Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5524-5532
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135552

Hans汉斯

异构网络下一种带松弛步的联邦学习算法

岳素云，谭雅欣

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年9月4日；录用日期：2023年10月16日；发布日期：2023年10月25日

摘　要

联邦学习是一个在保护用户隐私、数据安全方面具有显著优势的机器学习框架。针对联邦学习中各个设备间存在的数据异质性与系统异质性，许多学者提出了相应的解决方案，如FedProx算法等。考虑到松弛步可以起到提高收敛速度的作用，本文我们在FedProx的基础上引入松弛步，提出一种带松弛步的联邦学习算法FedProx + Relaxation。本文对FedProx + Relaxation的收敛性进行了理论分析，并通过数值实验展示了该算法的有效性与稳健性。通过数值实验，本文说明了FedProx + Relaxation相比于FedProx具有更加稳健的收敛效果。

关键词

联邦学习，松弛步，机器学习

A Federated Learning Algorithm with a Relaxation Step for Heterogeneous Networks

Suyun Yue, Yaxin Tan

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Sep. 4th, 2023; accepted: Oct. 16th, 2023; published: Oct. 25th, 2023

Abstract

Federated learning is a machine learning framework that has significant advantages in protecting user privacy and data security. Considering the data heterogeneity and system heterogeneity between various devices in federated learning, many scholars have proposed corresponding solutions, such as FedProx algorithm, etc. Considering that the relaxation step can play a role in im-

文章引用: 岳素云, 谭雅欣. 异构网络下一种带松弛步的联邦学习算法[J]. 运筹与模糊学, 2023, 13(5): 5524-5532.
DOI: 10.12677/orf.2023.135552

proving the speed of convergence, a relaxation step on the basis of the FedProx algorithm was introduced. And then a federated learning algorithm with a relaxation step, i.e., FedProx + Relaxation is proposed here. The convergence of the FedProx + Relaxation is analyzed theoretically, the efficiency and robustness of the algorithm are demonstrated by numerical experiments. Through the numerical experiment, it shows that the FedProx + Relaxation has more robust convergence than FedProx.

Keywords

Federated Learning, Relaxation Step, Machine Learning

1. 引言

传统的机器学习面临着隐私保护问题，在训练过程中，若将数据直接共享，就会有隐私泄露的风险。针对数据隐私问题，谷歌公司提出了一种联邦学习(Federated Learning)的算法框架。作为一种机器学习框架，它既能够一定程度地保护用户数据隐私，又使得用户数据能被人工智能系统更加高效地使用。

传统的机器学习建模是将数据样本集合到一个中心服务器再统一进行模型训练，然后利用获得的模型进行预测[1] [2] [3] [4]。联邦学习可以看作是一种数据储存在设备本地的分布式机器学习框架，所有设备在中心服务器的编排下协同训练模型，在训练过程中不直接共享数据，其目标是训练一个在大多数设备上表现良好的全局模型，在设备与中心服务器之间进行多轮模型更新，直到满足停机准则[5]。

目前，联邦学习面临的挑战之一就是异质性，其包含数据异质和系统异质两个方面。数据异质性往往是因为不同设备上的数据分布不一致导致，而系统异质主要是因为联邦网络中设备硬件，网络连接状况以及电源状态的差异导致。考虑到这两种异质性，Li 等人[6]提出了 FedProx 算法，其通过在局部求解问题上增加邻近项来控制本地模型与全局模型差异程度，从而处理数据异质的问题；同时，该算法还允许参与模型更新的设备进行不同轮数的本地更新，从而处理系统异质的问题。

松弛法是一种加速迭代方法，起源于变分思想，其利用当前迭代点和前一迭代点的凸组合来对新的迭代点进行更新，这一方法对数值计算中的各种问题都可起到加速收敛的作用。除此之外，还有超松弛法、群松弛法、逐次超松弛法等改进的松弛方法[7]。进一步地，松弛方法已被应用到许多方面，如基于机器学习和线性规划松弛的近似混合整数规划解[8]，非齐次马尔可夫跳跃系统 H∞滤波的广义松弛技术[9]以及基于改进临近丛法的拉格朗日松弛短期热液调度[10]等。为加速 FedProx 算法，我们在该算法的基础上添加松弛步，提出了异构网络下一种带松弛步的联邦学习算法，从而使得算法更加稳健，且迭代中步长参数的选择范围更大。针对该算法，本文给出了其收敛性的理论分析，并进一步地通过数值实验展示了该算法的有效性和稳健性。

2. 预备知识

在这一节中，我们介绍本文研究的问题，用到的定义以及引理。文中所用到的符号定义如下：$\mathbf{R}^d$ 表示欧几里得空间，$\langle\cdot,\cdot\rangle$ 表示向量内积，$\|\cdot\|$ 表示欧氏范数，∇ 表示梯度。我们用 $[m]$ 来简记集合 $\{1,2,\cdots,m\}$，

