Advances in Applied Mathematics 应用数学进展, 2026, 15(5), 72-87
Published Online May 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.155209

Hans汉斯

异质稳健分布式支持向量回归

高国庆

青岛大学数学与统计学院，山东 青岛

收稿日期：2026年4月12日；录用日期：2026年5月6日；发布日期：2026年5月13日

摘　要

在分布式算法中，由于本地数据生成机制的异构性，在开发联邦学习方法时考虑个性化非常重要。在这项工作中，本文提出了一种个性化联邦学习方法来解决鲁棒回归问题。具体来说，通过求解具有稀疏融合惩罚的平滑支持向量回归损失来学习回归权重。此外，还设计了用于鲁棒稀疏回归的个性化联邦学习(PerFL-SVR)算法，以有效地解决联邦系统中的估计问题。

关键词

分布式，个性化，稳健回归，异质数据

Heterogeneous Robust Distributed Support Vector Regression

Guoqing Gao

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: April 12, 2026; accepted: May 6, 2026; published: May 13, 2026

Abstract

In distributed algorithms, due to the heterogeneity of local data generation mechanisms, it is crucial to consider personalization when developing federated learning methods. In this work, we propose a personalized federated learning approach to address the robust regression problem. Specifically, the regression weights are learned by optimizing a smoothed support vector regression loss function coupled with a sparse fusion penalty. Furthermore, a personalized federated learning algorithm for robust sparse regression, termed PerFL-SVR, is designed to effectively solve the estimation problem within federated systems.

文章引用: 高国庆. 异质稳健分布式支持向量回归[J]. 应用数学进展, 2026, 15(5): 72-87.
DOI: 10.12677/aam.2026.155209

Keywords

Distributed Computing, Personalization, Robust Regression, Heterogeneous Data

1. 引言

随着支持向量机(SVM)理论的不断成熟，Vapnik [1]将其推广至回归领域并提出了支持向量回归(SVR)。近年来，随着大数据技术的飞速发展，数据隐私保护已成为消费电子等关键领域的核心诉求。作为解决“数据孤岛”问题的有效途径，联邦学习(FL)通过分布式协作训练模型，在保障用户原始数据不离域的前提下展现了巨大的应用潜力。

然而，在现实应用场景中，本地数据生成机制的异构性(即非独立同分布，Non-IID)普遍存在。McMahan [2]等人提出的 FedAvg 方法虽然奠定了联邦学习的基础，但在处理统计异质性数据时，局部更新往往会出现“漂移”现象，导致全局模型性能显著下降。尽管现有研究分析了 FedAvg 在强凸和光滑问题上的收敛速率，但在传统的联邦学习框架下，训练出的单一全局模型难以有效适配差异巨大的局部分布，泛化能力受到严重制约。

为了克服统计异质性带来的挑战，研究者们开始探索个性化联邦学习(Personalized Federated Learning)方法。Liu [3]等人提出了联邦双目标双重模型(FedDODM)，通过独立模型分别优化个性化与隐式泛化目标；Sang [4]等人则提出了基于单一通信前提的 FedOM 方法以提升效率。尽管如此，如何在保障模型鲁棒性的同时，实现高维特征的有效稀疏恢复与跨机器的个性化知识共享，仍是一个亟待解决的问题。

针对上述挑战，本文提出了一种异质稳健分布式支持向量回归方法(PerFL-SVR)。该方法在平滑支持向量回归损失函数的基础上，引入了稀疏融合惩罚项，旨在兼顾模型的个性化表达与全局信息的有效利用，从而在异构且含噪声的联邦环境中实现更精确的稳健回归估计。

2. 平滑支持向量回归

在经典的支持向量回归模型中，使用 ε-不敏感方法，在估计函数周围对称形成一个管道，管内的点不计算损失，只对管外的点计算损失。经典支持向量回归是一个凸二次规划问题，基于支持向量机中的合页损失函数，在支持向量回归中也有相同解释的目标函数，但是该损失函数是非平滑的，因此本文考虑一个平滑的损失函数，基于平滑的损失函数，提出了存在拜占庭故障问题下的平滑稳健的分布式支持向量回归方法。

令数据集为 $\mathcal{D}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$。其中 $\boldsymbol{x}_i\in R^p$ 表示第 i 个样本的特征向量，$y_i\in R$ 为其响应变量。基于经典的支持向量回归，设超平面 $f(x)=x^{\mathrm{T}}\beta_1+\beta_0$，$\beta_1\in R^p$ 为权重向量，$\beta_0\in R$ 为偏置项。根据支持向量机的启发，支持向量回归还有另外一种解释，记残差 $r_i(\beta)=y_i-\boldsymbol{X}_i^{\top}\beta$，就是最小化以下目标函数

$$\frac{1}{n}\sum_{i=1}^{n}\left|r_i(\beta)\right|_{\varepsilon}+\frac{\lambda}{2}\left\|\beta_1\right\|_2^2 \tag{1}$$

其中 $\left|r_i(\beta)\right|_{\varepsilon}=\max\left(\left|r_i(\beta)\right|-\varepsilon,0\right)$，目标函数在 $\left|r_i(\beta)\right|=\varepsilon$ 处不可微，且 $\left|r_i(\beta)\right|$ 在零点处不可微，难以直接

使用基于梯度的优化方法。为获得可微的近似形式，本文根据 Horowitz [5]、Pang 等[6]和 Chen 等[7]采用的核平滑技术，针对(1)式的目标函数构建一个平滑的损失函数。

首先引入符号函数

$$\operatorname{sgn}(r)=\begin{cases}-1, & r\le 0\\ 1, & r>0\end{cases}$$

则 $|r|=r\cdot\operatorname{sgn}(r)$，进一步令 $s_i(\beta)=|r_i(\beta)|-\varepsilon$，则 $|r_i(\beta)|_\varepsilon=\max(s_i(\beta),0)$。为平滑 $\max(s,0)$，定义光滑近似函数 $w_h(s)=sQ(s/h)$，其中带宽参数 $h>0$，并取 $Q(\cdot)$ 为对示性函数 $\mathbb{I}_{\{u\ge 0\}}$ 的平滑逼近(可视为某核函数积分形式)，具体定义为

$$Q(u)=\begin{cases}0, & u\le -1\\ \frac{1}{2}+\frac{15}{16}\left(u-\frac{2}{3}u^3+\frac{1}{5}u^5\right), & -1\le u\le 1\\ 1, & u\ge 1\end{cases}$$

当 $h\to 0$ 时，$Q(s/h)\to\mathbb{I}_{\{s\ge 0\}}$，其中 $\mathbb{I}_{\{s\ge 0\}}=\begin{cases}0, & s<0\\ 1, & s\ge 0\end{cases}$，从而 $sQ\left(\frac{s}{h}\right)\to\max(s,0)$，因此 SVR 的经验风险最小化目标函数可用如下光滑形式近似

$$L(\beta)=\frac{1}{n}\sum_{i=1}^{n}w_h(s_i(\beta))+\frac{\lambda}{2}\|\beta_1\|_2^2 \tag{2}$$

