EFFICIENT NEWTON-TYPE FEDERATED LEARNING WITH NON-IID DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

The mainstream federated learning algorithms only communicate the first-order information across the local devices, i.e., FedAvg and FedProx. However, only using first-order information, these methods are often inefficient and the impact of heterogeneous data is yet not precisely understood. This paper proposes an efficient federated Newton method (FedNewton), by sharing both first-order and second-order knowledge over heterogeneous data. In general kernel ridge regression setting, we derive the generalization bounds for FedNewton and obtain the minimax-optimal learning rates. For the first time, our results analytically quantify the impact of the number of local examples, the data heterogeneity and the model heterogeneity. Moreover, as long as the local sample size is not too small and data heterogeneity is moderate, the federated error in FedNewton decreases exponentially in terms of iterations. Extensive experimental results further validate our theoretical findings and illustrate the advantages of FedNewton over the first-order methods.

1 INTRODUCTION

Owing to the great potential in privacy preservation and in lowering the computational costs, federated learning (FL) McMahan et al. (2017); Li et al. (2020a); Zhang et al. (2021) becomes a promising framework in processing large-scale tasks. However, federated learning is facing massive challenges from the heterogeneous data Zhao et al. (2018); Zhou et al. (2023); Ye et al. (2023), including both the data heterogeneity and the model heterogeneity. The data heterogeneity comes from that inputs across devices are usually sampled from heterogeneous distributions, while the model heterogeneity measures the response shift due to inconsistency between local models and the global model.

First-order approaches, including FedAvg McMahan et al. (2017) and FedProx Li et al. (2020a), share the first-order information rather than the data across devices and tolerate the heterogeneity in federated learning, while Newton-type FL methods Ghosh et al. (2020); Gupta et al. (2021); Safaryan et al. (2022); Islamov et al. (2023); Liu et al. (2023); Dal Fabbro et al. (2024); Li et al. (2023) utilized second-order information for updating federated model. To the best of our knowledge, most of existing learning guarantees for FL methods are derived in the context of optimization and focused on in-sample predictive errors only, i.e., the convergence analysis (optimization) of first-order FL Li et al. (2020b); Karimireddy et al. (2020); Pathak & Wainwright (2020); Glasgow et al. (2022) and Newton-type FL Ghosh et al. (2020); Safaryan et al. (2022); Qian et al. (2022). However, beyond the optimization, the generalization guarantees (out-sample predictive performance) are of great practical and theoretical interests for FL. Despite recent efforts and progress on the generalization for first-order algorithms Mohri et al. (2019); Yagli et al. (2020); Su et al. (2021); Yuan et al. (2022), the generalization guarantees for Newton-type FL algorithms remain elusive, especially on heterogeneous data and localized models. Therefore, a challenging problem in FL is *how to quantify the impact of heterogeneity from the generalization perspective?*

In this paper, motivated by sharing second-order information, we propose a second-order federated optimization method, named `FedNewton`. It approximates the global predictor on the entire data by utilizing the global gradient and local Hessians, improving the predictive accuracy in an efficient communications framework. We then study the statistical properties of `FedNewton`, and derive the generalization bounds with the minimax optimal rates. We conclude with experiments on simulated data and publicly available tasks that complement our theoretical results, exhibiting the computational and statistical benefits of our approach. Due to the length limit, we leave the experiment part in the appendix. We summarize our contributions as below:

1) On the algorithmic front. We propose a fast second-order federated learning algorithm, which improves the approximation of the centralized model while only requiring similar computational and communication costs as the first-order methods. The convergence of `FedNewton` is exponentially fast and a few communications, for example, $t \leq 2$, can approximate the global model well.

2) On the statistical front. To our best knowledge, in presence of both data heterogeneity and model heterogeneity, we present the optimal generalization guarantees for the first time. Our results further analytically quantify the impacts of the local sample size, the data heterogeneity, and the model heterogeneity. Especially, the federated error decreases exponentially fast in benign cases, i.e., a sufficient number of local examples and moderate data heterogeneity.

