Federated Rényi Fair Inference in Federated Heterogeneous System

Zhiyong Ma*[1] **Yuanjie Shi***[2] **Yan Yan**[2] **Jian Chen**[1]

[1]School of SSE, South China University of Technology University, Guangzhou, Guangdong, China
[2]School of EECS, Washington State University, Pullman, WA, USA

Abstract

Federated learning (FL) is a prominent distributed learning approach that addresses two major challenges: statistical heterogeneity (i.e., non-identically distributed data) and system heterogeneity (i.e., variability in communication and computation on each client). As FL is commonly applied in sectors such as commercial and financial, group disparities can emerge and cause harm. However, current fairness algorithms assume homogeneous data, which do not align with the FL context. The main challenge is estimating global fairness measures (e.g., Rényi or Pearson correlation) in an asynchronous, heterogeneous system. To address this, we propose the FedRényi algorithm, which regularizes fairness by Rényi correlation. For statistical heterogeneity, FedRényi aggregates local fairness statistics to estimate the global Rényi correlation with an estimation error bound of $O(1/\sqrt{n})$, where n is the total number of data samples. This theoretical result improves significantly over the previous result $O(1/\sqrt{K})$ with K clients. We further prove that FedRényi converges at the same rate as in the homogeneous setting. For system heterogeneity, FedRényi approximates missing client updates through weighted averaging over a nearest neighbor region, ensuring a non-expansive approximation error under non-convex conditions. Extensive experiments demonstrate that FedRényi achieves a promising fairness-accuracy trade-off, with at least 2% improvement over baselines.

1 INTRODUCTION

Federated learning (FL) is an effective paradigm for decentralized learning in large-scale datasets McMahan et al. [2017], Kairouz et al. [2021], allowing models to be trained on multiple clients without sharing raw data, thus preserving privacy Zhang et al. [2023]. Many FL works Karimireddy et al. [2020], Li et al. [2020b], Xu et al. [2023], Zhu et al. [2021a] have been proposed to address challenges in FL, such as *statistical heterogeneity*, where locally distributed data are non-identically distributed (non-IID), and *system heterogeneity*, which involves variability in communication and computational capabilities between clients, such as unparticipating clients Li et al. [2020a,b]. These methods make FL attractive and suitable for many real-world sectors, such as commercial Jain and Jerripothula [2023] and finance Long et al. [2020], Mammen [2021], where large institutions (e.g., banks) seek to mitigate predictor bias caused by group disparities Barocas et al. [2023].

Group fairness Barocas et al. [2023] is a commonly used approach to mitigate prediction bias against certain demographic groups Kleinberg et al. [2018], divided by sensitive attributes such as race or gender Dwork et al. [2012], Mehrabi et al. [2021]. Many methods have been proposed to promote group fairness Baharlouei et al. [2020], Woodworth et al. [2017] , but they are mostly designed for a centralized and homogeneous setting.

Some works have been proposed to improve group fairness in FL Zeng et al. [2021], Abay et al. [2020], Du et al. [2020], Zhang et al. [2020], Chu et al. [2021], all of which require empirical estimation of fairness measures (e.g., group disparities). Although the previously established estimation error of the fairness measure is in $O(1/\sqrt{K})$ with K clients Chu et al. [2021] under non-IID conditions, it is worse than the standard estimation error of $O(1/\sqrt{n})$ with n data samples in the centralized setting Mohri et al. [2018].

In addition, it is unclear how to adapt these methods to the system heterogeneity (e.g., with dropping clients or "stragglers" Li et al. [2020b]). Specifically, the empirical fairness measure based on partially participating clients can deviate significantly from the true fairness measure based on fully participating clients. The above two challenges raise a question: *How can we develop a federated fairness-enhanced al-*

gorithm with theoretical guarantees for fairness and convergence in both statistically and systemically heterogeneous settings?

To this end, we propose Federated Rényi Fair Inference (FedRényi) algorithm to promote group fairness in FL. We use the general and tractable Rényi correlation Rényi [2007], Baharlouei et al. [2020] as regularization to induce fairness globally across all clients. Specifically, to estimate the global Rényi correlation, we first compute the necessary local group-wise statistics (see Eq. (5) later) on each client, and then aggregate these local statistics following two federated weighting schemes (see Eq. (2) later) from clients into a global measure Mansour et al. [2020]. For any nonparticipating client in each communication round, FedRényi approximates its local statistics/model by weighted averaging over its neighbor clients based on their similarity measures.

We theoretically show that FedRényi guarantees the estimation error bound in $O(1/\sqrt{n})$ order, which improves significantly over the previous established one in $O(1/\sqrt{K})$ Chu et al. [2021] ($K \ll n$ usually in FL Kairouz et al. [2021]). Furthermore, we derive a convergence rate of FedRényi in $O(1/\epsilon^4)$ iteration complexity, matching the same order as the standard FL result Karimireddy et al. [2020]. Moreover, we show that the proposed approximation is non-expansive for certain non-convex loss functions Liu et al. [2021] with pre-trained model Tan et al. [2022], Weller et al. [2022], Tian et al. [2022], i.e., the non-increasing distance between the approximated and true local statistics/models within a communication round. Finally, we empirically evaluate our method on benchmark datasets, showing that FedRényi provides a promising trade-off performance between global accuracy and group fairness with at least 2% improvement of the harmonic mean of accuracy and fairness over baselines in most cases.

Contributions: Our key contributions are summarized:

- We propose FedRényi to promote group fairness in FL by using Rényi correlation as a regularization term. We develop an aggregation method to estimate the global Rényi statistics from local clients, and an approximation scheme to approximate local statistics/models based on similarity measures between clients.
- We theoretically prove that our FedRényi effectively provides a tight estimation error bound of $O(1/\sqrt{n})$. Based on the improved results, we further derive the same convergence rate $O(1/\epsilon^4)$ of FedRényi with the standard FL result Karimireddy et al. [2020]. In addition, the similarity-based approximation scheme is non-expansive (the distance between the approximated and true statistics/models is non-increasing) under mild conditions.
- Extensive experimental results verify the improved trade-off ability of FedRényi (at least 2% improvement in the harmonic mean of accuracy and fairness over baselines in most cases).

2 RELATED WORK

Fairness in FL. Gajane and Pechenizkiy [2017] systematically divides machine learning fairness into five types: group fairness, individual fairness, unconscious fairness, counterfactual fairness, and preference-based fairness. Many studies have examined the impact of group fairness in FL using metrics such as demographic parity and/or equality of opportunity Shi et al. [2021]. To mitigate group bias in heterogeneous settings, MWR Selialia et al. [2024] employs a heuristic approach that enhances fairness by using importance weighting and regularization to optimize the accuracy of the worst-performing group. As data statistics are allowed to be shared in FL (e.g., Shao et al. [2023], Zhu et al. [2021b], Jeong et al. [2018], Seo et al. [2016]), FairFed Ezzeldin et al. [2023] increases the aggregated weights of clients with small deviations between local and global fairness metrics or accuracy. Based on FairBatch Roh et al. [2021], FedFB Zeng et al. [2021] relies on statistical information about the performance of the client to adjust the minibatch sizes of each client in local update process to optimize the group-specific losses, thus imposing group fairness. FedFair Chu et al. [2021] estimates model fairness and incorporates this estimation as a loss function constraint. Its estimation bound is $O(1/\sqrt{K})$, which is especially large compared to centralized results (e.g., Mohri et al. [2018]) as $K \ll n$. However, both FedFB and FairFed lack theoretical analysis of estimation errors to justify their validity.

Heterogeneity in FL. In FL, heterogeneity is categorized into statistical and system heterogeneity. Statistical heterogeneity refers to variability in data distributions across clients, which impacts performance and convergence of FL algorithms. Methods like control variates to reduce variance Karimireddy et al. [2020] or proximal terms to stabilize training Li et al. [2020b] address these issues but lack theoretical guarantees. System heterogeneity refers to disparities in communication and computational capacities among clients, leading to inefficient training and potential client dropout. Studies analyze estimation errors due to these disparities. Sefidgaran et al. [2024] investigates the partial estimation error caused by inter-client estimation discrepancies. In this framework, Sefidgaran et al. [2024] investigates the impact of communication rounds on the estimation error in federated learning, finding that increased communication does not always improve performance. Hu et al. [2023] introduces a two-level distribution framework to analyze the full estimation error caused by inter- and intra-client estimation errors in FL. It establishes learning bounds for participating and nonparticipating clients (stragglers), respectively. However, a comprehensive evaluation of the total estimation error remains limited.

3 BACKGROUND AND MOTIVATION

Notations. Let $(x, y, s) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ be a data sample, where $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} = \{1, ..., C\}$ and $\mathcal{S} = \{1, ..., P\}$ represent the feature, classification label, and protected attribute spaces, respectively. Let $\mathcal{P}$ be the underlying distribution defined on $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$. Let X, Y, S be the random variables drawn from their respective distributions and x, y, s be the realization. K denotes the number of clients, and the set of all client indices is $[K]$. Let $\mathcal{I}_k$ be the index set for data on client k, n_k denote the number of data samples on k-th client, the total number of data is n and $n = \sum_{k=1}^{K} n_k$. Denote $n_{\min} := \min_k n_k$ as the minimum number of data samples across all clients. We define the distribution on client k by $\mathcal{P}_k$ and refer the data heterogeneity as $\mathcal{P}_{k_1} \neq \mathcal{P}_{k_2}$ where $k_1 \neq k_2$. Let X_k and S_k be feature and attribute variables on client k, respectively. Define $f_\theta(x, s)$ as the prediction function parameterized by θ on (x, s) and $\ell(\widehat{y}, y)$ as a loss function measured on the predicted label $\widehat{y}$ and true label y. Define $\mathbb{P}_c = P[f_\theta(X, S) = c]$ as the probability of the model predicting class c and $\rho = \min_{c \in C} \mathbb{P}_c$ as the smallest model prediction probability over all classes. Let $\mathbb{1}[\cdot]$ be an indicator function and $A \backsim B$ means that A and B are in the same order. Define τ as the total number of communication round and e as the index of communication round, respectively. Denote M as the number of local update iterations in one communication round e and m as the index of local iteration. Define the total number of iterations as T, where $T = \tau M$. Notations are summarized in Table 3.

Federated Learning. In FL, given K distributions $\mathcal{P}_1, ..., \mathcal{P}_K$ with sizes $\{n_k\}_{k=1}^K$ from client $k \in [K]$, one aims to learn a model that minimizes the overall loss $L(\theta)$:

$$\min_\theta L(\theta) = \sum_{k=1}^{K} \gamma_k \cdot \mathbb{E}_{(x,s,y)\sim\mathcal{P}_k}[\ell(f_\theta(x, s), y)], \quad (1)$$

where γ_k is the weights of client k. Two target weighting schemes are commonly considered, i.e., uniform over sample and uniform over client:

$$\gamma_k = \begin{cases} n_k/n, & \text{uniform over sample,} \\ 1/K, & \text{uniform over client.} \end{cases} \quad (2)$$

The choices of γ_k represent different schemes, which are both used in fairness-aware FL studies Mohri et al. [2019], Li et al. [2019], Lyu et al. [2020], Chu et al. [2021], Fan et al. [2021]. Since $L(\theta)$ is defined on population and is not accessible, an empirical FL objective is defined as:

$$\min_\theta \widehat{L}(\theta) := \sum_{k=1}^{K} \gamma_k \underbrace{\sum_{i=1}^{n_k} \frac{1}{n_k} \ell(f_\theta(x_{ki}, s_{ki}), y_{ki})}_{\widehat{L}_k(\theta)}, \quad (3)$$

where (x_{ki}, s_{ki}, y_{ki}) is i-th data sample on client k.

In fairness-aware FL, minimizing $\widehat{L}(\theta)$ alone can lead to group disparities across sensitive attributes S. To mitigate this, many fairness-aware FL studies Li et al. [2021], Chu et al. [2021] incorporate a fairness regularization term $\widehat{R}(\theta)$ with parameter λ into the objective function:

$$\min_\theta \widehat{L}(\theta) + \lambda \widehat{R}(\theta).$$

In this paper, we consider using Rényi correlation as the fairness regularization term.

Rényi Fair Inference. Rényi correlation measures the correlation between two random variables, ranging from 0 (independent) to 1 (strictly dependent). Unlike Pearson correlation, which captures linear relationships Zafar et al. [2017], Rényi correlation indicates high order dependencies Baharlouei et al. [2020]. It is also computationally tractable, compared to very expensive mutual information Song et al. [2019]. Rényi correlation between two random variables $A \in \{A_1, \cdots A_a\}$ and $B \in \{B_1, \cdots B_b\}$ is defined as:

$$\rho_R(A, B) = \sup_{f,g} \mathbb{E}[f(A)g(B)]$$
$$\text{s.t. } \mathbb{E}[f(A)] = \mathbb{E}[g(B)] = 0, \mathbb{E}[f^2(A)] = \mathbb{E}[g^2(B)] = 1.$$

Following Witsenhausen [1975], Baharlouei et al. [2020], Rényi correlation $\rho_R(a, b)$ is the second largest singular value of the matrix Q, where each element $q_{ij} = \frac{\mathbb{P}[A=A_i, B=B_j]}{\sqrt{\mathbb{P}[A=A_i]\mathbb{P}[B=B_j]}}$ for $1 \leq i \leq a$ and $1 \leq j \leq b$. The main idea of Rényi fair inference Baharlouei et al. [2020] is to minimize the correlation between predictions and sensitive attributes. Following Baharlouei et al. [2020], we can also re-formulate the squared term $\rho^2(A, B)$ as follows:

$$\rho^2(A, B) = \max_{\mathbf{v} \perp \mathbf{v}_1, \|\mathbf{v}\|^2 \leq 1} \mathbf{v}^\top Q^\top Q \mathbf{v}, \quad (4)$$

where $\mathbf{v}_1 = (\sqrt{\mathbb{P}[B = B_1]}, ..., \sqrt{\mathbb{P}[B = B_b]}) \in \mathbb{R}^b$ is the right singular vector associated with the largest singular value of Q.

In *centralized* machine learning, where samples are i.i.d. and accessible, the estimation error for statistics (e.g., the Rényi correlation) is bounded by $O(1/\sqrt{n})$ with high probability Mohri et al. [2018]. However, it remains unclear how accurately these statistics can be estimated in FL under *decentralized* and *heterogeneous* settings. Specifically, the following challenges arise: First, server is not allowed to access attribute data to calculate Rényi correlation in FL. This restriction complicates the global estimation of fairness measures; Second, clients often have vastly different data distributions due to statistical heterogeneity. This variation makes it difficult to precisely estimate the Rényi correlation across all devices. Third, due to varying computational capabilities and communication resources, the server may not receive updates from some clients (i.e., stragglers). This variability complicates the aggregation of local statistics and model updates. To address these challenges, we propose FedRényi algorithm.

4 FEDERATED RÉNYI-REGULARIZED LEARNING

In this section, we detail the design and theoretical analysis of the FedRényi algorithm. We begin by formulating the federated Rényi-regularized objective function. Next, we introduce the synchronous variant of FedRényi and present its theoretical analysis, highlighting improved estimation error bounds and convergence guarantees. Finally, we propose the asynchronous variant and analyze its approximation and estimation errors.

4.1 FEDERATED RÉNYI-REGULARIZED OBJECTIVE

We use the squared Rényi correlation as a regularization term combined with the federated loss $\widehat{L}(\theta)$ aforementioned in (3). Since all elements of the matrix Q in (4) are defined in population and unavailable to compute, instead, we aggregate the local statistics to estimate Q based on a fixed model θ. We denote this empirically aggregated estimation as $\widehat{Q}_\theta \in \mathbb{R}^{C\times P}$ (recall we have C classes and P attributes). For $1 \le c \le C, 1 \le p \le P$, the each entry in $\widehat{Q}_\theta$ is defined as $\widehat{q}_{cp} := \frac{\hat{j}(c,p)\cdot\hat{r}(p)}{\sqrt{\hat{u}(c)\cdot\hat{r}(p)}}$, where:

$$\begin{aligned}
\hat{j}(c,p) &= \sum_{k=1}^{K} \gamma_k \underbrace{\widehat{\mathbb{P}}[f_{\theta_k}(X_k,S_k)=c|S_k=p]}_{=\bar{j}_k(c,p)},\\
\hat{r}(p) &= \sum_{k=1}^{K} \gamma_k \underbrace{\widehat{\mathbb{P}}[S_k=p]}_{=\bar{r}_k(p)},\\
\hat{u}(c) &= \sum_{k=1}^{K} \gamma_k \underbrace{\widehat{\mathbb{P}}[f_{\theta_k}(X_k,S_k)=c]}_{=\bar{u}_k(c)},
\end{aligned} \tag{5}$$

where $\widehat{\mathbb{P}}[X_k \in D] := \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} \mathbb{1}[x_{ki} \in D]$ represents the empirical probability that X_k in any measurable set D.

Therefore, following (4), we define $\widehat{H}(\theta,\mathbf{v}) := \mathbf{v}^\top \widehat{Q}_\theta^\top \widehat{Q}_\theta \mathbf{v}$, and we formulate the federated Rényi-regularized objective:

$$\min_\theta \left\{ \widehat{L}(\theta) + \max_{\mathbf{v}\perp\hat{\mathbf{v}}_1, \|\mathbf{v}\|^2\le 1} \lambda \widehat{H}(\theta,\mathbf{v}) \right\}, \tag{6}$$

where $\hat{\mathbf{v}}_1 = \left(\sqrt{\hat{r}(1)}, ..., \sqrt{\hat{r}(P)}\right) \in \mathbb{R}^P$. The counterpart of $\widehat{H}(\theta,\mathbf{v})$ defined in population instead of empirical level is denoted by $H(\theta,\mathbf{v})$, as $\widehat{L}(\theta)$ in (3) and $L(\theta)$ in (1)

4.2 SYNCHRONOUS FEDRÉNYI

4.2.1 Algorithm Design.

To solve problem (6) without violating privacy constraints in FL, We propose the FedRényi algorithm (summarized in Algorithm 1). Specifically, we first initialize θ_0^0, $\mathbf{v}^0$. Then we compute $\bar{r}_k(p)$ and aggregate $\hat{r}(p)$ (see Line 1 to Line 3). During each communication round e, clients update the local model $\theta^e_{k,m+1}$ for $m \in 0,\cdots,M-1$(see Line 6 to Line 7). After completing local updates, each client calculates $\bar{j}^e_k(c,p)$ and $\bar{u}^e_k(c)$ (see Line 9 to Line 13), and then uploads these statistics and local model $\theta^e_{k,M}$ to server (see Line 14). For synchronous FedRényi (Option I), the server aggregates the global model θ^{e+1}, global statistics $\hat{j}^{e+1}(c,p)$, and $\hat{u}^{e+1}(c)$ (see Line 16 to 18). Next, we compute matrix $\widehat{Q}^{e+1}_\theta$ and then apply SVD method to calculate $\mathbf{v}^{e+1}$ (see Line 26 and 27). Finally, the server broadcasts the global model θ^{e+1} and the fairness component $\mathbf{v}^{e+1}$ to each client (see Line 28).

