算模型

我们用 c_n 表示第 n 个 UE 执行一个数据样本所需的 CPU 周期数，这可以在离线下测量 [7]，称为先验。由于所有的样本 x_i $y_i, i \in D_n$ 具有相同的大小（即位数），因此第 n 个 UE 运行一次局部迭代所需的 CPU 周期数为 $c_n D_n$。用 f_n 表示第 n 个 UE 的 cpu 周期频率。则一次局部迭代计算的 CPU 能耗可以表示为 [8]:

$$E_n^{cmp}(f_n) = \sum_{i=1}^{c_n D_n} \frac{n}{2} f_n^2 = \frac{n}{2} c_n D_n f_n^2$$

其中 $\frac{n}{2}$ 为第 n 个 UE 的计算芯片组的有效电容系数。此外，第 n 个 UE 的每次局部迭代的计算时间为 $\frac{c_n D_n}{f_n}, \forall n$。我们用 $f \in R_n$ 表示 f_n 的向量。

3.2.3 通信模型

在 FEDL 中，针对终端的通信阶段，我们考虑了终端的分时多址协议。我们注意到，这种分时模式没有限制，因为其他方案，如 OFDMA，也可以应用于 FEDL。第 n 个 UE 的可实现传输速率（nats/s）定义如下:

$$r_n = Bln(1 + \frac{h_n p_n}{N_0})$$

其中 B 是带宽，N_0 是背景噪声，P_n 是发射功率，H_n 是信道增益，进一步假设 H_n 在 FEDL 学习时间内是常数，那么第 n 个 UE 所分配的通信时间的比例则记为 T_n ，UE 的参数 W_n 与梯度 ∇J_n 大小记为 S_n ，并假设它们的大小在整个 FEDL 学习过程中是恒定的，则每个 UE 的传输速率为：

$$r_n = \frac{s_n}{\tau_n}$$

这被证明是最节能的传输策略 [9]，为了在持续时间 T_n 内发送 S_n ，则 UE 能量消耗表达式如下：

$$E_n^{com}(\tau_n) = \tau_n p_n = \tau_n p_n(\frac{s_n}{\tau_n})$$

3.2.4 问题制定

用 E_{glob} 定义每次全局迭代所有 UE 的总能耗，表示为：

$$E_{glob}(f, \tau, \theta) := \sum_{n=1}^{N} E_n^{com}(\tau_n) + log(\frac{1}{\theta} E_n^{cmp}(f_n))$$

然后，我们考虑一个优化问题 FEDL，如下所示：

$$minimize_{f,\tau,\theta,T_{com},T_{cmp}} K(\theta)[E_{glob}(f, \tau, \theta) + kT_{glob}(T_{cmp}, T_{com}, \theta)]$$

subject to :

$$\sum_{n=1}^{N} \tau_n \leq T_{com}$$
$$\max_n \frac{c_n D_n}{f_n} = T_{cmp}$$
$$f_n^{min} \leq f_n \leq f_n^{max}, \forall n \in N$$
$$p_n^{min} \leq p_n(\frac{s_n}{\tau_n}) \leq p_n^{max}, \forall n \in N$$
$$0 \leq \theta \leq 1$$

最小化 UE 的能耗和联邦学习时间是相互冲突的。例如，UE 可以通过始终设置最低频率级别来节省能量，但这肯定会增加学习时间。因此，为了在能量成本和学习时间之间取得平衡，在目标中使用的权重 k（焦耳/秒）作为 FEDL 愿意为减少一个单位的学习时间而承担的额外能量成本，它捕获了 UE 的能量成本和联邦学习时间之间的帕累托最优权衡。例如，当大多数 UE 都插入时，则 UE 的能源成本不是主要关注的问题，因此 k 可以很大。

3.3 解决方案

FEDL 方案：FEDL 全局最优解可以拆分为若干子问题的组合解，这个定理的证明很简单，其思想是利用 KKT 条件来寻找 FEDL 的驻点。然后，可以将 KKT 条件方程分解成若干组，每一组都与子问题的 KKT 条件完全匹配，可以分别求解其子问题的唯一闭合解。