2 Problem Setup

In a standard framework of federated learning, there is a global parameter server and m local computational clients. On the j-th local machine $\forall j \in [m]$, the local data $\mathfrak{D}_j = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{i=1}^{|\mathfrak{D}_j|}$ is drawn from a local distribution ρ_j on the joint space $\mathcal{X} \times \mathcal{Y}$. The total sample $\mathcal{D} = \bigcup_{j=1}^{m} \mathfrak{D}_j$ is the disjoint union of local data and corresponds to a global distribution ρ. For any local devices $j, k \in [m]$ and $j \neq k$, data distributions are identical $\rho_j = \rho_k = \rho$ in the homogeneous setting (iid data), while data distributions are distinct $\rho_j \neq \rho_k$ in the heterogeneous case (non-iid data).

We base our analysis on the standard non-parametric regression setup and assume that the target solution f^* belongs to a reproducing kernel Hilbert space (RKHS) induced by a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Mercer's theorem guarantees the kernel function admits an implicit feature mapping $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_K$ and the norm by $\| \cdot \|_K$. The predictor can be stated as $f_{\mathcal{D},\lambda}(\boldsymbol{x}) = \langle \boldsymbol{w}_{\mathcal{D},\lambda}, \phi(\boldsymbol{x}) \rangle$ where $\boldsymbol{w}_{\mathcal{D},\lambda}$ minimizes the objective on the entire data $\mathcal{D}$

$$\operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{H}_K} \left\{ \frac{1}{2|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left(f(\boldsymbol{x}_i) - y_i\right)^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_K^2 \right\}, \tag{1}$$

where $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$, and $\lambda > 0$ is the regularity parameter. The above regression problem, known as Kernel Ridge Regression (KRR), admits a closed-form solution

$$\boldsymbol{w}_{\mathcal{D},\lambda} = (\boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{\Phi}_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{y}_{\mathcal{D}}, \tag{2}$$

where $\boldsymbol{\Phi}_{\mathcal{D}} = \frac{1}{\sqrt{|\mathcal{D}|}} \left[\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_{|\mathcal{D}|})\right]^T \in \mathbb{R}^{|\mathcal{D}|} \times \mathcal{H}_K$ are feature mappings on the training set $\mathcal{D}$ and $\boldsymbol{y}_{\mathcal{D}} = \frac{1}{\sqrt{|\mathcal{D}|}} \left(y_1, \cdots, y_{|\mathcal{D}|}\right)^\top$ are the corresponding labels.

By averaging the local models, the simplest federated method only communicates once, known as Distributed Kernel Ridge Regression (DKRR) with the closed-form solution

$$\bar{\boldsymbol{w}}_{\mathcal{D},\lambda} = \sum_{j=1}^{m} p_j (\boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{\Phi}_{\mathfrak{D}_j} + \lambda I)^{-1} \boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{y}_{\mathfrak{D}_j},$$

where p_j is the weight of the j-th local model, which is usually set $p_j = |\mathfrak{D}_j|/|\mathcal{D}|$. Note that, $\boldsymbol{\Phi}_{\mathfrak{D}_j} = \frac{1}{\sqrt{|\mathfrak{D}_j|}} \left[\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_{|\mathfrak{D}|_j})\right]^T \in \mathbb{R}^{|\mathfrak{D}_j|} \times \mathcal{H}_K$ are local feature mappings and $\boldsymbol{y}_{\mathfrak{D}_j} = \frac{1}{\sqrt{|\mathfrak{D}_j|}} \left(y_1, \cdots, y_{|\mathfrak{D}_j|}\right)^\top$ are labels on the j-th local train set $\mathfrak{D}_j = \left\{(\boldsymbol{x}_{ij}, y_{ij})\right\}_{i=1}^{|\mathfrak{D}_j|}, \quad \forall j \in [m]$.