Algorithm 1 FedRényi Algorithm

1: Initialize θ_0^0, $\mathbf{v}^0$ and hyperparameter λ, M, J and η on server and clients
2: Each client $k \in K$ compute $\bar{r}_k(p)$ following Eq. (5) and upload $\bar{r}_k(p)$ and n_k
3: Server aggregate $\hat{r}(p) = \sum_{k=1}^K \gamma_k \bar{r}_k(p)$ and $\hat{\mathbf{v}}_1 = [\sqrt{\hat{r}(1)}, ..., \sqrt{\hat{r}(P)}]$
4: **for** $e = 0, \ldots, \tau-1$ **do**
5: **for** $m \in \{0, \cdots, M-1\}$ **do**
6: Each client k compute $Q_{\theta^e_{k,m}}$
7: $\theta^e_{k,m+1} = \theta^e_{k,m} - \eta\partial_\theta(\widehat{L}_k(\theta^e_{k,m}) + \lambda\hat{H}(\theta^e_{k,m},\mathbf{v}))$
8: **end for**
9: **for** $c \in \{1,\ldots,C\}$ **do**
10: **for** $p \in \{1,\ldots,P\}$ **do**
11: Compute $\{\bar{j}^e_k(c,p), \bar{u}^e_k(c)\}$ following Eq. (5)
12: **end for**
13: **end for**
14: Upload $\{\bar{j}^e_k(c,p), \bar{u}^e_k(c)\}$ and $\theta^e_{k,M}$
15: **Option I Synchronous FedRényi**:
16: $\theta_0^{e+1} = \sum_{k=1}^K \gamma_k \theta^e_{k,M}$
17: $\hat{j}^{e+1}(c,p) = \sum_{k=1}^K \gamma_k \bar{j}^e_k(c,p)$
18: $\hat{u}^{e+1}(c) = \sum_{k=1}^K \gamma_k \bar{u}^e_k(c)$
19: **Option II Asynchronous FedRényi**:
20: Find stragglers sets I^{e+1}, where $|I^{e+1}| = \widetilde{K}^{e+1}$
21: Find neighbor set $Rob_\zeta(\widetilde{k})$ for stragglers $\widetilde{k} \in I^{e+1}$
22: Approximate the $\widetilde{\theta}^e_{\widetilde{k},M}, \widetilde{j}^e_{\widetilde{k}}(c,p), \widetilde{u}^e_{\widetilde{k}}(c)$ for all stragglers $\widetilde{k} \in I^{e+1}$ by Algorithm 2
23: $\theta^{e+1} = \sum_{k=1}^{K-\widetilde{K}^{e+1}} \gamma_k \theta^e_{k,M} + \sum_{\widetilde{k}=1}^{\widetilde{K}^{e+1}} \gamma_{\widetilde{k}} \widetilde{\theta}^e_{\widetilde{k},M}$
24: $\hat{j}^{e+1}(c,p) = \sum_{k=1}^{K-\widetilde{K}^{e+1}} \gamma_k \bar{j}^e_k(c,p) + \sum_{\widetilde{k}=1}^{\widetilde{K}^{e+1}} \gamma_{\widetilde{k}} \widetilde{j}^e_{\widetilde{k}}(c,p)$
25: $\hat{u}^{e+1}(c) = \sum_{k=1}^{K-\widetilde{K}^{e+1}} \gamma_k \bar{u}^e_k(c) + \sum_{\widetilde{k}=1}^{\widetilde{K}^{e+1}} \gamma_{\widetilde{k}} \widetilde{u}^e_{\widetilde{k}}(c)$
26: Compute $\hat{Q}^{e+1}_\theta$ where each entry $\hat{q}^{e+1}_{c,p} = \frac{\hat{j}^{e+1}(c,p)\cdot\hat{r}(p)}{\sqrt{\hat{u}^{e+1}(c)\cdot\hat{r}(p)}}$
27: $\mathbf{v}^{e+1} \leftarrow \arg\max_{\mathbf{v}\perp\hat{\mathbf{v}}_1}[\widehat{L}(\theta^{e+1}) + \lambda\widehat{H}(\theta^{e+1},\mathbf{v})]$
28: Broadcast θ^{e+1} and $\mathbf{v}^{e+1}$ to all clients $k \in K$
29: **end for**

4.2.2 Theoretical analysis

In this part, we first study the estimation error for synchronous FedRényi from empirically aggregated $\widehat{H}(\theta, \mathbf{v})$ to population $H(\theta, \mathbf{v})$ under two weighting schemes in Theorem 1. Then, we discuss the convergence guarantee of synchronous FedRényi in Proposition 1. Particularly, we show how the estimation error bound derived in Theorem 1 benefits the convergence guarantee.

We first prove that, under mild conditions, the estimation error is bounded by $O(1/\sqrt{n})$, which significantly improves upon the prior result $O(1/\sqrt{K})$ reported in Chu et al. [2021], leading to improved accuracy and stability in fairness estimation.

Theorem 1. *(Estimation error of Rényi regularization for synchronous FedRényi) Suppose $j_{min} \backsim u_{min} \backsim r_{min} = O(1)$ and $n_{min} \backsim \frac{n}{K\log(K)}$. When $\gamma_k = \frac{n_k}{n}$ or $\frac{1}{K}$, for any global model θ and $\delta \in (0,1)$, the following inequality holds:*

$$\mathbb{P}\Big[\widehat{H}(\theta, \mathbf{v}) - H(\theta, \mathbf{v}) \leq O\big(1/\sqrt{n}\big)\big|\theta\Big] \geq 1 - \delta.$$

Remark 1. *The above theorem shows that for any fixed global model θ, the estimation error between population $H(\theta, \mathbf{v})$ and empirically aggregated $\widehat{H}(\theta, \mathbf{v})$ is bounded by $O(1/\sqrt{n})$ with high probability in two schemes. Compared with previous results on the order of $O(1/K)$, FedRényi significantly reduces estimation error, achieving a bound that is comparable to the standard estimation error in centralized settings Mohri et al. [2018].*

Then, we analyze the convergence guarantee of synchronous FedRényi to achieve ϵ-stationary solution. We first denote $F(\theta) = L(\theta) + \max_{\mathbf{v}\perp\mathbf{v}_1, \|\mathbf{v}\|^2\leq 1} \lambda H(\theta)$, and $\widehat{F}(\theta) = \widehat{L}(\theta) + \max_{\mathbf{v}\perp\mathbf{v}_1, \|\mathbf{v}\|^2\leq 1} \lambda \widehat{H}(\theta)$. Then, we consider the uniform-over-client weighting scheme following Karimireddy et al. [2020], i.e., $\gamma_k = 1/K$. We highlight that the convergence analysis for the empirical measure $\mathbb{E}[\|\nabla \hat{F}(\theta)\|^2]$ computed on infinite samples follows the same framework of previous work Karimireddy et al. [2020], Li et al. [2020b]. Our main target of the convergence analysis for the population measure $\mathbb{E}[\|\nabla F(\theta)\|^2]$ is more difficult to achieve while considering the impact of the estimation error between $H(\theta, \mathbf{v})$ and $\widehat{H}(\theta, \mathbf{v})$.

Proposition 1. *(Convergence of synchronous FedRényi) Suppose $\eta \leq O(1/M)$ and $L_k(\theta)$ satisfies (G_L, B_L)-bounded gradient dissimilarity, where $\frac{1}{K}\sum_{k=1}^K \|\frac{\partial L_k(\theta)}{\partial \theta}\|^2 \leq G_L^2 + B_L^2 \|\frac{\partial L(\theta)}{\partial \theta}\|^2$. If $\|\frac{\partial L(\theta)}{\partial \theta}\|^2, \|\frac{\partial Q_\theta}{\partial \theta}\|^2, \|\frac{\partial \widehat{Q}_\theta}{\partial \theta}\|^2, \|\frac{\partial \mathbf{v}_\theta}{\partial \theta}\|^2$ are bounded by $\bar{G}$ for all θ, Then, $H(\theta, \mathbf{v})$ satisfies (G_H, B_H)-bounded gradient dissimilarity, where G_H is $\frac{C\bar{G}(\rho+C)}{\rho^2}$ and B_H is 1. $F(\theta)$ also satisfies (G_F, B_F)-bounded gradient dissimilarity, where $B_F = 2B_L^2$ and $G_F = 2G_L^2 + (4\lambda - 2B_L^2\lambda^2)\frac{C\bar{G}(\rho+C)}{\rho^2} + 4B_L^2\lambda\cdot\sqrt{\frac{C\bar{G}^2(\rho+C)}{\rho^2}}$. Thus, FedRényi algorithm achieves $\mathbb{E}[\|\nabla \widehat{F}(\theta_T)\|^2] \leq \epsilon$ and $\mathbb{E}[\|\nabla F(\theta_T)\|^2] \leq \epsilon + O\Big(\frac{1}{n} + \max_\theta \big\|\frac{\partial \widehat{Q}_\theta}{\partial \theta} - \frac{\partial Q_\theta}{\partial \theta}\big\|^2\Big)\Big)$ when $T \geq O(1/\epsilon^2)$.*

Remark 2. *The above proposition shows that the empirical version of synchronous FedRényi algorithm could converge to ϵ-stationary solution and the population version could converge to approximate ϵ-stationary solution with gap $O\Big(\frac{1}{n} + \max_\theta \big\|\frac{\partial \widehat{Q}_\theta}{\partial \theta} - \frac{\partial Q_\theta}{\partial \theta}\big\|^2\Big)\Big)$, while the former is at the same order of previous work Karimireddy et al. [2020], Li et al. [2020b]. From Theorem 1, we know that the estimation error is bounded by $O(1/\sqrt{n})$, which is explicitly present in the above result. Thus, the improved estimation can benefit not only the fairness guarantee but also the convergence.*

4.3 ASYNCHRONOUS FEDRÉNYI

4.3.1 Algorithm Design

In asynchronous FL, stragglers may fail to provide timely updates, which can compromise accurate estimation of global Rényi correlation. To overcome this challenge, following Zhang et al. [2021], Wang et al. [2021], we assume that clients with similar empirical prediction distributions also have comparable data distributions, making the nearest-neighbor approximation a reasonable strategy for estimating missing updates. Consequently, even if a client fails to upload its model and local statistics within the communication threshold, its contribution to the global fairness measure can still be estimated reliably using its robust neighbors.

Specifically, we first identify a robust neighbor set for each straggler. Define $Rob_\zeta(\widetilde{k})$ is the neighbor set of straggler $\widetilde{k}$, where $Rob_\zeta(\widetilde{k}) := \{k' : \|\omega\bar{u}_{k'}(c) - \bar{u}_{\widetilde{k}}(c)\| \leq \zeta, k' \in [K], \forall c \text{ and } \forall \omega \in (0,1)\}$. Next, we compute the similarity between clients. Let $dist(\cdot,\cdot)$ denote the Euclidean distance, which represents the dissimilarity. The similarity between the k-th and k'-th client is then quantified by weights $W_{k,k'}$, where larger weights indicate greater similarity. Based on the local statistics uploaded in the first communication round, for a straggler $\widetilde{k}$, we define:

$$W_\theta^{\widetilde{k},k'} = \exp\Big(\frac{-dist(\theta^0_{\widetilde{k},M}, \theta^0_{k',M})}{\rho}\Big), , k' \in [K] \tag{7}$$

$$W_j^{\widetilde{k},k'} = \exp\Big(\frac{-dist(\bar{j}^0_{\widetilde{k},M}(c,p), \bar{j}^0_{k',M}(c,p))}{\rho}\Big), k' \in Rob_\zeta(\widetilde{k}),$$

$$W_u^{\widetilde{k},k'} = \exp\Big(\frac{-dist(\bar{u}^0_{\widetilde{k},M}(c), \bar{u}^0_{k',M}(c)}{\rho}\Big), k' \in Rob_\zeta(\widetilde{k}),$$

where ρ is the temperature parameter.

Then the server approximates the model parameter $\widetilde{\theta}^e_{\widetilde{k},M}$ and statistics $\widetilde{j}^e_{\widetilde{k},M}(c,p)$ and $\widetilde{u}^e_{\widetilde{k},M}(c)$ for each straggler $\widetilde{k}$. We summarize the approximation method in Algorithm 2.

Algorithm 2 Localized Approximation

1: **Input**: $\{\bar{j}^e_{k,M}(c,p), \bar{u}^e_{k,M}(c), \theta^e_{k,M}\}$ for all non-straggler clients $k \in [K]\backslash I^{e+1}$, and temperature parameter ρ.
2: Compute $W_{\widetilde{k},k'}$ for all stragglers following Eq. (7).
3: For each straggler $\widetilde{k}$, $\widetilde{\theta}^e_{\widetilde{k},M} = \frac{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{\widetilde{k},k'} \theta^e_{k',M}}{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{\widetilde{k},k'}}$.
4: $\widetilde{j}^e_{\widetilde{k},M}(c,p) = \frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W_j^{k,k'} \bar{j}^e_{k,M}(c,p)}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W_j^{k,k'}}$.
5: $\widetilde{u}^e_{\widetilde{k},M}(c) = \frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W_u^{k,k'} \bar{u}^e_{k,M}(c)}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W_u^{k,k'}}$.

After approximating local models and statistics for stragglers, we integrate these approximations into an asynchronous FedRényi algorithm to compute the global Rényi correlation (see Option II in Algorithm 1). In each round $e+1$, the server identifies the set of stragglers I^{e+1} (with $|I^{e+1}| = \widetilde{K}^{e+1}$) as those clients that either fail to return their local model and statistics within the communication threshold or have a local update timestamp below M (see Line 20). For each straggler $\widetilde{k}$, the server selects its robust neighbor set $Rob_\zeta(\widetilde{k})$ (see Line 21). Next, the server approximates the missing statistics $\widetilde{j}^e_{\widetilde{k},M}(c,p)$ and $\widetilde{u}^e_{\widetilde{k},M}(c)$, , as well as the model $\widetilde{\theta}^e_{\widetilde{k},M}$ for each straggler $\widetilde{k}$ by Algorithm 2 (see Line 20 and 22). Finally, global Rényi regularization statistics $j^{e+1}(c,p)$, $u^{e+1}(c)$, and the global model θ^{e+1} are aggregated (see Line 23 to 25).

4.3.2 Theoretical Analysis

In this part, we first analyze the approximation error of each straggler from actual local statistics $\bar{j}^e_{\widetilde{k},M}(c,p)$, $\bar{u}^e_{\widetilde{k},M}(c)$ and model $\theta^e_{\widetilde{k},M}$ to approximated statistics $\widetilde{j}^e_{\widetilde{k},M}(c,p)$, $\widetilde{u}^e_{\widetilde{k},M}(c)$ and model $\widetilde{\theta}^e_{\widetilde{k},M}$ in Proposition 2. Then, in Theorem 2, we analyze the estimation error of the asynchronous FedRényi algorithm, explicitly accounting for the approximation error.

Before analyzing the approximation error, we first assume the following assumptions:

Assumption 1. *(β-co-coercive condition of $\nabla F(\theta)$) For all clients and any model θ, the gradient of $F(\theta)$ satisfies β-co-coercive condition with $\beta \geq \frac{\eta}{2}$ if:*

$$\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle \geq \beta \|\nabla F(\theta_1) - \nabla F(\theta_2)\|^2.$$

Assumption 2. *(L-Lipschitz) $\widehat{\mathbb{P}}[f_\theta(X_k, S_k) = c]$ is L-lipschitz on model θ such that $|\mathbb{P}[f_\theta(X_k, S_k) = c] - \mathbb{P}[f_{\theta'}(X_k, S_k) = c]| \leq L\|\theta - \theta'\|$.*

With the two above assumptions, we can bound the approximation error of each straggler by the following proposition:

Proposition 2. *(Approximation error of each straggler in asynchronous FedRényi) Define $\max_{k,k'\in[K]} \|\theta^e_{k,0} - \theta^e_{k',0}\| = \varepsilon^e_0$. Suppose that Assumption 1 and 2 hold. Then, for each communication round e, the approximation errors of model and local statistics on stragglers $\widetilde{k}$ are upper bounded as follows:*

$$\begin{aligned} &\|\widetilde{\theta}^e_{\widetilde{k},M} - \theta^e_{\widetilde{k},M}\| \leq \varepsilon^e_0, \\ &|\widetilde{j}^e_{\widetilde{k},M}(c,p) - \bar{j}^e_{\widetilde{k},M}(c,p)| \leq L\varepsilon^e_0 + \zeta, \\ &|\widetilde{u}^e_{\widetilde{k},M}(c) - \bar{u}^e_{\widetilde{k},M}(c)| \leq L\varepsilon^e_0 + \zeta. \end{aligned} \tag{8}$$

Remark 3. *In the above result, the localized approximation for the stragglers shows non-expansion behavior (1st line), i.e., the error after running s stages is not larger than the error at the 1st iteration. In practice, we could set a smaller learning rate η (due to Assumption 1) and use pre-trained model to decrease the approximation error ε_0. Besides, using the pre-trained model, which is common in FL Tan et al. [2022], Weller et al. [2022], Tian et al. [2022], could make the Assumption 1 easy to hold.*

Then, we study how approximation error influences the estimation error of the asynchronous FedRényi algorithm. To explicitly investigate this impact, we define $\widetilde{Q}^{e+1}_\theta \in \mathbb{R}^{C\times P}$ as the global empirical matrix at communication round $e+1$, and $\widetilde{\mathbf{v}}^{e+1}$ is its corresponding second largest singular vector. Therefore, the empirical objective function in asynchronous setting could be rewritten as $\widehat{H}(\theta^{e+1}, \widetilde{\mathbf{v}}^{e+1}) = (\widetilde{\mathbf{v}}^{e+1})^\top (\widetilde{Q}^{e+1}_\theta)^\top \widetilde{Q}^{e+1}_\theta \widetilde{\mathbf{v}}^{e+1}$. Our goal is to study the estimation between $\widehat{H}(\theta^{e+1}, \widetilde{\mathbf{v}}^{e+1})$ and $H(\theta^{e+1}, \mathbf{v}^{e+1})$.

Theorem 2. *(Estimation error of Rényi regularization for asynchronous FedRényi) Suppose $j_{min} \backsim u_{min} \backsim r_{min} = O(1)$ and $n_{min} \backsim \frac{n}{K\log(K)}$. When $\gamma_k = \frac{n_k}{n}$ or $\frac{1}{K}$, for any communication round e, any global model θ^{e+1} and $\delta \in (0,1)$, we have the following inequality holds:*

$$\begin{aligned} &\mathbb{P}\Big[\widehat{H}(\theta^{e+1}, \widetilde{\mathbf{v}}^{e+1}) - H(\theta^{e+1}, \mathbf{v}^{e+1}) \\ &\quad \leq O\big(1/\sqrt{n} + (L\varepsilon^e_0 + \zeta)^2\big) \Big| \theta^{e+1}\Big] \geq 1 - \delta. \end{aligned}$$

Remark 4. *The above theorem shows that, with high probability, the estimation error for asynchronous FedRényi between population $H(\theta^{e+1}, \mathbf{v}^{e+1})$ and empirically aggregated $\widehat{H}(\theta^{e+1}, \widetilde{\mathbf{v}}^{e+1})$ in two consecutive communication rounds is bounded by $O(1/\sqrt{n} + (L\varepsilon^e_0 + \zeta)^2)$ and follows a stage-wise recurrence. For each communication round e, the dynamic estimation error is bounded by fixed term $O(1/\sqrt{n})$ and the largest distance of model at the beginning of each stage ε^e_0. During communication, $\varepsilon^{e+1}_0 \leq \varepsilon^e_M \leq \varepsilon^e_0$ (see Section B.2.3 and Equation (24) in Appendix). Thus, the estimation error decreases as the communication progress. Unlike previous works Sefidgaran et al. [2024], Hu et al. [2023], our global estimation error studies the inter-client and intra-client estimation error of all clients, including stragglers and participating clients.*

Table 1: Experimental results of all methods with the heterogeneous setting ($Dir = 0.5$) on four datasets. For ACC, FR, and HM, higher values indicate better performance. Accuracy, fairness, and harmonic mean are denoted by ACC, FR, and HM, respectively. The best results are in **bold**. The mean and standard deviation of 20 results with better HM for each method under different hyperparameter settings are presented. Comparing FedRényi with other baselines, there exist at least 2% improvements of ACC, FR and HM over three datasets (ADULT, DRUG, DUTCH).

	FedAvg	**FedProx**	**Scaffold**	**FedFair**	**FL-FairBatch**	**FedFB**	**FairFed**	**FedRényi** $(1/K)$	**FedRényi** (n_k/n)
ADULT									
ACC	0.62±0.12	0.61±0.12	0.56±0.20	0.51±0.07	0.64±0.00	0.65±0.00	0.62±0.17	**0.67±0.03**	0.65±0.04
FR	0.87±0.1	0.88±0.11	0.88±0.13	0.84±0.17	0.91±0.02	0.92±0.03	0.77±0.16	**0.94±0.04**	**0.94±0.04**
HM	0.72±0.11	0.72±0.11	0.68±0.16	0.63±0.10	0.75±0.00	0.76±0.00	0.69±0.16	**0.78±0.03**	0.77±0.04
COMPAS									
ACC	0.66±0.01	0.66±0.01	0.47±0.12	0.62±0.03	0.67±0.01	0.67±0.01	0.62±0.03	**0.68±0.01**	**0.68±0.01**
FR	0.79±0.03	0.79±0.03	**0.82±0.10**	0.79±0.10	0.78±0.02	0.75±0.03	0.79±0.10	0.81±0.02	0.82±0.01
HM	0.72±0.01	0.72±0.01	0.60±0.11	0.69±0.05	0.72±0.01	0.71±0.01	0.69±0.05	0.72±0.03	**0.73±0.02**
DRUG									
ACC	0.67±0.02	0.67±0.01	0.66±0.01	0.67±0.02	0.66±0.00	0.66±0.00	0.50±0.08	0.68±0.01	**0.69±0.01**
FR	0.86±0.02	0.86±0.02	0.82±0.06	0.86±0.02	0.84±0.00	0.85±0.00	0.77±0.10	**0.96±0.03**	0.96±0.02
HM	0.75±0.02	0.75±0.01	0.73±0.02	0.75±0.02	0.74±0.00	0.74±0.00	0.61±0.09	**0.80±0.01**	**0.80±0.01**
DUTCH									
ACC	0.81±0.01	0.80±0.01	0.60±0.12	0.61±0.16	0.81±0.01	0.69±0.05	0.62±0.13	**0.83±0.01**	**0.83±0.01**
FR	0.64±0.08	0.63±0.09	0.84±0.18	0.65±0.35	0.66±0.06	0.92±0.04	0.78±0.25	0.94±0.04	**0.96±0.04**
HM	0.72±0.02	0.7±0.02	0.7±0.14	0.63±0.22	0.73±0.02	0.79±0.04	0.69±0.17	0.88±0.02	**0.89±0.02**

5 NUMERICAL EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Hyperparameters and Dataset. In this paper, we use several combinations of hyperparameters (λ, ρ, $T\&M$, α, and Dir) to train FL models. We use ADULT, COMPAS, DRUG, and DUTCH datasets, which are widely studied benchmarks for fairness evaluation in FL. They vary in size and demographic attributes, allowing comprehensive fairness analysis. More details are provided in Appendix C.1.