$\mathbb{E}[\cdot]$表示数学期望。

2.1. 本文问题

联邦学习涉及到对一个优化问题的求解，通过对该优化问题的求解得到联邦学习中模型的参数，该优化问题可抽象为如下形式

$$\min_{\omega} f(\omega)=\sum_{k=1}^{N} p_k F_k(\omega)=\mathbb{E}_k\left[F_k(\omega)\right], \tag{1}$$

其中 N 是设备总数，概率 $p_k \geq 0$ 且 $\sum_{k=1}^{N} p_k = 1$。通常，在联邦学习中每个设备拥有的数据分布 D_k 是不同的，局部目标函数 $F_k(\omega) := \mathbb{E}_{x_k \sim D_k}\left[f_k(\omega; x_k)\right]$ 是非凸的，其用来度量各设备的局部期望风险。若设备 k 有 n_k 个可用样本，则可令 $p_k = \frac{n_k}{n}$，其中 $n = \sum_{k=1}^{N} n_k$ 为样本总数。

2.2. 相关定义与引理

在算法中，通过对内循环中子问题的非精确求解，可以灵活的调整每个设备的局部计算量与通信量。我们在下面正式介绍这种非精确解的概念。

定义 **1** (γ-非精确解)令函数 $h(\omega;\omega_0) = F(\omega) + \frac{\mu}{2}\|\omega-\omega_0\|^2$ 且参数 $\gamma \in [0,1]$。如果

$$\left\|\nabla h(\omega^*;\omega_0)\right\| \leq \gamma \left\|\nabla h(\omega_0;\omega_0)\right\|,$$

其中 $\nabla h(\omega;\omega_0) = \nabla F(\omega) + \mu(\omega-\omega_0)$，则我们就定义 ω^* 为 $\min_{\omega} h(\omega;\omega_0)$ 的一个 γ-非精确解。明显地，上述定义中 γ 越小求解精确度越高。

为分析算法收敛性，我们需要量化局部目标函数与全局目标函数间的差异程度。Li 等人[6]提出了一种联邦网络中设备差异性的度量准则，即 B-局部差异，其定义如下所示。该度量准则在文献[11] [12]中均有应用。

定义 **2** (B-局部差异)我们称 F_k 在 ω 处满足 B-局部差异，如果局部函数 F_k 在 ω 处满足以下条件

$$\mathbb{E}_k\left[\left\|\nabla F_k(\omega)\right\|^2\right] \leq B^2\left\|\nabla f(\omega)\right\|^2.$$

进一步地，$\forall \left\|\nabla f(\omega)\right\|^2 \neq 0$，我们定义 $B(\omega) = \sqrt{\dfrac{\mathbb{E}_k\left\|\nabla F_k(\omega)\right\|^2}{\left\|\nabla f(\omega)\right\|^2}}$.

引理 1 给出光滑函数的相关性质，其在算法收敛性证明中起到了重要作用，具体证明可以参考([13]，引理 1.2.3)。

引理 1 (下降引理[13])若函数 $f: \mathbf{R}^d \to \mathbf{R}$ 且梯度 ∇f 是 L-Lipschitz 连续的，其中 $L > 0$，则我们有

$$\left|f(y)-f(x)-\left\langle \nabla f(x), y-x\right\rangle\right| \leq \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

3. 算法与收敛分析

本节，我们将给出所提算法 FedProx + Relaxation，并对其进行收敛分析。

3.1. 联邦优化算法

首先我们给出 FedProx + Relaxation 的具体迭代格式如下所示：

算法 1: FedProx + Relaxation $\left(K,T,\mu,\alpha,\gamma,\omega^0,N,p_k,k=1,2,\cdots,N\right)$

输入： $K,T,\mu,\alpha\in(0,1),\gamma,\omega^0,N,p_k,k=1,2,\cdots,N,$

1 **for** $t=0,1,\cdots,T-1$ **do**

2 选择 K 个设备的子集 S_t (每个设备 k 被选中的概率为 p_k),

3 中心服务器将 ω^t 发送给所有选中的设备，

4 每个被选中的设备 $k\in S_t$ 计算一个 γ_k^t 非精确解 ω_k^{t+1}，即

$$\omega_k^{t+1}\approx\arg\min_{\omega}h_k\left(\omega;\omega^t\right)=\arg\min_{\omega}\left\{F_k\left(\omega\right)+\frac{\mu}{2}\left\|\omega-\omega^t\right\|^2\right\}$$

并将结果发送给中心服务器，

5 中心服务器计算 $\omega^{t+1}=\alpha\omega^t+(1-\alpha)\frac{1}{K}\sum_{k\in S_t}\omega_k^{t+1}$，并发送给所有设备。

6 **end for**

该算法与 FedProx 算法的主要区别为在第 5 迭代步中引入了凸松弛的思想，利用了前一迭代步的全局模型信息。为证明算法的收敛性，我们对优化问题中的目标函数进行一些假设。

假设 **1** (光滑，有界) $\forall k\in[m]$，设备 k 的局部目标函数 $F_k(\cdot)$ 是非凸可微且 L-光滑的，即

$$\left\|\nabla F_k(u)-\nabla F_k(v)\right\|\le L\|u-v\|,\ \forall\ u,v\in\mathbf{R}^d.$$

此外目标函数 $f(\cdot)$ 的最优值是有下界的，满足 $f^*=\min_{\omega}f(\omega)>-\infty$.

