The Aggregation–Heterogeneity Trade-off in Federated Learning

Xuyang Zhao XUYANGZHAO@PKU.EDU.CN
Huiyuan Wang HUIYUAN.WANG@PKU.EDU.CN
Wei Lin WEILIN@MATH.PKU.EDU.CN
School of Mathematical Sciences and Center for Statistical Science, Peking University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Conventional wisdom in machine learning holds that the more data you train your model on, the better the model can perform. Accordingly, a plethora of federated learning methods have been developed to aggregate as many local samples as possible. Contrary to this belief, this paper shows that aggregation of more data is not necessarily beneficial in the presence of heterogeneity, and reveals a fundamental trade-off between aggregation and heterogeneity in federated learning. We consider a general family of weighted M-estimators that interpolate between FedAvg and the local estimator, in which an aggregation rule is determined by the weights of local samples. We derive an upper bound for the estimation error of the weighted M-estimators, which decomposes into a bias term induced by heterogeneity and a variance term influenced by aggregation. A measure of heterogeneity, the federated smoothness β, is introduced to simplify the general result. As an important consequence, the optimal aggregation rule for each local device is to aggregate only its $\lfloor K^{2\beta/(2\beta+1)}/(n/\sigma^2)^{1/(2\beta+1)} \rfloor \vee 1$ closest neighbors among the K devices, where n is the local sample size and σ^2 is the noise variance. Moreover, we show that our estimator, termed FedKNN, attains the minimax optimal rate over a certain parameter space characterized by β. This optimal procedure depends crucially on the neighboring structure among devices in terms of the proximity of local parameters. Finally, we prove that without such prior knowledge no estimator can achieve a convergence rate faster than $O(\sigma^2/n)$ and hence adaptation is impossible.

Keywords: federated learning, optimal aggregation, k-nearest neighbors

1. Introduction

Federated learning is an emerging machine learning paradigm in which models are trained on data that are locally possessed by individual devices (McMahan et al., 2017; Li et al., 2020a). From the statistical perspective, one major advantage of federated learning is that data from other devices can be *aggregated* to produce improved results for each device's local task. To be specific, we consider Federated Averaging (FedAvg) (McMahan et al., 2017), a baseline algorithm for federated learning. For the task of parameter estimation, FedAvg outputs a single global parameter estimate by solving the optimization problem

$$\widehat{\theta}^{(\mathrm{avg})} = \arg\min_{\theta} \frac{1}{N} \sum_{k=1}^{K} n_k \widehat{L}_k(\theta)$$

in a distributed manner, where n_k is the local sample size of device k, $N = \sum_{k=1}^{K} n_j$ is the total sample size, and $\widehat{L}_k(\theta)$ is the empirical risk for device k. FedAvg aggregates data from all devices after weighting them by local sample size and yields a global model with the estimate $\widehat{\theta}^{(\mathrm{avg})}$. This allows the algorithm to utilize a much larger sample size N instead of n_k, greatly reducing the variance in parameter estimation when K is large.

However, federated learning algorithms such as FedAvg suffers from potential distribution shifts among local devices, which is known as the issue of *heterogeneity* (Li et al., 2020c; Chen et al., 2021; Li et al., 2020a). For parameter estimation problems, this means that the true parameters of local devices may be different from each other. In the presence of heterogeneity, for estimating the local parameters, aggregating data from other devices introduces bias owing to parameter discrepancy. Consequently, aggregating more local samples would lead to a larger bias, which may partially or fully offset the benefit of variance reduction due to sample size increase.

A fundamental *aggregation–heterogeneity trade-off* thus exists in federated learning: borrowing more information from other devices by aggregation tends to increase the effective sample size and improve performance for each local device, but meanwhile would introduce more heterogeneity into the aggregated sample and diminish the accuracy. This can be viewed as a special form of the bias–variance trade-off, which plays a central role in the analysis of machine learning algorithms. Although a considerable amount of recent effort has been devoted to the related topic of personalized federated learning (Li et al., 2020b; Wang et al., 2019; Mansour et al., 2020; Jiang et al., 2019; Wang et al., 2022), a precise learning-theoretic characterization of this trade-off is still lacking. This gap poses further obstacles to answering the important question: *How should one optimally aggregate data from local devices in heterogeneous federated learning?* Here, optimality is viewed from the perspective of improving each device's local task.

It is clear that optimal aggregation should depend on the degree of heterogeneity. To illustrate, consider two extreme cases. In the completely homogeneous case where all devices have the same distribution, aggregating data from other devices does no harm to the local task. In this case, all local samples should be aggregated in order to maximally reduce variance, and hence FedAvg may be an optimal algorithm. At the other extreme where the true parameters of any two local devices are far apart, pooling any two local samples will introduce a huge bias, which tends to negate the benefit of variance reduction. In this case, the local estimator that uses only the device's own data would be the best choice. In the intermediate situation between these two extremes, it is possible that neither FedAvg nor the local estimator is optimal, and one should carefully balance aggregation and heterogeneity to achieve the best performance.

Among prior theoretical work in federated learning, Chen et al. (2021) seems the most closely related to our work. They considered FedAvg and the local estimator from a minimax perspective, and uncovered a phase transition phenomenon: when their proposed measure of heterogeneity, R, is larger than a certain threshold, FedAvg is minimax optimal; otherwise, the local estimator is minimax optimal. Roughly speaking, the quantity R describes the variance of the true parameters, which is a *global* measure of heterogeneity and cannot capture the local dissimilarity relationships among devices. Therefore, it is not suitable for studying the aggregation–heterogeneity trade-off and the corresponding optimal aggregation problem. Consider a simple example, where K devices are divided evenly into two clusters. Within each cluster, the devices have the same true parameters, while the parameters of the two clusters are far away from each other. When $K \to \infty$, it is easy to see that both the FedAvg algorithm run on all devices and the local estimator are suboptimal, since they are inferior to FedAvg run on each cluster.

In this paper, we investigate the aggregation–heterogeneity trade-off and the related optimal aggregation problem in federated learning. We focus on the parametric M-estimation framework, where the task of each local device is to estimate its own parameters. We are interested in analyzing the estimation performance when both the local sample size n and the number of devices K may increase to infinity. To this end, we consider a general family of weighted M-estimators that interpolate between FedAvg and the local estimator, in which an aggregation rule is determined by the weights of local samples. By establishing an upper bound for the estimation error of the weighted M-estimators, we reveal and precisely quantify an aggregation–heterogeneity trade-off. In order to simplify the expressions and find an optimal aggregation rule, we introduce a novel quantity, the *federated smoothness* β, to measure the degree of heterogeneity among devices. We further develop an estimator through a k-nearest neighbors procedure, termed FedKNN, which strikes an optimal balance between aggregation and heterogeneity. Finally, we show that FedKNN is minimax optimal over a certain parameter space characterized by β, whereas the two baseline methods, FedAvg and the local estimator, are both suboptimal. Our contributions are summarized as follows.

- We consider a family of weighted M-estimators in Section 3, in which an aggregation rule is determined by the weights of local samples. Two baseline methods, FedAvg and the local estimator, are contained in this family. We study how the choice of weights and the degree of heterogeneity affect the estimation error, which demonstrates the aggregation–heterogeneity trade-off.

- We propose the *federated smoothness* β in Section 4.1 to measure the degree of heterogeneity among devices. Roughly speaking, β measures the rate at which the parameter distance between neighboring devices decays to 0. A larger β indicates a lower degree of heterogeneity. Compared with other existing heterogeneity measures, β captures the *local* dissimilarity relationships among devices and is more appropriate for studying the trade-off.

- We construct an estimator, FedKNN, using a k-nearest neighbors procedure in Section 4.2. We specialize our general result to this estimator and give a more concise characterization of the aggregation–heterogeneity trade-off. By an optimal choice of neighbor size, FedKNN achieves the optimal trade-off, with estimation error of order

$$\left(\frac{\sigma^2}{nK}\right)^{\frac{2\beta}{2\beta+1}} \wedge \frac{\sigma^2}{n}.$$

- By establishing information-theoretic lower bounds in Section 4.3, we show that the convergence rate of FedKNN is minimax optimal over the parameter space defined by β. We also show that both FedAvg and the local estimator are suboptimal for this parameter space.

- FedKNN is constructed with prior knowledge of the neighboring structure among devices. Without such prior information, it may be desirable to develop an adaptive algorithm to detect the neighboring structure in a data-driven manner. However, in Section 5 we show that in this case no algorithm can improve the local rate σ^2/n and hence adaptation is impossible. When n is fixed, even if $K \to \infty$, no adaptive algorithm can reduce the estimation error to 0.

1.1. Related Work

Personalized federated learning. There is a growing body of work on personalized federated learning, which aims at personalizing global models for individual devices to address the issue of heterogeneity. In particular, Fallah et al. (2020) pointed out the need to fine-tune the initial shared model on local data and presented a personalized variant of FedAvg. Li et al. (2020b) proposed FedProx, which uses ℓ_2 regularization to combine local models. Other regularization-based methods also exist (Hanzely et al., 2020; Acar et al., 2021). Mansour et al. (2020) proposed to first cluster similar devices and then apply federated averaging to each cluster. Similar ideas were explored by Ghosh et al. (2020). From a transfer learning perspective, Wang et al. (2019) proposed to first train a global model and then fine-tune the parameters of the global model on local devices. Jiang et al. (2019) borrowed ideas from model-agnostic meta learning to deal with personalized federated learning problems. Smith et al. (2017) proposed MOCHA to produce personalized solutions in the context of multi-task learning. For a more comprehensive review, see Kulkarni et al. (2020).

Theory of heterogeneous federated learning. Theoretical development of heterogeneous federated learning is relatively limited. Deng et al. (2020) proposed an algorithm for personalized federated learning with learning-theoretic guarantees. By encoding heterogeneity information through a graph, Wang et al. (2022) proposed a fused-lasso-based approach with statistical guarantees. Their estimation error bounds depend on the structure and fidelity measure of the graph. Chen et al. (2021) studied regimes of heterogeneity where either FedAvg or the local estimator is minimax optimal.