Measurement. We use the accuracy (ACC), fairness score (FR), and harmonic mean (HM) of ACC and FR to measure the performance. To evaluate global accuracy (ACC) in FL, we compute the local accuracy of each client and aggregate them using either n_k/n for uniform data or $1/K$ uniform client settings across clients. To measure how unfair a model is, Donini et al. [2018] propose DEO by extending the equal opportunity (EOD) Hardt et al. [2016] as follows: $|\mathbb{P}(f_\theta(X,S)|S=1,Y=1) - \mathbb{P}(f_\theta(X,S)|S=0,Y=1)|$, and the FR is extended by $FR = 1 - DEO(f_\theta)$. For ACC, FR, and HM, higher values indicate better performance.

Data Distribution Setting. To simulate the IID and non-IID data distribution setting, we build quantity skew and control heterogeneity levels through Dir following Li et al. [2022], Ezzeldin et al. [2023], Lee et al. [2023]. Smaller Dir indicates a more imbalanced scenario about data quantity across clients, and $Dir = +\infty$ represents the uniform case.

Baselines. We include general FL methods (FedAvg McMahan et al. [2017], FedProx Li et al. [2020b], Scaffold Karimireddy et al. [2020]) and fairness-aware FL baselines (FedFair Chu et al. [2021], FL-FairBatch Roh et al. [2021], FedFB Zeng et al. [2021], FairFed Ezzeldin et al. [2023]) to compare fairness trade-offs in FL settings. Following Baharlouei et al. [2020], Chu et al. [2021], Zeng et al. [2021], Ezzeldin et al. [2023], we use the logistic regression model as backbone. More details are shown in Appendix C.1. The FedRényi code is available at https://github.com/AllenMa97/Federated-Renyi.

5.2 EXPERIMENT RESULT

FedRényi consistently outperforms baseline methods. The main experimental results are summarized in Table 1. Since hyperparameter affect the performance of algorithms, we select top 20 results (with better HM) for each method and report their mean and standard deviaition to ensure reliable comparisons. In ADULT, DRUG, and DUTCH, FedRényi outperforms other algorithms. Although FedRényi does not outperform all baselines in COMPAS, its HM rank second with a small gap from the highest (≤ 0.1). These results demonstrate the effectiveness of FedRényi. More detailed results for all datasets are supplemented in the Table 9, 10, 11, and 12.

FedRényi effectively balances accuracy and fairness by adjusting λ. To further examine the trade-off, we adjust the regularization coefficient λ within $\{0.1, 0.5, 1, 5, 1000\}$ and visually present some experimental results in Figure 1. As shown in Figure 1, the FR of FedRényi becomes larger as λ enlarges at the heterogeneous setting and the ACC increases as λ becomes smaller. More experimental results on four datasets are presented in Figure 6.

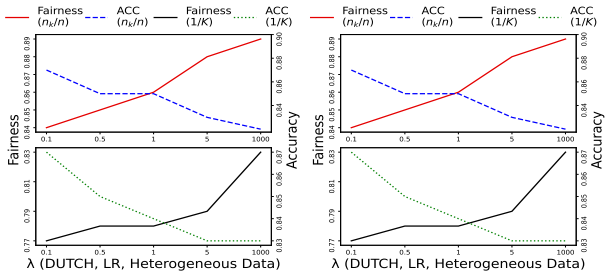

Figure 1: The accuracy and fairness trade-off adjusting via λ of FedRényi in DUTCH with heterogeneous and isomorphic setting. We could observe that the fairness increase and accuracy decrease with a larger λ value.

FedRényi takes the optimal trade-off between accuracy and fairness. A comparison of ACC and FR across different algorithms is shown in Figure 2. Only the top 5 results (with better HM value) of each algorithm will be plotted, and some methods show less than 5 points are caused by overlap. Intuitively, red and yellow scatter points (FedRényi) are closer to the optimal corner than others in most cases. Besides, these scatters approximately form several curves, exhibiting the trade-off ability between ACC and FR. More results on four datasets are presented in Figure 7.

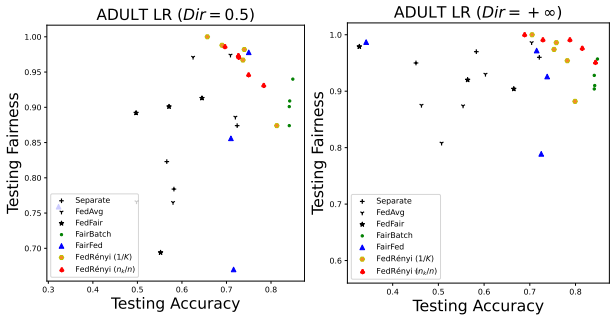

Figure 2: The ACC and FR trade-off on ADULT of all methods with two distribution settings. FedRényi performs closer to optimal and approximately forms trade-off curves from the most accurate and least fair to the least accurate and most fair.

FedRényi converges. To study the convergence behavior, training loss at different communication rounds on four datasets are illustrated in Figure 3. The training loss of synchronous FedRényi decreases as communication proceeds

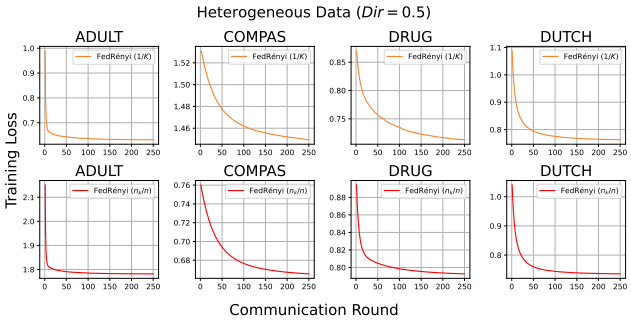

Figure 3: The training loss of FedRényi under heterogeneity data settings, which verify that FedRényi converges to a stable range after a certain number of rounds.

Table 2: The HM and the average approximation errors across all stragglers of FedRényi with different α. These approximation errors are measured by the L2 distance between the approximation values and the actual values on stragglers.

Dir=0.5 λ=1	**Drop α: 0%**	**Drop α: 30%**	**Drop α: 50%**
(T, I) =(100,4)	**HM/j Err./ u Err./θ Err.**	**HM/j Err./ u Err./θ Err.**	**HM/j Err./ u Err./θ Err.**
COMPAS			
Asyn. (n_k/n)	0.77/0/ 0/0	0.73/0.03/ 0.01/0.92	0.71/0.01/ 0.02/1.41
Asyn. ($1/K$)	0.75/0/ 0/0	0.76/0.04/ 0.01/0.27	0.78/0.01/ 0.02/0.34
Syn. (n_k/n)	0.76/0/ 0/0	0.76/0/ 0/0	0.76/0/ 0/0
DRUG			
Asyn. (n_k/n)	0.74/0/ 0/0	0.74/0.01/ 0.01/0.25	0.75/0.08/ 0.03/0.40
Asyn. ($1/K$)	0.73/0/ 0/0	0.72/0.01/ 0.02/0.29	0.73/0.09/ 0.02/0.37
Syn. (n_k/n)	0.74/0/ 0/0	0.74/0/ 0/0	0.74/0/ 0/0

and becomes stable at around 50 rounds, verifying the convergence property. More results on other datasets are shown in Figure 8 and 9.

Asynchronous FedRényi maintains stable HM performance and effectively controls estimation errors. To verify the performance with the asynchronous FedRényi, we build experiments and simulate different communication thresholds by controlling the proportion of straggler α. As shown in Table 2, the asynchronous FedRényi not only maintains stable HM performance but also achieves effective approximation error control. More experimental results on other datasets are presented in Appendix C.3.

Assumption 1 is empirically valid. To empirically demonstrate the validity of Assumption 1, we conduct experiments on the DUTCH dataset under a heterogeneous ($Dir = 0.5$)

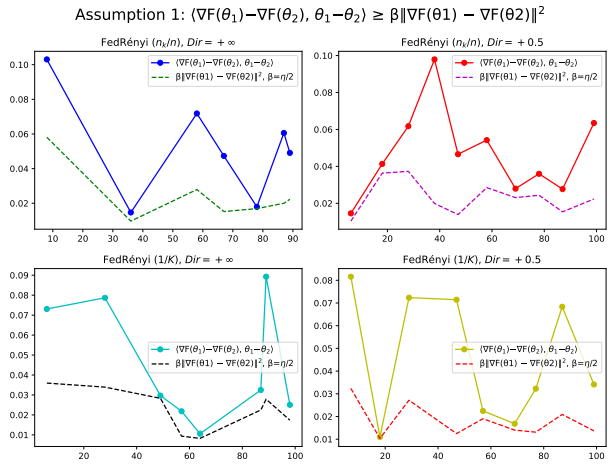

Figure 4: Verification of the co-coercivity assumption (Assumption 1) on DUTCH under $Dir = 0.5$ and $Dir = +\infty$, with uniform over data ($\lambda = n_k/n$) and uniform over client ($\lambda = n_k/n$) settings and $\lambda = 1$. The X-axis represents training iterations, while the Y-axis shows the values corresponding to each side of the co-coercivity inequality: (i) $\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle$, the left side of the inequality in Assumption 1; (ii) $\beta \|\nabla F(\theta_1) - \nabla F(\theta_2)\|^2$ with $\beta = \eta/2$, the right side of the inequality in Assumption 1.

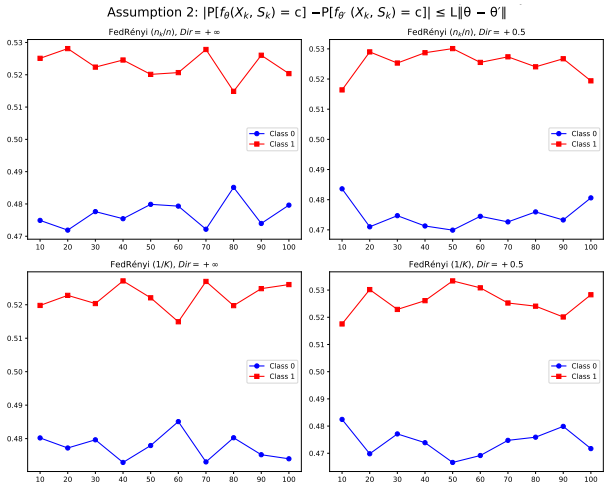

Figure 5: Verification of Lipschitz assumption (Assumption 2) on DUTCH dataset under a $Dir = 0.5$ and $Dir = +\infty$, with uniform over data ($\lambda = n_k/n$) and uniform over client ($\lambda = n_k/n$) settings and $\lambda = 1$. The X-axis denotes training iterations, and the Y-axis represents the predicted probability for each class.

and isomorphism Data ($Dir = +\infty$) distribution setting, with uniform over data ($\gamma = n_k/n$) and uniform over client ($\gamma = n_k/n$) settings and $\lambda = 1$. Specifically, in Figure 4, the X-axis represents training iterations, while the Y-axis shows the values corresponding to each side of the co-coercivity inequality: (i) $\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle$, the left-hand side of the inequality in Assumption 1; (ii) $\beta \|\nabla F(\theta_1) - \nabla F(\theta_2)\|^2$ with $\beta = \eta/2$, the right-hand side of the inequality in Assumption 1. At each iteration, we randomly select 5 clients and record their gradients and parameter vectors. Then we compute $\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle$ and $\beta \|\nabla F(\theta_1) - \nabla F(\theta_2)\|^2$ values for these clients, following the inequality structure in Assumption 1. Next, we plot the average across all selected clients over each iteration. As shown in Figure 4, in most iterations, the line of $\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle$ consistently lies above the line of $\beta \|\nabla F(\theta_1) - \nabla F(\theta_2)\|^2$, indicating that the inequality holds, which verifying the validity of the co-coercivity assumption (i.e., $\langle \nabla F(\theta_1) - \nabla F(\theta_2), \theta_1 - \theta_2 \rangle \geq \beta \|\nabla F(\theta_1) - \nabla F(\theta_2)\|^2$) in practice.

Assumption 2 is empirically valid. We plot Figure 5 to empirically demonstrate the validity of Assumption 2, which assumes that the change in predicted class probabilities is Lipschitz continuous with respect to model parameters θ. We adopt the same setup: DUTCH dataset, $\lambda = 1$, under both a heterogeneous ($Dir = 0.5$) and isomorphism data ($Dir = +\infty$) distribution settings and two uniform over data ($\gamma = n_k/n$) and uniform over client ($\gamma = n_k/n$) weights settings. Each subplot in Figure 5 visualizes the prediction probability of Class 0 (blue) and Class 1 (red) over training iterations. The X-axis denotes training iterations, and the Y-axis represents the predicted probability for each class. To compute these probabilities, at each iteration, we compute the local predicted class probabilities over all clients and then compute the average over clients to obtain the global prediction probability. As shown in Figure 5, the predicted probabilities for both classes evolve smoothly and do not exhibit sharp fluctuations throughout the iterations. This consistent behavior across multiple configurations empirically supports the Lipschitz continuity assumption with respect to θ.

6 CONCLUSION

We propose FedRényi, a federated fairness-aware algorithm that enhances group fairness in decentralized heterogeneous systems under two weighting schemes. FedRényi addresses data heterogeneity by aggregating local empirical statistics to estimate global Rényi correlation, with an estimation error of $O(1/\sqrt{n})$, matching centralized learning bounds and improving upon prior estimation error bounds. FedRényi algorithm reduces the expected squared gradient norm to $O(\epsilon + 1/n)$ with an iteration complexity of $O(1/\epsilon^4)$. For system heterogeneity, asynchronous FedRényi uses weighted averaging over a nearest neighbor region to approximate stragglers, with a non-increasing approximation error over a communication round. Our experiments on multiple benchmark datasets clearly demonstrate that FedRényi could achieve better accuracy and fairness trade-off over prior FL fairness baselines with at least 2% improvement.

References

Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *ICLR*, 2020.

Hao Ban and Kaiyi Ji. Fair resource allocation in multitask learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *NeurIPS*, volume 31, 2018.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints, 2020.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P Friedlander, Changxin Liu, and Yong Zhang. Improving fairness for data valuation in federated learning. *arXiv preprint arXiv:2109.09046*, 2021.

Elaine Fehrman, Awaz K. Muhammad, Evgeny M. Mirkes, Vincent Egan, and Alexander N. Gorban. The five factor model of personality and evaluation of drug consumption risk. pages 231–242, 2017.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. 2016.

Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *The Eleventh International Conference on Learning Representations*, 2023.

Shreyansh Jain and Koteswar Rao Jerripothula. Federated learning for commercial image sources. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6534–6543, 2023.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. page 202–207, 1996.

Jeff Larson, Marjorie Roswell, and Vaggelis Atlidakis. Compas analysis. URL https://github.com/propublica/compas-analysis/. 2016.

Royson Lee, Minyoung Kim, Da Li, Xinchi Qiu, Timothy Hospedales, Ferenc Huszár, and Nicholas Donald Lane. Fedl2p: Federated learning to personalize. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=FM81CI68Iz.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *IEEE International Conference on Data Engineering*, 2022.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

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 I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020b. URL https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated learning: privacy and incentive*, pages 240–254. Springer, 2020.

Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer, 2020.

Priyanka Mary Mammen. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*, 2021.

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 *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Alfred Rényi. *Foundations of probability*. Courier Corporation, 2007.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021.

Milad Sefidgaran, Romain Chor, Abdellatif Zaidi, and Yijun Wan. Lessons from generalization error analysis of federated learning: You may communicate less often! In *Forty-first International Conference on Machine Learning*, 2024.

Khotso Selialia, Yasra Chandio, and Fatima M Anwar. Mitigating group bias in federated learning for heterogeneous devices. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1043–1054, 2024.

Hyowoon Seo, Jihong Park, Seungeun Oh, Mehdi Bennis, and Seong-Lyun Kim. Federated knowledge distillation. 2016.

Jiawei Shao, Zijian Li, Wenqiang Sun, Tailin Zhou, Yuchang Sun, Lumin Liu, Zehong Lin, and Jun Zhang. A survey of what to share in federated learning: perspectives on model utility, privacy leakage, and communication efficiency. *arXiv preprint arXiv:2307.10655*, 2023.

Yuxin Shi, Han Yu, and Cyril Leung. A survey of fairness-aware federated learning. *arXiv preprint arXiv:2111.01872*, 2021.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35:19332–19344, 2022.

Yuanyishu Tian, Yao Wan, Lingjuan Lyu, Dezhong Yao, Hai Jin, and Lichao Sun. Fedbert: When federated learning meets pre-training. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–26, 2022.

Yuanli Wang, Joel Wolfrath, Nikhil Sreekumar, Dhruv Kumar, and Abhishek Chandra. Accelerated training via device similarity in federated learning. In *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*, pages 31–36, 2021.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. Pretrained models for multilingual federated learning. *arXiv preprint arXiv:2206.02291*, 2022.

Hans S Witsenhausen. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

Chenhao Xu, Youyang Qu, Yong Xiang, and Longxiang Gao. Asynchronous federated learning on heterogeneous devices: A survey. *Computer Science Review*, 50:100595, 2023.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060. IEEE, 2020.

Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34:10092–10104, 2021.

Xinwen Zhang, Yihan Zhang, Tianbao Yang, Richard Souvenir, and Hongchang Gao. Federated compositional deep auc maximization. *arXiv preprint arXiv:2304.10101*, 2023.

Ligeng Zhu, Hongzhou Lin, Yao Lu, Yujun Lin, , and Song Han. Delayed gradient averaging: Tolerate the communication latency in federated learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021a.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021b.

Federated Rényi Fair Inference in Federated Heterogeneous System (Supplementary Material)

Zhiyong Ma*[1] **Yuanjie Shi***[2] **Yan Yan**[2] **Jian Chen**[1]

[1]School of SSE, South China University of Technology University, Guangzhou, Guangdong, China
[2]School of EECS, Washington State University, Pullman, WA, USA

A MATHEMATICAL NOTATIONS

Table 3: **Key notations used in this paper.**

Notation	Meaning
$X \in \mathcal{X}$	Input feature
$Y \in \mathcal{Y}$	The ground-truth label
$S \in \mathcal{S}$	Sensitive attributes
$\mathcal{P}$	The underlying distribution defined on $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$
(X, Y, S)	The data sample drawn from a distribution $\mathcal{P}$
(x, y, s)	The realization of data sample
C	Total number of classes
P	Total number of attributes
k	Index of client
X_k	Feature variables on client k
S_k	Attribute variables on client k
$\mathcal{P}_k$	Distribution on client k
f_θ	The prediction function parameterised by θ
n_k	The number of data examples for client k
$n_{\min} := \min_k n_k$	The minimal number of data samples across all clients
$\mathbb{1}[\cdot]$	An indicator function
$\mathbb{P}_c = P[f_\theta(X, S) = c]$	The probability of the model predicting class c
$\rho = \min_{c \in C} \mathbb{P}_c$	The smallest model prediction probability over all classes.
τ	Total number of communication rounds
e	Index of communication rounds
M	Total number of local updates
m	Index of local updates
T	Total number of iterations

B THEORETICAL ANALYSIS

In this section, we prove the theoretical results in this paper.

B.1 PROOF FOR SECTION 4.2.2

In this section, we prove Theorem 1 and Proposition 1 in Section 4.2.2.