假设 **2** (差异有界性)对于 $\varepsilon>0$，存在 B_ε，使得对于任意点 $\omega\in\left\{\omega\middle|\left\|\nabla f(\omega)\right\|^2>\varepsilon\right\}$，满足 $B(\omega)\le B_\varepsilon$。

3.2. 收敛性分析

基于以上假设，我们将给出 FedProx + Relaxation 算法的收敛性分析结果。为简便起见，我们在证明过程中取 $\gamma_k^t=\gamma$。在此之前我们将首先给出相关引理，引理 2 证明与文献[6]中类似，为保证文章的完整性，本文给出相关证明过程。

引理 **2** 令假设 1~2 成立，$\nabla^2F_k\succeq-L\_I$，其中 $L\_>0$，且 $\bar{\mu}:=\mu-L\_>0$。设 ω^t 不是问题(1)的一个稳定点，且局部函数 F_k 在 ω^t 上是 B-局部差异的，即满足 $B\left(\omega^t\right)\le B$。则我们有

$$\left\|\bar{\omega}^{t+1}-\omega^t\right\|\le\frac{B(1+\gamma)}{\bar{\mu}}\left\|\nabla f\left(\omega^t\right)\right\|,\tag{2}$$

$$\left\|M_{t+1}\right\|\le\left(\frac{L(1+\gamma)}{\bar{\mu}}+\gamma\right)B\left\|\nabla f\left(\omega^t\right)\right\|,\tag{3}$$

其中 $\bar{\omega}^{t+1}:=E_k\left[\omega_k^{t+1}\right]$，$M_{t+1}=-\nabla f\left(\omega^t\right)-\mu\left(\bar{\omega}^{t+1}-\omega^t\right)$。

证明考虑到算法中 ω_k^{t+1} 的更新为子问题的一个 γ-非精确解，则我们定义误差变量

$$e_k^{t+1}=\nabla F_k\left(\omega_k^{t+1}\right)+\mu\left(\omega_k^{t+1}-\omega^t\right),\tag{4}$$

且该误差变量满足

$$\left\|e_k^{t+1}\right\|\le\gamma\left\|\nabla F_k\left(\omega^t\right)\right\|.\tag{5}$$

对公式(4)取期望，并结合 $\bar{\omega}^{t+1}$ 的定义，我们有 $\mathbb{E}_k\left[e_k^{t+1}\right]=\mathbb{E}_k\left[\nabla F_k\left(\omega_k^{t+1}\right)\right]+\mu\left(\bar{\omega}^{t+1}-\omega^t\right)$，进一步整理

可得

$$\bar{\omega}^{t+1}-\omega^t=-\frac{1}{\mu}\mathbb{E}_k\left[\nabla F_k\left(\omega_k^{t+1}\right)\right]+\frac{1}{\mu}\mathbb{E}_k\left[e_k^{t+1}\right].$$

由于 $\bar{\mu}:=\mu-L_-\geq 0$，故可知 h_k 是 $\bar{\mu}$ -强凸的，则有

$$\begin{aligned}\bar{\mu}\left\|\omega_k^{t+1}-\omega^t\right\|^2&\leq\left\langle\omega_k^{t+1}-\omega^t,\nabla F_k\left(\omega_k^{t+1}\right)+\mu\left(\omega_k^{t+1}-\omega^t\right)-\nabla F_k\left(\omega^t\right)\right\rangle\\&\leq\left\langle\omega_k^{t+1}-\omega^t,e_k^{t+1}-\nabla F_k\left(\omega^t\right)\right\rangle\leq\left\|\omega_k^{t+1}-\omega^t\right\|\left\|e_k^{t+1}-\nabla F_k\left(\omega^t\right)\right\|\\&\leq\left\|\omega_k^{t+1}-\omega^t\right\|\left(\left\|e_k^{t+1}\right\|+\left\|\nabla F_k\left(\omega^t\right)\right\|\right)\leq(1+\gamma)\left\|\omega_k^{t+1}-\omega^t\right\|\left\|\nabla F_k\left(\omega^t\right)\right\|,\end{aligned}$$