More generally, the idea of borrowing strength from related tasks to improve performance on a target task is well established in statistics and widely used in areas such as multi-task learning (Jacob et al., 2008; Hanneke and Kpotufe, 2022; Du et al., 2021), meta-analysis (Cai et al., 2022; Maity et al., 2022), and transfer learning (Cai and Wei, 2021; Li et al., 2022a; Kpotufe and Martinet, 2021). The intuition of federated smoothness and FedKNN comes from the k-nearest neighbors (KNN) method in nonparametric regression (Györfi et al., 2002); see Madrid Padilla et al. (2020) and Demirkaya et al. (2022) for some recent developments.

2. Preliminaries

We begin with introducing some notation used throughout this paper.

2.1. Notation

For $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$. We use $\|\cdot\|$ to denote the ℓ_2-norm of a vector in Euclidean space. For two nonnegative sequences a_n and b_n, we denote $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \leq Cb_n$, and $a_n \gtrsim b_n$ or $b_n = O(a_n)$ if $a_n \geq Cb_n$, for some constant $C > 0$ independent of n. For $a, b \in \mathbb{R}$, we use $a \wedge b$ and $a \vee b$ to denote $\min\{a, b\}$ and $\max\{a, b\}$, respectively. For two distributions P and Q defined on the same space, we use $\|P - Q\|_{\mathrm{TV}}$ to denote their total variation distance.

2.2. Problem Setup

Suppose that there are K devices, where device k holds n_k i.i.d. samples $\{Z_i^{(k)}\}_{i=1}^{n_k}$ drawn from some unknown distribution P_k. At the local device k, the samples are drawn independently from the distribution $P_k, k = 1, \ldots, K$. However, we assume P_ks to be different from each other. For some loss function $\ell(\cdot, \cdot)$, define the expected and empirical risks for device k by $L_k(\theta) = \mathbb{E}_{Z^{(k)} \sim P_k} \ell(\theta, Z^{(k)})$ and $\widehat{L}_k(\theta) = n_k^{-1} \sum_{i=1}^{n_k} \ell(\theta, Z_i^{(k)})$, respectively. The task of device $k \in [K]$ is to estimate its local parameter, which is defined by

$$\theta_k^* = \arg\min_{\theta} L_k(\theta).$$

We directly equate distribution shifts with differences of local parameters, since the distributions and the parameters are in one-to-one correspondence as long as the parameters are identifiable. Hence, the degree of heterogeneity is determined by the distances $\|\theta_k^* - \theta_j^*\|$ for $k \neq j$.

We now state the assumptions needed for our theoretical analysis, which are standard and commonly adopted in the federated learning literature (Li et al., 2020c; Chen et al., 2021; Li et al., 2022b; Wang et al., 2022).

Assumption 1 (Loss function) *The loss function $\ell(\cdot, z)$ is strongly convex with parameter $\mu > 0$ and smooth with parameter $\eta > 0$ for any z; that is, it is differentiable and satisfies*

$$\ell(\theta_1, z) \geq \ell(\theta_2, z) + \nabla_\theta \ell(\theta_2, z)^\top (\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2$$

and

$$\|\nabla_\theta \ell(\theta_1, z) - \nabla_\theta \ell(\theta_2, z)\| \leq \eta \|\theta_1 - \theta_2\|$$

for any θ_1 and θ_2.

Assumption 2 (Gradient noise) *The variance of the gradient is bounded by σ^2, that is,*

$$\mathrm{Var}(\nabla_\theta \ell(Z^{(k)}, \theta)) \leq \sigma^2$$

for every θ and $k \in [K]$.

Assumption 3 (Local sample size) *The local sample sizes $n_k = n$ for all $k \in [K]$.*

Assumption 1 imposes regularity conditions on the loss function. As a direct consequence, the local risk functions $L_k(\theta)$ are μ-strongly convex and η-smooth, and hence the true parameters θ_k^* are well-defined and unique. Assumption 2 implies that, for supervised problems, the variance of local noise is uniformly bounded. It can be relaxed to the case where different devices have different noise sizes. Assumption 3 will be needed in Sections 4 and 5 to simplify our results. Different local sample sizes can be easily accommodated by considering the minimum or maximum sample size.

3. General Weighted M-Estimators

In this section, we present a family of weighted M-estimators with varying degrees of aggregation and establish estimation bounds for these estimators. We first consider the two baseline methods, FedAvg and the local estimator, which estimate θ_k^* by

$$\widehat{\theta}_k^{(\mathrm{avg})} = \arg\min_\theta \frac{1}{K}\sum_{j=1}^K \widehat{L}_j(\theta)$$

and

$$\widehat{\theta}_k^{(\mathrm{loc})} = \arg\min_\theta \widehat{L}_k(\theta),$$

respectively. Intuitively, these two estimators are two extremes of aggregation: $\widehat{\theta}_k^{(\mathrm{avg})}$ aggregates all K devices, whereas $\widehat{\theta}_k^{(\mathrm{loc})}$ aggregates none. To study the situation between these two extremes, we consider the family of *weighted M-estimators*, which includes FedAvg and the local estimator as special cases. Specifically, given any weight matrix $W = (w_{kj}) \in \mathbb{R}^{K \times K}$ with $\sum_{j=1}^{K} w_{kj} = 1$ and $w_{kj} \geq 0$ for $k, j \in [K]$, the corresponding weighted M-estimator is defined by

$$\widehat{\theta}_k^{(W)} = \arg\min_{\theta} \left\{ \widehat{L}_{\mathbf{w}}(\theta) \equiv \sum_{j=1}^{K} w_{kj} \widehat{L}_j(\theta) \right\}.$$

When W is the identity matrix I_K, $\widehat{\theta}_k^{(W)}$ reduces to the local estimator, which assigns zero weights to any devices other than itself. When $w_{kj} = n_j / \sum_{\ell=1}^{K} n_\ell$ for all k, j, $\widehat{\theta}^{(W)}$ reduces to FedAvg, in which case different local devices share the same weight vector. Setting W to other values represents other aggregation rules. Let $\mathbf{w}_k = (w_{k1}, \ldots, w_{kK})^T$ be the weight vector of device k. Intuitively, the more evenly distributed the weights $\mathbf{w}_k$ are, the more information from other devices are aggregated. We remark that each device can have its specific weights $\mathbf{w}_k$, so that instead of obtaining a single global model, the weighted M-estimator outputs K possibly different parameters for personalized inference.

The following theorem is our main result on the estimation error of $\widehat{\theta}_k^{(W)}$, which characterizes how aggregation and heterogeneity together affect the estimation.

Theorem 1 *Under Assumptions 1 and 2, for each $k \in [K]$, we have*

$$\mathbb{E}\|\widehat{\theta}_k^{(W)} - \theta_k^*\|^2 \leq \left(\frac{16\eta^2}{\mu^2} + 4 \right) \sum_{j=1}^{K} w_{kj} \|\theta_k^* - \theta_j^*\|^2 + \frac{8\sigma^2}{\mu^2} \sum_{j=1}^{K} \frac{w_{kj}^2}{n_j}. \tag{1}$$

The error bound in (1) breaks down into two terms, which are of order $\sum_{j=1}^{K} w_{kj} \|\theta_j^* - \theta_k^*\|^2$ and $\sum_{j=1}^{K} w_{kj}^2 \sigma^2 / n_j$. The first term is the weighted average of squared distances between parameters, which originates from the heterogeneity among the K devices. To make it smaller, each device k should avoid aggregating data from irrelevant devices and instead concentrate its weights w_{kj} on devices whose parameters are close to its own. The second term is caused by the variance of data aggregated by device k. By the Cauchy–Schwarz inequality, to reduce this term, each device k should make its weights w_{kj} more evenly distributed. In general, these two terms cannot be minimized at the same time, and therefore an *aggregation–heterogeneity trade-off* arises.

The optimal weights that minimize the bound (1) are determined by the degree of heterogeneity and the local noise level simultaneously. Consider the following two extreme cases. For simplicity, we assume $n_k = n$ and $\sigma_k^2 = 1$ for all k. The first case is the homogeneous setting, where $\theta_j^* = \theta_k^*$ for all j, k. In this case, the first term in (1) vanishes and we need only minimize the second term. By the Cauchy–Schwarz inequality, the optimal weights for each k are $\mathbf{w}_k = (1/K, \ldots, 1/K)^T$. Thus, the optimal weighted M-estimator reduces to FedAvg whose estimation error is of order $1/(nK)$. This greatly improves the local rate $1/n$ when K is large. The second case is that the parameters of any two devices are far apart; that is, for all $j \neq k$, $\|\theta_j^* - \theta_k^*\| \geq c$ for some constant $c \gtrsim 1/\sqrt{n}$. Then the optimal weights for each k are approximately given by $w_{kk} = 1$ and $w_{kj} = 0$ for $j \neq k$, which minimize the first term in (1). These optimal weights correspond to the local estimator, and the optimal estimation error is of order $1/n$. In this setting, aggregation does not bring any improvement over the local estimator owing to the high heterogeneity.

We remark that Theorem 1 is similar in spirit to Theorem 3 of Ben-David et al. (2010), which also considered the minimization of combined empirical risk from different domains and provided a general learning bound. However, their analysis is confined to classification problems and the case of two domains, a source domain and a target domain, which is not well suited to federated learning scenarios where the number of devices may be large and even tend to infinity (e.g., Internet of Things (Atzori et al., 2010)).

4. Optimal Aggregation Method

Generally, it is difficult to obtain a closed-form solution for the optimal weights in (1) if we do not impose any assumption on the heterogeneity terms $\{\|\theta_k^* - \theta_j^*\|\}_{j \neq k}$. In order to investigate the optimal weights and the resulting estimator, in Section 4.1 we propose *federated smoothness* to measure the degree of heterogeneity. Using this measure, in Section 4.2 we specialize our general result to a more concise version and construct an optimal estimator, FedKNN, using a k-nearest neighbors (KNN) procedure. In Section 4.3 we give minimax lower bounds and show that FedKNN is indeed minimax optimal over the parameter space defined by federated smoothness.