3. PerFL-SVR 算法

将 n 个样本的索引集合 $1,\cdots,n$ 划分为 L 个两两不交的子集 $\mathcal{M}_{l=1}^{L}$，约定第 l 个机器持有索引集合 $\mathcal{M}_l$ 对应的样本，$l=1,2,\cdots,L$，每个机器含有 m 个样本，$m=n/L$。即

$$\mathcal{M}_l=\{(x_{li},y_{li})\}_{i=1}^{m},l=1,\cdots,L。$$

考虑跨客户端数据异质性，假设每个机器对应一套线性模型

$$y_{li}=x_{li}^{\mathrm{T}}\beta_l^*+\epsilon_{li}$$

其中 $\beta_l^*=(\beta_{l1}^*,\beta_{l2}^*,\cdots,\beta_{lp}^*)$ 为第 l 个机器的真实回归系数，允许不同机器之间取值不同。误差项满足

$$E(\epsilon_{li}\mid x_{li})=0,\quad E(|\epsilon_{li}|^{1+\delta})<\infty$$

对某个 $(\delta>0)$ 成立。由于异质性，各机器协变量的协方差结构 $E(x_{li}x_{li}^{\top})$ 以及随机误差 ϵ_{li} 也可能存在差异。

为在保持个性化的同时利用跨机器共享信息，引入“逐坐标”的相似性惩罚。具体地，对任意坐标 $d\in[p]$，在不同机器之间对 $|\beta_{ld}-\beta_{l'd}|$ 施加成对惩罚，从而允许不同坐标具有不同的聚类结构；当所有坐标共享同一分组时，该结构退化为传统的整向量聚类联邦学习情形。相似性正则写为

$$\sum_{d=1}^{p}\sum_{1\le l<l'\le L}p_{\lambda_1}(|\beta_{ld}-\beta_{l'd}|),$$

其中 $p_{\lambda_1}(\cdot)$ 为后续指定的正则函数，$\lambda_1>0$ 控制“个性化–共享”的权衡。

为实现高维特征选择与稀疏恢复，对各机器系数再加入稀疏惩罚

$$\sum_{d=1}^{p}\sum_{l=1}^{L}p_{\lambda_2}(|\beta_{ld}|),$$

其中 $\lambda_2>0$。将所有机器参数向量化为 $\beta=vec[(\beta_1,\beta_2,\cdots,\beta_L)]\in R^{pL}$。为了后续对公式进行简单化，令：

$$\mathcal{H}(\beta)=\frac{1}{L}\sum_{l=1}^{L}\frac{1}{m}\sum_{i=1}^{m}w_h\left(s_{li}(\beta)\right)$$

综上，稀疏且稳健的个性化支持向量回归估计量定义为

$$\hat{\beta}=\arg\min_{\beta\in R^{pL}}\mathcal{H}(\beta)+\sum_{d=1}^{p}\sum_{1\le l<l'\le L}p_{\lambda_1}\left(\left|\beta_{ld}-\beta_{l'd}\right|\right)+\sum_{d=1}^{p}\sum_{l=1}^{L}p_{\lambda_2}\left(\left|\beta_{ld}\right|\right). \tag{3}$$

其中 $p_\lambda(\cdot)$ 表示正则化惩罚项，λ 为正则化参数向量。本文设 $\lambda=(\lambda_1,\lambda_2)$，且 $\lambda_1>0,\lambda_2>0$，分别控制相似性惩罚与稀疏惩罚的强度。为了实现稀疏恢复，本文采用了两种凹正则化函数：SCAD 和 MCP。这两种正则化项相比传统的 ℓ_1 正则化项，具有更好的稀疏恢复性能，并能够减少估计偏差。SCAD 正则化函数定义如下：

$$p_\lambda^{\mathrm{SCAD}}(x)=\begin{cases}\lambda|x|, & |x|\le\lambda\\ \dfrac{\omega\lambda|x|-x^2/2-\lambda^2/2}{\omega-1}, & \lambda<|x|\le\omega\lambda\\ \dfrac{\lambda^2(\omega+1)}{2}, & |x|>\omega\lambda\end{cases}$$

其中 $\omega>2$。MCP 正则化函数定义为：

$$p_\lambda^{\mathrm{MCP}}(x)=\begin{cases}\lambda|x|-\dfrac{x^2}{2\omega}, & |x|\le\omega\lambda\\ \dfrac{\lambda^2\omega}{2}, & |x|>\omega\lambda\end{cases}$$

其中 $\omega>1$。

为了衡量模型之间的差异性，引入了一个矩阵 $\mathbf{\Omega}=E\otimes I_p$，其中 $E=\left(e_i-e_j,i<j\right)^\top\in R^{\frac{L(L-1)}{2}\times L}$，$\{e_i\}\subset R^L$ 为标准基向量，I_p 是 $p\times p$ 的单位矩阵，$\otimes$ 是 Kronecker 积。对于任意正定矩阵 H，用 $|x|_H^2=x^\top Hx$ 来表示在 H 度量下的范数。对于一个 $p\times p$ 矩阵 Z，用 $\varsigma_{\min}(Z)$，$\varsigma_{\max}(Z)$ 表示 Z 的最小特征值和最大特征值。令 $I[\cdot]$ 为指示函数。

令 $A=\left[\Omega^\top,I_{p\times L}\right]^\top$ 且 $\delta=A\beta$，式(3)可以重新表述为一个约束优化问题：

$$\min_\beta\mathcal{H}(\beta)+h_\lambda(\delta),\ \text{s.t. } A\beta=\delta,$$

其中 $h_\lambda(\delta)=\sum_d p_{\lambda^*}\left(|\delta_d|\right)$ 为正则化项。其中，λ^* 如果 $d>pL(L-1)/2$，则为 λ_2，否则为 λ_1。

上述问题可以通过基于 ADMM (交替方向乘子法)算法的增广拉格朗日法来求解，增广拉格朗日法函数为

$$\mathcal{L}_\rho(\beta,\delta,\gamma)=\mathcal{H}(\beta)+h_\lambda(\delta)+\langle\gamma,A\beta-\delta\rangle+\frac{\rho}{2}|A\beta-\delta|_2^2,$$

其中，γ 是拉格朗日乘子，ρ 是惩罚参数。在具体的迭代中，β、δ 和 γ 可以按如下方式依次更新：

$$\beta^{(t+1)}=\arg\min_{\beta\mathcal{L}_\rho}\left(\beta,\delta^{(t)},\gamma^{(t)}\right) \tag{4}$$

$$\delta^{(t+1)}=\arg\min_{\delta\mathcal{L}_\rho}\left(\beta^{(t+1)},\delta,\gamma^{(t)}\right),$$

$$\gamma^{(t+1)}=\gamma^{(t)}+\rho\left(A\beta^{(t+1)}-\delta^{(t+1)}\right) \tag{5}$$