其中第二个不等式是根据误差变量的定义(4)得到，最后一个不等式是根据(5)式所得。整理上式，我们有

$$\left\|\omega_k^{t+1}-\omega^t\right\|\leq\frac{(1+\gamma)\left\|\nabla F_k\left(\omega^t\right)\right\|}{\bar{\mu}}\tag{6}$$

根据 $\bar{\omega}^{t+1}$ 的定义，即 $\bar{\omega}^{t+1}:=\mathbb{E}_k\left[\omega_k^{t+1}\right]$，结合三角不等式，我们有

$$\begin{aligned}\left\|\bar{\omega}^{t+1}-\omega^t\right\|&\leq\mathbb{E}_k\left[\left\|\omega_k^{t+1}-\omega^t\right\|\right]\leq\frac{(1+\gamma)}{\bar{\mu}}\mathbb{E}_k\left[\left\|\nabla F_k\left(\omega^t\right)\right\|\right]\\&\leq\frac{(1+\gamma)}{\bar{\mu}}\sqrt{\mathbb{E}_k\left[\left\|\nabla F_k\left(\omega^t\right)\right\|^2\right]}\leq\frac{B(1+\gamma)}{\bar{\mu}}\left\|\nabla f\left(\omega^t\right)\right\|,\end{aligned}$$

其中第二个不等式是由(6)式所得，第三个不等式是根据 Jensen's 不等式得到，最后一个不等式利用了 B-局部差异的定义，则(2)式得证。

根据 M_{t+1} 的定义，可得 $M_{t+1}=\mathbb{E}_k\left[\nabla F_k\left(\omega_k^{t+1}\right)-\nabla F_k\left(\omega^t\right)-e_k^{t+1}\right]$。由假设 1 可知

$$\left\|M_{t+1}\right\|\leq\mathbb{E}_k\left\|\nabla F_k\left(\omega_k^{t+1}\right)-\nabla F_k\left(\omega^t\right)-e_k^{t+1}\right\|\leq\mathbb{E}_k\left[L\left\|\omega_k^{t+1}-\omega^t\right\|+\left\|e_k^{t+1}\right\|\right].$$

根据公式(5-6)，我们有

$$\left\|M_{t+1}\right\|\leq\left(\frac{L(1+\gamma)}{\bar{\mu}}+\gamma\right)\mathbb{E}_k\left\|\nabla F_k\left(\omega^t\right)\right\|\leq\left(\frac{L(1+\gamma)}{\bar{\mu}}+\gamma\right)\sqrt{\mathbb{E}_k\left\|\nabla F_k\left(\omega^t\right)\right\|^2}\leq\left(\frac{L(1+\gamma)}{\bar{\mu}}+\gamma\right)B\left\|\nabla f\left(\omega^t\right)\right\|,$$

其中第二个不等式利用了 Jensen's 不等式，第三个不等式由 B-局部差异可得，则(3)式得证。证毕

引理 3 令假设 2 成立，$\nabla^2 F_k\succeq -L_- I$，其中 $L_->0$，且 $\bar{\mu}:=\mu-L_->0$。设 ω^t 不是问题(1)的一个稳定点，且局部函数 F_k 在 ω^t 上是 B-局部差异的，即满足 $B\left(\omega^t\right)\leq B$。则我们有

$$\mathbb{E}_{S_t}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2\leq\frac{2(1+\gamma)^2B^2}{K\bar{\mu}^2}\left(K\alpha^2+4(1-\alpha)^2\right)\left\|\nabla f\left(\omega^t\right)\right\|^2,\tag{7}$$

$$\mathbb{E}_{S_t}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|\leq\frac{\sqrt{2}(1+\gamma)B}{\sqrt{K}\bar{\mu}}\sqrt{K\alpha^2+4(1-\alpha)^2}\left\|\nabla f\left(\omega^t\right)\right\|,\tag{8}$$

其中 $\mathbb{E}_{S_t}[\cdot]$ 表示在第 t 次迭代关于所选设备集 S_t 的期望。

证明根据算法 1 的迭代公式可知

$$\begin{aligned}\mathbb{E}_{S_t}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2&=\mathbb{E}_{S_t}\left\|\alpha\left(\omega^t-\bar{\omega}^{t+1}\right)+(1-\alpha)\frac{1}{K}\sum_{k\in S_t}\left(\omega_k^{t+1}-\bar{\omega}^{t+1}\right)\right\|^2\\&\leq 2\alpha^2\left\|\omega^t-\bar{\omega}^{t+1}\right\|^2+2\mathbb{E}_{S_t}\left\|\frac{1-\alpha}{K}\sum_{k\in S_t}\left(\omega_k^{t+1}-\bar{\omega}^{t+1}\right)\right\|^2.\end{aligned}\tag{9}$$