4.1. Federated Smoothness

To measure the degree of heterogeneity, we introduce *federated smoothness* in the following definition. We use $\Theta^* = (\theta_1^*, \ldots, \theta_K^*)$ to denote the concatenated parameter.

Definition 1 (Federated smoothness) *Let $0 \leq \beta \leq \infty$. We say that Θ^* has federated smoothness β if there exists a fixed constant $C > 0$ such that*

$$\|\theta_k^* - \theta_{\pi(k,j)}^*\| \leq C \left(\frac{j}{K} \right)^{\beta} \tag{2}$$

for every $k, j \in [K]$, where $\pi : [K] \times [K] \to [K]$ maps (k, j) to the index of the device whose parameter is the jth closest to the parameter of device k.

In the definition of π, we break ties in an arbitrary way if there are any. Also, $\pi(k, 1) = k$ by convention. For Θ^* satisfying this definition, we will also say that Θ^* is β-federated-smooth. Note that for any fixed Θ^*, such β and C always exist, although they are not unique. We are particularly interested in the class of parameters satisfying federated smoothness for some fixed C and β when K is allowed to increase. Moreover, if (2) holds for some β, then it also holds for any $0 \leq \beta' < \beta$. Hence, our results can be interpreted as true for the largest such β.

The value of β measures the rate at which the distance between the parameters of neighboring devices decays to 0. A larger β indicates a lower degree of heterogeneity, which further leads to a smaller estimation error as we will see later. The role of π is to describe the neighboring structure among devices. Examples include federated learning applications where the devices are distributed geographically or over a social network and each device shares similar parameters only with their closest neighbors. In this section, we consider the case where π is given as prior knowledge.

To get a sense of the applicability of Definition 1, we consider a few examples with different levels of federated smoothness. As it turns out, our framework includes many conventional regimes studied in previous work.

Example 1 (Homogeneity) *Suppose that all of the K devices have the same parameters. Then Θ^* has federated smoothness $\beta = \infty$.*

Example 2 (Complete heterogeneity) *Suppose that there exists a constant $c > 0$ such that $\|\theta_k - \theta_j\| \geq c$ for any $k \neq j$. Then for any fixed C, the federated smoothness β of Θ^* tends to 0 when $K \to \infty$.*

Example 3 (Clustering structure) *Suppose that, for an even K, the K devices are divided into two groups with equal size. Furthermore, devices in the same group have the same parameters, and the distance between the parameters of the two groups is denoted by δ. Then the federated smoothness $\beta = \log(C/\delta)/\log 2$.*

Example 4 (Hölder smooth function) *Consider any function $f\colon [0,1] \to \mathbb{R}^d$ satisfying $f'(0) = f^{(2)}(0) = \cdots = f^{(\lfloor\beta\rfloor-1)}(0) = 0$ and $\|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)\| \leq C|x-y|^{\beta-\lfloor\beta\rfloor}$ for any $x, y \in [0,1]$ and some $\beta > 0$. Suppose that $\theta^*_{\pi(k,j)} = f(j/K)$. Then applying Taylor's expansion gives $\|\theta^*_{\pi(k,j)} - \theta^*_k\| \leq C(j/K)^\beta$.*

Other heterogeneity measures have been considered in the literature. Defining $F^*_k = \widehat{L}_k(\widehat{\theta}^{(\mathrm{loc})}_k)$ and $F^* = K^{-1}\sum_{k=1}^K \widehat{L}_k(\widehat{\theta}^{(\mathrm{avg})})$, Li et al. (2020c) proposed using $\Gamma = F^* - K^{-1}\sum_{k=1}^K F^*_k$ to measure heterogeneity. If all the devices have the same distributions, their empirical risks are approximately the same and hence $\Gamma \approx 0$. Otherwise, Γ may be large since the minimizers of local empirical risks $\widehat{L}_k$ are different. Alternatively, Chen et al. (2021) suggested using the variance of local parameters, $R = K^{-1}\sum_{k=1}^K \|\theta^*_k - K^{-1}\sum_{j=1}^K \theta^*_j\|^2$, to measure the degree of heterogeneity.

The common intuition behind these measures is that they all depict heterogeneity from a *global* point of view. They focus on the degree of heterogeneity across all K devices, but do not reflect the local dissimilarity relationships among devices. In comparison, our proposed federated smoothness is a *local* quantity in that it characterizes the decay rate of the distances between local parameters. Therefore, it is more appropriate for analyzing the aggregation–heterogeneity trade-off.

4.2. FedKNN: Optimal Weights via KNN

With the notion of federated smoothness in mind, we now focus on a particular set of weighted M-estimators. Specifically, we determine the weight matrix $W^{(m)} = (w^{(m)}_{kj})$ by the following KNN procedure:

$$w^{(m)}_{k,\pi(k,j)} = \begin{cases} \dfrac{1}{m} & \text{if } 1 \leq j \leq m, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $m \in [K]$ is the number of neighbors (including the device itself) to be aggregated. We denote by $\widehat{\theta}^{(m)}$ the estimator obtained with $W^{(m)}$ and call it *federated k-nearest neighbors* (FedKNN).

FedKNN builds on the simple intuition that each device should only aggregate data from its closest neighbors. Accordingly, it uses hard thresholding to determine the weights: it assigns equal weights to the m closest devices and zero weights to the others. Here, m directly specifies the level of aggregation, so that a larger m means that more data from other devices are aggregated.

Using Theorem 1 and Definition 1, we obtain bounds for the estimation error of FedKNN as well as the optimal choice of m for β-federated-smooth parameters.

Proposition 1 *Under Assumptions 1–3, if Θ^* is β-federated-smooth for some $0 < \beta < \infty$, then for each $k \in [K]$, the FedKNN estimator $\widehat{\theta}^{(m)}_k$ satisfies*

$$\mathbb{E}\|\widehat{\theta}^{(m)}_k - \theta^*_k\|^2 \leq C_1 \frac{m^{2\beta}}{K^{2\beta}} + \frac{C_2\sigma^2}{mn}, \tag{4}$$

where $C_1 = C(16\eta^2 + 4\mu^2)2^{2\beta+1}/\{\mu^2(2\beta+1)\}$ *and* $C_2 = 8/\mu^2$. *In particular, if we take* $m = m_* \equiv \lfloor (C_2\sigma^2/C_1)^{1/(2\beta+1)}K^{2\beta/(2\beta+1)}/n^{1/(2\beta+1)} \rfloor \vee 1$, *then the above upper bound attains its minimum over the choice of* m*:*

$$\mathbb{E}\|\widehat{\theta}_k^{(m)} - \theta_k^*\|^2 \le C_\beta \left(\frac{\sigma^2}{nK}\right)^{2\beta/(2\beta+1)} \wedge \frac{8\sigma^2}{\mu^2 n}, \tag{5}$$

where $C_\beta = C_2^{2\beta/(2\beta+1)}C_1^{1/(2\beta+1)}$.

The two terms in (4) correspond to the two terms in (1). From (4) we see a clearer trade-off between aggregation and heterogeneity: a larger m indicates more aggregation from other devices, which increases the first term and decreases the second term. The value m_* balances the two terms optimally and attains the minimum of (4) over the choice of m. It can be interpreted as the *effective number of devices*, which characterizes how many other devices can optimally improve on the local estimation and leads to the optimal estimation error of order $\sigma^2/(m_*n)$.

For a fixed β, m_* decreases with n and increases with K. This is because if K is fixed, a larger n implies that the estimation error of the local estimator is smaller, and hence the criterion on whether to aggregate data from other devices is stricter, which leads to a smaller m_*. On the other hand, if n is fixed, a larger K means that there are more devices with smaller bias, and hence one should aggregate data from more devices, so that m_* increases. When $\beta \to \infty$, it is easy to check that $m_* \to K$ and $\mathbb{E}\|\widehat{\theta}_k - \theta_k^*\|^2 \to \sigma^2/(nK)$, which means that almost all devices are useful. Indeed, $\beta = \infty$ corresponds to the homogeneous setting, where the estimation error is exactly $O(\sigma^2/(nK))$.

When $K \gtrsim n^{1/(2\beta)}$, we have $m_* \gtrsim 1$ and the term $1/(nK)^{2\beta/(2\beta+1)}$ dominates the bound (5). In this regime, there exist aggregation methods that outperform the local estimator. Indeed, it is easy to check that in this case $\|\theta^*_{\pi(k,2)} - \theta^*_k\| \lesssim 1/\sqrt{n}$, so that there exists at least another device that can improve on the estimation of θ_k^*. Even when the local sample size n is fixed, FedKNN is still consistent as long as the number of devices $K \to \infty$.

We remark that FedKNN is inspired by the classical KNN methods in nonparametric statistics (Györfi et al., 2002). Indeed, we can naturally make the analogy between federated learning and nonparametric regression: each device corresponds to a covariate point, the local parameter θ_k^* corresponds to the conditional mean of the response, and the local variance $O(\sigma^2/n)$ corresponds to the variance of the response. The role of π is to define the neighboring structure among devices. From this point of view, β can be related to the smoothness of the regression function in nonparametric regression, which is the reason why we call it federated *smoothness*.

Our FedKNN estimator bears some resemblance to the rank-based procedure considered by Hanneke and Kpotufe (2022). However, the settings of Hanneke and Kpotufe (2022) and our work differ remarkably in terms of learning scenarios and the source of heterogeneity. Accordingly, the procedures for exploiting the neighboring structure and their performance bounds are also very different. Specifically, heterogeneity in their work stems from distribution shifts, and the ranking information reflects discrepancies between excess risks under different distributions. In contrast, heterogeneity in our setting is due to varying parameters, and the neighboring structure is defined through distances between local parameters.