The solution of KRR equation 2 can be rewritten in the Newton's method form

$$\boldsymbol{w}_{\mathcal{D},\lambda} = \boldsymbol{w} - \boldsymbol{H}_{\mathcal{D},\lambda}^{-1} \boldsymbol{g}_{\mathcal{D},\lambda}. \tag{3}$$

where the gradient and Hessian matrix are defined as

$$\begin{aligned} \boldsymbol{g}_{\mathcal{D},\lambda} &:= (\boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{\Phi}_{\mathcal{D}} + \lambda I)\boldsymbol{w} - \boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{y}_{\mathcal{D}}, \\ \boldsymbol{H}_{\mathcal{D},\lambda} &:= (\boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{\Phi}_{\mathcal{D}} + \lambda I). \end{aligned}$$

Algorithm 1 Federated Learning with Newton Method (`FedNewton`)

```
Input: Local training data subset $\mathfrak{D}_j$, $\forall j \in [m]$. Feature mapping $\phi : \mathcal{X} \to \mathbb{R}^M$.
Output: The global estimator $\bar{\boldsymbol{w}}^T_{\mathcal{D},\lambda}$.
Local machines: Compute feature mapping $\boldsymbol{\Phi}_{\mathfrak{D}_j}$, $\boldsymbol{H}_{\mathfrak{D}_j,\lambda} = (\boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{\Phi}_{\mathfrak{D}_j} + \lambda I)$, $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}$ and $\boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{y}_{\mathfrak{D}_j}$ for any $j \in [m]$.
Local machines: Initialize the local estimators by $\boldsymbol{w}^0_{\mathfrak{D}_j,\lambda} = \boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1} \boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{y}_{\mathfrak{D}_j}$ and upload them to the global server ($\uparrow$).
Global server: Initialize the solution by $\bar{\boldsymbol{w}}^0_{\mathcal{D},\lambda} = \sum_{j=1}^m p_j \boldsymbol{w}^0_{\mathfrak{D}_j,\lambda}$, and send it to the local nodes ($\downarrow$).
for $t = 1$ to $T$ do
    Local machines: Compute local gradients $\boldsymbol{g}^{t-1}_{\mathfrak{D}_j,\lambda} = \boldsymbol{H}_{\mathfrak{D}_j,\lambda} \bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} - \boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{y}_{\mathfrak{D}_j}$ and upload them to global server ($\uparrow$).
    Global server: Compute the global gradient $\boldsymbol{g}^{t-1}_{\mathfrak{D},\lambda} = \sum_{j=1}^m p_j \boldsymbol{g}^{t-1}_{\mathfrak{D}_j,\lambda}$ and send it to local nodes ($\downarrow$).
    Local machines: Compute the local updates $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}\, \boldsymbol{g}^{t-1}_{\mathfrak{D},\lambda}$ and upload it to the global server ($\uparrow$).
    Global server: Update the global estimator $\bar{\boldsymbol{w}}^t_{\mathcal{D},\lambda} = \bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} - \sum_{j=1}^m p_j \boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}\, \boldsymbol{g}^{t-1}_{\mathfrak{D},\lambda}$ and communicate it to local machines ($\downarrow$).
end for
```

From equation 3, the global gradient $\boldsymbol{g}_{\mathcal{D},\lambda}$ and Hessian $\boldsymbol{H}_{\mathcal{D},\lambda}$ is the key to achieving the centralized model $\boldsymbol{w}_{\mathcal{D},\lambda}$. Note that, since the fact $\boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{\Phi}_{\mathcal{D}} = \sum_{j=1}^m p_j \boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{\Phi}_{\mathfrak{D}_j}$ for data partition $\mathcal{D} = \bigcup_{j=1}^m \mathfrak{D}_j$, one can easily obtain the following property for the global gradient and global Hessian.

Proposition 1 (Partitonability). *If the loss is squared loss, the global gradient and Hessian matrix consist of the local ones, i.e.* $\boldsymbol{g}_{\mathcal{D},\lambda} = \sum_{j=1}^m p_j \boldsymbol{g}_{\mathfrak{D}_j,\lambda}$ *and* $\boldsymbol{H}_{\mathcal{D},\lambda} = \sum_{j=1}^m p_j \boldsymbol{H}_{\mathfrak{D}_j,\lambda}$.