考虑到设备间的相互独立性，我们有

$$\begin{aligned}\mathbb{E}_{S_t}\left\|\frac{1-\alpha}{K}\sum_{k\in S_t}\left(\omega_k^{t+1}-\bar{\omega}^{t+1}\right)\right\|^2 &\le \frac{(1-\alpha)^2}{K}\mathbb{E}_k\left[\left\|\omega_k^{t+1}-\bar{\omega}^{t+1}\right\|^2\right]\\ &\le \frac{(1-\alpha)^2}{K}\mathbb{E}_k\left[2\left\|\omega_k^{t+1}-\omega^t\right\|^2+2\left\|\omega^t-\bar{\omega}^{t+1}\right\|^2\right]\\ &\le \frac{(1-\alpha)^2}{K}\mathbb{E}_k\left[2\left(\frac{(1+\gamma)\left\|\nabla F_k\left(\omega^t\right)\right\|}{\bar{\mu}}\right)^2+2\left(\frac{B(1+\gamma)}{\bar{\mu}}\left\|\nabla f\left(\omega^t\right)\right\|\right)^2\right]\\ &\le \frac{4(1-\alpha)^2(1+\gamma)^2}{K\bar{\mu}^2}B^2\left\|\nabla f\left(\omega^t\right)\right\|^2.\end{aligned}$$

其中第三个不等式是根据公式(2)和(6)所得，最后一个不等式利用了 B-局部差异的定义。由上式及公式(2)，我们对公式(9)整理可得

$$\begin{aligned}\mathbb{E}_{S_t}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2 &\le 2\alpha^2\left(\frac{B(1+\gamma)}{\bar{\mu}}\left\|\nabla f\left(\omega^t\right)\right\|\right)^2+\frac{8(1-\alpha)^2(1+\gamma)^2}{K\bar{\mu}^2}B^2\left\|\nabla f\left(\omega^t\right)\right\|^2\\ &=\frac{2(1+\gamma)^2B^2}{K\bar{\mu}^2}\left(K\alpha^2+4(1-\alpha)^2\right)\left\|\nabla f\left(\omega^t\right)\right\|^2.\end{aligned}$$

进一步地，根据 Jensen's 不等式，我们有

$$\mathbb{E}_{S_t}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|\le\sqrt{\mathbb{E}_{S_t}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2}=\frac{\sqrt{2}(1+\gamma)B}{\sqrt{K}\bar{\mu}}\sqrt{K\alpha^2+4(1-\alpha)^2}\left\|\nabla f\left(\omega^t\right)\right\|.$$ 证毕

下面我们给出算法 1 在期望意义下目标函数的梯度可达到 $O(1/T)$ 的次线性收敛率。

定理 1 若引理 2 的条件成立，选取算法 1 中的参数 α,μ,K 和 γ，使得

$$\begin{aligned}\rho=\Bigg[&\frac{1-\gamma B}{\mu}-\frac{LB(1+\gamma)}{\mu\bar{\mu}}-\frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}-\frac{L(1+\gamma)^2B^2}{K\bar{\mu}^2}\left(K\alpha^2+4(1-\alpha)^2\right)\\ &-\left(\frac{2LB(1+\gamma)}{\bar{\mu}}+1\right)\frac{\sqrt{2}(1+\gamma)B}{\sqrt{K}\bar{\mu}}\sqrt{K\alpha^2+4(1-\alpha)^2}\Bigg]>0\end{aligned}$$

成立，则有

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f\left(\omega^t\right)\right\|^2\right]\le\frac{f\left(\omega^0\right)-f^*}{T\rho},$$

其中 $f^*=\min_{\omega}f(\omega)>-\infty$。

证明一方面，由于 f 是 L-光滑的，则根据下降引理 1 可知

$$\begin{aligned}f\left(\bar{\omega}^{t+1}\right)&\le f\left(\omega^t\right)+\left\langle\nabla f\left(\omega^t\right),\bar{\omega}^{t+1}-\omega^t\right\rangle+\frac{L}{2}\left\|\bar{\omega}^{t+1}-\omega^t\right\|^2\\ &=f\left(\omega^t\right)+\left\langle\nabla f\left(\omega^t\right),-\frac{1}{\mu}\nabla f\left(\omega^t\right)-\frac{1}{\mu}M_{t+1}\right\rangle+\frac{L}{2}\left\|\bar{\omega}^{t+1}-\omega^t\right\|^2\\ &\le f\left(\omega^t\right)-\frac{1}{\mu}\left\|\nabla f\left(\omega^t\right)\right\|^2+\frac{1}{\mu}\left\|\nabla f\left(\omega^t\right)\right\|\left\|M_{t+1}\right\|+\frac{L}{2}\left(\frac{1+\gamma}{\bar{\mu}}\right)^2B^2\left\|\nabla f\left(\omega^t\right)\right\|^2,\end{aligned}$$