4.3. Minimax Lower Bounds

Although FedKNN is only a subset of weighted M-estimators, our next result indicates that it indeed achieves minimax optimality over the parameter space determined by β. To derive the minimax lower bounds, we consider a mean estimation problem. Suppose that the data are generated by

$$Z_i^{(k)} = \theta_k^* + \varepsilon_i^{(k)}, \quad k \in [K],\, i \in [n_k], \tag{6}$$

where $\theta_k^* \in \mathbb{R}^d$ and $\varepsilon_i^{(k)}$ are i.i.d. Gaussian with mean 0 and variance $\sigma^2 I_d$. Denote the concatenated parameter matrix by $\Theta^* = (\theta_1^*, \ldots, \theta_K^*) \in \mathbb{R}^{d\times K}$. We consider the parameter space

$$\mathcal{P}(\beta, K, C, \pi) \equiv \left\{ \Theta^* \in \mathbb{R}^{d\times K} : \|\theta^*_{\pi(k,i)} - \theta^*_k\|_2 \le C \left(\frac{i}{K}\right)^{\beta} \text{ for all } k, i \in [K] \right\}.$$

Note that in addition to β, K and C, the parameter space also depends on the neighboring structure π, which is given as prior information. The loss function is taken to be ℓ_2 loss, that is, $\ell(\theta, z) = \|\theta - z\|^2$. It is easy to check that Assumptions 1 and 2 are satisfied. We denote the parameter space by $\mathcal{P}_\beta$ or $\mathcal{P}(\beta, \pi)$ for simplicity when there is no confusion. For the estimation error of Θ in $\mathcal{P}_\beta$, we have the following information-theoretic lower bound.

Theorem 2 (Minimax lower bound) *Under Assumption 3, if $\beta < \infty$, $K \ge 3$ and $(nK)/\sigma^2 \ge (\log 2)2^{2\beta-1}/C^2$, then for each $k \in [K]$ we have*

$$\inf_{\widehat{\theta}_k} \sup_{\Theta^* \in \mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta^*_k\| \ge C'_\beta \left(\frac{\sigma^2}{nK}\right)^{2\beta/(2\beta+1)} \wedge \frac{\sigma^2}{n} \tag{7}$$

for some constant $C'_\beta > 0$ depending only on β and C, where $\widehat{\theta}_k$ denotes any measurable function of samples from all local devices.

While the pair (C, β) is not unique in Definition 1, it is important to have the class $\mathcal{P}_\beta$ depend on a fixed C, so that β is well defined. The lower bound (7) matches the upper bound (5), which shows that FedKNN with neighborhood size m_* is minimax optimal. This result illustrates that the federated smoothness β precisely characterizes the degree of heterogeneity and the extent to which information from other devices can improve on the local estimation. The proof is an application of the celebrated La Cam method, by carefully constructing two parameters $\Theta^{(1)}$ and $\Theta^{(2)}$ in $\mathcal{P}$ such that $\|\theta_k^{(1)} - \theta_k^{(2)}\|$ is relatively large while $\{\|\theta_j^{(2)} - \theta_j^{(2)}\|\}_{j\neq k}$ are as small as possible.

When $K^{2\beta} \lesssim n$, the minimax rate is σ^2/n, which agrees with the local estimation rate without aggregation. This is because in this case the local variance σ^2/n is relatively small so that the bias introduced by aggregation dominates the estimation error. Therefore, no devices other than device k contain information that can further improve the local estimation error σ^2/n. When $K^{2\beta} \gtrsim n$, owing to the restriction of β-federated-smoothness, other devices also contain information useful for estimating θ_k, the amount of which depends on the magnitude of K and β. As a result, the minimax rate is improved to $\{\sigma^2/(nK)\}^{2\beta/(2\beta+1)}$.

Moreover, the following result suggests that in our heterogeneous setting both FedAvg and the local estimator are suboptimal since their worst-case errors are larger than the minimax rate (7).

Theorem 3 (Suboptimality of FedAvg and local estimators) *Consider the Gaussian mean estimation problem specified in* (13)*. Under Assumption 3, there exists an integer L independent of K and n such that, for every $K \geq L$ and $n \geq L$, there exists $k \in [K]$ and π such that*

$$\sup_{\Theta^* \in \mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k^{(\mathrm{loc})} - \theta_k^*\|^2 \geq \frac{\sigma^2 d}{n} \tag{8}$$

and

$$\sup_{\Theta^* \in \mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k^{(\mathrm{avg})} - \theta_k^*\|^2 \geq \frac{C}{6^{\beta+1}}. \tag{9}$$

The suboptimality of FedAvg and the local estimator stems from the fact that they are two extremes of the aggregation–heterogeneity trade-off. The bound (8) for the local estimator is intuitive: since the local estimator does not aggregate any other devices, its rate is typically $O(\sigma^2/n)$, which has nothing to do with heterogeneity but also does not benefit from aggregation. On the other hand, FedAvg aggregates all K devices, which introduces too much heterogeneity. Thus, the benefit of aggregation is dominated by the harm of heterogeneity, which leads to an $O(1)$ worst-case error (9). Notably, the rates of these two methods do not involve K, so that their estimation errors will not be reduced by increasing the number of devices. As a result, in the setting where the local sample size n is relatively small and the number of devices K is relatively large, they underperform FedKNN substantially.

We remark that the estimation errors of these two baseline methods have been studied by Chen et al. (2021). Over a certain parameter space determined by their proposed heterogeneity measure $R = \sigma^2 K^{-1} \sum_{k=1}^K \|\theta_k^* - K^{-1}\sum_{j=1}^K \theta_j^*\|$, they showed that if $R \lesssim \sigma^2/n$, then FedAvg is minimax optimal; otherwise, the local estimator is minimax optimal. This seems to contradict the suboptimality result of Theorem 3. The main reason for this discrepancy is that we measure the degree of heterogeneity in a different way. As discussed above, their quantity R is a global heterogeneity measure, which cannot characterize the aggregation–heterogeneity trade-off. Consequently, their result can be viewed as a phase transition from one extreme to the other. In contrast, our federated smoothness β is a local heterogeneity measure, which gives rise to the trade-off and allows the estimation rate to vary continuously with respect to the degree of heterogeneity β.

5. Impossibility of Adaptation

The construction of the FedKNN procedure and the minimax result in the previous section rely critically on the knowledge of the neighboring structure π. In the absence of such prior information, it is desirable to develop an adaptive algorithm to determine which devices should be federated in a data-driven manner. For example, we can first estimate π using the local estimators $\{\widehat{\theta}_k^{(\mathrm{loc})}\}_{k=1}^K$, and then perform the FedKNN procedure based on the estimated $\widehat{\pi}$. One might hope that there would exist some adaptive algorithm to achieve the minimax rate (5). Unfortunately, if there is no information beyond the samples, for such data-driven procedures the estimation error of π, which is typically of order σ^2/n, dominates and adaptation is impossible. Indeed, by proving a minimax lower bound for the case where π is unknown, we will show that no adaptive algorithm can actually improve the local estimation rate $O(\sigma^2/n)$.

We are still concerned with the mean estimation problem introduced in Section 4.3. However, instead of the parameter space $\mathcal{P}(\beta, K, C, \pi)$ studied above, we consider the following parameter space, which does not take π as its argument:

$$\widetilde{\mathcal{P}}(\beta, K, C) := \left\{ \Theta^* \in \mathbb{R}^{d\times K} \,:\, \exists \pi \text{ s.t. } \|\theta^*_{\pi(k,i)} - \theta^*_k\|_2 \le C \left(\frac{i}{K}\right)^{\beta} \text{ for all } k, i \in [K] \right\}.$$

We denote it by $\widetilde{\mathcal{P}}_\beta$ for simplicity when there is no confusion. No prior knowledge of π means that each parameter in $\widetilde{\mathcal{P}}_\beta$ still satisfies the β-federated smoothness, but has its specific neighboring structure π, which is not given. Indeed, this parameter space is the union of $\mathcal{P}(\beta, \pi)$ defined before, that is, $\widetilde{\mathcal{P}}_\beta = \cup_\pi \mathcal{P}(\beta, \pi)$, where the union is taken over all possible π. If π is known as prior information, we only need to consider the worst-case error over some specific $\mathcal{P}(\beta, \pi)$; otherwise, we must consider the worst case over all possible π, or $\widetilde{\mathcal{P}}_\beta$.

For this larger parameter space, we have the following minimax lower bound.

Theorem 4 *Under Assumption 3, there exists an integer L independent of K and n such that for any $K \ge L$, $n \ge L$ and $0 < \beta < \infty$, we have, for any $k \in [K]$,*

$$\inf_{\widehat{\theta}_k} \sup_{\Theta^* \in \widetilde{\mathcal{P}}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_*\|^2 \ge \frac{(\log 2)\sigma^2}{2n}. \tag{10}$$

Compared with Theorem 2, no knowledge of π means that the minimax analysis is performed over a larger set, that is, $\widetilde{\mathcal{P}}_\beta$ instead of $\mathcal{P}(\beta, \pi)$ specified by some known π. That is why the lower bound $O(1/n)$ here is much larger than that in (7). If we want to estimate π from scratch, its estimation error is already $O(\sigma^2/n)$. The most frustrating conclusion of Theorem 4 is that when the local sample size n is fixed, even if K increases to infinity and federated smoothness is satisfied for some $\beta < \infty$, the estimation error of any adaptive estimator does not vanish asymptotically. This suggests the necessity of prior knowledge about the neighboring structure: we must at least be able to estimate π at a faster rate than $O(\sigma^2/n)$.

More technically, an important property used in the proof of Theorem 4 is that $\widetilde{\mathcal{P}}_\beta$ is closed under permutation. Then we can apply Le Cam's method by considering two concatenated parameters $\Theta^{(1)}$ and $\Theta^{(2)}$ in $\widetilde{\mathcal{P}}_\beta$ such that they differ only by swapping the local parameters of two specific devices k and j. In this case, data from devices $l \neq k, j$ do not provide any information that can distinguish between $\theta_k^{(1)}$ and $\theta_k^{(2)}$, which results in the large rate $O(\sigma^2/n)$.

We remark that similar results have been obtained for multi-task learning by Hanneke and Kpotufe (2022). For a multi-task classification problem where all data distributions induce the same optimal classifier, they showed that no adaptive algorithm attains the minimax convergence rate unless the ranking of distributional discrepancy is known a priori. At a high level, our messages about the impossibility of adaptation and necessity of knowledge of π are the same as theirs, while we are working in a quite different setting.