拉格朗日乘子的更新是直接的。

δ 的更新：参数 δ 通过 SCAD 和 MCP 定义的阈值算子以确定性的方式进行更新，具体方式如下所示。首先，定义软阈值规则为

$$ST(x,\lambda)=\operatorname{sgn}(x)(|x|-\lambda)I[|x|-\lambda>0]$$

令 $T_{\lambda,\rho}$ 表示阈值算子，定义为：

$$T_{\lambda,\rho}(y)=\arg\min_{x\in R}\left\{p_\lambda(|x|)+\frac{\rho}{2}(x-y)^2\right\},$$

其中 ρ 是常数。然后，SCAD 正则化的阈值算子定义为：

$$T_{\lambda,\rho}^{\mathrm{SCAD}}(y)=\begin{cases}ST(y,\lambda/\rho), & |y|\le\lambda+\dfrac{\lambda}{\rho}\\ \dfrac{(\omega\rho-\rho)ST\left(y,\dfrac{\omega\rho}{\omega\rho-\rho}\right)}{\omega\rho-\rho-1}, & \lambda+\dfrac{\lambda}{\rho}<|y|\le\lambda\omega\\ y, & |y|>\lambda\omega.\end{cases}$$

对于 MCP 正则化，阈值算子 $T_{\lambda,\rho}^{\mathrm{MCP}}(y)$ 定义为：

$$T_{\lambda,\rho}^{\mathrm{MCP}}(y)=\begin{cases}\dfrac{\omega\rho}{\omega\rho-1}ST(y,\lambda/\rho), & |y|\le\lambda\omega\\ y, & |y|>\lambda\omega\end{cases}$$

因此，$\delta^{(t+1)}$ 的第 d 元素更新为：

$$\delta_d^{(t+1)}=T_{\lambda,\rho}^{*}\left(A\beta^{(t+1)}+\rho^{-1}\gamma^{(t)}\right)\tag{6}$$

β 的更新：β 的更新并非一项简单的任务。通过利用平方损失，Ma 等(2019) [8]和 Yang、Yan 和 Huang (2019) [9]显式地求解了式(4)。因此，考虑了 ADMM 的线性化方法，并定义了如下的近似目标函数：

$$\begin{aligned}\hat{L}\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)=&\mathcal{H}\left(\beta^{(t)}\right)+\nabla\mathcal{H}\left(\beta^{(t)}\right)^{\top}\left(\beta-\beta^{(t)}\right)\\&+\frac{\nu}{2}\left|\beta-\beta^{(t)}\right|_H^2+\left\langle\gamma,A\beta-\delta\right\rangle+\frac{\rho}{2}\left|A\beta-\delta\right|_2^2,\end{aligned}$$

其中 $\nu>0$ 是正定矩阵 H 的一个参数。$\nabla\mathcal{H}(\beta)$ 来表示函数 $\mathcal{H}$ 对模型权重 β 的梯度。通过 $\hat{L}\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)$ 对 β 进行最小化，本文得到：

$$\beta^{(t+1)}=\left(\nu^{-1}H+\rho A^{\top}A\right)^{-1}\left(\nu^{-1}H\beta^{(t)}-\nabla\mathcal{H}\left(\beta^{(t)}\right)+\rho A^{\top}\delta^{(t)}-A^{\top}\gamma^{(t)}\right),$$

其中，$\nu^{-1}H+\rho A^{\top}A$ 是一个 $Lp\times Lp$ 矩阵。当数据单元数目或变量数目较多时，计算 $\nu^{-1}H+\rho A^{\top}A$ 是非常耗费计算和存储的。为了避免这个问题，本文使用 $H=rI-\rho\nu A^{\top}A$，并且 $r\ge\rho\nu\cdot\varsigma_{\max}\left(A^{\top}A\right)+1$，在该条件下 H 为正定矩阵。得到：

$$\beta^{(t+1)}=r^{-1}H\beta^{(t)}-r^{-1}\nu\left[\nabla\mathcal{H}\left(\beta^{(t)}\right)-\rho A^{\top}\delta^{(t)}+A^{\top}\gamma^{(t)}\right]\tag{7}$$

为了设计一个有效的联邦机器——服务器系统算法，本文考虑了两个方面：(a) 通信效率；(b) 系统异质性。为了实现通信效率，本文设计了最小化机器和服务器之间消息大小的方案。对于一个包含大量机器、并且具有不同计算能力和网络条件的联邦系统，服务器不可能在每次通信中与所有机器进行通信。

为了适应系统异质性，本文设计了一种允许每轮低机器参与率的算法。

本文设计了在联邦系统中传递模型权重的方案。更新公式(7)可以分解为：

$$\beta^{(t+1)} = \tilde{\beta}^{(t)} - r^{-1}\nu\nabla\mathcal{H}\left(\beta^{(t)}\right),$$

其中，

$$\tilde{\beta}^{(t)} = r^{-1}H\beta^{(t)} - r^{-1}\nu\left[-\rho A^{\top}\delta^{(t)} + A^{\top}\gamma^{(t)}\right] \tag{8}$$

需要注意的是，$\delta^{(t)}$ 和 $\gamma^{(t)}$ 被用于权重相似性学习，因此它们被存储在服务器中。因此，在上传 $\beta^{(t)}$ 到服务器后，$\tilde{\beta}^{(t)}$ 可以在服务器上计算。梯度 $\nabla\mathcal{H}\left(\beta^{(t)}\right)$ 需要本地数据，因此它应在机器上计算。对于第 l 个机器，其模型权重位于 $\beta^{(t)}$ 的第 $1+(l-1)p$ 到 lp 坐标上，即：

$$\beta_l^{(t)} = \left(\beta_{1+(l-1)p}^{(t)}, \cdots, \beta_{lp}^{(t)}\right)^{\top}.$$

然后，模型权重可以在本地更新为：

$$\beta_l^{(t+1)} = \tilde{\beta}_l^{(t)} - r^{-1}\nu\nabla\mathcal{H}_l\left(\beta_l^{(t)}\right), \tag{9}$$

其中 $\mathcal{H}_l\left(\beta_l^{(t)}\right) = \frac{1}{m}\sum_{i=1}^{m} w_h\left(s_{li}\left(\beta\right)\right)$。

为了适应系统异质性，本文在每轮通信中允许部分机器参与。参与率取决于系统条件，例如机器数量、计算能力和网络状况。如果机器数量较少，本文可以允许所有机器参与；而当机器数量较多且网络状况较差时，本文将选择部分机器进行更新，以提高通信效率。本文通过随机采样机器的方式来处理系统异质性，确保每次与服务器的通信时，都能有效地选择参与的机器。

本文更新机器的个性化模型权重，并将所有信息汇总到服务器，以研究本地权重之间的相似性。服务器端的计算非常简单且快速，因为它们是一步更新，只需进行矩阵乘法即可。

PerFL-SVR 为第 t 轮与服务器的通信选择 $\left|S^{(t)}\right|$ 个机器。服务器持有并更新 $\delta^{(t)}$、$\gamma^{(t)}$ 和 $\tilde{\beta}^{(t)}$。