其中等式是根据 M_{t+1} 的定义所得，第二个不等式是根据公式(2)及 Schwarz 不等式所得。进一步，根据公式(3)，我们有

$$\begin{aligned} f\left(\bar{\omega}^{t+1}\right) &\le f\left(\omega^t\right)-\frac{1}{\mu}\left\|\nabla f\left(\omega^t\right)\right\|^2+\frac{B}{\mu}\left(\frac{L(1+\gamma)}{\bar{\mu}}+\gamma\right)\left\|\nabla f\left(\omega^t\right)\right\|^2+\frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}\left\|\nabla f\left(\omega^t\right)\right\|^2 \\ &= f\left(\omega^t\right)-\left(\frac{1-\gamma B}{\mu}-\frac{LB(1+\gamma)}{\mu\bar{\mu}}-\frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}\right)\left\|\nabla f\left(\omega^t\right)\right\|^2. \end{aligned} \tag{10}$$

另一方面，根据下降引理 1，我们有

$$f\left(\omega^{t+1}\right)\le f\left(\bar{\omega}^{t+1}\right)+\left\langle\nabla f\left(\bar{\omega}^{t+1}\right),\omega^{t+1}-\bar{\omega}^{t+1}\right\rangle+\frac{L}{2}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2.$$

对上式两边关于 S_t 取期望，并利用 Schwarz 不等式，可得

$$\begin{aligned} \mathbb{E}_{S_t}\left[f\left(\omega^{t+1}\right)\right] &\le f\left(\bar{\omega}^{t+1}\right)+\mathbb{E}_{S_t}\left[\left\|\nabla f\left(\bar{\omega}^{t+1}\right)\right\|\cdot\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|+\frac{L}{2}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2\right] \\ &= f\left(\bar{\omega}^{t+1}\right)+\mathbb{E}_{S_t}\left[\left(\left\|\nabla f\left(\bar{\omega}^{t+1}\right)-\nabla f\left(\omega^t\right)+\nabla f\left(\omega^t\right)\right\|+\frac{L}{2}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|\right)\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|\right] \\ &\le f\left(\bar{\omega}^{t+1}\right)+\mathbb{E}_{S_t}\left[\left(L\left\|\bar{\omega}^{t+1}-\omega^t\right\|+\left\|\nabla f\left(\omega^t\right)\right\|+\frac{L}{2}\left\|\omega^{t+1}-\omega^t\right\|+\frac{L}{2}\left\|\omega^t-\bar{\omega}^{t+1}\right\|\right)\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|\right] \\ &\le f\left(\bar{\omega}^{t+1}\right)+\mathbb{E}_{S_t}\left[\left(2L\left\|\bar{\omega}^{t+1}-\omega^t\right\|+\left\|\nabla f\left(\omega^t\right)\right\|\right)\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|+\frac{L}{2}\left\|\omega^{t+1}-\bar{\omega}^{t+1}\right\|^2\right]. \end{aligned}$$

其中第二个不等式利用了三角不等式及 f 的 L-光滑性。根据引理 2-3，我们整理上式可得

$$\begin{aligned} \mathbb{E}_{S_t}\left[f\left(\omega^{t+1}\right)\right] \le f\left(\bar{\omega}^{t+1}\right)+\Bigg[&\left(\frac{2LB(1+\gamma)}{\bar{\mu}}+1\right)\frac{\sqrt{2}(1+\gamma)B}{\sqrt{K}\bar{\mu}}\sqrt{K\alpha^2+4(1-\alpha)^2} \\ &+\frac{L(1+\gamma)^2B^2}{K\bar{\mu}^2}\left(K\alpha^2+4(1-\alpha)^2\right)\Bigg]\left\|\nabla f\left(\omega^t\right)\right\|^2. \end{aligned} \tag{11}$$

联立公式(10)和(11)，我们得出算法 1 在前后步迭代点上目标函数值之间的关系，即

$$\begin{aligned} \mathbb{E}_{S_t}\left[f\left(\omega^{t+1}\right)\right] \le f\left(\omega^t\right)-\Bigg[&\frac{1-\gamma B}{\mu}-\frac{LB(1+\gamma)}{\mu\bar{\mu}}-\frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}-\frac{L(1+\gamma)^2B^2}{K\bar{\mu}^2}\left(K\alpha^2+4(1-\alpha)^2\right) \\ &-\left(\frac{2LB(1+\gamma)}{\bar{\mu}}+1\right)\frac{\sqrt{2}(1+\gamma)B}{\sqrt{K}\bar{\mu}}\sqrt{K\alpha^2+4(1-\alpha)^2}\Bigg]\left\|\nabla f\left(\omega^t\right)\right\|^2. \end{aligned}$$

在上述公式中，对 t 从 $0\sim T-1$ 进行累和并移项可得

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f\left(\omega^t\right)\right\|^2\right]\le\frac{f\left(\omega^0\right)-f^*}{T\rho},$$