6. Discussion and Future Work

In this paper, we have considered parameter estimation in heterogeneous federated learning from a theoretical perspective. We have derived estimation bounds for a family of weighted M-estimators, which reveals a fundamental aggregation–heterogeneity trade-off. We have introduced federated smoothness to quantify the degree of heterogeneity among devices and have constructed a k-nearest neighbors estimator, FedKNN, to optimally balance the trade-off. We have given minimax lower bounds, from which we show that FedKNN is minimax optimal and that the two baseline methods, FedAvg and the local estimator, are both suboptimal since they ignore the aggregation–heterogeneity trade-off. Finally, we have demonstrated the necessity of prior knowledge about the neighboring structure by proving the impossibility of adaptation.

We point out some directions for future research. The first is the optimization aspects of FedKNN. Although FedKNN can be conducted in a distributed manner by simply regarding each device as a center and performing FedAvg multiple times, this trivial procedure incurs computation and communication costs m times those of FedAvg, where m is the number of aggregated neighbors. How to more efficiently implement the optimization is of practical interest.

Second, it remains to investigate the settings where π is inaccurate or β is unknown. We have shown that the lack of knowledge of π leads to a pessimistic $O(\sigma^2/n)$ rate. In practice, we may have partial knowledge about π or be given an inaccurate version of it. How the accuracy of π affects the estimation error is an interesting question. Also, adaptation to unknown β would be possible by following Lepski's method (Lepskii, 1991).

Finally, extensions to structured heterogeneity settings would be worthwhile. In this paper and most related work in federated learning, heterogeneity is simply encoded in the Euclidean distance between parameters. In more complex scenarios, however, heterogeneity may be structured. For example, there may be some components of the local parameters that are shared, while the rest are device-specific. Both the shared parts and the specific parts can be high-dimensional or very complex. How to borrow strength from multiple devices to estimate the shared components without being harmed by the specific parts is a direction for future work.

Acknowledgments

This research was supported by Beijing Natural Science Foundation grant Z190001 and National Natural Science Foundation of China grants 12171012 and 12292981.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Luigi Atzori, Antonio Iera, and Giacomo Morabito. The Internet of Things: A survey. *Computer Networks*, 54(15):2787–2805, 2010.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, 117(540): 2105–2119, 2022.

Shuxiao Chen, Qinqing Zheng, Qi Long, and Weijie J Su. A theorem of the alternative for personalized federated learning. *arXiv preprint arXiv:2103.01901*, 2021.

Emre Demirkaya, Yingying Fan, Lan Gao, Jinchi Lv, Patrick Vossler, and Jingbo Wang. Optimal nonparametric inference with two-scale distributional nearest neighbors. *Journal of the American Statistical Association*, 2022.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 19586–19597, 2020.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 2304–2315, 2020.

Laurent Jacob, Jean-philippe Vert, and Francis Bach. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems*, volume 21, pages 745–752, 2008.

Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 794–797, 2020.

O. V. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, 35(3):454–466, 1991.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society, Series B*, 84(1):149–173, 2022a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2020c.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local SGD. In *Conference on Learning Theory*, pages 1613–1661, 2022b.

Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela M Witten. Adaptive nonparametric regression with the K-nearest neighbour fused lasso. *Biometrika*, 107(2):293–310, 2020.

Subha Maity, Yuekai Sun, and Moulinath Banerjee. Meta-analysis of heterogeneous data: Integrative sparse regression in high-dimensions. *Journal of Machine Learning Research*, 23(198):1–50, 2022.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4427–4437, 2017.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.

Huiyuan Wang, Xuyang Zhao, and Wei Lin. Heterogeneous federated learning on a graph. *arXiv preprint arXiv:2209.08737*, 2022.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Appendix A. Upper Bounds

A.1. Proof of Theorem 1

Proof For simplicity, we omit the superscript W throughout this proof. For each $k \in [K]$, define $L_k(\theta) = \sum_{j=1}^{K} w_{kj}\mathbb{E}[\ell(\theta, Z^{(j)})]$, $\widehat{L}_k(\theta) = \sum_{j=1}^{K} \frac{w_{kj}}{n_j}\sum_{i=1}^{n_j} \ell(\theta, Z_i^{(j)})$ and $\widetilde{\theta}_k = \arg\min_{\theta\in\Omega} L_k(\theta)$. By the strong convexity of ℓ and the definition of $\widehat{\theta}_k$, we have

$$0 \geq \widehat{L}_k(\widehat{\theta}_k) - \widehat{L}_k(\widetilde{\theta}_k) \geq \nabla\widehat{L}(\widetilde{\theta}_k)^\top(\widehat{\theta}_k - \widetilde{\theta}_k) + \frac{\mu}{2}\|\widetilde{\theta}_k - \widehat{\theta}_k\|^2,$$

which implies

$$\|\widehat{\theta}_k - \widetilde{\theta}_k\|^2 \le \frac{4}{\mu^2}\|\nabla \widehat{L}_k(\widetilde{\theta}_k)\|^2.$$

Also, by definition, $\mathbb{E}[\nabla \widehat{L}_k(\widetilde{\theta}_k)] = 0$. Taking expectation on both sides yields

$$\begin{aligned}\mathbb{E}\|\widehat{\theta}_k - \widetilde{\theta}_k\|^2 &\le \frac{4}{\mu^2}\mathbb{E}\|\nabla \widehat{L}_k(\widetilde{\theta}_k)\|^2 = \frac{4}{\mu^2}\mathrm{Var}(\nabla \widehat{L}_k(\widetilde{\theta}_k)) \\ &= \frac{4}{\mu^2}\sum_{j=1}^{K}\frac{w_{kj}^2}{n_j}\mathrm{Var}(\nabla \ell(\widetilde{\theta}_k, Z^{(j)})) \\ &\le \frac{4\sigma^2}{\mu^2}\sum_{j=1}^{K}\frac{w_{kj}^2}{n_j}.\end{aligned} \tag{11}$$

By the strong convexity again we have

$$\left\|\widetilde{\theta}_k - \sum_{j=1}^{K} w_{kj}\theta_j^*\right\|^2 \le \frac{4}{\mu^2}\left\|\nabla L_k\left(\sum_{j\in[K]} w_{kj}\theta_j^*\right)\right\|^2.$$

Therefore, by the triangle inequality,

$$\begin{aligned}\|\widetilde{\theta}_k - \theta_k^*\|^2 &\le 2\left\|\widetilde{\theta}_k - \sum_{j\in[K]} w_{kj}\theta_j^*\right\|^2 + 2\left\|\sum_{j\in[K]} w_{kj}\theta_j^* - \theta_k^*\right\|^2 \\ &\le \frac{8}{\mu^2}\left\|\nabla L_k\left(\sum_{j\in[K]} w_{kj}\theta_j^*\right)\right\|^2 + 2\left\|\sum_{j\in[K]} w_{kj}\theta_j^* - \theta_k^*\right\|^2.\end{aligned}$$

For the first term, using the smoothness of ℓ gives

$$\begin{aligned}&\frac{8}{\mu^2}\left\|\nabla L_k\left(\sum_{j\in[K]} w_{kj}\theta_j^*\right)\right\|^2 \\ &\quad = \frac{8}{\mu^2}\left\|\sum_{j\in[K]} w_{kj}\nabla\left\{E\ell\left(\sum_{j\in[K]}\mathbf{w}_j\theta_j^*, Z^{(j)}\right) - E\ell(\theta_j^*, Z^{(j)})\right\}\right\|^2 \\ &\quad \le \frac{8}{\mu^2}\left(\sum_{j\in[K]} w_{kj}\left\|\nabla\left\{E\ell\left(\sum_{j\in[K]}\mathbf{w}_j\theta_j^*, Z^{(j)}\right) - E\ell(\theta_j^*, Z^{(j)})\right\}\right\|\right)^2 \\ &\quad \le \frac{8\eta^2}{\mu^2}\left(\sum_{j\in[K]} w_{kj}\left\|\sum_{j\in[K]} w_{kj}\theta_j^* - \theta_j^*\right\|\right)^2 \\ &\quad \le \frac{8\eta^2}{\mu^2}\sum_{j\in[K]} w_{kj}\left\|\sum_{j\in[K]} w_{kj}\theta_j^* - \theta_j^*\right\|^2,\end{aligned}$$

where the last formula is obtained by the Cauchy–Schwarz inequality. Then by the property of weighted average,

$$\begin{aligned}\|\widetilde{\theta}_k - \theta_k^*\|^2 &\leq \left(\frac{8\eta^2}{\mu^2} + 2\right) \sum_{j\in[K]} w_{kj} \left\| \sum_{j\in[K]} w_{kj}\theta_j^* - \theta_j^* \right\|^2 \\ &\leq \left(\frac{8\eta^2}{\mu^2} + 2\right) \sum_{j\in[K]} w_{kj} \|\theta_k^* - \theta_j^*\|^2.\end{aligned} \tag{12}$$

Combining (11) and (12) yields

$$\mathbb{E}\|\widehat{\theta}_k - \theta_k^*\|^2 \leq \left(\frac{16\eta^2}{\mu^2} + 4\right) \sum_{j=1}^K w_{kj}\|\theta_k^* - \theta_j^*\|^2 + \frac{8\sigma^2}{\mu^2} \sum_{j=1}^K \frac{w_{kj}^2}{n_j},$$

which completes the proof. ■

A.2. Proof of Proposition 1

Proof Recall that the KNN weight matrix is defined by

$$w_{k,\pi(k,j)}^{(m)} = \begin{cases} \dfrac{1}{m} & \text{if } 1 \leq j \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

Substituting it into the first term in (1) and using the definition of β, we obtain

$$\begin{aligned}\sum_{j=1}^K w_{kj}^{(m)} \|\theta_j^* - \theta_k^*\|^2 &= \sum_{j=1}^m \frac{1}{m} \|\theta_k^* - \theta_j^*\|^2 \\ &\leq \frac{C}{m} \sum_{j=1}^m \frac{j^{2\beta}}{K^{2\beta}} \leq \frac{C}{mK^{2\beta}} \int_1^{m+1} x^{2\beta} dx \\ &\leq \frac{2^{2\beta+1}C}{2\beta+1} \frac{m^{2\beta}}{K^{2\beta}}.\end{aligned}$$