B.1.1 Proof of Theorem 1

Theorem 3. *(Theorem 1 restated, Estimation error of Rényi regularization for synchronous FedRényi) Suppose $j_{min} \backsim u_{min} \backsim r_{min} = O(1)$ and $n_{min} \backsim \frac{n}{K\log(K)}$. When $\gamma_k = \frac{n_k}{n}$ or $\frac{1}{K}$, for any global model θ and $\delta \in (0,1)$, the following inequality holds:*

$$\mathbb{P}\Big[\widehat{H}(\theta,\mathbf{v}) - H(\theta,\mathbf{v}) \leq O\big(1/\sqrt{n}\big)\big|\theta\Big] \geq 1-\delta.$$

Proof. of (Theorem 1)

Before proving Theorem 1, we first present some technical lemmas. Lemma 1 and 2 provides the estimation error of general random variables for $\gamma_k = n_k/n$ and $\gamma_k = 1/K$ respectively. Lemma 3 shows that the estimation errors from Lemma 1 and 2 can be transferred to each entry of matrix $\hat{Q}_\theta$. Lemma 4 bounds the norm of matrix Q_θ

Lemma 1. *(mean-of-sum for $\gamma_k = \frac{n_k}{n}$) For any distribution $\mathcal{P}_k$ on different clients, denoting $V_k = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} V_{k,i}$, then the condition $\gamma_k = \frac{n_k}{n}$ gives*

$$\mathbb{P}\left\{\left|\mathbb{E}[V] - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} V_{k,i}\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}\right\} \geq 1-\delta. \tag{9}$$

Lemma 2. *(mean-of-mean for $\gamma_k = \frac{1}{K}$) For any distribution $\mathcal{P}_k$ on different clients, define $n_{min} := \min_{k=1,\dots,K} n_k$ as the minimal number of data samples across different clients, and $\mu_{min} := \min_k V_k$, then the condition $\gamma_k = \frac{1}{K}$ gives*

$$\mathbb{P}\left\{\left|\frac{1}{K}\sum_{k=1}^{K} V_k - \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} V_k\right]\right| \leq \sqrt{\frac{\log(2K/\delta)\cdot\log(4/\delta)}{K n_{min}\mu_{min}}}\right\} \geq 1-\delta. \tag{10}$$

Lemma 3. *Suppose $|j(c,p) - \hat{j}(c,p)| \leq \epsilon$, $|u(c) - \hat{u}(c)| \leq \epsilon$ and $|r(p) - \hat{r}(p)| \leq \epsilon$. Under $\epsilon = O\Big(\frac{1}{\sqrt{n}}\Big)$ and $\hat{j}(c,p) \backsim \hat{u}(c) \backsim \hat{r}(p) = \Omega\big(\frac{1}{\sqrt{n}}\big)$, the following inequality holds:*

$$\frac{j(c,p)r(p)}{\sqrt{u(c)r(p)}} \leq \frac{\hat{j}(c,p)\hat{r}(p)}{\sqrt{\hat{u}(c)\hat{r}(p)}} + O(\epsilon). \tag{11}$$

Lemma 4. *Recall that $\rho = \min_{c\in C}\mathbb{P}_c$ as the smallest model prediction probability over all classes. Then, the norm of matrix Q_θ is bounded : $\|Q_\theta\| \leq \sqrt{\frac{C}{\rho}}$.*

Now we begin to prove Theorem 1. Given fixed global model θ, according to Lemma 3, the estimation error between each entry of matrix Q_θ (i.e., q_{cp}) and $\widehat{Q}_\theta$ (i.e., $\widehat{q}_{cp}$) is bounded by $O(\epsilon)$, where ϵ is the estimation error bound in Lemma 1 or 2. Thus, we define $\varepsilon \in \mathbb{R}^{C\times P}$ as an estimation error matrix and $\varepsilon := Q_\theta - \hat{Q}_\theta$, where each entry $\varepsilon_{c,p} = \epsilon$ is the same.

First, we prove the estimation error of Rényi when $\gamma_k = \frac{n_k}{n}$:

$$
\begin{aligned}
\widehat{H}(\theta, \mathbf{v}) &= \mathbf{v}^\top \hat{Q}_\theta^\top \hat{Q}_\theta \mathbf{v} \\
&= \mathbf{v}^\top (\hat{Q}_\theta^\top \hat{Q}_\theta) \mathbf{v} \\
&\leq \mathbf{v}^\top \Big[(Q_\theta + \varepsilon)^\top (Q_\theta + \varepsilon) \Big] \mathbf{v} \\
&= \mathbf{v}^\top \Big[Q_\theta^\top Q_\theta + \varepsilon^\top \hat{Q}_\theta + Q_\theta^\top \varepsilon + \varepsilon^\top \varepsilon \Big] \mathbf{v} \\
&\leq \mathbf{v}^\top \Big[Q_\theta^\top Q_\theta + 2\|\varepsilon\| \cdot \|\hat{Q}_\theta\| + \|\varepsilon\|^2 \Big] \mathbf{v} \\
&\leq \mathbf{v}^\top \Big[Q_\theta^\top Q_\theta + 2(C \cdot P)^2 \cdot \epsilon \cdot \sqrt{\frac{C}{\rho}} + (C \cdot P)^2 \cdot \epsilon^2 \Big] \mathbf{v} \\
&= \mathbf{v}^\top Q_\theta^\top Q_\theta \mathbf{v} + \mathbf{v}^\top \Big[2(C \cdot P)^2 \cdot \epsilon \cdot \sqrt{\frac{C}{\rho}} + (C \cdot P)^2 \cdot \epsilon^2 \Big] \mathbf{v} \\
&\leq \mathbf{v}^\top Q_\theta^\top Q_\theta \mathbf{v} + \|\mathbf{v}\| \cdot \Big[(2(C \cdot P)^2 \cdot \epsilon \cdot \sqrt{\frac{C}{\rho}} + (C \cdot P)^2 \cdot \epsilon^2 \Big] \cdot \|\mathbf{v}\| \\
&\leq \mathbf{v}^\top Q_\theta^\top Q_\theta \mathbf{v} + \Big[2(C \cdot P)^2 \cdot \epsilon \cdot \sqrt{\frac{C}{\rho}} + (C \cdot P)^2 \cdot \epsilon^2 \Big],
\end{aligned}
$$

Where the second, third, and fourth inequality is due to Cauchy–Schwarz inequality.

Let ϵ equals to the result in Lemma 1, the following inequality holds:

$$
\mathbb{P}\Big[\widehat{H}(\theta, \mathbf{v}) - H(\theta, \mathbf{v}) \leq C^2 P^2 \big(\frac{\log(2/\delta)}{2n} + 2\sqrt{\frac{C \log(2/\delta)}{2n\rho}} \big) \Big| \theta \Big] \geq 1 - \delta.
$$

Therefore, we have that:

$$
\mathbb{P}\Big[\widehat{H}(\theta, \mathbf{v}) - H(\theta, \mathbf{v}) \leq \tilde{O}\Big(\frac{1}{\sqrt{n}} \Big) \Big| \theta \Big] \geq 1 - \delta, \text{ when } \gamma_k = n_k / n. \tag{12}
$$

Now we begin to prove estimation error when $\gamma_k = \frac{1}{K}$.

Let the estimation error of $\hat{j}_k(c, p)$, $\hat{u}_k(c)$ and $\hat{r}_k(p)$ be ϵ_j, ϵ_u and ϵ_r respectively.

Under the condition of $j_{\min} \backsim u_{\min} \backsim r_{\min} = O(1)$, according to lemma 2, the following holds:

$$
\begin{aligned}
\epsilon_j &= \sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min} j_{\min}}} = \sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min}}}, \\
\epsilon_u &= \sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min} u_{\min}}} = \sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min}}}, \\
\epsilon_r &= \sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min} r_{\min}}} = \sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min}}}.
\end{aligned}
$$

Thus, the estimation errors of $\hat{j}_k(c, p)$, $\hat{u}_k(c)$ and $\hat{r}_k(p)$ are the same: $\sqrt{\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min}}}$.

Then, the following inequality holds:

$$
\mathbb{P}\Bigg[\widehat{H}(\theta, \mathbf{v}) - H(\theta, \mathbf{v}) \leq C^2 P^2 \Big(\frac{\log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min}} + 2\sqrt{\frac{C \log(2K/\delta) \cdot \log(4/\delta)}{K n_{\min} \rho}} \Big) \Bigg| \theta \Bigg] \geq 1 - \delta.
$$

Due to $n_{\min} \backsim \frac{n}{K\log(K)}$, we have that:

$$\mathbb{P}\Big[\widehat{H}(\theta,\mathbf{v}) - H(\theta,\mathbf{v}) \leq \tilde{O}\Big(\frac{1}{\sqrt{n}}\Big)\Big|\theta\Big] \geq 1-\delta, \text{ where } \gamma_k = 1/K. \tag{13}$$

Combining Equation (12) and (13), we have that:

$$\mathbb{P}\Big[\widehat{H}(\theta,\mathbf{v}) - H(\theta,\mathbf{v}) \leq \tilde{O}\Big(\frac{1}{\sqrt{n}}\Big)\Big|\theta\Big] \geq 1-\delta. \tag{14}$$

Thus, the proof of Theorem 1 is finished.

□

B.1.2 Proof of Lemma 1

Proof. (of Lemma 1) Define $X_{k,i} = \mathbb{1}[\mathcal{E}_{k,i}]$ for an event $\mathcal{E}_{k,i}$. This event $\mathcal{E}$ can be instantiated with the components contained in Q_θ. For example, it can be defined by the case where for a data (x,y,s), $s=0$, $f_{\theta_k,t}(x)=1$, or $s=1$ and $f_{\theta_k,t}(x)=1$.

Define $V_{k,i} := X_{k,i}$. Then we apply Hoeffding's inequality across n data samples as follows

$$\begin{aligned}
&\mathbb{P}\Bigg\{\Bigg|\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} V_{k,i} - \mathbb{E}\Bigg[\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} V_{k,i}\Bigg]\Bigg| \leq t\Bigg\} \\
\geq& 1 - 2\exp\Big(-\frac{2t^2}{\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}(1)^2}\Big) \\
\geq& 1-\delta,
\end{aligned}$$

or equivalently

$$\mathbb{P}\Bigg\{\Bigg|\mathbb{E}[V] - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} V_{k,i}\Bigg| \leq \sqrt{\frac{\log(2/\delta)}{2n}}\Bigg\} \geq 1-\delta. \tag{15}$$

□

B.1.3 Proof of Lemma 2

Proof. (of Lemma 2) Define $X_{k,i} = \mathbb{1}[\mathcal{E}_{k,i}]$ for an event $\mathcal{E}_{k,i}$. This event $\mathcal{E}$ can be instantiated with the components contained in Q_θ. For example, it can be defined by the case where for a data (x,y,s), $s=0$, $f_{\theta_k,t}(x)=1$, or $s=1$ and $f_{\theta_k,t}(x)=1$.

For fixed $k \in \{1,...,K\}$, applying Chernoff bound, we have

$$\mathbb{P}\Bigg\{\Bigg|\sum_{i\in\mathcal{I}_k} X_i - \mathbb{E}\Bigg[\sum_{i\in\mathcal{I}_k} X_i\Bigg]\Bigg| \geq \epsilon\mathbb{E}\Bigg[\sum_{i\in\mathcal{I}_k} X_i\Bigg]\Bigg\} \leq \exp\Bigg(-\frac{\epsilon^2\mathbb{E}[\sum_{i\in\mathcal{I}_k} X_{k,i}]}{3}\Bigg), \tag{16}$$

or equivalently,

$$\mathbb{P}\Bigg\{\Bigg|\sum_{i\in\mathcal{I}_k} X_i - \mathbb{E}\Bigg[\sum_{i\in\mathcal{I}_k} X_i\Bigg]\Bigg| \leq \epsilon\mathbb{E}\Bigg[\sum_{i\in\mathcal{I}_k} X_i\Bigg]\Bigg\} \geq 1 - \exp\Bigg(-\frac{\epsilon^2\mathbb{E}[\sum_{i\in\mathcal{I}_k} X_{k,i}]}{3}\Bigg). \tag{17}$$

Define $V_k := \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} X_{k,i}$. Let $\mathcal{G}_k$ be a *good* event such that $(1-\epsilon)\mathbb{E}[V_k] \le V_k \le (1+\epsilon)\mathbb{E}[V_k]$. From the above result, we know that $\mathbb{P}\{\mathcal{G}_k\} \ge 1 - \exp\left(-\frac{\epsilon^2\mathbb{E}[\sum_{i\in\mathcal{I}_k} X_{k,i}]}{3}\right)$. Furthermore, denote $\mathcal{G} = \bigcap_{k=1}^K \mathcal{G}_k$. Therefore,

$$\mathbb{P}\{\mathcal{G}\} \ge \prod_{k=1}^K \left(1 - \exp\Big(-\frac{\epsilon^2\mathbb{E}[\sum_{i\in\mathcal{I}_k} X_{k,i}]}{3}\Big)\right) \ge 1 - \sum_{k=1}^K \exp\Big(-\frac{\epsilon^2\mathbb{E}[\sum_{i\in\mathcal{I}_k} X_{k,i}]}{3}\Big) \ge 1 - K\cdot\exp\Big(-\frac{\epsilon^2 n_{\min}\mu_{\min}}{3}\Big). \tag{18}$$

Then we apply Hoeffding inequality across K clients as follows

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{k=1}^K V_k - \mathbb{E}\left[\sum_{k=1}^K V_k\right]\right| \le t\right\}\\
=&\mathbb{P}\{\mathcal{G}\}\cdot\mathbb{P}\left\{\left|\sum_{k=1}^K V_k - \mathbb{E}\left[\sum_{k=1}^K V_k\right]\right| \le t\middle|\mathcal{G}\right\} + (1-\mathbb{P}\{\mathcal{G}\})\cdot\mathbb{P}\left\{\left|\sum_{k=1}^K V_k - \mathbb{E}\left[\sum_{k=1}^K V_k\right]\right| \le t\middle|\overline{\mathcal{G}}\right\}\\
\ge&\left(1 - K\cdot\exp\Big(-\frac{\epsilon^2 n_{\min}\mu_{\min}}{3}\Big)\right)\cdot\left(1 - 2\exp\Big(-\frac{2t^2}{\sum_{k=1}^K(2\epsilon\mathbb{E}[V_k])^2}\Big)\right)\\
\ge&1 - K\cdot\exp\Big(-\frac{\epsilon^2 n_{\min}\mu_{\min}}{3}\Big) - 2\exp\Big(-\frac{2t^2}{\sum_{k=1}^K(2\epsilon\mathbb{E}[V_k])^2}\Big).
\end{aligned}$$

If $\epsilon \ge \sqrt{\frac{3\log(2K/\delta)}{n_{\min}\mu_{\min}}}$, then $K\cdot\exp\Big(-\frac{\epsilon^2 n_{\min}\mu_{\min}}{3}\Big) \le \delta/2$. On the other hand, if $t \ge \sqrt{\frac{K\log(2K/\delta)\cdot\log(4/\delta)}{n_{\min}\mu_{\min}}}$, then $2\exp\Big(-\frac{2t^2}{\sum_{k=1}^K(2\epsilon\mathbb{E}[V_k])^2}\Big) \le \delta/2$.

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{k=1}^K V_k - \mathbb{E}\left[\sum_{k=1}^K V_k\right]\right| \le t\right\}\\
\ge&\mathbb{P}\left\{\left|\sum_{k=1}^K V_k - \mathbb{E}\left[\sum_{k=1}^K V_k\right]\right| \le \sqrt{\frac{K\log(2K/\delta)\cdot\log(4/\delta)}{n_{\min}\mu_{\min}}}\right\}\\
\ge&1-\delta,
\end{aligned}$$

or equivalently

$$\mathbb{P}\left\{\left|\frac{1}{K}\sum_{k=1}^K V_k - \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K V_k\right]\right| \le \sqrt{\frac{\log(2K/\delta)\cdot\log(4/\delta)}{Kn_{\min}\mu_{\min}}}\right\} \ge 1-\delta. \tag{19}$$

□

B.1.4 Proof of Lemma 3

Proof. (of Lemma 3)

Lemma 1 and 2 show the estimation error of single statistics $\hat{u}(c)$ and $\hat{j}(c,p)$. Now we study how to transfer the estimation error of each entry of the matrix $\hat{Q}_\theta$ that $\widehat{q}_{cp} := \frac{\hat{j}(c,p)\cdot\hat{r}(p)}{\sqrt{\hat{u}(c)\cdot\hat{r}(p)}}$:

$$
\begin{aligned}
(\frac{jr}{\sqrt{ur}})^2 \leq & \frac{(\hat{j}+\epsilon)^2(\hat{r}+\epsilon)^2}{(\hat{u}-\epsilon)(\hat{r}-\epsilon)} = \frac{(\hat{j}^2+\epsilon^2+2\hat{j}\epsilon)(\hat{r}^2+\epsilon^2+2\hat{r}\epsilon)}{\hat{u}\hat{r}-\epsilon(\hat{r}+\hat{u})+\epsilon^2} \\
= & \frac{\hat{j}^2\hat{r}^2+\hat{j}^2\epsilon^2+2\hat{j}^2\hat{r}\epsilon+\epsilon^2\hat{r}^2+\epsilon^4+2\hat{r}\epsilon^3+2\hat{j}\hat{r}^2\epsilon+2\hat{j}\epsilon^3+4\hat{j}\hat{r}\epsilon^2}{\hat{u}\hat{r}-\epsilon(\hat{r}+\hat{u})+\epsilon^2} \\
\leq & \frac{\hat{j}^2\hat{r}^2+\hat{j}^2\epsilon^2+2\hat{j}^2\hat{r}\epsilon+\epsilon^2\hat{r}^2+\epsilon^4+2\hat{r}\epsilon^3+2\hat{j}\hat{r}^2\epsilon+2\hat{j}\epsilon^3+4\hat{j}\hat{r}\epsilon^2+\epsilon(\hat{r}+\hat{u})-\epsilon^2}{\hat{u}\hat{r}} \\
= & \frac{\hat{j}^2\hat{r}^2}{\hat{u}\hat{r}} + O(\epsilon),
\end{aligned}
$$

According to lemma 5, the last inequality is held due to ϵ is upper bounded by the order of $\tilde{O}(\frac{1}{\sqrt{n}})$ while $\hat{j}(c,p)$, $\hat{u}(c)$ and $\hat{r}(p)$ are all the empirical probabilities between 0 and 1.

□

Lemma 5. *For any positive number v, p, q and $0 < v \leq p \leq q - v$, if $v \leq q^2/(2p)$, then the following inequality holds:*

$$
\frac{p}{q-v} - \frac{p+v}{q} \leq 0. \tag{20}
$$

Proof. (of Lemma 5)

From the condition $v \leq q^2/(2p)$ and $0 < v \leq p \leq q - v$, we have following hold:

$$
v \leq q^2/(2p) \leq q^2/(p+v). \tag{21}
$$

Inequality 21 could be transferred as follows:

$$
(p+v)v \leq vq \leq q^2. \tag{22}
$$

Then, the following inequality hold:

$$
\begin{aligned}
& \frac{p}{q-v} - \frac{p+v}{q} \\
= & \frac{pq-(p+v)(q-v)}{(q-v)q} \\
= & \frac{pq-pq+pv-qv+v^2}{(q-v)q} \\
= & \frac{v(p-q+v)}{(q-v)q} \\
= & \frac{v(p+v)-vq}{q^2-vq} \\
\leq & 0,
\end{aligned}
$$

where the last inequality is due to the inequality 22.

Thus, Lemma 5 is proved. □

B.1.5 Proof of Lemma 4

Proof. (of Lemma 4)

$$
\begin{aligned}
\|Q_\theta\| =& \sqrt{\sum_{c=1}^{C}\sum_{p=1}^{P}\frac{j(c,p)\cdot r(p)}{\sqrt{u(c)\cdot r(p)}}} \\
\leq& \sqrt{\sum_{c=1}^{C}\sum_{p=1}^{P}\frac{j(c,p)\cdot r(p)}{u(c)}} \\
=& \sqrt{\sum_{c=1}^{C}\sum_{p=1}^{P}\frac{r(p)}{u(c)}} \\
=& \sqrt{\sum_{c=1}^{C}\frac{1}{u(c)}} \\
\leq& \sqrt{\frac{C}{\rho}},
\end{aligned}
$$

where the first inequality is due to $j(c,p) \leq 1$, and the second inequality is due to the definition of ρ.