更新过程：机器上传本地模型 $\beta_l^{(t)}$ 并下载所需的 $\tilde{\beta}^{(t)}$ 信息，以进行下次本地更新。机器使用 $\tilde{\beta}^{(t)}$ 和本地梯度计算 $\beta_l^{(t+1)}$。注意，$\beta_l^{(t+1)} = \left(\beta_1^{(t+1)}, \cdots, \beta_L^{(t+1)}\right)^{\top}$。对于选择的机器 $l \in S^{(t)}$，系数 $\beta_l^{(t)}$ 被更新为 $\beta_l^{(t+1)}$，而对于未选择的机器 $l' \notin S^{(t)}$，本文有 $\beta_l^{(t+1)} = \beta_{l'}^{(t)}$。因此，本文只在算法中具体说明了本地权重的更新规则。初始估计量 $\beta_l^{(0)}$ 是使用本地数据计算的：

$$\beta_l^{(0)} = \arg\min_{\beta \in R^p} \mathcal{H}_l\left(\beta\right) + \sum_{j=1}^{p} \lambda_2\left(\left|\beta_d\right|\right). \tag{10}$$

4. 数值模拟

本文考虑 L 个数据单元，每个数据单元从线性模型生成数据

$$y_l = X_l\beta_l^{*} + \epsilon_l, ,$$

其中 $l = 1, \cdots, L$。$X_l \in R^{m\times p}$ 为设计矩阵，$\beta_l^{*} \in R^p$ 为客户端 l 的真实回归系数，$\epsilon_l \in R^m$ 为随机误差向量。本文令设计矩阵的行向量 x_{li} 独立同分布于多元正态分布 $\mathcal{N}\left(0, \Sigma\right)$，其中协方差矩阵 Σ 采用等相关结构：

$$\Sigma_{jj} = 1, \quad \Sigma_{jk} = 0.3\left(j \neq k\right),$$

即 $\mathrm{cov}\left(x_{lij}, x_{lik}\right) = 0.3\ \left(j \neq k\right)$。$x_{lij}$ 表示客户端 l 的第 i 个样本在第 j 个特征上的取值。为刻画跨客户端异质性，将 L 个客户端划分为两组：

第一组包含 $\{1,\cdots,\lfloor L/2\rfloor\}$ 个客户端，第二组包含剩余客户端。两组对应的真实系数分别设为

$$\beta_A^* = (1,3,0,\cdots,0)^\top,$$

$$\beta_B^* = (2,3,0,\cdots,0)^\top,$$

从而两组在前两个坐标上存在差异，其余坐标为零以体现稀疏结构。进一步假设各客户端样本量相同。

对线性化 ADMM 中的调参 $\nu, \rho^{(0)}$ 与秩参数 r 的选取，本文参考 ADMM 的常用设定(Ma and Huang 等[10]；Lu 等[11])，固定 $\nu = 1$、$\rho^{(0)} = 2$。秩参数 r 取为 $r = \rho^{(0)}\nu \cdot \varsigma_{\max}\left(A^\top A\right) + \max\{\nu/2, 1\}$ 中 $\varsigma_{\max}(\cdot)$ 最大特征值。对于 SCAD 与 MCP 正则化的形状参数(本文记为 ω)，取 $\omega = 3$。

为系统评估所提出算法的性能，本文设置三组实验：

(a) 重尾噪声鲁棒性：在重尾或含离群点的噪声设定下评估算法的稳健性；

(b) 调参与收敛性分析：考察调参对收敛行为与性能的影响，包括参数 ε 以及正则化参数 λ_1, λ_2 等；

(c) 部分参与机制影响：研究每轮通信仅部分客户端参与时，对性能与通信效率的影响。

调参策略。正则化参数采用网格搜索，并使用信息准则(如 BIC)选择最优组合。为降低搜索维度，令 $\lambda_1 = r_\lambda \lambda_2$，并取 $r_\lambda \in \{3,6,9\}$ 的离散候选集合；在每个 r_λ 下，对 λ_2 做网格搜索并计算 BIC，最终选择使 BIC 最小的 $\left(\lambda_1^*, \lambda_2^*\right)$。

本文提出的 PerFL-SVR 旨在应对现实场景中常见的重尾噪声与极端值污染。本组实验用于比较在不同噪声分布下，PerFL-SVR 相较于基于 ℓ_2 损失的方法的实际表现差异。

对比方法。除 PerFL-SVR 外，我们引入基于 ℓ_2 损失的个性化联邦学习方法作为对照，记为 PerFL-ℓ_2，其更新过程与 PerFL-SVR 类似，但将平滑支持向量回归损失的梯度替换为平方损失的梯度，从而不具备对重尾噪声的鲁棒性。

噪声分布设置。柯西噪声：第一组客户端 $\epsilon_{li} \sim \text{Cauchy}(1)$，第二组客户端 $\epsilon_{li} \sim \text{Cauchy}(1.5)$。

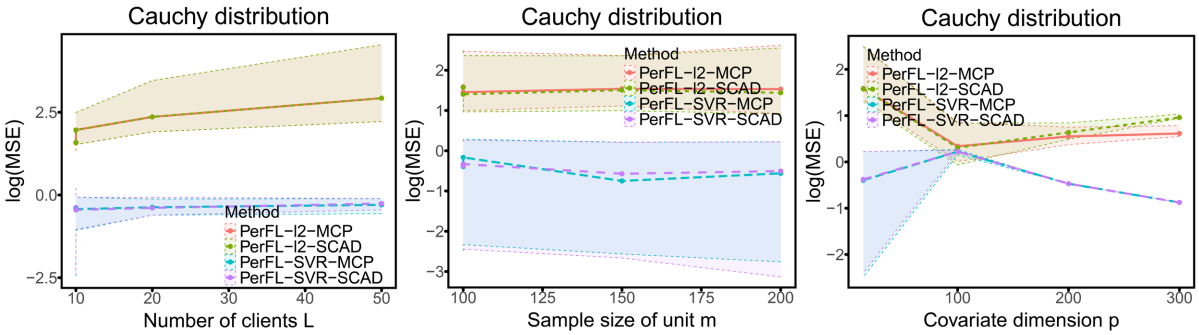

Figure 1. Performance comparison of algorithms under the Cauchy distribution
图 1. Cauchy 分布下算法性能比较

图 1 中展示了噪声分布 Cauchy 下，不同方法的均方误差对数值 log(MSE)随客户端数 L，单客户端样本量 m，协变量维度 p 的变化规律。图中阴影带表示模拟结果的四分位距(IQR)，即覆盖中间 50%重复实验的区间，用以刻画方法在重复实验下的稳健性与波动程度。