For the second term in (1), we have

$$\sum_{j=1}^K \frac{(w_{kj}^{(m)})^2}{n_j} = \sum_{j=1}^m \frac{1}{m^2 n_j} = \frac{1}{mn},$$

where the last step comes from Assumption 3. Combining the two terms gives the overall bound

$$\mathbb{E}\|\widehat{\theta}_k^{(m)} - \theta_k^*\|^2 \leq \left(\frac{16\eta^2}{\mu^2} + 4\right) \frac{2^{2\beta+1}C}{2\beta+1} \frac{m^{2\beta}}{K^{2\beta}} + \frac{8\sigma^2}{\mu^2 mn} = C_1 \frac{m^{2\beta}}{K^{2\beta}} + \frac{C_2\sigma^2}{mn},$$

where $C_1 = (16\eta^2 + 4\mu^2)2^{2\beta+1}C/\{\mu^2(2\beta+1)\}$ and $C_2 = 8/(\mu^2)$. This bound, as a function of m, is minimized by choosing $m = m^*$, where

$$m_* = \lfloor (C_2\sigma^2/C_1)^{1/(2\beta+1)} K^{2\beta/(2\beta+1)} / n^{1/(2\beta+1)} \rfloor \vee 1.$$

If $m^* = 1$, we only use samples of the device k to estimate θ_k^* (i.e., local estimator), in which case $\mathbb{E}\|\widehat{\theta}_k^{(m_*)} - \theta_k^*\|^2 \le 8\sigma^2/(\mu^2 n)$. Thus, the resulting rate scales as

$$\mathbb{E}\|\widehat{\theta}_k^{(m_*)} - \theta_k^*\|^2 \le C_\beta \left(\frac{\sigma^2}{nK}\right)^{2\beta/(2\beta+1)} \wedge \frac{8\sigma^2}{\mu^2 n},$$

where $C_\beta = C_2^{2\beta/(2\beta+1)} C_1^{1/(2\beta+1)}$. ■

Appendix B. Lower Bounds

B.1. Problem Setup

To derive the minimax lower bounds, we consider a mean estimation problem. Suppose that the data are generated by

$$Z_i^{(k)} = \theta_k^* + \varepsilon_i^{(k)}, \quad k \in [K],\, i \in [n_k], \tag{13}$$

where $\theta_k^* \in \mathbb{R}^d$ and $\varepsilon_i^{(k)}$ are i.i.d. Gaussian with mean 0 and variance $\sigma^2 I_d$. Denote the concatenated parameter matrix by $\Theta^* = (\theta_1^*, \ldots, \theta_K^*) \in \mathbb{R}^{d\times K}$. The loss function is taken to be ℓ_2 loss, that is, $\ell(\theta, z) = \|\theta - z\|^2$. It is easy to check that Assumptions 1 and 2 are satisfied.

B.2. Lower Bound with Known Neighboring Structure

Proof [Proof of Theorem 2] Recall that we consider the parameter space

$$\mathcal{P}(\beta, K, C, \pi) \equiv \left\{\Theta^* \in \mathbb{R}^{d\times K} : \|\theta^*_{\pi(k,i)} - \theta_k^*\|_2 \le C\left(\frac{i}{K}\right)^\beta \text{ for all } k, i \in [K]\right\}.$$

Using Le Cam's method (Wainwright, 2019), for any $k \in [K]$ and any two parameters $\Theta^{(1)} = (\theta_1^{(1)}, \ldots, \theta_K^{(1)})$ and $\Theta^{(2)} = (\theta_1^{(2)}, \ldots, \theta_K^{(2)})$ in $\mathcal{P}_\beta$, we have

$$\inf_{\widehat{\theta}_k} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \ge \frac{\|\theta_k^{(1)} - \theta_k^{(2)}\|^2}{8}\big(1 - \|P^{(1)} - P^{(2)}\|_{\text{TV}}\big), \tag{14}$$

where $P^{(1)}$ and $P^{(2)}$ denote the sample distributions of $\Theta^{(1)}$ and $\Theta^{(2)}$, respectively. For completeness, we will prove (14) in Lemma 5. For the total variation distance, we further have the bound

$$\|P^{(1)} - P^{(2)}\|_{\text{TV}}^2 \le \frac{1}{4}\left\{\exp\left(\frac{n}{\sigma^2}\sum_{j=1}^K \|\theta_j^{(1)} - \theta_j^{(2)}\|^2\right) - 1\right\}, \tag{15}$$

whose proof is given in Lemma 6.

Next, we proceed to prove the theorem by considering two different cases.

(i) First, we consider the case where $(\log 2)\sigma^2 K^{2\beta}/(4C^2 n) \ge 1$. Let $m \in [K/2]$ be a value to be specified later. Choose any $\mathcal{S} \subset [K]$ such that $k \in \mathcal{S}$ and $|\mathcal{S}| = m$. Then for $\delta_m = C(m/K)^\beta$, there exist $\Theta^{(1)} = (\theta_1^{(1)}, \dots, \theta_K^{(1)})$ and $\Theta^{(2)} = (\theta_1^{(2)}, \dots, \theta_K^{(2)})$ in $\mathcal{P}_\beta$ such that

$$\begin{aligned} \theta_j^{(1)} &= \theta_j^{(2)} = \theta_0 \quad \text{for any } j \in [K] - \mathcal{S}, \\ \theta_j^{(1)} &= \theta_1 \quad \text{for any } j \in \mathcal{S}, \\ \theta_j^{(2)} &= \theta_2 \quad \text{for any } j \in \mathcal{S}, \\ \|\theta_1 - \theta_2\| &= 2\|\theta_1 - \theta_0\| = 2\|\theta_2 - \theta_0\| = 2\delta_m. \end{aligned}$$

The existence of parameters satisfying these conditions is proved in (1) of Lemma 7. By (15), for these two parameters we have

$$\|P^{(1)} - P^{(2)}\|_{\mathrm{TV}}^2 \le \frac{1}{4}\left\{\exp\left(4\frac{mn}{\sigma^2}\delta_m^2\right) - 1\right\}.$$

Now we choose an $m \in [K/2]$ such that $\|P^{(1)} - P^{(2)}\|_{\mathrm{TV}} \le 1/2$. To this end, we need $4mn\delta_m^2 = 4C^2 m^{2\beta+1} n / K^{2\beta} \le (\log 2)\sigma^2$, that is,

$$m \le \left(\frac{\log 2}{4C^2} \cdot \frac{K^{2\beta}}{n/\sigma^2}\right)^{1/(2\beta+1)}.$$

Since

$$\frac{\log 2}{4C^2} \cdot \frac{\sigma^2 K^{2\beta}}{n} \ge 1,$$

we can directly set the value of m to

$$m_* \equiv \left\lfloor \left(\frac{\log 2}{4C^2}\right)^{1/(2\beta+1)} \frac{K^{2\beta/(2\beta+1)}}{(n/\sigma^2)^{1/(2\beta+1)}} \right\rfloor,$$

which is a positive integer. Also, the assumption $(nK)/\sigma^2 \ge \log 2 \cdot 2^{2\beta-1}/C^2$ ensures that $m_* \le K/2$. Thus, the aforementioned existence of parameters holds for m_*. In this case,

$$\delta_{m_*}^2 = C^2 \left(\frac{m_*}{K}\right)^{2\beta} \ge \frac{C^2}{4^\beta} \left(\frac{\log 2}{4C^2} \cdot \frac{\sigma^2}{nK}\right)^{2\beta/(2\beta+1)}.$$

Substituting δ_{m_*} into (14), we obtain

$$\inf_{\widehat{\theta}_k} \sup_{\Theta \in \mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \ge \frac{\delta_{m_*}^2}{4} \ge \frac{C^2}{4^{\beta+1}} \left(\frac{\log 2}{4C^2}\right)^{2\beta/(2\beta+1)} \left(\frac{\sigma^2}{nK}\right)^{2\beta/(2\beta+1)}. \tag{16}$$

(ii) Next, we consider the case $(\log 2)\sigma^2 K^{2\beta}/(4C^2 n) \le 1$. By (4) of Lemma 7, for

$$\delta = \sqrt{\frac{\log 2}{4} \cdot \frac{\sigma^2}{n}} \le \frac{C}{K^\beta},$$

there exist $\Theta^{(1)}$ and $\Theta^{(2)}$ such that

$$\|\theta_k^{(1)} - \theta_k^{(2)}\| = 2\delta$$

and

$$\|\theta^{(1)}_{\pi(k,j)} - \theta^{(2)}_{\pi(k,j)}\| = 0$$

for $j \geq 2$. For these two parameters, by (15),

$$\|P^{(1)} - P^{(2)}\|_{\mathrm{TV}} \leq \frac{1}{2}.$$

Substituting these pieces into (14) gives

$$\inf_{\widehat{\theta}_k} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \geq \frac{\delta^2}{4} = \frac{\log 2}{16} \cdot \frac{\sigma^2}{n}.$$

(iii) Since $(\log 2)\sigma^2 K^{2\beta}/(4C^2 n) \geq 1$ is equivalent to

$$\left(\frac{\sigma^2}{nK}\right)^{\frac{2\beta}{2\beta+1}} \leq \left(\frac{\log 2}{4C^2}\right)^{\frac{1}{2\beta+1}} \frac{\sigma^2}{n},$$

combining the above two cases gives

$$\begin{aligned}
\inf_{\widehat{\theta}_k} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 &\geq \min\left\{ \frac{C^2}{4^{\beta+1}}\left(\frac{\log 2}{4C^2}\right)^{\frac{2\beta}{2\beta+1}}, \frac{\log 2}{16}\left(\frac{\log 2}{4C^2}\right)^{-\frac{1}{2\beta+1}} \right\} \\
&\quad \times \min\left\{ \left(\frac{\sigma^2}{nK}\right)^{\frac{2\beta}{2\beta+1}}, \left(\frac{\log 2}{4C^2}\right)^{\frac{1}{2\beta+1}} \frac{\sigma^2}{n} \right\} \\
&\geq \min\left\{ \frac{C^2}{4^{\beta+1}}\left(\frac{\log 2}{4C^2}\right)^{\frac{2\beta}{2\beta+1}}, \frac{\log 2}{16}\left(\frac{\log 2}{4C^2}\right)^{-\frac{1}{2\beta+1}} \right\} \\
&\quad \times \min\left\{ \left(\frac{\log 2}{4C^2}\right)^{\frac{1}{2\beta+1}}, 1 \right\} \\
&\quad \times \min\left\{ \left(\frac{\sigma^2}{nK}\right)^{\frac{2\beta}{2\beta+1}}, \frac{\sigma^2}{n} \right\}.
\end{aligned}$$