Remark 1 (Computation of local inverse Hessian). *The compute of the inverse of local Hessians* $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}$ *is time consuming* $\mathcal{O}(|\mathfrak{D}_j|M^2 + M^3)$*, which is a common problem in second-order optimization Bottou et al. (2018). There are many classic work to reduce the time complexity of the inverse of Hessian, i.e. BFGS Broyden (1970), L-BFGS Liu & Nocedal (1989), inexact Newton Dembo et al. (1982), Gauss-Newton Schraudolph (2002) and Newton sketch Pilanci & Wainwright (2017). Those techniques can be used to improve the efficiency of* `FedNewton`*, but it is beyond the scope of this paper. We focus on theoretical novelties and leave further computational improvements in the future.*

Remark 2 (Feature mapping instead of kernel methods). *Without loss of generality, we assume the feature mappings are finite dimensional* $\phi : \mathcal{X} \to \mathbb{R}^M$*, which covers a wide range of generalized linear models, for example neural networks Neal (1995); Jacot et al. (2018), kernel methods Vapnik (2000), random features Rahimi & Recht (2007); Le et al. (2013); Yang et al. (2014), and random sketching Woodruff et al. (2014); Yang et al. (2017).*

3 Federated Learning with Newton Method

Motivated by recent gradient-based distributed learning Wang et al. (2018); Lin et al. (2020), we propose a Newton-type federated learning method to quantity the impact of data heterogeneity and model heterogeneity. Using Proposition 1, the exact Federated Newton's method communicate local Hessians $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}$ for computing the global Hessian matrix equation 3 whose the communication complexity is $\boldsymbol{O}(M^2)$, which is infeasible in federated learning. To reduce communication costs, we propose `FedNewton` that approximates the Newton's updates with the global gradient and local Hessian matrices, such that

$$\boldsymbol{H}_{\mathcal{D},\lambda}^{-1} \boldsymbol{g}_{\mathcal{D},\lambda} \approx \sum_{j=1}^m p_j \boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1} \boldsymbol{g}_{\mathcal{D},\lambda}. \tag{4}$$

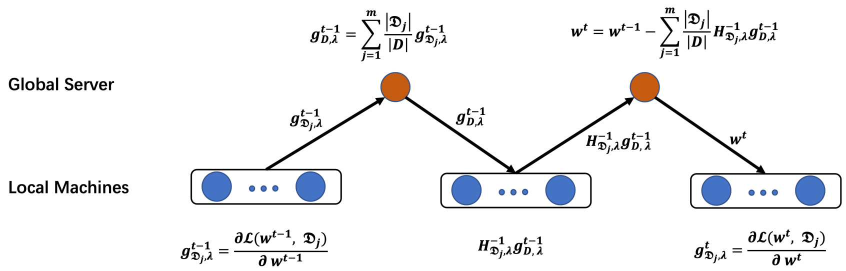
$g_{D,\lambda}^{t-1} = \sum_{j=1}^{m} \frac{|\mathfrak{D}_j|}{|D|} g_{\mathfrak{D}_j,\lambda}^{t-1}$
$w^t = w^{t-1} - \sum_{j=1}^{m} \frac{|\mathfrak{D}_j|}{|D|} H_{\mathfrak{D}_j,\lambda}^{-1} g_{D,\lambda}^{t-1}$
Global Server
$g_{\mathfrak{D}_j,\lambda}^{t-1}$
$g_{D,\lambda}^{t-1}$
$H_{\mathfrak{D}_j,\lambda}^{-1} g_{D,\lambda}^{t-1}$
w^t
Local Machines
$g_{\mathfrak{D}_j,\lambda}^{t-1} = \frac{\partial \mathcal{L}(w^{t-1}, \mathfrak{D}_j)}{\partial w^{t-1}}$
$H_{\mathfrak{D}_j,\lambda}^{-1} g_{D,\lambda}^{t-1}$
$g_{\mathfrak{D}_j,\lambda}^{t} = \frac{\partial \mathcal{L}(w^{t}, \mathfrak{D}_j)}{\partial w^{t}}$

Figure 1: The computations and communications in the t-th iteration for `FedNewton`.