其中 $f^*=\min_{\omega}f(\omega)>-\infty$。证毕

4. 数值实验

本节将通过数值实验来说明 FedProx + Relaxation 的收敛效果。我们采用与文献[14]类似的方法来生

成实验数据，对于每个设备 k，我们使用模型 $y=\arg\max\left(\text{softmax}\left(Wx+b\right)\right)$，$x\in\mathbf{R}^{60},W\in\mathbf{R}^{10\times60},b\in\mathbf{R}^{10}$ 来生成样本 $\left(X_k,Y_k\right)$。我们设置 $W_k\sim\mathcal{N}\left(u_k,1\right),b_k\sim\mathcal{N}\left(u_k,1\right),u_k\sim\mathcal{N}\left(0,a\right)$，$x_k\sim\mathcal{N}\left(v_k,\Sigma\right)$，其中协方差矩阵 Σ 是对角阵，且有 $\Sigma_{j,j}=j^{-1.2}$；均值向量 v_k 中的每个元素服从 $\mathcal{N}\left(B_k,1\right),B_k\sim\mathcal{N}\left(0,b\right)$ 的分布；a 控制局部模型之间的差异；b 控制每台设备上的本地数据与其他设备上的本地数据的差异程度。我们分别令 $\left(a,b\right)=\left(0.5,0.5\right),\left(1,1\right)$ 来生成两组异构的分布式数据集。实验的目标是学习全局 W 和 b。每个设备上的数据按 8:2 的比例被随机分为训练集和测试集。

我们在 Tensorflow 中实现所提的 FedProx + Relaxation 算法和现有的 FedProx 算法。这两个算法的局部求解器都是随机梯度算法(SGD)，且在每轮迭代中采用均匀抽样的方式选择设备。网络中设置了 30 个设备，每个设备上的样本数均遵循幂律，我们将每轮选定的设备数量固定为 10。对于所有合成数据实验，学习率选定为 0.0001。

图 1 中横坐标为外循环迭代次数，对应算法中的 t，纵坐标为全局损失函数。显然图 1 中的曲线下降越快代表算法求解速度越快。图 2 中横坐标为外循环迭代次数，对应算法中的 t，纵坐标为预测准确率。显然图 2 中曲线上升越快越高代表算法求得的模型的预测准确率越高。观察图中曲线可知，FedProx-Relaxation 相较于 FedProx 在求解该问题时更加稳健，同时也具有更高的预测准确性。

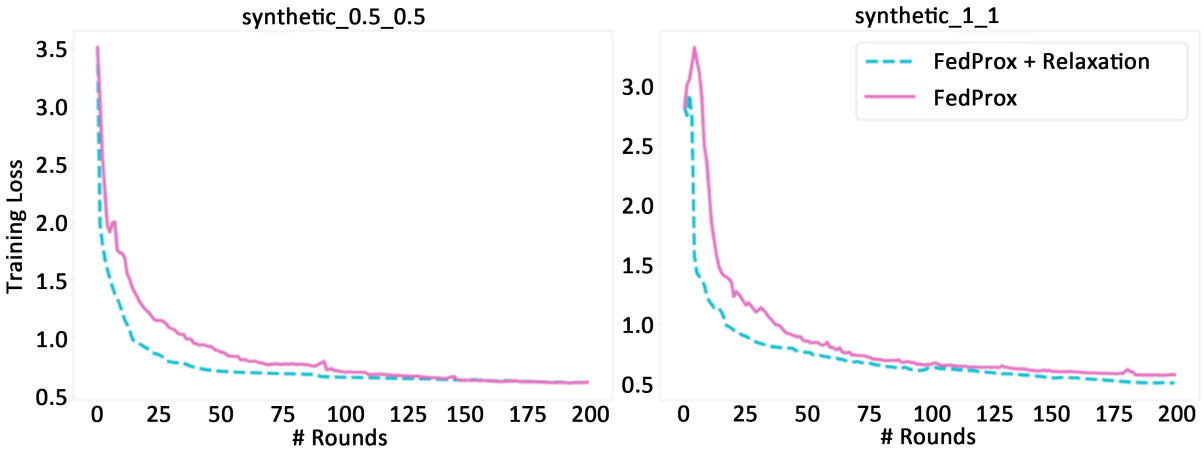

Figure 1. Convergence comparison between FedProx + relaxation algorithm and FedProx algorithm
图 1. FedProx + Relaxation 算法与 FedProx 算法的收敛效果对比