□

B.1.6 Proof of Proposition 1

Proposition 3. *(Proposition 1 restated, convergence of FedRényi) Suppose $\eta \leq O(1/M)$ and $L_k(\theta)$ satisfies (G_L, B_L)-bounded gradient dissimilarity, where $\frac{1}{K}\sum_{k=1}^{K}\|\frac{\partial L_k(\theta)}{\partial\theta}\|^2 \leq G_L^2 + B_L^2\|\frac{\partial L(\theta)}{\partial\theta}\|^2$. If $\|\frac{\partial L(\theta)}{\partial\theta}\|^2, \|\frac{\partial Q_\theta}{\partial\theta}\|^2, \|\frac{\partial \widehat{Q}_\theta}{\partial\theta}\|^2, \|\frac{\partial \mathbf{v}_\theta}{\partial\theta}\|^2$ are bounded by $\bar{G}$ for all θ, Then, $H(\theta, \mathbf{v})$ satisfies (G_H, B_H)-bounded gradient dissimilarity, where G_H is $\frac{C\bar{G}(\rho+C)}{\rho^2}$ and B_H is 1. $F(\theta)$ also satisfies (G_F, B_F)-bounded gradient dissimilarity, where $B_F = 2B_L^2$ and $G_F = 2G_L^2 + (4\lambda - 2B_L^2\lambda^2)\frac{C\bar{G}(\rho+C)}{\rho^2} + 4B_L^2\lambda\cdot\sqrt{\frac{C\bar{G}^2(\rho+C)}{\rho^2}}$. Thus, FedRényi algorithm achieves $\mathbb{E}[\|\nabla\widehat{F}(\theta_T)\|^2] \leq \epsilon$ and $\mathbb{E}[\|\nabla F(\theta_T)\|^2] \leq \epsilon + O\Big(\frac{1}{n} + \max_\theta\big\|\frac{\partial\widehat{Q}_\theta}{\partial\theta} - \frac{\partial Q_\theta}{\partial\theta}\big\|^2\big)\Big)$ when $T \geq O(1/\epsilon^2)$.*

Proof. (of Proposition 1)

Before we prove that $F(\theta)$ satisfies (G_F, B_F)-bounded gradient dissimilarity, we first prove that $H(\theta, \mathbf{v})$ satisfies (G_H, B_H)-bounded gradient dissimilarity:

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{\partial H_k(\theta_k, \mathbf{v})}{\partial\theta_k}\right\|^2 \\
=&\frac{1}{K}\sum_{k=1}^{K}\left\|Q_{\theta_k}\cdot\mathbf{v}_\theta\cdot\Big[\frac{\partial Q_{\theta_k}}{\partial\theta_k}\cdot\mathbf{v}_\theta + \frac{\partial\mathbf{v}_\theta}{\partial\theta}\cdot Q_{\theta_k}\Big]\right\|^2 \\
=&\frac{1}{K}\sum_{k=1}^{K}\|Q_{\theta_k}\|^2\cdot\|\mathbf{v}_\theta\|^2\cdot\left[\left\|\frac{\partial Q_{\theta_k}}{\partial\theta_k}\right\|^2\cdot\|\mathbf{v}_\theta\|^2 + \left\|\frac{\partial\mathbf{v}_\theta}{\partial\theta}\right\|^2\cdot\|Q_{\theta_k}\|^2\right] \\
\leq&\frac{1}{K}\sum_{k=1}^{K}\Big(\frac{C}{\rho}\Big)\cdot\left[\bar{G} + \bar{G}\cdot\Big(\frac{C}{\rho}\Big)\right] \\
=&\frac{C\bar{G}(\rho+C)}{\rho^2} \\
\leq&\frac{C\bar{G}(\rho+C)}{\rho^2} + \left\|\frac{\partial H(\theta,\mathbf{v})}{\partial\theta}\right\|^2,
\end{aligned}
$$

where the first inequality is due to the assumption that $\|\frac{\partial Q_\theta}{\partial \theta}\|^2, \|\frac{\partial \mathbf{v}_\theta}{\partial \theta}\|^2$ are bounded by $\bar{G}$, $\|\mathbf{v}_\theta\|^2 \leq 1$, and the last inequality is due to the non-negativity of $\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\|^2$.

Therefore, $H(\theta, \mathbf{v})$ satisfies (G_H, B_H)-bounded gradient dissimilarity, where G_H is $\frac{C\bar{G}(\rho+C)}{\rho^2}$ and B_H is 1.

Then, we begin to prove that $F(\theta)$ satisfies (G, B)-bounded gradient dissimilarity:

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{\partial F_k(\theta_k)}{\partial \theta_k}\right\|^2 \\
=&\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{\partial\big[L_k(\theta_k)+\lambda H_k(\theta_k,\mathbf{v})\big]}{\partial \theta_k}\right\|^2 \\
\leq&\frac{2}{K}\sum_{k=1}^{K}\left\|\frac{\partial L_k(\theta_k)}{\partial \theta_k}\right\|^2+\frac{2\lambda}{K}\sum_{k=1}^{K}\left\|\frac{\partial H_k(\theta_k,\mathbf{v})}{\partial \theta_k}\right\|^2 \\
\leq&2G_L^2+2B_L^2\cdot\left\|\frac{\partial L(\theta)}{\partial \theta}\right\|^2+2\lambda\cdot\frac{C\bar{G}(\rho+C)}{\rho^2}+2\lambda\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\|^2 \\
=&2B_L^2\cdot\left\|\frac{\partial L(\theta)}{\partial \theta}\right\|^2+2B_L^2\lambda^2\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\|^2+4B_L^2\lambda\left\langle\frac{\partial L(\theta)}{\partial \theta},\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\rangle \\
&+2G_L^2+2\lambda\frac{C\bar{G}(\rho+C)}{\rho^2}+2(\lambda-B_L^2\lambda^2)\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\|^2-4B_L^2\lambda\left\langle\frac{\partial L(\theta)}{\partial \theta},\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\rangle \\
=&2B_L^2\left\|\frac{\partial F(\theta)}{\partial \theta}\right\|^2+2G_L^2+2\lambda\frac{C\bar{G}(\rho+C)}{\rho^2}+2(\lambda-B_L^2\lambda^2)\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\|^2-4B_L^2\lambda\left\langle\frac{\partial L(\theta)}{\partial \theta},\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\rangle \\
\leq&2B_L^2\left\|\frac{\partial F(\theta)}{\partial \theta}\right\|^2+2G_L^2+2\lambda\frac{C\bar{G}(\rho+C)}{\rho^2}+2(\lambda-B_L^2\lambda^2)\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\|^2+4B_L^2\lambda\left\|\frac{\partial L(\theta)}{\partial \theta}\right\|\cdot\left\|\frac{\partial H(\theta,\mathbf{v})}{\partial \theta}\right\| \\
=&2B_L^2\left\|\frac{\partial F(\theta)}{\partial \theta}\right\|^2+2G_L^2+2\lambda\frac{C\bar{G}(\rho+C)}{\rho^2}+2(\lambda-B_L^2\lambda^2)\left\|Q_\theta\cdot\mathbf{v}_\theta\cdot\Big[\frac{\partial Q_\theta}{\partial \theta}\cdot\mathbf{v}_\theta+\frac{\partial \mathbf{v}_\theta}{\partial \theta}\cdot Q_\theta\Big]\right\|^2 \\
&+4B_L^2\lambda\left\|\frac{\partial L(\theta)}{\partial \theta}\right\|\cdot\left\|Q_\theta\cdot\mathbf{v}_\theta\cdot\Big[\frac{\partial Q_\theta}{\partial \theta}\cdot\mathbf{v}_\theta+\frac{\partial \mathbf{v}_\theta}{\partial \theta}\cdot Q_\theta\Big]\right\| \\
\leq&2B_L^2\left\|\frac{\partial F(\theta)}{\partial \theta}\right\|^2+2G_L^2+(4\lambda-2B_L^2\lambda^2)\frac{C\bar{G}(\rho+C)}{\rho^2}+4B_L^2\lambda\cdot\sqrt{\frac{C\bar{G}^2(\rho+C)}{\rho^2}},
\end{aligned}
$$

where the first inequality is due to $(a+b)^2 \leq 2a^2+2b^2$, and the second inequality is due to the (G_L, B_L)-bounded gradient dissimilarity condition of $L(\theta)$ and $H(\theta, \mathbf{v})$, the third inequality is due to $-\|a\|\cdot\|b\| \leq \langle a, b\rangle$, the last inequality is due to the assumption that $\|\frac{\partial L(\theta)}{\partial \theta}\|^2, \|\frac{\partial Q_\theta}{\partial \theta}\|^2, \|\frac{\partial \mathbf{v}_\theta}{\partial \theta}\|^2$ are bounded by $\bar{G}$.

Therefore, $F(\theta)$ satisfies (G_F, B_F)-bounded gradient dissimilarity, where $B_F = 2B_L^2$ and $G_F = 2G_L^2 + (4\lambda - 2B_L^2\lambda^2)\frac{C\bar{G}(\rho+C)}{\rho^2} + 4B_L^2\lambda\cdot\sqrt{\frac{C\bar{G}^2(\rho+C)}{\rho^2}}$.

Then, we begin to study the convergence of FedRényi. We first decompose $\|\nabla F(\theta)\|^2$ as follows

$$
\begin{aligned}
\|\nabla F(\theta)\|^2 =&\|\nabla F(\theta)-\nabla\widehat{F}(\theta)+\nabla\widehat{F}(\theta)\|^2 \leq 2\|\nabla\widehat{F}(\theta)\|^2+2\|\nabla\widehat{F}(\theta)-\nabla F(\theta)\|^2 \\
=&2\|\nabla\widehat{F}(\theta)\|^2+2\lambda^2\left\|\frac{\partial\widehat{H}(\theta,\widehat{\mathbf{v}}^*)}{\partial \theta}-\frac{\partial H(\theta,\mathbf{v}^*)}{\partial \theta}\right\|^2,
\end{aligned}
$$

where the first inequality is due to $(a+b)^2 \leq 2a^2+2b^2$.

Define $\mathbf{v}^* = \arg\max_{\mathbf{v}\perp\mathbf{v}_1, \|\mathbf{v}\|^2\leq 1} L(\theta) + \lambda H(\theta)$, $\widehat{\mathbf{v}}^* = \arg\max_{\mathbf{v}\perp\mathbf{v}_1, \|\mathbf{v}\|^2\leq 1} L(\theta) + \lambda\widehat{H}(\theta)$. Then we upper bound the

last term as follows.

$$
\begin{aligned}
&\left\|\frac{\partial \widehat{H}(\theta,\widehat{\mathbf{v}}^*)}{\partial\theta}-\frac{\partial H(\theta,\mathbf{v}^*)}{\partial\theta}\right\|^2 = 4\left\|\widehat{Q}_\theta(\widehat{\mathbf{v}}^*(\widehat{\mathbf{v}}^*)^\top)\cdot\frac{\partial\widehat{Q}_\theta}{\partial\theta}-Q_\theta(\mathbf{v}^*(\mathbf{v}^*)^\top)\cdot\frac{\partial Q_\theta}{\partial\theta}\right\|^2\\
=&\left\|\widehat{Q}_\theta\cdot(\widehat{\mathbf{v}}^*(\widehat{\mathbf{v}}^*)^\top-\mathbf{v}^*(\mathbf{v}^*)^\top)\cdot\frac{\partial\widehat{Q}_\theta}{\partial\theta}+\widehat{Q}_\theta\cdot(\mathbf{v}^*(\mathbf{v}^*)^\top)\cdot(\frac{\partial\widehat{Q}_\theta}{\partial\theta}-\frac{\partial Q}{\partial\theta})+(\widehat{Q}_\theta-Q_\theta)\cdot(\mathbf{v}^*(\mathbf{v}^*)^\top)\cdot\frac{\partial Q_\theta}{\partial\theta}\right\|^2\\
\le&2\|\widehat{Q}_\theta\|^2\cdot\|\widehat{\mathbf{v}}^*(\widehat{\mathbf{v}}^*)^\top-\mathbf{v}^*(\mathbf{v}^*)^\top\|^2\cdot\left\|\frac{\partial\widehat{Q}_\theta}{\partial\theta}\right\|^2+4\|\widehat{Q}_\theta\|^2\cdot\|\mathbf{v}^*(\mathbf{v}^*)^\top\|^2\cdot\left\|\frac{\partial\widehat{Q}_\theta}{\partial\theta}-\frac{\partial Q}{\partial\theta}\right\|^2\\
&+4\|\widehat{Q}_\theta-Q_\theta\|^2\cdot\|\mathbf{v}^*(\mathbf{v}^*)^\top\|^2\cdot\left\|\frac{\partial Q_\theta}{\partial\theta}\right\|^2\\
\le&2\frac{C}{\rho}\cdot\|\widehat{\mathbf{v}}^*(\widehat{\mathbf{v}}^*)^\top-\mathbf{v}^*(\mathbf{v}^*)^\top\|^2\cdot\bar{G}+4\frac{C}{\rho}\cdot\left\|\frac{\partial\widehat{Q}_\theta}{\partial\theta}-\frac{\partial Q}{\partial\theta}\right\|^2+4\|\widehat{Q}_\theta-Q_\theta\|^2\cdot\bar{G}\\
\le&O\left(\frac{1}{n}+\left\|\frac{\partial\widehat{Q}_\theta}{\partial\theta}-\frac{\partial Q_\theta}{\partial\theta}\right\|^2\right),
\end{aligned}
$$

where the inequality is due to $(a+b)^2\le 2a^2+2b^2$, and the second inequality is due to $\|\frac{\partial Q_\theta}{\partial\theta}\|^2\le\bar{G}, \|\frac{\partial\widehat{Q}_\theta}{\partial\theta}\|^2\le\bar{G}, \|\mathbf{v}\|\le 1, \|\widehat{\mathbf{v}}^*-\mathbf{v}^*\|\le O(\sqrt{\frac{1}{n}}), \|\widehat{Q}_\theta-Q_\theta\|\le O(\sqrt{\frac{1}{n}})$, and Lemma 4.

As a result, we have

$$
\|\nabla F(\theta)\|^2\le 2\|\nabla\widehat{F}(\theta)\|^2+O\left(\frac{1}{n}+\left\|\frac{\partial\widehat{Q}_\theta}{\partial\theta}-\frac{\partial Q_\theta}{\partial\theta}\right\|^2\right).
$$

Finally, the convergence result is the immediate from Theorem I in Karimireddy et al. [2020]. □

B.2 PROOF FOR SECTION 4.3.2

In this subsection, we prove Proposition 2 and Theorem 2 in Section 4.3.2.

B.2.1 Proof of Proposition 2

Proposition 4. *(Proposition 2 restated, approximation error of each straggler in asynchronous FedRényi) Define* $\max_{k,k'\in[K]}\|\theta^e_{k,0}-\theta^e_{k',0}\|=\varepsilon^e_0$. *Suppose that Assumption 1 and 2 hold. Then, for each communication round* e*, the approximation errors of model and local statistics on each stragglers* $\widetilde{k}$ *are upper bounded:*

$$
\begin{aligned}
&\|\widetilde{\theta}^e_{\widetilde{k},M}-\theta^e_{\widetilde{k},M}\|\le\varepsilon^e_0,\\
&|\widetilde{j}^e_{\widetilde{k},M}(c,p)-\bar{j}^e_{\widetilde{k},M}(c,p)|\le L\varepsilon^e_0+\zeta,\\
&|\widetilde{u}^e_{\widetilde{k},M}(c)-\bar{u}^e_{\widetilde{k},M}(c)|\le L\varepsilon^e_0+\zeta.
\end{aligned}
\tag{23}
$$

Proof. (of Proposition 2)

Before proving Proposition 2, we show the following technical lemma:

Lemma 6. *(Non-expansiveness of SGD under β-co-coercive condition)*

$$
\|SGD(x)-SGD(y)\|\le\|x-y\|.
$$

Now we begin to prove Proposition 2.

We first bound the model distance between $\theta^e_{k,M}$ and $\theta^e_{k',M}$ on arbitrary two different clients $k, k' \in [K], k \neq k'$ by following inequality:

$$\|\theta^e_{k,M} - \theta^e_{k',M}\| = \|\theta^e_{k,M-1} - \eta\nabla F(\theta^e_{k,M-1}) - \theta^e_{k',M-1} + \eta\nabla F(\theta^e_{k',M-1})\| \leq \|\theta^e_{k,M-1} - \theta^e_{k',M-1}\| \leq \varepsilon^e_{M-1} \leq \varepsilon^e_0, \tag{24}$$

where the first inequality is due to Assumption 1 and Lemma 6, the second inequality is due to the definition of ε^e_t, the last inequality is due to Lemma 6.

Then, we begin to bound the approximation error between $\theta^e_{\widetilde{k},M}$ and $\widetilde{\theta}^e_{k,M}$:

$$\begin{aligned}
&\|\widetilde{\theta}^e_{\widetilde{k},M} - \theta^e_{\widetilde{k},M}\| \\
=&\|\frac{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{k,k'}\theta^e_{k',M}}{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{k,k'}} - \theta^e_{k,M}\| \\
\leq&\frac{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{k,k'}\|\theta^e_{k',M} - \theta^e_{k,M}\|}{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{k,k'}} \\
\leq&\frac{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{k,k'}}{\sum_{k'=1}^{K-\widetilde{K}^{e+1}} W_\theta^{k,k'}}\varepsilon^e_0 \\
=&\varepsilon^e_0,
\end{aligned}$$

where the first inequality is due to the triangle inequality, and the second inequality is due to inequality 24.

Next, we bound the approximation error between of local statistics $\widetilde{j}^e_{\widetilde{k},M}(c,p)$ and $\widetilde{u}^e_{\widetilde{k},M}(c)$. Recall that $\bar{u}^e_{\widetilde{k},M}(c)$ and $\bar{j}^e_{\widetilde{k},M}(c,p)$ are the empirical probability and empirical conditional probability of $\widehat{\mathbb{P}}[f_\theta(X_k,S_k)=c]$, we could study the approximation error through $\widehat{\mathbb{P}}[f_{\theta^{e+1}_{k,0}}(X_k,S_k)=c]$:

$$\begin{aligned}
&\left|\frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k',0}}(X'_k,S'_k)=c\big]}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} - \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X_k,S_k)=c\big]\right| \\
\leq&\frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\left|\widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k',0}}(X'_k,S'_k)=c\big] - \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X_k,S_k)=c\big]\right|}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} \\
=&\frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\left|\widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k',0}}(X'_k,S'_k)=c\big] - \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X'_k,S'_k)=c\big] + \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X'_k,S'_k)=c\big] - \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X_k,S_k)=c\big]\right|}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} \\
\leq&\frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\left|\widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k',0}}(X'_k,S'_k)=c\big] - \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X'_k,S'_k)=c\big]\right|}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} \\
&+ \frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\left|\widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X'_k,S'_k)=c\big] - \widehat{\mathbb{P}}\big[f_{\theta^{e+1}_{k,0}}(X_k,S_k)=c\big]\right|}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} \\
\leq&L\cdot\frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\left\|\theta^{e+1}_{k',0} - \theta^{e+1}_{k,0}\right\|}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} + \frac{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}\zeta}{\sum_{k'=1}^{k'\in Rob_\zeta(\widetilde{k})} W^{k,k'}} \\
\leq&L\cdot\varepsilon^e_0 + \zeta,
\end{aligned}$$

Where the first and second inequality is due to triangle inequality, and the third inequality is due to Assumption 2 and definition of robust neighbor set $Rob_\zeta(\widetilde{k})$ that $Rob_\zeta(\widetilde{k}) = \{k' : \|\omega\bar{u}_{k'}(c) - \bar{u}_{\widetilde{k}}(c)\| \leq \zeta, k' \in [K], \forall c \text{ and } \forall\omega \in (0,1)\}$.

Combining the above inequality and definition of $j^e_{\widetilde{k},M}(c,p)$ and $u^e_{\widetilde{k},M}(c)$, we can bound the approximation error between $\widetilde{j}^e_{\widetilde{k},M}(c,p)$ and $j^e_{\widetilde{k},M}(c,p)$, $\widetilde{u}^e_{\widetilde{k},M}(c)$ and $u^e_{\widetilde{k},M}(c)$, respectively:

$$|\widetilde{j}^e_{\widetilde{k},M}(c,p) - \bar{j}^e_{\widetilde{k},M}(c,p)| \leq L\varepsilon^e_0 + \zeta,$$
$$|\widetilde{u}^e_{\widetilde{k},M}(c) - \bar{u}^e_{\widetilde{k},M}(c)| \leq L\varepsilon^e_0 + \zeta.$$

□

B.2.2 Proof of Lemma 6

Proof. (of Lemma 6)

$$\begin{aligned}
\|SGD(x) - SGD(y)\|^2 =& \|(x - \eta g(x)) - (y - \eta g(y))\|^2 \\
=& \|x-y\|^2 + \eta^2\|g(x) - g(y)\|^2 - 2\eta(x-y)^\top(g(x) - g(y)) \\
\leq& \|x-y\|^2 + \eta^2\|g(x) - g(y)\|^2 - 2\eta\beta\|g(x) - g(y)\|^2 \\
=& \|x-y\|^2 + \eta(\eta - 2\beta)\|g(x) - g(y)\|^2 \\
\leq& \|x-y\|^2,
\end{aligned}$$

where the first inequality is due to β-co-coercive condition of $g(x)$, and the last inequality is due to $\eta \leq 2\beta$. □

B.2.3 Proof of Theorem 2

Theorem 4. *(Theorem 2 restated, estimation error of Rényi regularization for asynchronous FedRényi) Suppose $j_{min} \backsim u_{min} \backsim r_{min} = O(1)$ and $n_{min} \backsim \frac{n}{K\log(K)}$. When $\gamma_k = \frac{n_k}{n}$ or $\frac{1}{K}$, for any communication round e, any global model θ^{e+1} and $\delta \in (0,1)$, we have the following inequality holds:*

$$\mathbb{P}\Big[\widehat{H}(\theta^{e+1}, \widetilde{\mathbf{v}}^{e+1}) - H(\theta^{e+1}, \mathbf{v}^{e+1}) \leq O\big(1/\sqrt{n} + (L\varepsilon^e_0 + \zeta)^2\big)\Big|\theta^{e+1}\Big] \geq 1 - \delta.$$

Proof. (of Theorem 2)

Recall that the $H(\theta^{e+1}, \mathbf{v}^{e+1}) = (\mathbf{v}^{e+1})^\top {Q^{e+1}_\theta}^\top Q^{e+1}_\theta \mathbf{v}^{e+1}$, where $\mathbf{v}^{e+1}$ is the second largest singular vector of Q^{e+1}_θ.