因此，这个唯一的驻点也是 FEDL 的全局最优解。通常，每个 UE 往往有两个独立的处理器：一个用于移动应用的 CPU，另一个用于无线电控制功能的基带 CPU。通过它们可以揭示通信成本比计算成本高多少，从而确定最佳局部精度水平。

将 FEDL 问题分解成三个子问题，分别解决问题 1 和问题 2，将问题 1 和问题 2 的解用来求解问题 3。

SUB1：

$$minimize_{f,T_{cmp}} \sum_{n=1}^{N} E_n^{cmp}(f_n) + kT_{cmp}$$

subject to:

$$\frac{c_n D_n}{f_n} \leq T_{cmp}, \forall n \in N$$

$$f_n^{min} \leq f_n \leq f_n^{max}, \forall n \in N$$

SUB2:

$$min_{\tau, T_{com}} \sum_{n=1}^{N} E_n^{com}(\tau_n) + kT_{com}$$

subject to:

$$\sum_{n=1}^{N} \tau_n \leq T_{com}$$
$$p_n^{min} \leq p_n(\frac{s_n}{\tau_n}) \leq p_n^{max}, \forall n$$

SUB3:

$$minimize_n K(\theta)[E_{glob}(f^*, \tau^*, \theta) + kT_{glob}(T_{cmp}^*, T_{com}^*, \theta)]$$

subject to:

$$0 \leq \theta \leq 1$$

关于 FEDL 伪代码如下所示 (即如何将 UE 分为三组子问题：N_1 由一组总是运行其最大频率的瓶颈 UE 构成、N_2 由以最小频率也能在计算截止日期之前完成任务的强 UE 组构成、N_3 则是在其可行集的内部具有最佳频率的 UE 组)。

Algorithm 1 Finding $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ in Lemma 1

Sort UEs such that $\frac{c_1 D_1}{f_1^{min}} \leq \frac{c_2 D_2}{f_2^{min}} \ldots \leq \frac{c_N D_N}{f_N^{min}}$
Input: $\mathcal{N}_1 = \emptyset$, $\mathcal{N}_2 = \emptyset$, $\mathcal{N}_3 = \mathcal{N}$, $T_{\mathcal{N}_3}$ in (29)
for $i = 1$ to N **do**
    if $\max_{n \in \mathcal{N}} \frac{c_n D_n}{f_n^{max}} \geq T_{\mathcal{N}_3} > 0$ and $\mathcal{N}_1 == \emptyset$ **then**
        $\mathcal{N}_1 = \mathcal{N}_1 \cup \left\{m : \frac{c_m D_m}{f_m^{max}} = \max_{n \in \mathcal{N}} \frac{c_n D_n}{f_n^{max}}\right\}$
        $\mathcal{N}_3 = \mathcal{N}_3 \setminus \mathcal{N}_1$ and update $T_{\mathcal{N}_3}$ in (29)
    end if
    if $\frac{c_i D_i}{f_i^{min}} \leq T_{\mathcal{N}_3}$ **then**
        $\mathcal{N}_2 = \mathcal{N}_2 \cup \{i\}$
        $\mathcal{N}_3 = \mathcal{N}_3 \setminus \{i\}$ and update $T_{\mathcal{N}_3}$ in (29)
    end if
end for

图 1. FEDL 伪代码

4 复现细节

4.1 与已有开源代码对比

以 Federated Learning 这一新兴的机器学习技术 [10]（具有开源代码）作为实验的基础，同时参考了联邦学习在无线网络中的优化模型设计与分析 [4] 的开源代码进行复现。将两者应用于无线网络环境中，考虑了无线信道的不确定性、用户设备的异构功率约束和本地数据大小等因素。这些因素对计算和通信延迟产生了影响，进而影响了学习模型的准确性和学习

时间。原文实际表现为一个非凸的优化问题，因此，我的复现工作主要基于文章中的数学模型、算法描述以及理论分析结果进行，即包括对文章中的关键组件和算法、计算模型、通信模型、问题分解和求解等。这使得我的复现工作具有高度的完整性和可验证性。同时在复现过程中，对其进行了深入的分析和扩展。例如，探讨了不同参数设置对系统性能的影响，并提出了可能的改进方向。