Let

$$C'_\beta = \min\left\{ \frac{C^2}{4^{\beta+1}}\left(\frac{\log 2}{4C^2}\right)^{\frac{2\beta}{2\beta+1}}, \frac{\log 2}{16}\left(\frac{\log 2}{4C^2}\right)^{-\frac{1}{2\beta+1}} \right\} \times \min\left\{ \left(\frac{\log 2}{4C^2}\right)^{\frac{1}{2\beta+1}}, 1 \right\},$$

which depends only on β and C. Then we finally obtain

$$\inf_{\widehat{\theta}_k} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \geq C'_\beta \min\left\{ \left(\frac{\sigma^2}{nK}\right)^{2\beta/(2\beta+1)}, \frac{\sigma^2}{n} \right\}.$$

■

B.3. Lower Bounds for FedAvg and Local Estimators

Proof [Proof of Theorem 3] We prove (8) and (9) in (1) and (2), respectively.

(1) For this problem instance, it's easy to show that the local estimator is the local empirical mean, i.e., $\widehat{\theta}_k^{(\mathrm{loc})} = \frac{1}{n}\sum_{i=1}^n Z_i^{(k)}$. Then for any Θ and any k, its estimation error is

$$\mathbb{E}\|\widehat{\theta}_k^{(\mathrm{loc})} - \theta_k\|^2 = \frac{1}{n}\mathbb{E}\|Z^{(k)} - \theta_k\|^2 = \frac{\sigma^2 d}{n},$$

which proves (8).

(2) For the FedAvg, taking derivative we obtain its explicit solution is

$$\widehat{\theta}_k^{(\mathrm{avg})} = \frac{1}{nK}\sum_{j=1}^K\sum_{i=1}^n Z_i^{(j)}$$

for each $k \in [K]$. Therefore, its estimation error is

$$\begin{aligned}\mathbb{E}\|\widehat{\theta}_k^{(\mathrm{avg})} - \theta_k\|^2 &= \frac{1}{nK^2}\sum_{j=1}^K \mathbb{E}\|Z^{(j)} - \theta_j\|^2 + \left\|\frac{1}{K}\sum_{j=1}^K \theta_j - \theta_k\right\|^2 \\ &= \frac{\sigma^2 d}{nK} + \left\|\frac{1}{K}\sum_{j=1}^K \theta_j - \theta_k\right\|^2 \\ &\geq \left\|\frac{1}{K}\sum_{j=1}^K \theta_j - \theta_k\right\|^2.\end{aligned}$$

By (3) of Lemma 7, for each $K \geq 3$ and $0 < \beta < \infty$, there exists π and $\Theta \in \mathcal{P}_\beta$ such that $\|\frac{1}{K}\sum_{j=1}^K \theta_j - \theta_k\|^2 \geq \frac{C}{6^{\beta+1}}$, which completes the proof directly.

■

B.4. Lower Bound with Unknown Neighboring Structure

Proof [Proof of Theorem 4] Recall that the parameter space now is

$$\widetilde{\mathcal{P}}(\beta, K, C) := \left\{\Theta^* \in \mathbb{R}^{d\times K} : \exists\, \pi \text{ s.t. } \|\theta^*_{\pi(k,i)} - \theta^*_k\|_2 \leq C\left(\frac{i}{K}\right)^\beta \text{ for all } k, i \in [K]\right\},$$

which does not specify π in advance.

We still use Le Cam's method as in (14). For some $j \neq k$, we consider two parameters $\Theta^{(1)}, \Theta^{(2)} \in \widetilde{\mathcal{P}}_\beta$ satisfying

$$\begin{aligned}&\|\theta_k^{(1)} - \theta_j^{(1)}\| = \frac{\sqrt{\log 2}\sigma}{\sqrt{n}}, \\ &\theta_k^{(1)} = \theta_j^{(2)}, \quad \theta_k^{(2)} = \theta_j^{(1)}, \\ &\theta_l^{(1)} = \theta_l^{(2)} \quad \text{for } l \neq k, j.\end{aligned}$$

The existence of such parameters is given in (2) of Lemma 7. Applying (14) to these two parameters yields

$$\inf_{\widehat{\theta}} \sup_{\Theta\in\mathcal{P}} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \geq \frac{\sigma^2 \log 2}{n}\left[1 - \frac{1}{2}\big(e^{\log 2} - 1\big)^{1/2}\right] = \frac{(\log 2)\sigma^2}{2n}, \tag{17}$$

which completes the proof. ■

B.5. Technical Lemmas

This section consists of technical lemmas used in the proofs of our main results.

Lemma 5 *For any two parameters $\Theta^{(1)}$ and $\Theta^{(2)}$ in $\mathcal{P}(\beta, K, C, \pi)$, let $P_n^{(1)}$ and $P_n^{(2)}$ be the corresponding distributions of $\{Z_i^{(k)}\}_{k\in[K], i\in[n]}$, respectively. Then we have*

$$\inf_{\widehat{\theta}} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \geq \frac{\|\theta_k^{(1)} - \theta_k^{(2)}\|^2}{8}\left[1 - \|P^{(1)} - P^{(2)}\|_{\mathrm{TV}}\right],$$

Proof Let $\delta = \frac{1}{2}\|\theta_k^{(1)} - \theta_k^{(2)}\|$. By Markov's inequality, for any Θ we have

$$\mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \geq \delta^2 P(\|\widehat{\theta}_k - \theta_k\| \geq \delta).$$

Therefore, since $\Theta^{(1)}, \Theta^{(2)} \in \mathcal{P}_\beta$, we obtain

$$\begin{aligned}
\inf_{\widehat{\theta}_k} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 &\geq \inf_{\widehat{\theta}_k} \sup_{\Theta\in\{\Theta^{(1)},\Theta^{(2)}\}} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \\
&\geq \delta^2 \sup_{\Theta\in\{\Theta^{(1)},\Theta^{(2)}\}} P(\|\widehat{\theta}_k - \theta_k\| \geq \delta) \\
&\geq \frac{\delta^2}{2}\Big(P_{\Theta^{(1)}}(\|\widehat{\theta}_k - \theta_k^{(1)}\| \geq \delta) + P_{\Theta^{(2)}}(\|\widehat{\theta}_k - \theta_k^{(2)}\| \geq \delta)\Big) \\
&\overset{\text{(i)}}{\geq} \frac{\delta^2}{2}\Big(P_n^{(1)}(\|\widehat{\theta}_k - \theta_k^{(2)}\| \leq \delta) + P_n^{(2)}(\|\widehat{\theta}_k - \theta_k^{(2)}\| \geq \delta)\Big) \\
&= \frac{\delta^2}{2}\left[1 - (P_n^{(1)}(\|\widehat{\theta}_k - \theta_k^{(2)}\| \geq \delta) + P_n^{(2)}(\|\widehat{\theta}_k - \theta_k^{(2)}\| \geq \delta))\right] \\
&\overset{\text{(ii)}}{\geq} \frac{\delta^2}{2}\left[1 - \|P_n^{(1)} - P_n^{(2)}\|_{\mathrm{TV}}\right],
\end{aligned}$$

where step (i) follows from $\|\theta_k^{(1)} - \theta_k^{(2)}\| = 2\delta$, and step (ii) follows from the definition of total variation norm. Therefore, we finally obtain

$$\inf_{\widehat{\theta}} \sup_{\Theta\in\mathcal{P}_\beta} \mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 \geq \frac{\|\theta_k^{(1)} - \theta_k^{(2)}\|^2}{8}\left[1 - \|P^{(1)} - P^{(2)}\|_{\mathrm{TV}}\right].$$

■

Lemma 6 *Let $P_n^{(1)}$ and $P_n^{(2)}$ be the distributions of $\{Z_i^{(k)}\}_{k\in[K],i\in[n]}$ with parameters $\Theta^{(1)}$ and $\Theta^{(2)}$, respectively. Then we have*

$$\|P_n^{(1)} - P_n^{(2)}\|_{\mathrm{TV}}^2 \le \frac{1}{4}\left[\exp\left(\frac{n}{\sigma^2}\sum_{j=1}^{K}\|\theta_j^{(1)} - \theta_j^{(2)}\|^2\right) - 1\right].$$

Proof Let $p^{(1)}$ and $p^{(2)}$ be the density functions of $P^{(1)}$ and $P^{(2)}$ with respect to the Lebesgue measure. By the Cauchy–Schwarz inequality we have

$$\begin{aligned}
\|P_n^{(1)} - P_n^{(2)}\|_{\mathrm{TV}}^2 &= \frac{1}{4}\left[\int_{\mathbb{R}^{dnK}} |p^{(1)}(z) - p^{(2)}(z)| dz\right]^2 \\
&= \frac{1}{4}\left[\int_{\mathbb{R}^{dnK}} \left|\frac{p^{(1)}(z)}{p^{(2)}(z)} - 1\right| p^{(2)}(z) dz\right]^2 \\
&\le \frac{1}{4}\int_{\mathbb{R}^{dnK}} \left(\frac{p^{(1)}(z)}{p^{(2)}(z)} - 1\right)^2 p^{(2)}(z) dz \\
&= \frac{1}{4}\left[\int_{\mathbb{R}^{dnK}} \frac{p^{(1)}(z)^2}{p^{(2)}(z)} dz - 1\right].
\end{aligned}$$