总体而言，对于客户端数，PerFL-SVR (MCP/SCAD)在噪声分布下均表现出更低的 log(MSE)及更窄的 IQR，相比之下，基于平方损失的 PerFL-ℓ_2 在重尾噪声下误差水平明显更高，且随 L 增大波动更明显。对于单客户端样本量，PerFL-SVR (MCP/SCAD)相较 PerFL-ℓ_2 显著降低 log(MSE)并缩小 IQR，且随 m 增加呈现更稳定的误差下降趋势，对于协变量维度，PerFL-ℓ_2 的 log(MSE)随 p 增大呈明显上升趋势，表明

高维场景下平方损失对极端观测更为敏感，误差累积效应更突出；同时其误差带宽也相对更大，反映估计不稳定性增强。相比之下，PerFL-SVR 的误差曲线整体更低，且随 p 增加并未出现同等幅度的劣化，说明 SVR 型稳健损失能够有效抑制重尾噪声下极端样本对高维估计的破坏，从而在高维重尾环境中保持较优的误差水平与稳定性。

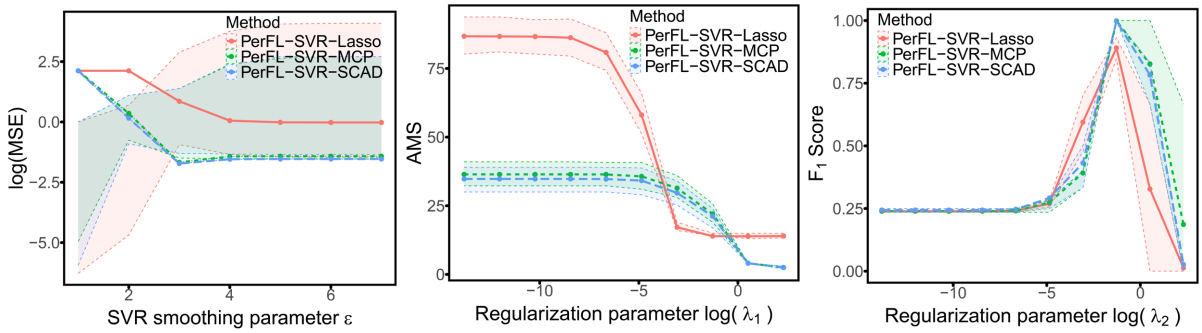

Figure 2. Sensitivity analysis of PerFL-SVR parameters
图 2. PerFL-SVR 参数敏感性分析

图 2 展示了 PerFL-SVR 在不同调参下的表现，其中第一行(左)图考察 SVR 平滑参数 ε 对估计误差 log(MSE)的影响，第一行(右)图考察融合正则参数 λ_1 对分组结构复杂度的影响，第二行的图考察稀疏正则参数 λ_2 对稀疏恢复性能的影响。总体来看，模型性能对调参较为敏感，合适的 $\varepsilon, \lambda_1, \lambda_2$ 能显著改善拟合精度与结构恢复，而不恰当的取值会导致误差上升或结构恢复退化，这与“个性化－共享”全衡与稀疏正则强度共同决定估计性质的直觉一致。

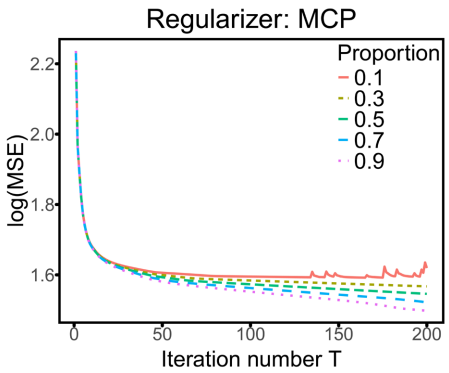

Figure 3. Convergence behavior under different communication ratios
图 3. 不同通信比例下的收敛表现

图 3 中给出了在 MCP 正则下，log(MSE)随迭代次数 T 的变化曲线，不同颜色/线型对应不同的参与比例(0.1, 0.3, 0.5, 0.7, 0.9)。

从图中可以观察到，随着迭代进行，各参与比例下的 log(MSE)均呈现明显下降趋势，并在一定迭代轮数后趋于平稳，表明所提出方法在部分参与机制下仍能有效收敛。进一步比较不同参与比例可见：参与比例越高，误差下降越快、收敛所需迭代轮数越少。例如，高参与比例的曲线在较少迭代内即可接近稳定区间，而低参与比例下降更缓慢，达到相近误差水平需要更多迭代轮数。

上述现象的原因在于：在每轮更新中，参与客户端数量越多，服务器端能够汇聚更充分的本地信息，从而更快修正全局共享量并推动模型向最优解方向迭代；相反，当参与比例较低时，每轮有效信息更新量减少，导致收敛速度在“迭代轮数”尺度上变慢。需要强调的是，尽管高参与比例在迭代轮数意义下收敛更快，低参与比例方案在实际系统中可能具有更低的通信开销，因此二者体现了收敛速度与通信成本之间的权衡。

5. 结论

基于本文的理论推导与数值模拟分析，可以得出以下结论：

该方法利用平滑支持向量回归损失函数，并引入 SCAD 或 MCP 稀疏融合惩罚项，在实现高维特征稀疏恢复的同时，兼顾了各个机器节点的模型个性化与全局信息的共享。在面对重尾柯西噪声及极端值污染时，PerFL-SVR 展现出显著的优势。相较于传统的基于 ℓ_2 损失的方法，PerFL-SVR 能够保持更低的均方误差和更窄的误差波动范围，且在协变量维度增加的高维场景下，依然能够有效抑制极端样本对模型估计的破坏。

6. 展望

当前方法本质上仍建立在线性回归框架下，模型本身对复杂非线性关系的刻画能力仍然有限，因此在存在显著非线性异质性的实际任务中，预测性能可能受到约束。再次，尽管平滑化有助于降低优化难度，但在加入稀疏惩罚项和融合惩罚项后，整体目标函数依然是复杂的非凸优化问题，现有基于线性化 ADMM 的求解策略在大规模联邦环境下仍可能面临通信代价高、参数调节敏感和收敛速度受限等问题。基于上述局限，未来工作可以从以下方向展开，可将该方法推广到非线性个性化联邦学习模型，例如结合核技巧的 SVR 框架或与神经网络模型相结合，以增强对复杂异质数据结构的表达能力。在计算层面，可以探索更高效的优化策略，如随机化近端方法、加速型一阶分布式算法或通信压缩机制，以降低联邦训练中的计算与通信负担。

参考文献

[1] Vapnik, V. (2013) The Nature of Statistical Learning Theory. Springer Science & Business Media.

[2] McMahan, B., *et al*. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*), Volume 54, 1273-1282.