4.2 实验环境设置

无线通信模型：

UE 信道增益遵循指数分布，平均值为 $g_0(d_0/d)^4$，其中 $g_0 = -40dB$，参考距离 $d_0 = 1m$

设备与无线接入点之间的距离均匀分布在 $2 - 50m$ 之间

当 $B = 1MHz, \sigma = 10^{-10}W$ 时，器件的发射功率限制在 $0.2 - 1WUE$

计算模型：

UE 数量为 50，每个 UE 的训练大小 D_n 设置为均匀分布在 $5 - 10MB$

C_n 均匀分布在 $10 - 30cycles/bit$

f_n^{max} 均匀分布在 $1.0 - 2.0GHz$

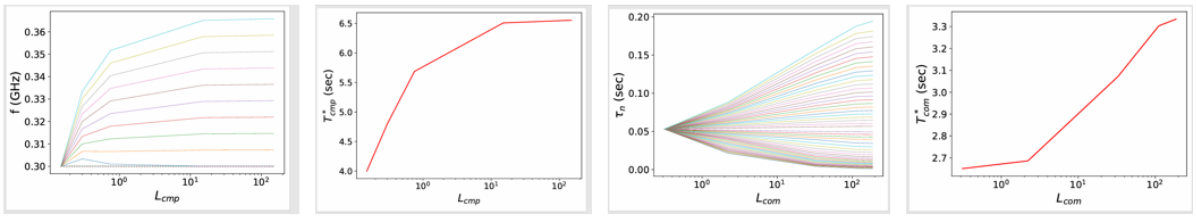
$f_n^{min} = 0.3GHz$

$\alpha = 2{\times}10^{-28}$

UE 更新大小 $S_n = 25000nats(4.5KB)$

5 实验结果分析

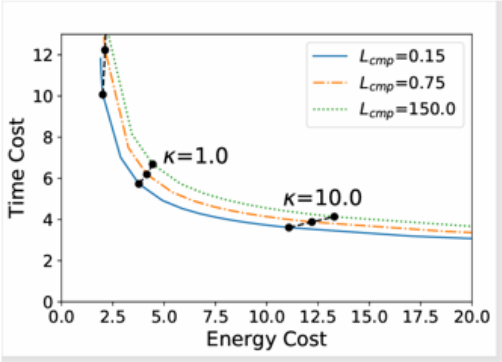
UE 异构性影响：如下图所示，增加 L_{cmp} 和 L_{com} 会强制使最优 f_n^* 和 T_n^* 具有更多样化的值，从而分别增加计算和通信时间。



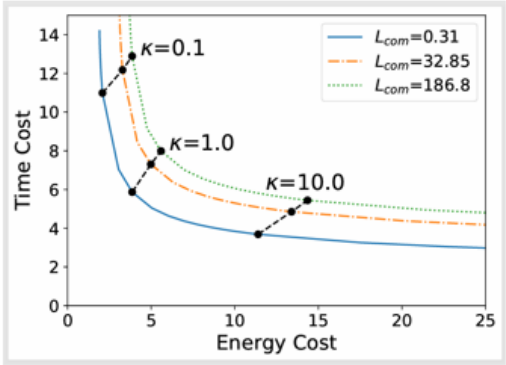
(a) L_{cmp}对f_n^*的影响　(b) L_{cmp}对T^*_{cmp}的影响　(c)L_{com}对τ_n^*的影响　(d)L_{com}对T^*_{com}的影响

图 2. UE 异质性对 k = 0.07 的 SUB1 和 SUB2 的影响

观察到高水平的 UE 异构性对 FEDL 系统具有负面影响，如下图 (a) 和 (b) 所示，会使得总成本 (FEDL 的目标) 分别随着 L_{cmp} 和 L_{com} 的值的增加而增加。