The global learner $\bar{f}_{\mathcal{D},\lambda}^t(\boldsymbol{x}) = \langle \bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^t, \phi(\boldsymbol{x}) \rangle$ is updated by

$$\bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^t = \bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^{t-1} - \sum_{j=1}^{m} p_j \boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1} \boldsymbol{g}_{\mathcal{D},\lambda}^{t-1}, \tag{5}$$

where $\bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^t$ is the model after t iterations and the global gradient is $\boldsymbol{g}_{\mathcal{D},\lambda}^{t-1} = \sum_{j=1}^m p_j \boldsymbol{g}_{\mathfrak{D}_j,\lambda}^{t-1}$ from Proposition 1. The approximation error between equation 3 and equation 5 is analyzed in Section 4. Without loss of generality, we present the details of `FedNewton` in Algorithm 1 and Figure 1, which includes two times communications as the first-order methods in per round. Note that, the algorithm uploads local Newton updates $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1} \boldsymbol{g}_{\mathcal{D},\lambda}^{t-1} \in \mathbb{R}^M$ instead of local inverse Hessians $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1} \in \mathbb{R}^{M \times M}$, reducing communication costs from $\boldsymbol{O}(M^2)$ to $\boldsymbol{O}(M)$.

Computational complexity analysis. With finite-dimensional feature mappings $\phi : \mathcal{X} \to \mathbb{R}^M$, we compute time complexity, space complexity, and communication complexity of `FedNewton`. The space complexity on the j-th local machine is $\mathcal{O}(|\mathfrak{D}_j|M + M^2)$ to store $\boldsymbol{\Phi}_{\mathfrak{D}_j}, \boldsymbol{H}_{\mathfrak{D}_j,\lambda}$ and $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}$, while the global server requires $\mathcal{O}(mM)$ space to store $\boldsymbol{g}_{\mathfrak{D}_j,\lambda}$ and $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1} \boldsymbol{g}_{\mathcal{D},\lambda}$. Before the iterations, the computations of $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}$ and $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}$ costs $\boldsymbol{O}(|\mathfrak{D}_j|M^2{+}M^3)$ time. In each iteration, the local time complexity is $\mathcal{O}(M^2)$ to compute local gradient $\boldsymbol{g}_{\mathfrak{D}_j,\lambda}$ and local Newton update $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}^{-1}\ \boldsymbol{g}_{\mathfrak{D},\lambda}$, while the time complexity on the global server is $\mathcal{O}(mM)$ to update the global gradient and estimator. Therefore, the total time complexity is $\mathcal{O}\left(\max_{j\in[m]} |\mathfrak{D}_j|M^2 + M^3 + M^2 t + mMt\right)$.

Remark 3 (Communication burdens)**.** *The per iteration communication costs of the proposed* `FedNewton` *are* 2 *times as compared to the first-order FL algorithms, e.g. FedAvg and FedProx, but the number of iterations for* `FedNewton` *is much fewer. The total communication complexity is $\mathcal{O}(Mt)$, the same as most first-order Federated algorithms. Notably, from Theorem 1 the iteration complexity is a linear convergence $t = \Omega(\log(1/\epsilon))$ where ϵ is the federated error, i.e.,* `FedNewton` *converges exponentially to the global estimator equation 2, while first-order federated algorithms requires a large number of communication rounds $t = \Omega(1/\epsilon)$ Su et al. (2021). Therefore,* `FedNewton` *cannot reduce the communication complexity for once communication as communication-efficient FL algorithms Sattler et al. (2019); Reisizadeh et al. (2020); Wu et al. (2022), but it significantly reduces the number of communication rounds, e.g.,* `FedNewton` *with $t \leq 2$ achieves good predictive performance in Section 7.*

Remark 4 (Beyond the squared loss)**.** *To quantify the impacts from local sample size, data heterogeneity and model heterogeneity, we apply the squared loss for* `FedNewton` *because it admits closed-form solutions and is convenient for the theoretical analysis. Nevertheless, the proposed algorithm* `FedNewton` *is not applies to a broad range of loss functions as long as they are twice differentiable to compute the gradient $\boldsymbol{g}_{\mathfrak{D}_j,\lambda}^{t-1}$ and the Hessian matrix $\boldsymbol{H}_{\mathfrak{D}_j,\lambda}$. If the Hessian is independent from the weights, the compute of local Hessians can be out of the loop, e.g. ReLU and the squared loss. However, if the Hessian is relevant to the weights, for example exponential loss functions and trigonometric loss functions, we should compute the local Hessians for all iterations, causing huge computational burdens. For other type loss functions, the weights can be initialized as $\bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^0 = \boldsymbol{0}$.*

4 Main Results

In this section, to explore the factors that affect performance, we derive the excess risk bounds for `FedNewton` in homogeneous settings and heterogeneous settings, respectively.