[3] Liu, H., Liu, S., Ji, J., Lin, Q., Chen, J. and Tan, K.C. (2024) Personalized Federated Learning with Enhanced Implicit Generalization. 2024 *International Joint Conference on Neural Networks* (*IJCNN*), Yokohama, 30 June-5 July 2024, 1-8. https://doi.org/10.1109/ijcnn60899.2024.10651164

[4] Sang, T., Chu, Z., Xuan, J., Zhang, X. and Li, X. (2025) Personalized Federated Learning in One-Shot: A Method for Heterogeneous Data Scenarios. *IEEE Internet of Things Journal*, **12**, 40415-40425. https://doi.org/10.1109/jiot.2025.3589414

[5] Horowitz, J.L. (1998) Bootstrap Methods for Median Regression Models. *Econometrica*, **66**, 1327-1351. https://doi.org/10.2307/2999619

[6] Pang, L., Lu, W. and Wang, H.J. (2012) Variance Estimation in Censored Quantile Regression via Induced Smoothing. *Computational Statistics & Data Analysis*, **56**, 785-796. https://doi.org/10.1016/j.csda.2010.10.018

[7] Chen, X., Liu, W. and Zhang, Y. (2019) Quantile Regression under Memory Constraint. *The Annals of Statistics*, **47**, 3244-3273. https://doi.org/10.1214/18-aos1777

[8] Ma, S., Huang, J., Zhang, Z. and Liu, M. (2020) Exploration of Heterogeneous Treatment Effects via Concave Fusion. *The International Journal of Biostatistics*, **16**, Article ID: 20180026. https://doi.org/10.1515/ijb-2018-0026

[9] Huang, F., Chen, S. and Huang, H. (2019) Faster Stochastic Alternating Direction Method of Multipliers for Nonconvex Optimization. *International Conference on Machine Learning*. *PMLR*, Long Beach, 9-15 June 2019, 2839-2848.

[10] Ma, S. and Huang, J. (2017) A Concave Pairwise Fusion Approach to Subgroup Analysis. *Journal of the American*

Statistical Association, **112**, 410-423. https://doi.org/10.1080/01621459.2016.1148039

[11] Lu, S., Lee, J., Razaviyayn, M. and Hong, M. (2021) Linearized ADMM Converges to Second-Order Stationary Points for Non-Convex Problems. *IEEE Transactions on Signal Processing*, **69**, 4859-4874. https://doi.org/10.1109/tsp.2021.3100976

附录

为建立算法的收敛性，下面给出若干基本假设。

假设 1. 函数 $\mathcal{H}(\beta)$ 的梯度是 Lipschitz 连续的，即存在常数 $\mu>0$，使得对任意 $\beta_1,\beta_2\in\mathbb{R}^{Lp}$，有

$$\left\|\nabla\mathcal{H}(\beta_1)-\nabla\mathcal{H}(\beta_2)\right\|_2\le\mu\left\|\beta_1-\beta_2\right\|_2.$$

假设 2. 定义

$$h_\lambda(\delta)=\sum_{j=1}^{Lp}p_\lambda\left(\left|\delta_j\right|\right).$$

正则化函数 $p_\lambda(\cdot)$ 满足以下条件：

1) $p_\lambda(\cdot)$ 关于原点对称，并在 $[0,+\infty)$ 上单调不减；
2) $p_\lambda(0)=0$；
3) $p'_\lambda(t)$ 除有限个点外存在且连续；
4) 存在常数 $c_\lambda>0$，使得对任意 $d\in\partial h_\lambda(\delta)$，均有 $\|d\|_2^2\le c_\lambda$；
5) 存在常数 $\kappa>0$，使得 $p_{\lambda,\kappa}(t):=p_\lambda(t)+\frac{\kappa}{2}t^2$ 为凸函数。

假设 3. 平滑支持向量回归经验风险函数 $\mathcal{H}(\beta)$ 满足假设 1；同时协变量满足

$$\sup_{l,i}\left\|\boldsymbol{X}_{li}\right\|_2^2<\infty.$$

假设 4. 矩阵 A 满列秩，且存在常数 $c_1,c_2>0$，满足

$$0<c_1\le\varsigma_{\min}\left(A^\top A\right)\le\varsigma_{\max}\left(A^\top A\right)\le c_2<\infty.$$

假设 5. 在第 t 轮，客户端参与集合记为 $S_t\subseteq\{1,\cdots,L\}$。设客户端 l 在第 t 轮被选中的概率为 $p_{l,t}$，并存在常数 $p>0$，使得对任意 l,t，有

$$p_{l,t}\ge p>0.$$

下面首先证明 β 更新的增广拉格朗日下降性质。

引理 1 (更新 β 的增广拉格朗日下降性质)。在假设 1 和假设 4 成立的条件下，设

$$H=rI-\rho^{(t)}\nu A^\top A.$$

若 $\beta^{(t+1)}$ 由上述更新公式生成，并且

$$r>\rho^{(t)}\nu\varsigma_{\max}\left(A^\top A\right)+\max\left(\frac{\nu\mu}{2},1\right),$$

则有

$$\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)},\delta^{(t)},\gamma^{(t)}\right)-\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)\le-c^{(t)}\|\beta^{(t+1)}-\beta^{(t)}\|_2^2,$$

其中

$$c^{(t)}=\frac{\varsigma_{\min}(H)}{\nu}+\frac{\rho^{(t)}\varsigma_{\min}\left(A^\top A\right)}{2}-\frac{\mu}{2}>0.$$

证明：由算法构造可知，对于每个参与客户端 $l\in S_t$，$\beta_l^{(t+1)}$ 是如下线性化增广拉格朗日子问题的极小点：

$$\hat{\mathcal{L}}\left(\beta,\delta^{(t)},\gamma^{(t)}\right)=\mathcal{H}\left(\beta^{(t)}\right)+\nabla\mathcal{H}\left(\beta^{(t)}\right)^{\top}\left(\beta-\beta^{(t)}\right)+\frac{1}{2\nu}\left\|\beta-\beta^{(t)}\right\|_H^2$$
$$+\left\langle\gamma^{(t)},A\beta-\delta^{(t)}\right\rangle+\frac{\rho^{(t)}}{2}\left\|A\beta-\delta^{(t)}\right\|_2^2.$$

若 $l\notin S_t$，则按算法定义有

$$\beta_l^{(t+1)}=\beta_l^{(t)}.$$

因此，整体上有一阶最优性条件

$$0=\nabla\mathcal{H}\left(\beta^{(t)}\right)+\nu^{-1}H\left(\beta^{(t+1)}-\beta^{(t)}\right)+A^{\top}\gamma^{(t)}+\rho^{(t)}A^{\top}\left(A\beta^{(t+1)}-\delta^{(t)}\right)\tag{11}$$

将式(11)两边左乘$\left(\beta^{(t)}-\beta^{(t+1)}\right)^{\top}$，得到

$$\begin{aligned}&\left\langle\nabla\mathcal{H}\left(\beta^{(t)}\right),\beta^{(t)}-\beta^{(t+1)}\right\rangle-\nu^{(-1)}\left|\beta^{(t+1)}-\beta^{(t)}\right|_H^2\\&+\left\langle\gamma^{(t)},A\beta^{(t)}-A\beta^{(t+1)}\right\rangle+\rho^{(t)}\left\langle A\beta^{(t+1)}-\delta^{(t)},A\beta^{(t)}-A\beta^{(t+1)}\right\rangle=0\end{aligned}\tag{12}$$