Before proving Theorem 2, we propose the following lemma to bound $\|\mathbf{v}^{e+1} - \widetilde{\mathbf{v}}^{e+1}\|$.

Lemma 7. *Define $\xi = \min|\lambda_2 - \lambda_3|, |\widetilde{\lambda}_2 - \widetilde{\lambda}_3|$, where $\lambda_1 \geq \cdots \geq \lambda_p$ and $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_p$ are singular values of matrix Q and $\widetilde{Q}$. Assume ξ is at the constant order. Suppose that $\|Q - \widetilde{Q}\| \leq \epsilon_Q$ and $\hat{j}(c,p) \backsim \hat{u}(c) \backsim \hat{r}(p) = \Omega\big(\frac{1}{\sqrt{n}}\big)$, the following inequality holds:*

$$\|\mathbf{v} - \widetilde{\mathbf{v}}\| \leq \sqrt{2}\frac{\epsilon_Q}{\xi}.$$

First, we study the approximation error between Q_θ and $\widetilde{Q}_\theta$.

Given fixed global model θ^{e+1} for arbitrary communication round $e+1$, For all participating client $k \in [K]\backslash I^{e+1}$, $\theta^{e+1}_{k,0} = \theta^{e+1} = \sum_{k=1}^{K-\widetilde{K}^{e+1}} \gamma_k \theta^e_{k,M} + \sum_{\widetilde{k}=1}^{\widetilde{K}^{e+1}} \gamma_{\widetilde{k}} \widetilde{\theta}^e_{\widetilde{k},M}$. For any straggler $\widetilde{k} \in I^{e+1}$, $\theta^{e+1}_{\widetilde{k},0} = \theta^e_{\widetilde{k},M}$. Thus, $\epsilon^{e+1}_0 \leq \epsilon^e_M$, where ϵ^e_M is bounded by Proposition 2.

Next, according to Lemma 3, the approximation error between each entry of matrix Q^{e+1}_θ and $\widetilde{Q}^{e+1}_\theta$ is bounded by $O(\epsilon^{e+1}_0)$, We define $\tilde{\varepsilon} \in \mathbb{R}^{C\times P}$ as an approximation error matrix and $\tilde{\varepsilon} = Q^{e+1}_\theta - \widetilde{Q}^{e+1}_\theta$, where each entry $\tilde{\varepsilon}_{c,p} = L\varepsilon^e_0 + \zeta$ is the same.

Now we start to prove Theorem 2.

$$
\begin{aligned}
&\widehat{H}(\theta^{e+1},\widetilde{\mathbf{v}}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
=&\widehat{H}(\theta^{e+1},\widetilde{\mathbf{v}}^{e+1})-\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
=&(\widetilde{\mathbf{v}}^{e+1})^\top(\widetilde{Q}_\theta^{e+1})^\top\widetilde{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&(\widetilde{\mathbf{v}}^{e+1})^\top\big[(\hat{Q}_\theta^{e+1}+\tilde{\varepsilon})^\top(\hat{Q}_\theta^{e+1}+\tilde{\varepsilon})\big]\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&(\widetilde{\mathbf{v}}^{e+1})^\top\big[(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}+2\|\tilde{\varepsilon}\|\cdot\|\hat{Q}_\theta^{e+1}\|+\|\tilde{\varepsilon}\|^2\big]\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&(\widetilde{\mathbf{v}}^{e+1})^\top\big[(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}+2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+}\\
=&(\widetilde{\mathbf{v}}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}+(\widetilde{\mathbf{v}}^{e+1})^\top\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}\\
&+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&(\widetilde{\mathbf{v}}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+}\\
=&(\widetilde{\mathbf{v}}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}+(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}-(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\mathbf{v}^{e+1}\\
&+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
=&((\widetilde{\mathbf{v}}^{e+1})^\top-(\mathbf{v}^{e+1})^\top)(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\widetilde{\mathbf{v}}^{e+1}+(\mathbf{v}^{e+1})^\top(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}(\widetilde{\mathbf{v}}^{e+1}-\mathbf{v}^{e+1})\\
&+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&\|(\widetilde{\mathbf{v}}^{e+1})^\top-(\mathbf{v}^{e+1})^\top\|\cdot\|(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\|\cdot\|\widetilde{\mathbf{v}}^{e+1}\|+\|(\mathbf{v}^{e+1})^\top\|\cdot\|(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\|\cdot\|\widetilde{\mathbf{v}}^{e+1}-\mathbf{v}^{e+1}\|\\
&+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
=&\|(\widetilde{\mathbf{v}}^{e+1})^\top-(\mathbf{v}^{e+1})^\top\|\cdot\|(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\|\cdot\Big(\|\widetilde{\mathbf{v}}^{e+1}\|+\|(\mathbf{v}^{e+1})^\top\|\Big)+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]\\
&+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&2\|(\widetilde{\mathbf{v}}^{e+1})^\top-(\mathbf{v}^{e+1})^\top\|\cdot\|(\hat{Q}_\theta^{e+1})^\top\hat{Q}_\theta^{e+1}\|+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&2\sqrt{2}\frac{L\varepsilon_0^e+\zeta}{\xi}\cdot\|\hat{Q}_\theta^{e+1}\|^2+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
\le&2\sqrt{2}\frac{L\varepsilon_0^e+\zeta}{\xi}\cdot(\frac{C}{\rho})+\big[2C^2P^2\cdot(L\varepsilon_0^e+\zeta)+(L\varepsilon_0^e+\zeta)^2\big]+\widehat{H}(\theta^{e+1},\mathbf{v})-H(\theta^{e+1},\mathbf{v}^{e+1})\\
=&2C^2P^2\cdot(L\varepsilon_0^e+\zeta)^2+(2C^2P^2+2\sqrt{2}\frac{C}{\rho\xi})\cdot(L\varepsilon_0^e+\zeta)+\widehat{H}(\theta^{e+1},\mathbf{v}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1}),
\end{aligned}
$$

where the second and third inequality is due to Cauchy–Schwarz inequality, the sixth inequality is due to $\|\mathbf{v}\|\le 1$ as $\mathbf{v}$ is the singular vector of matrix, the seventh inequality is due to Lemma 7, the last inequality is due to Lemma 4.

Then, combining the above inequality with Theorem 1, we have that:

$$
\mathbb{P}\left[\widehat{H}(\theta^{e+1},\widetilde{\mathbf{v}}^{e+1})-H(\theta^{e+1},\mathbf{v}^{e+1})\le O\left(\sqrt{\frac{1}{n}}+(L\varepsilon_0^e+\zeta)^2\right)\middle|\theta^{e+1}\right]\ge 1-\delta. \tag{25}
$$

□

B.2.4 Proof of Lemma 7

Proof. (of Lemma 7)

Define $\alpha=\Theta(\mathbf{v},\widetilde{\mathbf{v}})$ as the angle between vector $\mathbf{v}$ and $\widetilde{\mathbf{v}}$. Then, according to Davis–Kahan theorem Yu et al. [2015], the following inequality hold:

$$
\sin(\alpha)\le\frac{\|Q-\widetilde{Q}\|}{\xi}.
$$

Then we begin to bound $\|\mathbf{v} - \widetilde{\mathbf{v}}\|$:

$$
\begin{aligned}
&\|\mathbf{v} - \widetilde{\mathbf{v}}\| \\
=&\sqrt{\|\mathbf{v}\|^2 + \|\widetilde{\mathbf{v}}\|^2 - 2\mathbf{v}^\top \widetilde{\mathbf{v}}} \\
=&\sqrt{2\big(1 - \cos(\alpha)\big)} \\
=&\sqrt{2\big(1 - \sqrt{1 - \sin^2(\alpha)}\big)} \\
\leq&\sqrt{2\big(1 - \sqrt{1 - \frac{\|Q - \widetilde{Q}\|^2}{\xi^2}}\big)} \\
\leq&\sqrt{2\big(1 - \sqrt{1 - \frac{\epsilon_Q^2}{\xi^2}}\big)} \\
\leq&\sqrt{2 - 2\big(1 - \frac{\epsilon_Q^2}{\xi^2}\big)} \\
=&\sqrt{2}\frac{\epsilon_Q}{\xi},
\end{aligned}
$$

where the first inequality is due to the Davis–Kahan theorem, the third inequality is due to $\sqrt{1-x} \geq 1 - x$ when $0 \leq x \leq 1$ and $0 \leq \frac{\epsilon_Q^2}{\xi^2} \leq 1$.

□

C SUPPLEMENTARY NUMERICAL EXPERIMENTS

C.1 EXPERIMENTAL DETAILS

Dataset. To have an impartial experiment result, we conduct the test-bed in four widely used benchmark datasets, ADULT, COMPAS, DRUG, and DUTCH, following the setups of Chu et al. [2021], Du et al. [2020] and Donini et al. [2020]. Specifically, the ADULT dataset Kohavi [1996] contains $45,222$ instances, where the training and test part are two separated files consisting of $32K$ and $14K$ samples, respectively, and the training data is partitioned into 50 clients. The binary class label of each instance indicates whether a person's annual income exceeds $50K$ dollars. Following the settings of Hardt et al. [2016], we take gender as the sensitive attribute. The COMPAS Larson et al. and the DRUG Fehrman et al. [2017] dataset contain $5,278$ and $1,885$ data instances, respectively. Following the design of Chu et al. [2021], we uniformly sample $4,800$ and $1,600$ instances as the training data from COMPAS and DRUG, respectively, and then use the remaining part as the test dataset. The training data in the experiments is divided into 20 clients for the COMPAS dataset and 10 clients for the DRUG dataset. In COMPAS, the binary class label indicates whether the person is a recidivist or not, while in DRUG, it manifests whether the person abuses a volatile substance or not. Following Chu et al. [2021], we use the ('African-American', 'Caucasian') as the sensitive attribute in COMPAS and ('white','non-white') in DRUG. The DUTCH dataset collects personal information of the inhabitants in the Netherlands and the task is to classify the individual into high-income or low-income, with gender as the protected attribute. It contains $60,419$ data instances. We also sample 80% data to construct the training set and use the remaining part as the test dataset.

Hyperparameters. In this paper, several combinations of hyperparameters are adopted, including the regularization parameter $\lambda \in (\{0.1, 0.5, 1, 5, 1000\})$, temperature parameter ρ is 0.1, training rounds T and local updates iteration $M \in (\{(100, 10), (100, 4), (100, 2), (500, 50), (1000, 4)\})$, and proportion of straggler $\alpha \in (\{0, 0.3, 0.5\})$. Most experimental settings of baselines follow the configurations proposed by the original authors. The structure of the logistics regression model follows Baharlouei et al. [2020]. We fix the batch size as 64. We tune the optimization step size η in {5e-3, 2e-3, 1e-3, 5e-2, 2e-2, 1e-2, 0.5, 0.2, 0.1, 5, 2, 1}, and pick the optimal setting for each dataset by observing the average of top-20 HM values of the model trained by FedAvg. Then we set η for every experiment in our work on ADULT, COMPAS, DRUG, and DUTCH as {5, 0.1, 0.02, 0.1}. The client number of FL system on ADULT, COMPAS, DRUG, and DUTCH are as {50, 20, 10, 30} following Du et al. [2020], Chu et al. [2021], Ezzeldin et al. [2023]. The fraction of activate client in FL system is set as 0.4. The hyperparameter of regularization term μ in FedProx Li et al. [2020b] and Scaffold Karimireddy et al. [2020] are tuned in {0.01, 0.1, 0.5, 1, 2}, and we set $\mu = 1$ according to the optimal result. The hyperparameter used in FedFair Chu et al. [2021], LCO Chu et al. [2021], FL-FairBatch Roh et al. [2021], FedFB Zeng et al. [2021], and FairFed Ezzeldin et al. [2023] are following the setting proposed by the authors. Additionally, we follow the common stopping criteria in FL that each algorithm stops training when the number of training rounds T is reached.

Baselines. In this paper, we adopt the following state-of-the-art algorithms as our baselines, which are designed for the problems of heterogeneity and group fairness:

• Local: To observe the effect of server aggregation, we also adopt the local training setting, where each client updates their model by only local training.

• FedAvg McMahan et al. [2017]: the original FL algorithm which does not consider fairness for different demographic groups.

• FedProx Li et al. [2020b]: the representative FL algorithm for tackling the statistical and system heterogeneous problem by sloving an optimization object with regularization constraint. We compare the performance of FedProx to verify the effectiveness of FedRényi in tackling heterogeneous problem.

• Scaffold Karimireddy et al. [2020]: the FL algorithm for statistical heterogeneous problem by constructing regularization term with aggregated variate. We set Scaffold in our comparative group to observe the impact of statistical heterogeneity.

• FedFair Chu et al. [2021]: the cross-silo federated framework for group fairness by leveraging estimated statistics from participants. We use FedFair as a baseline to compare accuracy, fairness, and their trade-off.

• LCO Chu et al. [2021]: the local variant of FedFair for the locally optimization problem. We take LCO as our baseline for the same reason as FedFair.

• FedAvg+FairBatch (FL-FairBatch) Roh et al. [2021]: the enhancement of FedAvg that each client adopts the FairBatch to debias its local training data. To verify the effectiveness of FedRényi in tackling group fairness problem, we take this

algorithm as our baseline.

• FedFB Zeng et al. [2021]: an in-processing debiasing approach in FL based on FairBatch, where the server computes new weights for each client based on their statistics. Improving group fairness by leveraging the aggregated local statistics from each client, we compare this method with FedRényi.

• FairFedEzzeldin et al. [2023]: the federated framework which adjust the aggregated weights of clients according to the deviations between local and global fairness metrics or accuracy. We take this method as our baseline to compare the effect of different aggregation methods to solve the group fairness problem and optimize the accuracy of model.

Communication Simulation. To simulate the network latency in practice, we use the beta distribution to simulate the communication ability of each client. Specifically, we use the beta distribution generation package in Scipy. The hyperparameters of beta distribution (a and b) are set as 0.3 and 1, respectively. The results of the probability density function (PDF) are used to compute the latency. For PDF values that tend to be positively infinite, we trim them to 16 based on network programming in practice, where the network wait time has a specific upper limit. Each client has a communication delay of at least 1 second.

C.2 ADDITIONAL EXPERIMENTS FOR RESULTS IN MAIN PAPER

Effect of Regularization parameter λ. Next, we analyze how the regularization parameter λ affects the performance of the models trained by FedRényi in another three datasets. As the result shown in Figs. 6, for each dataset with two data distribution settings, we train and fine-tune the base models with λ in $\{0.1, 0.5, 1, 5, 1000\}$.

Among these empirical results, the fairness performance of FedRényi shows a growing trend in most cases, while the accuracy performance shows a tendency to decrease or unchanged on the contrary, especially on DUTCH, as shown in Figure 2 of the main text. Recalling Equation 6, the regularization term becomes more significant to the FL training object, when λ becomes larger, which might compromise accuracy. Therefore, these results further demonstrate that FedRényi can construct a trade-off between accuracy and fairness. However, the FedRényi performance does not always exhibit the expected trade-off. As shown in Figure 6, the accuracy of models does not always drop with λ decreasing. We speculate that the sensitive features in these datasets are easy to identify. Therefore, increasing λ might over-optimize the empirical target, making the ACC unstable. Overall, the defects do not obscure the fact that the FedRényi algorithm can balance the accuracy and fairness through different λ values by adjusting the value of λ and obtain the optimal results, according to different preferences.

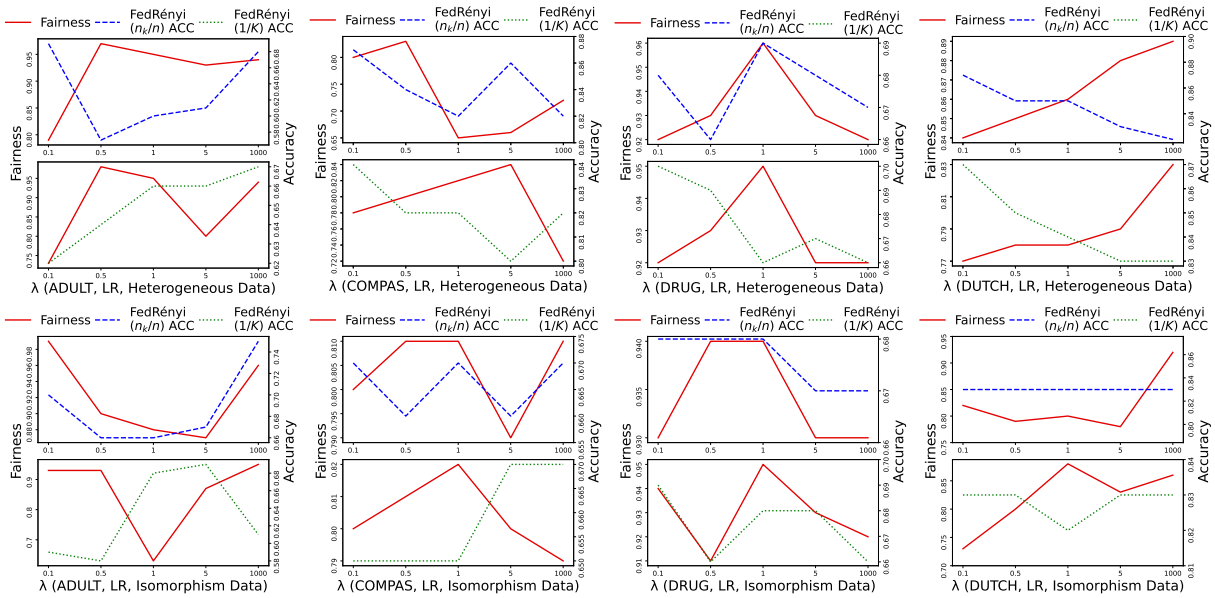

Figure 6: The effect of parameter λ in four dataset.

The Trade-off between Accuracy and Fairness. A comparison of testing ACC and FR about each algorithm is shown in Figure 7. Only the top-5 results (with better HM value) of each algorithm will be plotted, and some methods show less than 5 points are caused by overlap. Intuitively, red and yellow scatters (FedRényi results) get closer to the optimal corner than others in most cases. Besides, these scatters approximately form several curves, exhibiting the trade-off ability between ACC and FR. In particular, most baselines behave closely in these experiments, except the FairFed (blue). Some blue triangles tend to be towards the upper left, which means FairFed may over-emphasize fairness, thus penalizing the accuracy.

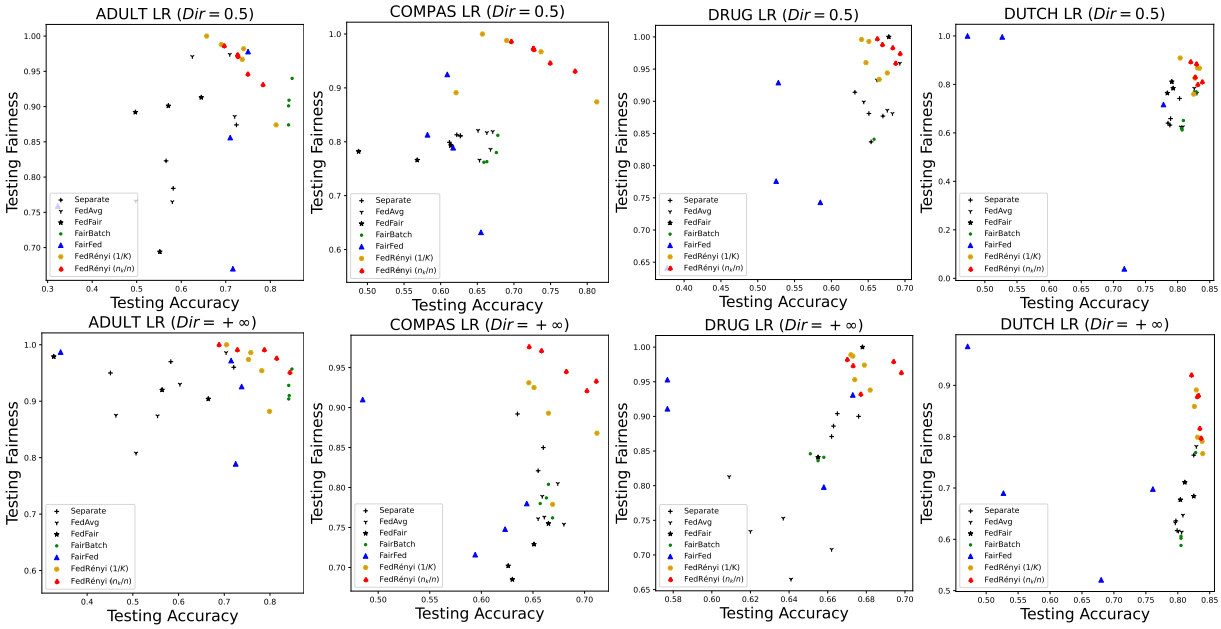

Figure 7: The ACC and FR trade-off in four datasets with two data distribution settings (Dir=0.5, Dir=+∞). Some methods show less than 5 points are caused by overlap. FedRényi performs closer to the optimal (the top right) and approximately forms trade-off curves from the bottom right (most accurate and least fair) to the top left (least accurate and most fair).