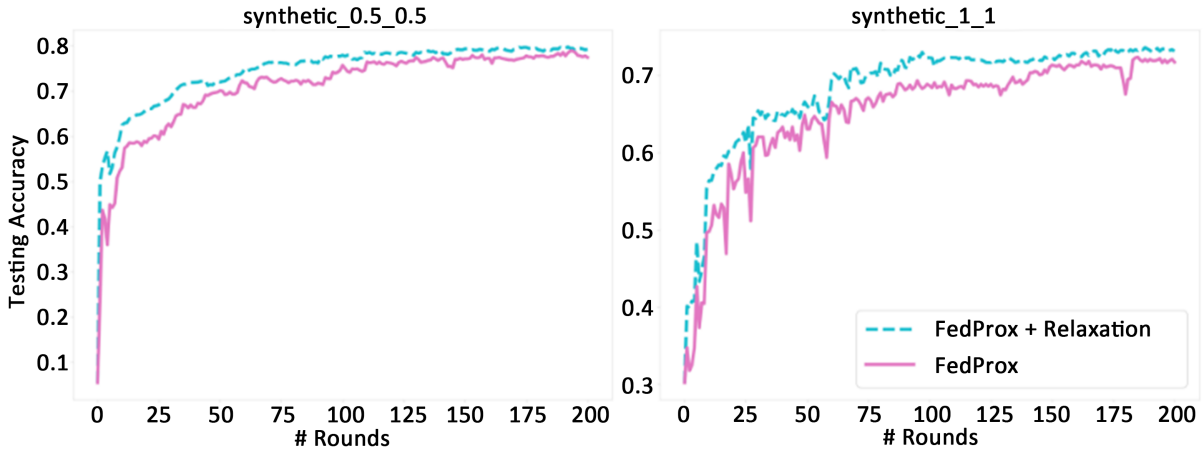

Figure 2. Prediction accuracy comparison between FedProx + relaxation algorithm and FedProx algorithm
图 2. FedProx + Relaxation 算法与 FedProx 算法的预测准确率对比

5. 总结

本文在 FedProx 基础上添加了松弛步，提出了一种带松弛步的联邦学习算法 FedProx + Relaxation。文中对算法进行了理论分析，证明了目标函数的梯度在期望意义下的次线性收敛率。同时，本文也通过数值实验展示了 FedProx + Relaxation 相对于原算法的改进效果。

参考文献

[1] Boyd, S., Parikh, N., Chu, E., *et al*. (2010) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016

[2] Dekel, O., Gilad-Bachrach, R., Shamir, O., *et al*. (2012) Optimal Distributed Online Prediction Using Mini-Batches. *Journal of Machine Learning Research*, **13**, 165-202.

[3] Richtárik, P. and Takáč, M. (2016) Distributed Coordinate Descent Method for Learning with Big Data. *The Journal of Machine Learning Research*, **17**, 2657-2681.

[4] Mitchell, T.M. (2003) Machine Learning. McGraw-Hill, New York.

[5] Sheller, M.J., Edwards, B., Reina, G.A., *et al*. (2020) Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, **10**, Article No. 12598. https://doi.org/10.1038/s41598-020-69250-1

[6] Li, T., Sahu, A., Sanjabi, M., *et al*. (2019) Federated Optimization for Heterogeneous Networks. *Proceedings of the* 1*st Adaptive & Multitask Learning Workshop*, Long Beach, 2020, 429-450.

[7] 林崇德. 中国成人教育百科全书: 数学·电脑[M]. 海口: 南海出版公司, 1994.

[8] Lin, X., Hou, Z.J., Ren, H., *et al*. (2019) Approximate Mixed-Integer Programming Solution with Machine Learning Technique and Linear Programming Relaxation. 2019 3*rd International Conference on Smart Grid and Smart Cities* (*ICSGSC*), Berkeley, 25-28 June 2019, 101-107. https://doi.org/10.1109/ICSGSC.2019.00-11

[9] Kim, S.H. (2019) Generalized Relaxation Techniques for Robust H∞ Filtering of Nonhomogeneous Markovian Jump Systems. *Applied Mathematics and Computation*, **347**, 542-556. https://doi.org/10.1016/j.amc.2018.10.075

[10] Yan, Z., Liao, S., Cheng, C., *et al*. (2021) Lagrangian Relaxation Based on Improved Proximal Bundle Method for Short-Term Hydrothermal Scheduling. *Sustainability*, **13**, 4706. https://doi.org/10.3390/su13094706

[11] Yin, D., Pananjady, A., Lam, M., *et al*. (2018) Gradient Diversity: A Key Ingredient for Scalable Distributed Learning. *International Conference on Artificial Intelligence and Statistics*, Playa Blanca, 2018, 1998-2007.

[12] Schmidt, M. and Roux, N.L. (2013) Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition. *Mathematics*, arXiv: 1308.6370, 2013.

[13] Nesterov, Y. (2004) Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Publishers, Amsterdam. https://doi.org/10.1007/978-1-4419-8853-9

[14] Shamir, O., Srebro, N. and Zhang, T. (2014) Communication Efficient Distributed Optimization Using an Approximate Newton-Type Method. *International Conference on Machine Learning*, Beijing, 2014, 1000-1008.