4.1 Notations and Assumptions

We consider a broader scenario for federated learning, where the local training sets contain both heterogenous inputs (covariate shift) $\mathfrak{D}_j \sim \rho_j$ and different responses (concept shift) $\boldsymbol{y}_{\mathfrak{D}_j} \sim \rho_j(y|\boldsymbol{x})$. The concept shift is represented as

$$f^*(\boldsymbol{x}) = \int_{\mathcal{Y}} y d\rho(y|\boldsymbol{x}),\, \boldsymbol{x} \in \mathcal{X}, \qquad f_j^*(\boldsymbol{x}) = \int_{\mathcal{Y}} y d\rho_j(y|\boldsymbol{x}),\, \boldsymbol{x} \in \mathcal{X},\, j \in [m], \tag{6}$$

where f_j^* is the underlying mechanism governing the true responses on the j-th worker. Give a $\boldsymbol{x} \in \mathcal{X}$ and $j, k, \in [m]$, the responses may be different $f_j^*(\boldsymbol{x}) \neq f_k^*(\boldsymbol{x})$ when $j \neq k$.

Definition 1 (Operators with feature mapping ϕ)**.** *Using the feature mapping $\phi : \mathcal{X} \to \mathcal{H}_K$, $\forall\, \boldsymbol{\beta} \in \mathcal{H}_K$, the covariance operators $C, C_j, C_{\mathcal{D}}, C_{\mathfrak{D}_j} : \mathcal{H}_K \to \mathcal{H}_K$ are defined as*

$$C\boldsymbol{\beta} = \int_X \langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle \phi(\boldsymbol{x}) d\rho_X(\boldsymbol{x}), \qquad C_{\mathcal{D}}\boldsymbol{\beta} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}_i)\rangle \phi(\boldsymbol{x}_i),\ \forall\ (\boldsymbol{x}_i, y_i) \in \mathcal{D},$$

$$C_j\boldsymbol{\beta} = \int_X \langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle \phi(\boldsymbol{x}) d\rho_j(\boldsymbol{x}), \qquad C_{\mathfrak{D}_j}\boldsymbol{\beta} = \frac{1}{|\mathfrak{D}_j|}\sum_{i=1}^{|\mathfrak{D}_j|} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}_i)\rangle \phi(\boldsymbol{x}_i),\ \forall\ (\boldsymbol{x}_i, y_i) \in \mathfrak{D}_j.$$

Note that, $C_{\mathcal{D}} = \boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{\Phi}_{\mathcal{D}}$, $C_{\mathfrak{D}_j} = \boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{\Phi}_{\mathfrak{D}_j}$ are the empirical covariance operators on $\mathcal{D}$ and $\mathfrak{D}_j$, while $C = \mathbb{E}_\rho[C_{\mathcal{D}}], C_j = \mathbb{E}_{\rho_j}[C_{\mathfrak{D}_j}]$ are their expected counterparts.

For the sake of readability, we provide some notations

$$\begin{aligned} &\mathcal{P}_{\mathfrak{D}_j,\lambda} := \|(C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_j + \lambda I)\|, \qquad &&\mathcal{R}_{\mathfrak{D}_j,\lambda} := \|(C_j + \lambda)^{-1}(C_j - C_{\mathfrak{D}_j})\|, \\ &\Delta_{\mathfrak{D}_j} := \|C - C_j\|, &&\Delta_{f_j} := \|f^* - f_j^*\|. \end{aligned}$$

The quantities $\mathcal{P}_{\mathfrak{D}_j,\lambda}$ and $\mathcal{R}_{\mathfrak{D}_j,\lambda}$ measure the similarity between the expected covariance operator and its empirical counterpart. From contraction inequalities for self-adjoint operators, a larger number of local samples $|\mathfrak{D}_j|$ leads to smaller $\mathcal{P}_{\mathfrak{D}_j,\lambda}$ and $\mathcal{R}_{\mathfrak{D}_j,\lambda}$. Note that, $\Delta_{\mathfrak{D}_j}$ measures the data heterogeneity on the expected covariance operator, while Δ_{f_j} measures the model heterogeneity on the true regressions.