另一方面，由假设 1 以及下降引理，有

$$\mathcal{H}\left(\beta^{(t+1)}\right)\le\mathcal{H}\left(\beta^{(t)}\right)+\left\langle\nabla\mathcal{H}\left(\beta^{(t)}\right),\beta^{(t+1)}-\beta^{(t)}\right\rangle+\frac{\mu}{2}\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2.\tag{13}$$

结合式(12)、式(13)以及恒等式

$$\left(a-b\right)^{\top}\left(b-c\right)=\frac{1}{2}\left(\left\|a-c\right\|_2^2-\left\|a-b\right\|_2^2-\left\|b-c\right\|_2^2\right),$$

可得

$$\begin{aligned}&\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)},\delta^{(t)},\gamma^{(t)}\right)-\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)\\&\le-\nu^{(-1)}\left|\beta^{(t+1)}-\beta^{(t)}\right|_H^2+\frac{\mu}{2}\left|\beta^{(t+1)}-\beta^{(t)}\right|_2^2-\frac{\rho^{(t)}}{2}\left|A\left(\beta^{(t+1)}-\beta^{(t)}\right)\right|_2^2\end{aligned}$$

再由

$$\left|\beta^{(t+1)}-\beta^{(t)}\right|_H^2\ge\varsigma_{\min}\left(H\right)\left|\beta^{(t+1)}-\beta^{(t)}\right|_2^2$$

和

$$\left\|\left(\beta^{(t+1)}-\beta^{(t)}\right)\right\|_2^2\ge\varsigma_{\min}\left(A^{\top}A\right)\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2,$$

即可推出

$$\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)},\delta^{(t)},\gamma^{(t)}\right)-\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)\le-c^{(t)}\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2.$$

引理得证。

下面说明对偶变量序列与增广拉格朗日函数序列的有界性。

引理 2 在假设 1~2 条件下，设 $\rho^{(t+1)}=\alpha\rho^{(t)}$，其中 $\alpha>1$。若 $\left\{\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)\right\}_{t=1}^{\infty}$ 由算法 3 生成，则序列 $\left\{\gamma^{(t)}\right\}$ 与 $\left\{\mathcal{L}^{(t)}\right\}$ 均有界，其中

$$\mathcal{L}^{(t)}:=\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right).$$

证明：由 δ 的更新最优性条件可得

$$0 \in \partial h_\lambda\left(\delta^{(t+1)}\right) - \gamma^{(t)} - \rho^{(t)}\left(A\beta^{(t+1)} - \delta^{(t+1)}\right). \tag{14}$$

再结合乘子更新公式

$$\gamma^{(t+1)} = \gamma^{(t)} + \rho^{(t)}\left(A\beta^{(t+1)} - \delta^{(t+1)}\right), \tag{15}$$

可得

$$\gamma^{(t+1)} \in \partial h_\lambda\left(\delta^{(t+1)}\right).$$

于是由假设 2，有

$$\left\|\gamma^{(t+1)}\right\|_2^2 \le c_\lambda,$$

故 $\gamma^{(t)}$ 有界。

下面证明 $\mathcal{L}^{(t)}$ 有界。注意到

$$\mathcal{L}_{\rho^{(t+1)}}\left(\beta^{(t+1)}, \delta^{(t+1)}, \gamma^{(t+1)}\right) - \mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)}, \delta^{(t+1)}, \gamma^{(t+1)}\right) = \frac{\rho^{(t+1)} - \rho^{(t)}}{2\left(\rho^{(t)}\right)^2}\left\|\gamma^{(t+1)} - \gamma^{(t)}\right\|_2^2, \tag{16}$$

以及

$$\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)}, \delta^{(t+1)}, \gamma^{(t+1)}\right) - \mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)}, \delta^{(t+1)}, \gamma^{(t)}\right) = \frac{1}{\rho^{(t)}}\left\|\gamma^{(t+1)} - \gamma^{(t)}\right\|_2^2. \tag{17}$$

又由于 $\delta^{(t+1)}$ 是 δ -子问题的极小点，有

$$\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)}, \delta^{(t+1)}, \gamma^{(t)}\right) - \mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)}, \delta^{(t)}, \gamma^{(t)}\right) \le 0. \tag{18}$$

再由引理 1，

$$\mathcal{L}_{\rho^{(t)}}\left(\beta^{(t+1)}, \delta^{(t)}, \gamma^{(t)}\right) - \mathcal{L}_{\rho^{(t)}}\left(\beta^{(t)}, \delta^{(t)}, \gamma^{(t)}\right) \le -c^{(t)}\left\|\beta^{(t+1)} - \beta^{(t)}\right\|_2^2. \tag{19}$$

将式(16)~(19)合并，得到

$$\begin{aligned}\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)} &\le \frac{\rho^{(t+1)} + \rho^{(t)}}{2\left(\rho^{(t)}\right)^2}\left\|\gamma^{(t+1)} - \gamma^{(t)}\right\|_2^2 - c^{(t)}\left\|\beta^{(t+1)} - \beta^{(t)}\right\|_2^2 \\ &\le \frac{\rho^{(t+1)} + \rho^{(t)}}{2\left(\rho^{(t)}\right)^2}c_\lambda - c^{(t)}\left\|\beta^{(t+1)} - \beta^{(t)}\right\|_2^2.\end{aligned} \tag{20}$$

对式(20)从 $t=0$ 到 T 求和，可得

$$\mathcal{L}^{(T+1)} - \mathcal{L}^{(0)} \le \sum_{t=0}^{T}\frac{\rho^{(t+1)} + \rho^{(t)}}{2\left(\rho^{(t)}\right)^2}c_\lambda - \sum_{t=0}^{T}c^{(t)}\left\|\beta^{(t+1)} - \beta^{(t)}\right\|_2^2. \tag{21}$$

由于 $\rho^{(t+1)} = \alpha\rho^{(t)}$，且 $\alpha > 1$，有

$$\sum_{t=0}^{\infty}\frac{\rho^{(t+1)} + \rho^{(t)}}{2\left(\rho^{(t)}\right)^2} = \frac{\alpha+1}{2\rho^{(0)}}\sum_{t=0}^{\infty}\alpha^{-t} < \infty.$$

因此$\left\{\mathcal{L}^{(t)}\right\}$上有界。另一方面，由

$$\mathcal{L}^{(t)}=\mathcal{H}\left(\beta^{(t)}\right)+h_{\lambda}\left(\delta^{(t)}\right)+\left\langle\gamma^{(t)},A\beta^{(t)}-\delta^{(t)}\right\rangle+\frac{\rho^{(t)}}{2}\left\|A\beta^{(t)}-\delta^{(t)}\right\|_2^2$$

以及

$$\left\langle\gamma^{(t)},A\beta^{(t)}-\delta^{(t)}\right\rangle=\frac{1}{\rho^{(t)}}\left\langle\gamma^{(t)},\gamma^{(t+1)}-\gamma^{(t)}\right\rangle,$$