By the definition of $Z_i^{(k)}$, $p^{(1)}(z) = \prod_{k=1}^{K}\prod_{i=1}^{n} p_k^{(1)}(z_i^{(k)})$, where $p_k^{(1)}$ is the density function of $\mathcal{N}(\theta_k^{(1)}, \sigma^2 I_d)$. A similar formula holds for $P_n(2)$. Then we further have

$$\begin{aligned}
\|P_n^{(1)} - P_n^{(2)}\|_{\mathrm{TV}}^2 &\le \frac{1}{4}\left[\int_{\mathbb{R}^{dnK}} \prod_{k=1}^{K}\prod_{i=1}^{n} \frac{p^{(1)}(z_i^{(k)})^2}{p^{(2)}(z_i^{(k)})} dz - 1\right] \\
&= \frac{1}{4}\left[\exp\left(\frac{n}{\sigma^2}\sum_{k=1}^{K}\|\theta_k^{(1)} - \theta_k^{(2)}\|^2\right) - 1\right],
\end{aligned}$$

which completes the proof. ■

Lemma 7 *The parameter spaces $\mathcal{P}(\beta, K, C, \pi)$ and $\widetilde{\mathcal{P}}(\beta, K, C)$ satisfy the following properties.*

(1) Suppose $K \ge 3$, $\frac{\log 2}{4C^2} \cdot 2^{2\beta+1} \le \frac{nK}{\sigma^2}$ and $\frac{\log 2}{4C^2} \cdot \frac{\sigma^2 K^{2\beta}}{n} \ge 1$. For any $m \in [K]$ and $\mathcal{S} \subset [K]$ with $|\mathcal{S}| = m$ and $m \le K/2$, there exist $\Theta^{(1)}, \Theta^{(2)} \in \mathcal{P}(\beta, K, C, \pi)$ such that

$$\begin{aligned}
\theta_j^{(1)} = \theta_j^{(2)} &= \theta_0 \quad \textit{for any } j \in [K] - \mathcal{S}, \\
\theta_j^{(1)} &= \theta_1 \quad \textit{for any } j \in \mathcal{S}, \\
\theta_j^{(2)} &= \theta_2 \quad \textit{for any } j \in \mathcal{S}, \\
\|\theta_1 - \theta_2\| = 2\|\theta_1 - \theta_0\| &= 2\|\theta_2 - \theta_0\| = 2\delta_m,
\end{aligned}$$

where $\delta_m = C(m/K)^\beta$.

(2) *Suppose $K \geq 3$ and $n \geq 4 \cdot 6^{2\beta}/C^2$. For any $j \neq k$, there exist $\Theta^{(1)}, \Theta^{(2)} \in \widetilde{\mathcal{P}}(\beta, K, C)$ such that*

$$\begin{gathered}\|\theta_k^{(1)} - \theta_j^{(1)}\| = \frac{2}{\sqrt{n}},\\ \theta_k^{(1)} = \theta_j^{(2)}, \quad \theta_k^{(2)} = \theta_j^{(1)},\\ \theta_l^{(1)} = \theta_l^{(2)} \quad \textit{for } l \neq k, j.\end{gathered}$$

(3) *For any $K \geq 3$ and $0 < \beta < \infty$, there exist π and $\Theta \in \mathcal{P}(\beta, K, C, \pi)$ such that for any $k \in [K]$,*

$$\left\| \frac{1}{K} \sum_{j=1}^{K} \theta_j - \theta_k \right\| \geq \frac{C}{6^{\beta+1}}.$$

(4) *For any $K \geq 3$ and any $\mathcal{P}(\beta, K, C, \pi)$, for any $\delta \leq C/K^\beta$ there exists $\Theta^{(1)}, \Theta^{(2)} \in \mathcal{P}(\beta, K, C, \pi)$ such that $\|\theta_k^{(1)} - \theta_k^{(2)}\| = 2\delta$ and $\theta_j^{(1)} = \theta_j^{(2)}$ for any $j \neq k$.*

Proof

(1) For any θ_1 and θ_2 such that $\|\theta_1 - \theta_2\| = 2\delta$, let θ_0 be their midpoint. Then for any $\mathcal{S} \subset [K]$, the corresponding two parameters can be specified by the above conditions. We only need to check that they are indeed in the parameter space $\mathcal{P}_\beta$. We first consider $\Theta^{(1)}$. For any $k \in \mathcal{S}$ and $j \leq m$, $\|\theta_k^{(1)} - \theta_{\pi(k,j)}^{(1)}\| = 0 \leq C(j/K)^\beta$. For $j > m$, $\|\theta_k^{(1)} - \theta_{\pi(k,j)}^{(1)}\| = \delta_m = C(m/K)^\beta \leq C(j/K)^\beta$. For any $k \in [K] - \mathcal{S}$, $m \leq K/2$ implies that $\|\theta_k^{(1)} - \theta_{\pi(k,j)}^{(1)}\| \leq \|\theta_l^{(1)} - \theta_{\pi(l,j)}^{(1)}\|$ for any $l \in \mathcal{S}$. Therefore, $\Theta^{(1)}$ satisfies the federated smoothness condition and hence $\Theta^{(1)} \in \mathcal{P}_\beta$. Similar arguments also hold for $\Theta^{(2)}$.

(2) We first show that $\widetilde{\mathcal{P}}(\beta, K, C)$ is closed under permutation. For any $\Theta = (\theta_1, \ldots, \theta_K) \in \widetilde{\mathcal{P}}_\beta$ with neighboring structure π and any permutation $\sigma : [K] \to [K]$ of $[K]$, define $\Theta^\sigma \in \mathbb{R}^{d \times K}$ by $\Theta^\sigma = (\theta_{\sigma(1)}, \ldots, \theta_{\sigma(K)})$. Further define $\pi^\sigma(k, j) = \pi(\sigma(k), j)$ which is also a neighboring structure. Then we have

$$\|\Theta_k^\sigma - \Theta_{\pi^\sigma(k,j)}^\sigma\| = \|\theta_{\sigma(k)} - \theta_{\pi(\sigma(k),j)}\| \leq C\frac{j^\beta}{K^\beta},$$

which implies that $\Theta^\sigma \in \mathcal{P}_\beta$ with neighboring structure $\pi(\sigma(k, j))$.

Therefore, we only need to prove there exists $\Theta^{(1)} \in \widetilde{\mathcal{P}}_\beta$ such that $\|\theta_k^{(1)} - \theta_j^{(1)}\| = 2/\sqrt{n}$, then $\Theta^{(2)}$ is immediately obtained by swapping $\theta_k^{(1)}$ and $\theta_j^{(1)}$ and keeping other parameters fixed.

We first fix any $\theta_k, \theta_j \in \mathbb{R}^d$ such that $\|\theta_k - \theta_j\| = 2/\sqrt{n}$ whose existence is obvious. For other $K-2$ devices, let $\mathcal{C}_1 \subset [K] \setminus \{k, j\}$ be any subset with $||\mathcal{C}_1| - (K/2 - 1)| < 1$ and let $\mathcal{C}_2 = [K] \setminus (\{k, j\} \cup \mathcal{C}_1)$. For any $l \in \mathcal{C}_1$, set $\theta_l = \theta_k$; for any $l \in \mathcal{C}_2$, set $\theta_l = \theta_j$. Then we construct $\Theta^{(1)}$ using these parameters by $\Theta^{(1)} = (\theta_1, \ldots, \theta_K)$ and let π by its neighboring structure. For any $k \in [K]$ and $j \in [K]$ such that $j < K/2$, $\|\Theta_k^{(1)} - \Theta_{\pi(k,j)}^{(1)}\| = 0 \leq (\frac{j-1}{K})^\beta$;

for $j \in [K]$ such that $j \geq K/2$, $\|\Theta_k^{(1)} - \Theta_{\pi(k,j)}^{(1)}\| = 2/\sqrt{n} \leq C(\frac{K/2-1}{K})^\beta \leq C(\frac{j-1}{K})^\beta$ by the assumptions $K \geq 3$ and $n \geq 4 \cdot 6^{2\beta}/C^2$. Therefore, $\Theta^{(1)} \in \widetilde{\mathcal{P}}(\beta, K, C)$.

(3) We still consider clustered parameters similar to (2). We arbitrarily choose $\mathcal{C}_1 \in [K]$ and $\mathcal{C}_2 = [K] \setminus \mathcal{C}_1$ such that $\big||\mathcal{C}_1| - (K/2 - 1)\big| < 1$. Then we consider K parameters such that $\theta_k = a$ for $k \in \mathcal{C}_1$ and $\theta_k = b$ for $k \in \mathcal{C}_2$, where $a, b \in \mathbb{R}^d$ satisfy $\|a - b\| = C/6^\beta$. For any $k \in [K]$ and $j \in [K]$ such that $j < K/2$, $\|\theta_k - \theta_{\pi(k,j)}\| = 0 \leq (\frac{j-1}{K})^\beta$; for $j \in [K]$ such that $j \geq K/2$, $\|\theta_k - \theta_{\pi(k,j)}\| \leq C/6^\beta \leq C \cdot (\frac{K/2-1}{K})\beta \leq (\frac{j-1}{K})^\beta$. Therefore, $\Theta = (\theta_1, \ldots, \theta_K) \in \mathcal{P}(\beta, K, C, \pi)$ for some π.

Moreover, for this parameter, for any $k \in [K]$,

$$\left\|\frac{1}{K}\sum_{j=1}^{K}\theta_j - \theta_k\right\| \geq \frac{K-1}{2K}\|a - b\| \geq \frac{1}{3}\|a - b\| \geq \frac{C}{6^{\beta+1}},$$

which completes the proof.

(4) For any a and b such that $\|a - b\| = 2\delta$, let c be their midpoint. Then we set $\theta_j^{(1)} = \theta_j^{(2)} = c$ for $j \neq k$, $\theta_k^{(1)} = a$ and $\theta_k^{(2)} = b$. Then it is easy to verify that all conditions are satisfied.

This completes the proof of the lemma. ■