We let $\|f\|_2 = \sqrt{\langle f, f\rangle} = \sqrt{\int_X |f(\boldsymbol{x})|^2 d\mathbb{P}(\boldsymbol{x})}$ denote the $L^2(\mathbb{P})$ norm and $L^2(\mathbb{P}) = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_2^2 < \infty\}$. Throughout this paper, we assume the outputs are bounded $|y| \leq B$ almost surely for some $B > 0$ and $\kappa := \|\phi(\boldsymbol{x})\|_K < \infty$ for any $\boldsymbol{x} \in \mathcal{X}$.

Assumption 1 (Federated capacity condition)**.** *For $\lambda \in (0, 1)$, we define the effective dimensions on the global distribution ρ and local distributions ρ_j, $\forall j \in [m]$ as*

$$\mathcal{N}(\lambda) = Tr(C(C + \lambda I)^{-1}),\ \mathcal{N}_j(\lambda) = Tr(C_j(C_j + \lambda I)^{-1}).$$

Assume there exists $Q > 0$ and $\gamma \in [0, 1]$, such that

$$\max\left(\mathcal{N}(\lambda), \mathcal{N}_1(\lambda), \cdots, \mathcal{N}_m(\lambda)\right) \leq Q^2 \lambda^{-\gamma}.$$

Assumption 2 (Source condition)**.** *Define the integral operators $L : L^2(\mathbb{P}) \to L^2(\mathbb{P})$,*

$$(Lg)(\cdot) = \int_X \langle \phi(\cdot), \phi(\boldsymbol{x})\rangle g(\boldsymbol{x}) d\rho_X(\boldsymbol{x}), \quad \forall\, g \in L^2(\mathbb{P}).$$

Assume there exists $R > 0$, $r > 0$, such that $\|L^{-r} f^*\| \leq R$. where the operator L^r denotes the r-th power of L as a compact and positive operator.

Capacity condition and source condition are standard assumptions in the optimal statistical learning for the KRR related literature Caponnetto & De Vito (2007); Smale & Zhou (2007); Rudi & Rosasco (2017); Lin & Cevher (2020); Liu et al. (2021). The effective dimensions $\mathcal{N}(\lambda)$ and $\mathcal{N}_j(\lambda)$ measure the capacities of the RKHS $\mathcal{H}_K$ on the global distribution ρ and the local distributions $\rho_j,\ \forall j \in [m]$.

Here, we modify the conventional capacity condition for federated learning to impose constraints on local estimators. Note that, for effective dimensions, it holds $1/2 \leq \max\left(\mathcal{N}(\lambda), \mathcal{N}_1(\lambda), \cdots, \mathcal{N}_m(\lambda)\right) \leq \kappa^2\lambda^{-1}$ Rudi et al. (2015). Assumption 1 reflects the variance of the estimator. A larger γ leads to a larger $\mathcal{H}_K$ and $\gamma = 1$ corresponds to the capacity independence case. Assumption 2 controls the bias of an estimator, which reflects the regularity of the estimator. The bigger r leads to the stronger regularity of the regression and the easier learning problem. The general settings $(r = 1/2, \gamma = 1)$ lead to $\boldsymbol{O}(1/\sqrt{|D|})$ convergence rates for KRR related approaches.

4.2 Error Decomposition

Theorem 1. *Let $f_{\mathcal{D},\lambda}, \bar{f}^t_{\mathcal{D},\lambda}, f^*$ be defined according to equation 2, equation 5 and equation 6. Then, the following error decomposition holds*

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\| \leq \underbrace{\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|}_{\textit{federated error}} + \underbrace{\|f_{\mathcal{D},\lambda} - f^*\|}_{\textit{centralized excess risk}}, \tag{7}$$

and the federated error for `FedNewton` *is bounded by:*

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \leq \Upsilon^t \left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^0_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K,$$

where $\Upsilon = \sum_{j=1}^m p_j \mathcal{P}_{\mathfrak{D}_j,\lambda}\left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left(1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right).$

In the above theorem, we decompose the excess risk for `FedNewton` into two parts: the federated error $\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|$ and the excess risk for the centralized KRR $\|f_{\mathcal{D},\lambda} - f^*\|$. Since the generalization analysis for $\|f_{\mathcal{D},\lambda} - f^*\|$ is standard Caponnetto & De Vito (2007); Smale & Zhou (2007), we focus on the federated error $\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|$.