再结合$\left\{\gamma^{(t)}\right\}$有界，可得$\left\{\mathcal{L}^{(t)}\right\}$下有界，从而$\left\{\mathcal{L}^{(t)}\right\}$有界。引理得证。 □

下面给出$\left\{\delta^{(t)}\right\}$有界性的结论。

命题 1 若当$\|\delta\|_2\to\infty$时有$h_{\lambda}(\delta)\to\infty$，则序列$\left\{\delta^{(t)}\right\}$有界。

证明：由引理 2 可知$\mathcal{L}^{(t)}$有界。由于增广拉格朗日函数中的其余各项均有界或可控，故$h_{\lambda}\left(\delta^{(t)}\right)$必有上界。若$\delta^{(t)}$无界，则由$h_{\lambda}(\delta)\to\infty$可知$h_{\lambda}\left(\delta^{(t)}\right)\to\infty$，与其有上界矛盾。因此$\delta^{(t)}$有界。

至此，可以给出算法的主收敛结论。

定理 1 (算法 1 的收敛性) 在假设 1~5 条件下，设$\rho^{(t+1)}=\alpha\rho^{(t)}$且$\alpha>1$。由算法生成的序列

$$\left\{\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)\right\}_{t=0}^{\infty}$$

是有界的。进一步地，存在一个收敛子序列

$$\left\{\left(\beta^{(t_k)},\delta^{(t_k)},\gamma^{(t_k)}\right)\right\}_{k=1}^{\infty},$$

其极限点$\left(\beta^*,\delta^*,\gamma^*\right)$满足约束优化问题的 KKT 条件。此外，还有

$$\lim_{t\to\infty}\left(\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2+\left\|\delta^{(t+1)}-\delta^{(t)}\right\|_2^2\right)=0.$$

若进一步考虑客户端随机参与机制，则有

$$\sum_{t=0}^{\infty}\mathbb{E}\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2<\infty.$$

证明：由命题 1，$\delta^{(t)}$有界；由引理 2，$\gamma^{(t)}$与$\mathcal{L}^{(t)}$有界。再由乘子更新公式(15)，

$$A\beta^{(t+1)}-\delta^{(t+1)}=\frac{1}{\rho^{(t)}}\left(\gamma^{(t+1)}-\gamma^{(t)}\right).$$

由于$\rho^{(t)}\to\infty$且$\gamma^{(t)}$有界，故

$$\left\|A\beta^{(t+1)}-\delta^{(t+1)}\right\|_2\to 0.$$

于是$A\beta^{(t)}$有界。又因为A满列秩，故$\beta^{(t)}$有界，从而

$$\left\{\left(\beta^{(t)},\delta^{(t)},\gamma^{(t)}\right)\right\}$$

整体有界。

接下来，由式(21)以及$\mathcal{L}^{(t)}$有界可得

$$0\le\sum_{t=0}^{\infty}c^{(t)}\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2<\infty.$$

因此

$$\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2 \to 0.$$

又由恒等式

$$\delta^{(t+1)}-\delta^{(t)}=\left(A\beta^{(t)}-\delta^{(t)}\right)-\left(A\beta^{(t+1)}-\delta^{(t+1)}\right)+A\left(\beta^{(t+1)}-\beta^{(t)}\right),$$

结合

$$\left\|A\beta^{(t)}-\delta^{(t)}\right\|_2 \to 0,\ \left\|A\beta^{(t+1)}-\delta^{(t+1)}\right\|_2 \to 0,\ \left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2 \to 0,$$

可知

$$\left\|\delta^{(t+1)}-\delta^{(t)}\right\|_2 \to 0.$$

由于整个序列有界，根据 Bolzano-Weierstrass 定理，存在收敛子序列

$$\left(\beta^{(t_k)},\delta^{(t_k)},\gamma^{(t_k)}\right)\to\left(\beta^*,\delta^*,\gamma^*\right).$$

下面证明该极限点满足 KKT 条件。

由 β 更新的一阶最优性条件(11)，沿收敛子序列取极限，并利用

$$\beta^{(t+1)}-\beta^{(t)}\to 0,\ A\beta^{(t+1)}-\delta^{(t)}\to 0,$$

可得

$$0=\nabla\mathcal{H}\left(\beta^*\right)+A^{\top}\gamma^*.$$

另一方面，由 δ 更新的最优性条件(14)，沿收敛子序列取极限可得

$$0\in\partial h_{\lambda}\left(\delta^*\right)-\gamma^*.$$

最后，由原始残差收敛于零，

$$A\beta^*-\delta^*=0.$$

因此，$\left(\beta^*,\delta^*,\gamma^*\right)$满足约束优化问题的 KKT 条件。

下面证明随机客户端参与下的期望平方可和性质。记 $\overline{\beta}^{(t+1)}$ 为第 t 轮在所有客户端均参与时由 β-子问题得到的虚拟更新。则式(20)可等价写为

$$\mathcal{L}^{(t+1)}-\mathcal{L}^{(t)}\le\frac{\rho^{(t+1)}+\rho^{(t)}}{2\left(\rho^{(t)}\right)^2}c_{\lambda}-c^{(t)}\sum_{l\in S_t}\left\|\overline{\beta}_l^{(t+1)}-\beta_l^{(t)}\right\|_2^2. \tag{22}$$

对客户端抽样取期望，结合假设 5，可得

$$\mathbb{E}\left[\mathcal{L}^{(t+1)}-\mathcal{L}^{(t)}\right]\le\frac{\rho^{(t+1)}+\rho^{(t)}}{2\left(\rho^{(t)}\right)^2}c_{\lambda}-c^{(t)}p\left\|\overline{\beta}^{(t+1)}-\beta^{(t)}\right\|_2^2.$$

又由于 $c^{(t)}\ge c^{(0)}$，且

$$\left\|\overline{\beta}^{(t+1)}-\beta^{(t)}\right\|_2^2\ge\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2,$$

故

$$\mathbb{E}\left[\mathcal{L}^{(t+1)}-\mathcal{L}^{(t)}\right]\le\frac{\rho^{(t+1)}+\rho^{(t)}}{2\left(\rho^{(t)}\right)^2}c_{\lambda}-c^{(0)}p\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2.$$

对上式从$t=0$到T求和并整理，得到

$$c^{(0)}p\sum_{t=0}^{T}\mathbb{E}\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2\le\mathcal{L}^{(0)}+\sum_{t=0}^{T}\frac{\rho^{(t+1)}+\rho^{(t)}}{2\left(\rho^{(t)}\right)^2}c_\lambda.$$

令$T\to\infty$，由于

$$\sum_{t=0}^{\infty}\frac{\rho^{(t+1)}+\rho^{(t)}}{2\left(\rho^{(t)}\right)^2}<\infty,$$

故有

$$\sum_{t=0}^{\infty}\mathbb{E}\left\|\beta^{(t+1)}-\beta^{(t)}\right\|_2^2<\infty.$$

定理得证。