Convergence. To further verify the convergence properties of FedRényi in the heterogeneous and the uniform (isomorphism) data distribution setting, we visually record the training loss at different communication rounds in Figure 8 and 9. Also, the testing ACC, testing FR of FedRényi at different communication rounds on four datasets are presented. Intuitively, the training losses of FedRényi decrease with increasing communication and become stable at around 25 rounds in ADULT and DUTCH, and at around 50 rounds in COMPAS and DRUG.

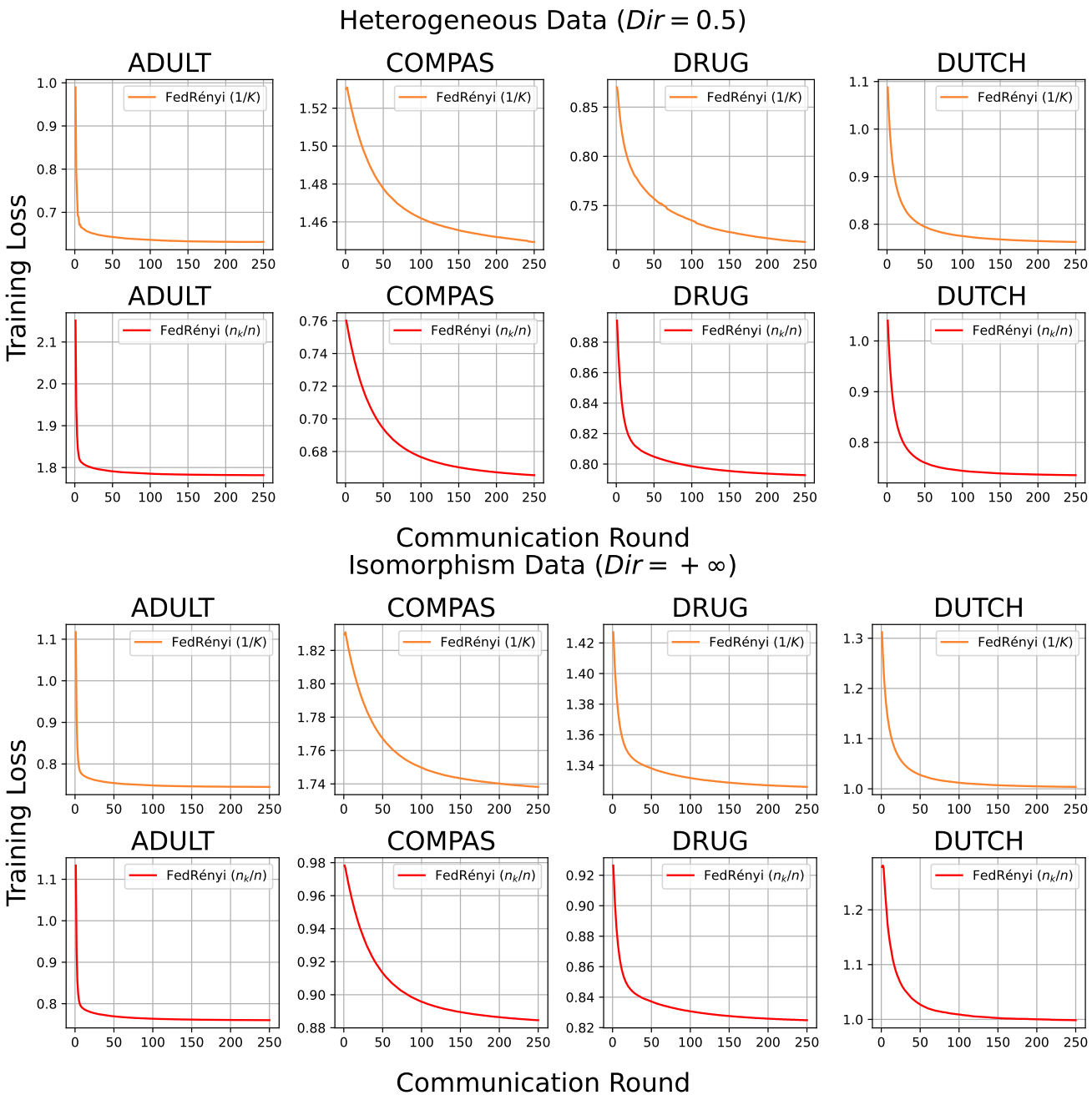

Figure 8: The illustration about the training loss of FedRényi under the uniform (Isomorphism) data settings on four datasets, which verify that FedRényi converges to a stable range after a certain number of rounds

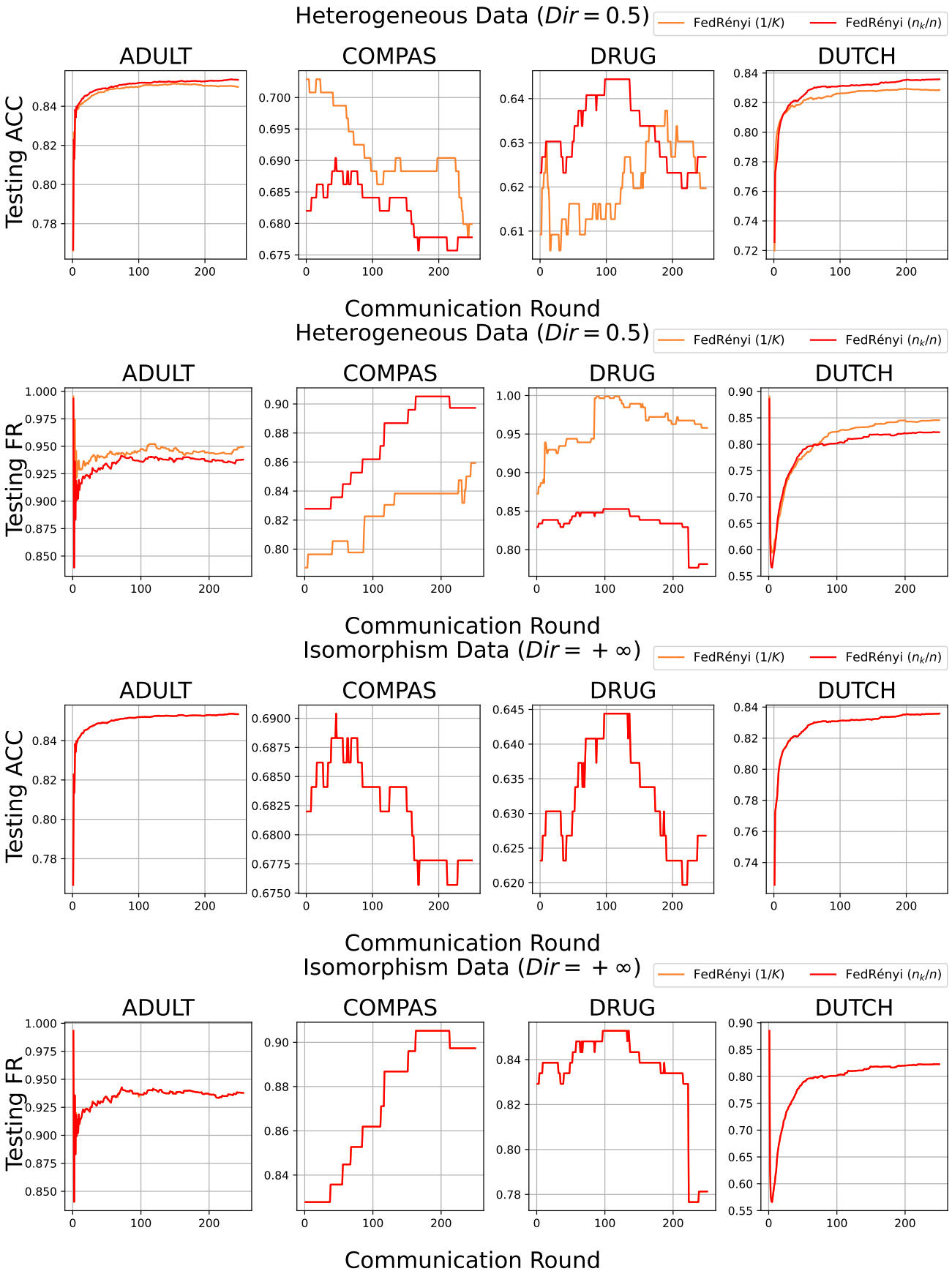

Figure 9: The illustration about the training ACC, and FR of FedRényi under the uniform (Isomorphism) data settings on four datasets, which verify that FedRényi converges to a stable range after a certain number of rounds

C.3 ROBUSTNESS EXPERIMENT

Compared to the general machine learning scenario, one of the implementation challenges of FL is client communication. In this paper, we adjust the step T and local updates interval M to investigate the effect of communication frequency. We also adjust the client dropping rate to simulate the situation where some clients are lost in the communication round. To verify the robustness of FedRényi about the condition change of communication, two groups of additional experiments are set up.

Robustness of total iteration T and local update iteration M. To evaluate the sensibility of FedRényi to communication frequency, we compare the HM of FedRényi with FedAvg and FairFed. It is well known that communication costs are especially expensive in FL. Comparing the result in different columns in Table 4, the performance of FedRényi with different γ_k settings improves as the training rounds T or the number of local updates (M) are increased in most case. For example, when the number of communication round is increasing with the same training epoch, e.g., (T:100, M:10) to (T:100, M:4) or (T:100, M:4) to (T:100, M:2), the accuracy of our proposed method is improved.

Comparing different results that testing with the same communication frequency, e.g., (T:100, M:10) and (T:500, M:50), the overall performances of FedRényi achieve outstanding stability, which proves the robustness to communication costs of our algorithm. In conclusion, the results of the supplementary experiment provide evidence of the FedRényi robustness specifically for different training epochs and communication rounds.

Table 4: The HM effect of different federated dropping rates on COMPAS and DRUG with Heterogeneous setting.

Method		**Step T**		**Local Updates M**	
		T=100, M=10	**T=100, M=4**	**T=100, M=2**	**T=500, M=50**
		Dir=0.5/+∞	**Dir=0.5/+∞**	**Dir=0.5/+∞**	**Dir=0.5/+∞**
COMPAS					
FedAvg		0.72/0.71	0.723/0.713	0.72/0.71	0.713/0.72
FairFed		0.68/0.71	0.73/0.65	0.64/0.68	0.69/0.63
FedRényi ($1/K$)	**λ=1000**	0.71/0.715	0.703/0.717	0.72/0.717	0.72/0.715
	λ=0.1	0.71/0.715	0.717/0.718	0.72/0.722	0.715/0.713
FedRényi (n_k/n)	**λ=1000**	0.712/0.713	0.718/0.717	0.72/0.71	0.725/0.722
	λ=0.1	0.718/0.715	0.722/0.72	0.723/0.72	0.718/0.715
DUTCH					
FedAvg		0.69/0.7	0.687/0.71	0.683/0.697	0.8/0.797
FairFed		0.69/0.73	0.51/0.59	0.75/0.69	0.64/0.63
FedRényi ($1/K$)	**λ=1000**	0.805/0.798	0.768/0.773	0.747/0.737	0.85/0.808
	λ=0.1	0.805/0.795	0.784/0.767	0.769/0.742	0.809/0.805
FedRényi (n_k/n)	**λ=1000**	0.805/0.795	0.77/0.772	0.733/0.73	0.807/0.83
	λ=0.1	0.812/0.807	0.787/0.768	0.775/0.74	0.853/0.84

Asynchronous affect. To verify the performance with the asynchronous FedRényi (Option II in Algorithm 1), we build experiments and simulate different communication thresholds to control the proportion of straggler α. Generally, as the proportion of stragglers (α) increases, the amount of algorithm available data will decrease significantly, resulting in degraded HM. As shown in Table 5 and 6, the asynchronous FedRényi not only performs stable HM but also does fairly well in bias control. When the asynchronous scheme is utilized in the training process of FedRényi, there exists a tolerable decline in HM (smaller than 0.06). These results demonstrate our method could accelerate the training process against stragglers with a small performance decline.

In addition, in order to explore the trade-off ability of FedRényi between FR and ACC under asynchronous settings, we set α as 0.3 and 0.5 respectively and record them in the form of scatter plots, as shown in Figure 10. Also, only the top-5 results (with better HM value) of each algorithm will be plotted, and some methods show less than 5 points are caused by overlap. As shown in Figure 10, the scatters of FedRényi approximately form several curves, exhibiting the trade-off ability between ACC and FR in most cases.

Table 5: The HM and the average approximation errors over stragglers with different α. These approximation errors are measured by the L2 distance between the approximation values and the corresponding target from stragglers.

Dir=0.5 λ=1	**Drop α: 0%**	**Drop α: 30%**	**Drop α: 50%**
(T, M) = (100, 4)	**HM/j Error/ u Error/θ Error**	**HM/j Error/ u Error/θ Error**	**HM/j Error/ u Error/θ Error**
ADULT			
Asynchronous FedRényi (n_k/n)	0.88/0/ 0/0	0.88/0.03/ 0.04/0.01	0.87/0.03/ 0.04/0.02
Asynchronous FedRényi ($1/K$)	0.88/0/ 0/0	0.88/0.03/ 0.04/0.01	0.87/0.03/ 0.04/0.02
Synchronous FedRényi (n_k/n)	0.88/0/ 0/0	0.88/0/ 0/0	0.88/0/ 0/0
COMPAS			
Asynchronous FedRényi (n_k/n)	0.77/0/ 0/0	0.73/0.03/ 0.01/0.92	0.71/0.01/ 0.02/1.41
Asynchronous FedRényi ($1/K$)	0.75/0/ 0/0	0.76/0.04/ 0.01/0.27	0.78/0.01/ 0.02/0.34
Synchronous FedRényi (n_k/n)	0.76/0/ 0/0	0.76/0/ 0/0	0.76/0/ 0/0
DRUG			
Asynchronous FedRényi (n_k/n)	0.74/0/ 0/0	0.74/0.01/ 0.01/0.25	0.75/0.08/ 0.03/0.40
Asynchronous FedRényi ($1/K$)	0.73/0/ 0/0	0.72/0.01/ 0.02/0.29	0.73/0.09/ 0.02/0.37
Synchronous FedRényi (n_k/n)	0.74/0/ 0/0	0.74/0/ 0/0	0.74/0/ 0/0
DUTCH			
Asynchronous FedRényi (n_k/n)	0.78/0/ 0/0	0.77/0.03/ 0.02/2.44	0.79/0.03/ 0.01/4.04
Asynchronous FedRényi ($1/K$)	0.74/0/ 0/0	0.74/0.03/ 0.02/2.47	0.78/0.03/ 0.01/4.01
Synchronous FedRényi (n_k/n)	0.78/0/ 0/0	0.78/0/ 0/0	0.78/0/ 0/0

Table 6: The comparison between FedProx and FedRényi with different straggler proportions in the heterogeneous setting.

Dir=0.5		**(T, M) = (100, 4), λ=1**	
Method	**Drop: 0%**	**Drop: 30%**	**Drop: 50%**
	ACC/FR/HM	**ACC/FR/HM**	**ACC/FR/HM**
ADULT			
FedProx	0.84/0.91/0.87	0.84/0.93/0.88	0.84/0.93/0.88
Asynchronous FedRényi (n_k/n)	**0.85/0.92/0.88**	**0.85/0.92/0.88**	**0.85/0.92/0.88**
Asynchronous FedRényi ($1/K$)	0.85/0.91/0.87	0.85/0.91/0.87	0.85/0.91/0.87
COMPAS			
FedProx	0.68/0.83/0.75	0.69/0.85/0.76	0.69/0.85/0.76
Asynchronous FedRényi (n_k/n)	0.68/0.89/0.77	0.66/0.81/0.73	0.67/0.76/0.71
Asynchronous FedRényi ($1/K$)	**0.70/0.80/0.75**	**0.69/0.86/0.76**	**0.68/0.91/0.78**
DRUG			
FedProx	0.63/0.84/0.72	0.63/0.84/0.72	0.62/0.88/0.73
Asynchronous FedRényi (n_k/n)	**0.66/0.84/0.74**	**0.65/0.84/0.74**	**0.67/0.85/0.75**
Asynchronous FedRényi ($1/K$)	0.61/0.93/0.73	0.61/0.89/0.72	0.64/0.85/0.73
DUTCH			
FedProx	0.81/0.59/0.68	0.81/0.63/0.71	0.81/0.66/0.73
Asynchronous FedRényi (n_k/n)	**0.82/0.75/0.78**	**0.82/0.77/0.79**	0.82/0.75/0.78
Asynchronous FedRényi ($1/K$)	0.82/0.67/0.74	0.82/0.67/0.74	0.82/0.74/0.78

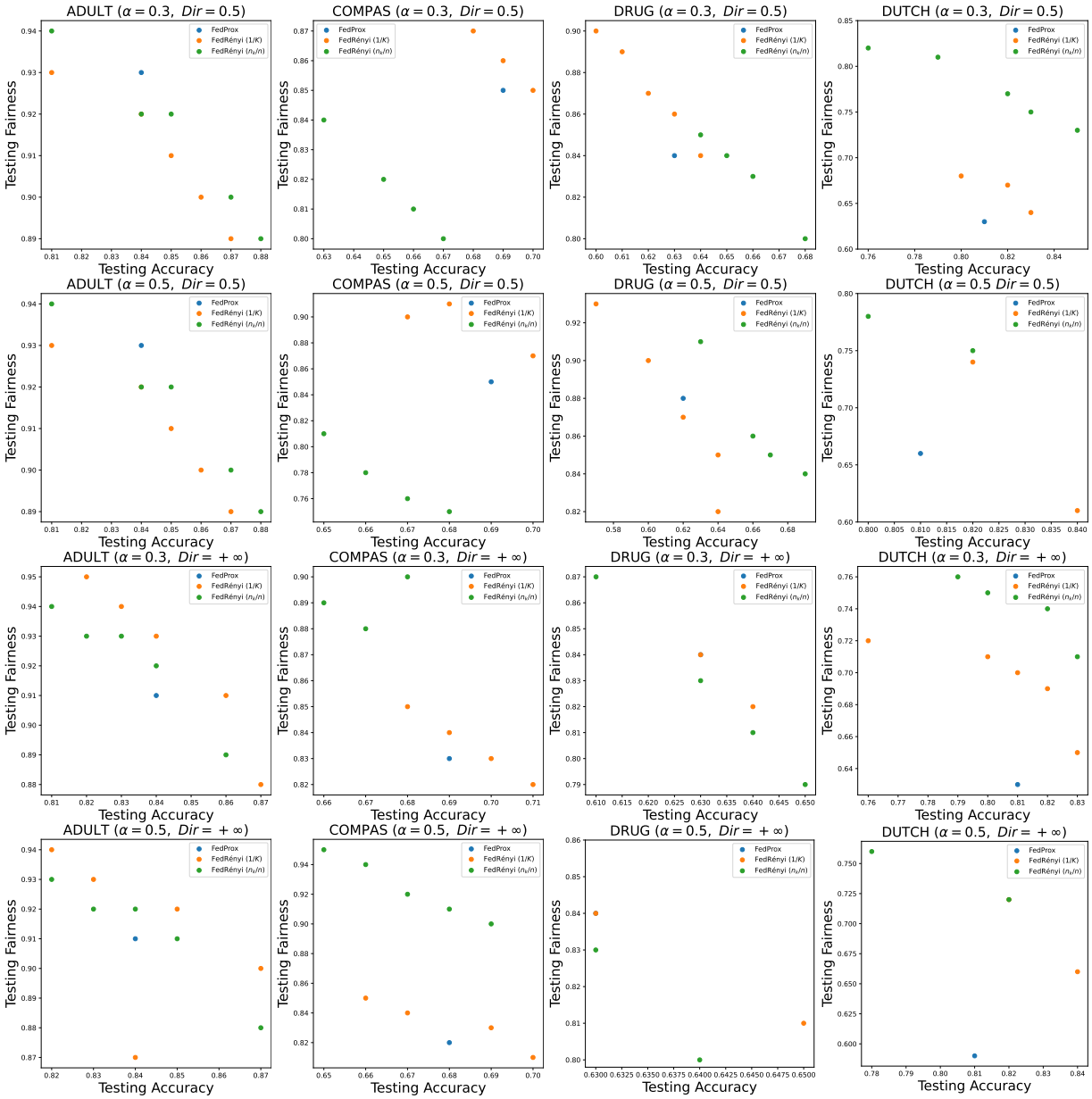

Figure 10: The ACC and FR trade-off in four datasets with two data distribution settings (Dir=0.5, Dir=+∞) and two proportion of straggler settings ($\alpha = 0.3$, $\alpha = 0.5$). Some methods show less than 5 points are caused by overlap.