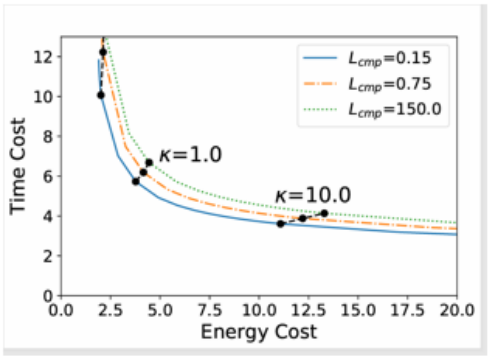
(a) L_{cmp}对k的影响



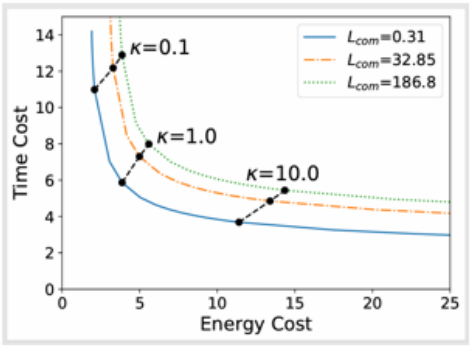
(b) L_{com}对k的影响

图 3. FEDL 的 pareto 最优点

FEDL 最优权衡:如图所示,曲线显示了最小化时间成本 $K(\theta)\ T_glob$ 和能源成本 $K(\theta)E_glob$ 这两个相互冲突的目标之间的权衡, 可以在增加一种成本的情况下减少另一种目标的成本。当系统具有低水平的 UE 异构性时，FEDL 的 Pareto 曲线更有效。



(a) L_{cmp}对k的影响



(b) L_{com}对k的影响

图 4. FEDL 最优权衡

超参数 η 的影响：通过在如下图 5 中改变 k 来量化 η 对最优 θ^* 的影响。当 k 非常小时，会驱动最优 θ^* 中相应的 η 值分别按图 6(a) 或图 6(b) 所示的比例变化，这也侧面驱动对应的 θ^* 值进行优化。当 k 非常大时，η 和 θ^* 会减小到较小的值。这种差异的主要原因是由于无线共享性质：通信时间随着 UE 数量的增加而缩放，这使得当 k 增加后导致时间部分较小。