Robustness of client dropping. In practice, a fraction of offline clients may drop out randomly during the communication stage (upload and download) in the training process. This proportion of dropped clients is known as the drop rate. In this paper, the setting of the proportion of stragglers α is [0, 0.3, 0.5], where 0 means no client is dropped. The effect of the proportion of stragglers is evaluated in four datasets as shown in Table 7. With different dropping rates, the ACC/FR/HM of FedRényi under two configurations both stay stable with tolerable fluctuation in most cases (0.05), which demonstrates the robustness of FedRényi to different dropping rates and fits for the FL practical implementation.

Table 7: The effect of different proportion of federated stragglers on four datasets with different heterogeneous settings

T:100 M:10	**ACC**		**FR**		**HM**	
Drop α	**0%/30%/50%**	**0%/30%/50%**	**0%/30%/50%**	**0%/30%/50%**	**0%/30%/50%**	**0%/30%/50%**
Method	**Dir = 0.5**	**Dir = +∞**	**Dir = 0.5**	**Dir = +∞**	**Dir = 0.5**	**Dir = +∞**
ADULT						
FedAvg	0.58/0.41/**0.62**	0.6/**0.55**/0.46	0.92/0.72/**0.97**	0.93/**0.87/0.88**	0.71/0.53/**0.76**	0.73/**0.68**/0.61
FedRényi ($1/K$)	**0.62/0.51**/0.5	**0.61**/0.52/0.51	**0.94/0.86**/0.91	**0.94**/0.84/0.85	**0.75/0.63**/0.64	**0.74**/0.62/0.63
FedRényi (n_k/n)	0.61/0.47/0.47	0.6/0.54/**0.54**	0.93/0.8/0.9	0.93/0.84/0.85	0.74/0.59/0.61	0.73/0.65/**0.66**
COMPAS						
FedAvg	0.66/**0.67**/0.66	**0.67**/0.66/**0.66**	0.8/0.75/**0.82**	0.76/**0.79**/0.76	0.72/0.71/**0.73**	0.71/**0.72**/0.7
FedRényi ($1/K$)	**0.67**/0.66/**0.67**	**0.67/0.67/0.66**	0.82/0.78/0.76	0.83/0.77/**0.78**	**0.74**/0.71/0.71	0.74/**0.72/0.71**
FedRényi (n_k/n)	**0.67**/0.66/**0.67**	**0.67/0.67/0.66**	**0.83/0.79**/0.76	**0.84**/0.77/0.77	**0.74/0.72**/0.71	**0.75**/0.71/**0.71**
DRUG						
FedAvg	0.64/**0.68/0.69**	**0.66**/0.62/0.63	0.77/0.89/**0.96**	0.71/0.65/0.63	0.7/**0.77/0.8**	**0.68**/0.63/0.63
FedRényi ($1/K$)	**0.68**/0.67/0.66	**0.68/0.66**/0.66	**0.94**/0.9/0.9	0.92/0.88/0.9	**0.79**/0.76/0.77	0.78/**0.76**/0.76
FedRényi (n_k/n)	**0.68**/0.66/0.67	**0.68/0.66/0.67**	**0.94**/0.89/0.9	**0.93/0.89/0.92**	**0.79**/0.76/0.77	**0.79/0.76/0.77**
DUTCH						
FedAvg	0.81/0.8/0.81	0.8/0.8/0.8	0.62/0.58/0.62	0.62/0.62/0.63	0.7/0.67/0.7	0.7/0.7/0.7
FedRényi ($1/K$)	**0.83/0.83/0.83**	**0.83/0.83/0.83**	0.84/0.77/**0.79**	**0.84/0.77/0.75**	**0.84**/0.8/**0.81**	**0.84/0.8/0.79**
FedRényi (n_k/n)	**0.83/0.83/0.83**	**0.83/0.83/0.83**	**0.86/0.82**/0.76	**0.84**/0.76/**0.75**	**0.84/0.82**/0.79	**0.84**/0.79/0.78

C.4 SCALABILITY EXPERIMENT

To evaluate the scalability to large-scale dataset of our method, we conduct the image classification on CelebA dataset Liu et al. [2015] following the setting in FairGrad Ban and Ji [2024]. The CelebA contains 202599 samples with 162770 for training, 19867 for validation, and 19962 for testing. Each image sample in the CelebA contains 40 binary attribute labels, and we focus on 2 attribution among 40 for binary classification. We take the 21-th attribute (gender) as the sensitive attribute, take the third attribute (attractive) as the classification label for each image. The training data of CelebA is partitioned into 20 client. We compare the performance of our method and FedAvg with 20% activate client as shown in Table 8.

Table 8: Performances of methods with the heterogeneous setting on CelebA. The Accuracy, fairness, and harmonic mean are denoted by AC, FR, and HM, respectively.

CelebA			
Dir=0.5, λ=1	**Drop α: 0%**	**Batch Size=256**	**9-layers CNN**
(T, M) = (100, 10)	**ACC**	**FR**	**HM**
FedAvg	0.702±0.005	1.0±0.0	0.825±0.003
FedRényi (n_k/n)	0.715±0.001	1.0±0.0	0.834±0.001
FedRényi ($1/K$)	0.716±0.001	1.0±0.0	0.834±0.001

C.5 RESULT SUMMARY

To evaluate the performance of each algorithm on four different datasets, we construct several sets of experiments with different levels of heterogeneity. To avoid the unfair comparison caused by random factors in practice and hyperparameter settings, we record the mean and standard deviation of the top-20 performances (with better HM values) about different algorithms.

As shown in experimental results (see Table 9-12), FedRényi outperforms other baseline methods and shows a satisfactory level of ACC, FR, and HM in most cases, under different heterogeneity scenarios. These results support the effectiveness of our proposed method.

Table 9: Performances of methods with the heterogeneous setting on ADULT. The Accuracy, fairness, and harmonic mean are denoted by AC, FR, and HM, respectively. A smaller Dir indicates a more heterogeneous distribution across clients. Dir=+∞ represents the uniform data distribution setting.

ADULT	Method	Dir=0.5	Dir=1	Dir=8	Dir=+∞
ACC	**Local**	0.56±0.12	0.67±0.06	**0.67±0.06**	0.63±0.12
	FedAvg	0.62±0.12	0.59±0.12	0.59±0.12	0.6±0.10
	FedProx	0.61±0.12	0.63±0.07	0.63±0.07	0.58±0.12
	Scaffold	0.56±0.20	0.62±0.06	0.56±0.20	0.56±0.14
	FedFair	0.51±0.07	0.58±0.02	0.52±0.08	0.5±0.13
	LCO	0.52±0.01	0.59±0.04	0.52±0.07	0.5±0.13
	FL-FairBatch	0.64±0.00	0.64±0.00	0.64±0.00	0.64±0.00
	FedFB	0.65±0.00	0.65±0.01	0.64±0.00	0.64±0.00
	FairFed	0.62±0.17	0.64±0.21	0.64±0.21	0.63±0.17
	FedRényi ($1/K$)	**0.67±0.03**	**0.69±0.04**	0.67±0.03	**0.68±0.03**
	FedRényi (n_k/n)	0.65±0.04	0.65±0.03	0.65±0.04	**0.68±0.03**
FR	**Local**	0.87±0.07	0.93±0.02	0.93±0.02	**0.97±0.01**
	FedAvg	0.87±0.1	0.89±0.12	0.89±0.12	0.91±0.07
	FedProx	0.88±0.11	0.84±0.06	0.84±0.06	0.76±0.14
	Scaffold	0.88±0.13	0.85±0.05	0.84±0.06	0.79±0.11
	FedFair	0.84±0.17	0.86±0.08	0.85±0.11	0.85±0.16
	LCO	0.86±0.07	0.89±0.11	0.87±0.04	0.84±0.11
	FL-FairBatch	0.91±0.02	0.93±0.02	0.93±0.02	0.92±0.02
	FedFB	0.92±0.03	0.93±0.02	0.93±0.02	0.92±0.02
	FairFed	0.77±0.16	**0.95±0.07**	**0.95±0.07**	0.92±0.08
	FedRényi ($1/K$)	**0.94±0.04**	0.94±0.05	0.95±0.05	0.92±0.04
	FedRényi (n_k/n)	**0.94±0.04**	0.95±0.03	0.92±0.05	0.93±0.05
HM	**Local**	0.68±0.09	0.78±0.03	0.78±0.03	0.76±0.02
	FedAvg	0.72±0.11	0.71±0.12	0.71±0.12	0.72±0.08
	FedProx	0.72±0.11	0.72±0.06	0.72±0.06	0.66±0.13
	Scaffold	0.68±0.16	0.72±0.05	0.67±0.09	0.66±0.12
	FedFair	0.63±0.10	0.69±0.03	0.65±0.09	0.63±0.14
	LCO	0.65±0.02	0.71±0.06	0.65±0.05	0.63±0.12
	FL-FairBatch	0.75±0.00	0.76±0.00	0.76±0.00	0.75±0.00
	FedFB	0.76±0.00	0.77±0.01	0.76±0.00	0.75±0.00
	FairFed	0.69±0.16	0.76±0.11	0.76±0.11	0.75±0.11
	FedRényi ($1/K$)	**0.78±0.03**	**0.8±0.04**	**0.79±0.04**	0.78±0.03
	FedRényi (n_k/n)	0.77±0.04	0.77±0.03	0.76±0.04	**0.79±0.04**

Table 10: Performances of methods with the heterogeneous setting on COMPAS. The Accuracy, fairness, and harmonic mean are denoted by AC, FR, and HM, respectively. A smaller Dir indicates a more heterogeneous distribution across clients. Dir=+∞ represents the uniform data distribution setting.

COMPAS	Method	Dir=0.5	Dir=1	Dir=8	Dir=+∞
ACC	**Local**	0.62±0.01	0.64±0.01	0.65±0.01	0.65±0.01
	FedAvg	0.66±0.01	**0.67±0.01**	0.66±0.01	0.66±0.01
	FedProx	0.66±0.01	**0.67±0.01**	**0.67±0.01**	**0.67±0.00**
	Scaffold	0.47±0.12	0.48±0.13	0.45±0.14	0.50±0.13
	FedFair	0.62±0.03	0.57±0.04	0.62±0.03	0.59±0.06
	LCO	0.59±0.03	0.58±0.06	0.56±0.07	0.56±0.06
	FL-FairBatch	0.67±0.01	0.67±0.00	0.67±0.00	0.66±0.00
	FedFB	0.67±0.01	**0.67±0.01**	0.67±0.0	0.66±0.01
	FairFed	0.62±0.03	0.57±0.04	0.62±0.03	0.59±0.06
	FedRényi ($1/K$)	**0.68±0.01**	0.66±0.02	0.66±0.01	0.66±0.01
	FedRényi (n_k/n)	**0.68±0.01**	0.65±0.01	0.66±0.01	0.66±0.01
FR	**Local**	**0.81±0.01**	0.80±0.04	0.79±0.03	0.80±0.00
	FedAvg	0.79±0.03	0.77±0.03	0.78±0.02	0.77±0.02
	FedProx	0.79±0.03	0.79±0.03	0.78±0.03	0.77±0.02
	Scaffold	0.82±0.10	0.74±0.10	0.71±0.05	0.81±0.06
	FedFair	0.79±0.10	0.91±0.05	0.70±0.11	0.79±0.07
	LCO	**0.85±0.09**	0.83±0.06	**0.87±0.04**	**0.90±0.05**
	FL-FairBatch	0.78±0.02	0.79±0.01	0.79±0.01	0.78±0.01
	FedFB	0.75±0.03	0.74±0.01	0.76±0.01	0.74±0.01
	FairFed	0.79±0.10	**0.91±0.05**	0.70±0.11	0.79±0.07
	FedRényi ($1/K$)	0.81±0.02	0.81±0.03	0.77±0.05	0.82±0.02
	FedRényi (n_k/n)	0.82±0.01	0.81±0.01	0.80±0.06	0.81±0.02
HM	**Local**	0.70±0.01	0.71±0.02	0.71±0.01	0.72±0.00
	FedAvg	0.72±0.01	0.72±0.01	0.72±0.01	0.71±0.01
	FedProx	0.72±0.01	0.73±0.01	0.72±0.01	0.72±0.00
	Scaffold	0.60±0.11	0.58±0.11	0.55±0.07	0.62±0.08
	FedFair	0.69±0.05	0.70±0.04	0.66±0.05	0.68±0.06
	LCO	0.70±0.04	0.68±0.06	0.68±0.05	0.69±0.05
	FL-FairBatch	0.72±0.01	0.73±0.00	0.73±0.00	0.72±0.00
	FedFB	0.71±0.01	0.70±0.01	0.71±0.00	0.7±0.01
	FairFed	0.69±0.05	0.70±0.04	0.66±0.05	0.68±0.06
	FedRényi ($1/K$)	0.72±0.03	**0.73±0.02**	0.71±0.02	**0.73±0.01**
	FedRényi (n_k/n)	**0.73±0.02**	0.71±0.01	**0.72±0.02**	**0.73±0.01**

Table 11: Performances of methods with the heterogeneous setting on DRUG. The Accuracy, fairness, and harmonic mean are denoted by AC, FR, and HM, respectively. A smaller Dir indicates a more heterogeneous distribution across clients. Dir=+∞ represents the uniform data distribution setting.

DRUG	Method	Dir=0.5	Dir=1	Dir=8	Dir=+∞
ACC	**Local**	0.65±0.01	0.66±0.01	0.66±0.02	0.67±0.01
	FedAvg	0.67±0.02	0.67±0.02	0.67±0.02	0.64±0.02
	FedProx	0.67±0.01	0.65±0.02	0.67±0.01	0.66±0.01
	Scaffold	0.66±0.01	0.55±0.13	0.62±0.04	0.54±0.13
	FedFair	0.67±0.02	0.67±0.02	0.67±0.02	0.64±0.02
	LCO	0.49±0.05	0.60±0.06	0.49±0.11	0.65±0.01
	FL-FairBatch	0.66±0.00	0.66±0.00	0.66±0.00	0.66±0.00
	FedFB	0.66±0.00	0.66±0.00	0.66±0.00	0.66±0.00
	FairFed	0.50±0.08	0.63±0.04	0.56±0.12	0.62±0.04
	FedRényi $(1/K)$	0.68±0.01	0.68±0.01	**0.69±0.01**	0.68±0.01
	FedRényi (n_k/n)	**0.69±0.01**	**0.69±0.01**	**0.69±0.01**	**0.69±0.01**
FR	**Local**	0.88±0.03	0.89±0.03	0.87±0.05	0.89±0.01
	FedAvg	0.86±0.02	0.85±0.01	0.87±0.01	0.85±0.03
	FedProx	0.86±0.02	0.85±0.01	0.87±0.01	0.85±0.03
	Scaffold	0.82±0.06	0.84±0.01	0.87±0.06	0.83±0.05
	FedFair	0.86±0.02	0.85±0.01	0.87±0.01	0.85±0.03
	LCO	0.93±0.03	0.85±0.07	0.83±0.11	0.86±0.04
	FL-FairBatch	0.84±0.00	0.84±0.00	0.84±0.01	0.84±0.00
	FedFB	0.85±0.00	0.84±0.00	0.84±0.00	0.84±0.00
	FairFed	0.77±0.10	0.84±0.11	0.81±0.06	0.90±0.06
	FedRényi $(1/K)$	**0.96±0.03**	0.95±0.03	**0.96±0.02**	**0.96±0.02**
	FedRényi (n_k/n)	0.96±0.02	**0.96±0.02**	**0.96±0.02**	**0.96±0.02**
HM	**Local**	0.75±0.01	0.76±0.01	0.75±0.03	0.76±0.01
	FedAvg	0.75±0.02	0.75±0.01	0.76±0.01	0.73±0.02
	FedProx	0.75±0.01	0.74±0.01	0.76±0.01	0.74±0.01
	Scaffold	0.73±0.02	0.66±0.02	0.72±0.05	0.65±0.07
	FedFair	0.75±0.02	0.75±0.01	0.76±0.01	0.73±0.02
	LCO	0.64±0.04	0.70±0.06	0.62±0.11	0.74±0.02
	FL-FairBatch	0.74±0.00	0.74±0.00	0.74±0.00	0.74±0.00
	FedFB	0.74±0.00	0.74±0.00	0.74±0.00	0.74±0.00
	FairFed	0.61±0.09	0.72±0.06	0.66±0.08	0.73±0.05
	FedRényi $(1/K)$	**0.80±0.01**	0.79±0.01	**0.80±0.01**	**0.80±0.01**
	FedRényi (n_k/n)	**0.80±0.01**	**0.80±0.01**	**0.80±0.01**	**0.80±0.01**

Table 12: Performances of methods with the heterogeneous setting on DUTCH. The Accuracy, fairness, and harmonic mean are denoted by AC, FR, and HM, respectively. A smaller Dir indicates a more heterogeneous distribution across clients. Dir=+∞ represents the uniform data distribution setting.

DUTCH	Method	Dir=0.5	Dir=1	Dir=8	Dir=+∞
ACC	**Local**	0.79±0.01	0.80±0.01	0.81±0.01	0.80±0.01
	FedAvg	0.81±0.01	0.81±0.01	0.81±0.01	0.81±0.01
	FedProx	0.80±0.01	0.81±0.01	0.81±0.01	0.81±0.01
	Scaffold	0.60±0.12	0.57±0.13	0.55±0.13	0.57±0.13
	FedFair	0.61±0.16	0.62±0.08	0.61±0.15	0.61±0.13
	LCO	0.62±0.03	0.67±0.03	0.61±0.05	0.61±0.13
	FL-FairBatch	0.81±0.01	0.81±0.01	0.81±0.01	0.81±0.01
	FedFB	0.69±0.05	0.74±0.05	0.74±0.04	0.60±0.10
	FairFed	0.62±0.13	0.70±0.07	0.75±0.06	0.61±0.12
	FedRényi ($1/K$)	**0.83±0.01**	0.83±0.00	0.83±0.00	0.83±0.00
	FedRényi (n_k/n)	0.83±0.01	**0.83±0.01**	**0.83±0.01**	**0.83±0.01**
FR	**Local**	0.67±0.04	0.67±0.05	0.65±0.06	0.65±0.07
	FedAvg	0.64±0.08	0.65±0.08	0.66±0.07	0.66±0.06
	FedProx	0.63±0.09	0.67±0.06	0.65±0.07	0.63±0.08
	Scaffold	0.84±0.18	0.87±0.13	0.87±0.16	0.84±0.16
	FedFair	0.65±0.35	0.71±0.12	0.62±0.19	0.72±0.19
	LCO	0.65±0.35	0.73±0.01	0.64±0.01	0.72±0.11
	FL-FairBatch	0.66±0.06	0.66±0.06	0.66±0.07	0.64±0.07
	FedFB	0.92±0.04	0.72±0.21	0.69±0.20	0.87±0.23
	FairFed	0.78±0.25	0.80±0.18	0.75±0.11	0.80±0.20
	FedRényi ($1/K$)	0.94±0.04	0.93±0.03	0.94±0.03	0.94±0.04
	FedRényi (n_k/n)	**0.96±0.04**	**0.95±0.04**	**0.94±0.04**	**0.96±0.04**
HM	**Local**	0.73±0.02	0.73±0.02	0.72±0.02	0.72±0.02
	FedAvg	0.72±0.02	0.72±0.02	0.73±0.02	0.73±0.02
	FedProx	0.70±0.02	0.73±0.02	0.72±0.02	0.71±0.02
	Scaffold	0.70±0.14	0.69±0.13	0.67±0.14	0.68±0.14
	FedFair	0.63±0.22	0.66±0.10	0.61±0.17	0.66±0.15
	LCO	0.63±0.06	0.70±0.01	0.62±0.02	0.66±0.12
	FL-FairBatch	0.73±0.02	0.73±0.02	0.73±0.02	0.72±0.02
	FedFB	0.79±0.04	0.73±0.08	0.71±0.07	0.71±0.14
	FairFed	0.69±0.17	0.75±0.10	0.75±0.08	0.69±0.15
	FedRényi ($1/K$)	0.88±0.02	0.88±0.00	0.88±0.00	0.88±0.00
	FedRényi (n_k/n)	**0.89±0.02**	**0.89±0.02**	**0.88±0.02**	**0.89±0.02**