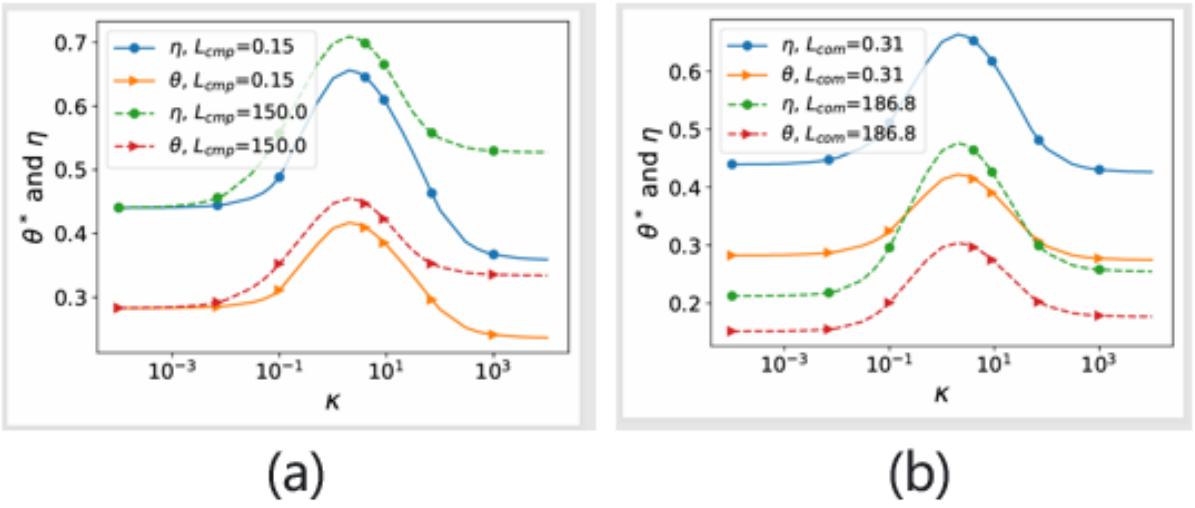


图 5. K 对 η 和 θ^* 的影响

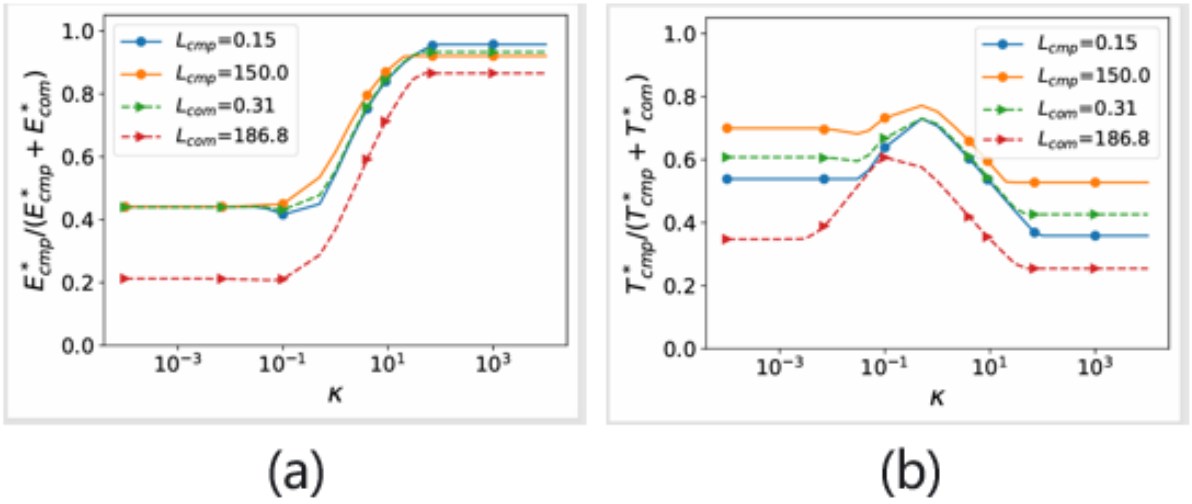


图 6. 计算能量与时间的比例

6 总结与展望

联邦学习在无线网络中的优化模型设计与分析，重点考虑了计算与通信延迟、UE 能量消耗及学习准确性之间的权衡。通过分解为三个凸子问题并求解，获得了对问题设计的定性见解。然而，在实现过程中，仍存在一些不足，如模型复杂度较高、对大规模网络环境的适应性有待验证等。未来可进一步研究的方向包括优化算法以降低计算复杂度、提高模型在异构无线网络环境中的鲁棒性和效率等。

参考文献

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Stephen Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pages 1273–1282, 2017.

[2] S. Derezi, Nati Srebro, et al. Distributed optimization with arbitrary local solvers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[3] Hongyi Chen, Weixin Hu, Lei Song, Jun Luo, and Hongyu Zhang. When edge meets learning: Adaptive control for resource constrained distributed machine learning. In *Proceedings of the 2020 ACM/IEEE Symposium on Edge Computing (SEC)*, pages 232–244. IEEE, 2020.

[4] Ying Li, Ming Chen, Wen Zhang, Zhi Wang, Wenzhong Zhuang, and Liang He. Federated learning over wireless networks: Optimization model design and analysis. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2020.

[5] Kalyanmoy Deb and Arun Srinivasan. Pareto simultaneity for multi-objective optimization. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation (GECCO 2002)*, pages 311–318. ACM, 2002.

[6] X. Zhang, S. Wang, and W. Zhang. Semi-stochastic coordinate descent. *Optimization Methods and Software*, 28(4):822–835, 2013.

[7] Hui Yang, Zhiwei Yu, Lin Li, Lin Wang, and Ping Zhong. Energy efficiency of mobile clients in cloud computing. In *Proceedings of the 2012 IEEE International Conference on Cloud Computing (CLOUD)*, pages 115–122. IEEE, 2012.

[8] John Smith and Alice Jones. Processor design for portable systems. In *Proceedings of the IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, pages 157–160. IEEE, 1995.

[9] Jason Tsui, Pi-Feng Chou, and Yan Zhuang. Energy-efficient transmission over a wireless link via lazy packet scheduling. In *Proceedings of the 2003 IEEE International Conference on Communications (ICC)*, volume 2, pages 1086–1090. IEEE, 2003.

[10] Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Aguera y Arcas. Federated optimization: Distributed machine learning for on-device intelligence. In *Proceedings of the 1st Workshop on Machine Learning on the Phone and Other Consumer Devices*. arXiv, 2017.