From Theorem 1, we find that the value of Υ determines the effectiveness of multiple iterations. If $\Upsilon \geq 1$, `FedNewton` with multiple communications is worse than oneshot federated learning (DKRR). However, when $\Upsilon < 1$, the federated error decreases exponentially and the rate of convergence is referred to as *linear convergence* in the optimization literature Bottou et al. (2018). The quantities $\mathcal{P}_{\mathfrak{D}_j,\lambda}$ and $\mathcal{R}_{\mathfrak{D}_j,\lambda}$ measure the similarity between $C_{\mathfrak{D}_j}$ and C_j where those quantities decrease as the local sample size $|\mathfrak{D}_j|$ increases. Because Υ is proportional to $\mathcal{P}_{\mathfrak{D}_j,\lambda}$, $\mathcal{P}_{\mathfrak{D}_j,\lambda}$ and $\Delta_{\mathfrak{D}_j}$, the *linear convergence* requires both a sufficient number of local examples $|\mathfrak{D}_j|$ and moderate data heterogeneity $\Delta_{\mathfrak{D}_j}$. If $t = 0$, the above error bound degrades into that for DKRR $\|\bar{f}_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \leq \left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^0_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K$.

Theorem 2. *Under Assumption 2, with a high probability $1-\delta$, $\forall \delta \in (0,1)$, the federated error can be bounded*

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \lesssim \Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}}\left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right) \cdot \left(\left(\frac{1}{|\mathfrak{D}_j|\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{|\mathfrak{D}_j|}}\right)\log\frac{2}{\delta} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}\right).$$

Theorem 2 illustrates the key factors that affect the federated error: the discrepancy between expected and empirical covariance operators $\mathcal{R}_{\mathfrak{D}_j,\lambda}$, the covariate shift $\Delta_{\mathfrak{D}_j}$, and the model heterogeneity Δ_{f_j}. The smaller these factors, the smaller the federated error. The federated error results from three parts: distributed error $\frac{1}{\sqrt{\lambda}|\mathfrak{D}_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|\mathfrak{D}_j|}}$, covariate shift $\Delta_{\mathfrak{D}_j}/\lambda$ and concept shift Δ_{f_j}. Specifically, as the increase of local sample size, the distributed error decreases. However, the concept shifts Δ_{f_j} is a constant and it will dominate the federated error when model heterogeneity Δ_{f_j} is large. In the case $\Upsilon < 1$, iterators can reduce the federated error, alleviating the entire federated error term.

4.3 Homogeneous Setting

Theorem 3. *Let $\delta \in (0, 1/3]$, $\lambda = |\mathcal{D}|^{\frac{-1}{2r+\gamma}}$ and $2r + \gamma \geq 1$. Under Assumptions 1, 2, if $\Delta_{\mathfrak{D}_j} = 0$ and $\Delta_{f_j} = 0$, with the probability at least $1 - 3\delta$, it holds*

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \lesssim \Upsilon^t \sum_{j=1}^m p_j \aleph_j \, \log^2 \frac{2}{\delta} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}} \log \frac{2}{\delta}.$$

Here, $\aleph_j$ and Υ have different values w.r.t local sample size

$$\aleph_j = \begin{cases} |\mathfrak{D}_j|^{-2}|\mathcal{D}|^{\frac{1.5}{2r+\gamma}}, & \textit{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \\ |\mathfrak{D}_j|^{-1.5}|\mathcal{D}|^{\frac{1+0.5\gamma}{2r+\gamma}}, & \textit{if } |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\ |\mathfrak{D}_j|^{-1}|\mathcal{D}|^{\frac{1+\gamma}{4r+2\gamma}}, & \textit{if } |\mathcal{D}|^{\frac{1}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}} \\ |\mathcal{D}|^{\frac{-r}{2r+\gamma}}, & \textit{if } |\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}}, \end{cases}$$

and $\Upsilon = 2\sum_{j=1}^m p_j \mathcal{P}_{\mathfrak{D}_j,\lambda}\mathcal{R}_{\mathfrak{D}_j,\lambda}$ holds

$$\begin{cases} \Upsilon \geq 1, & \textit{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\ \Upsilon \lesssim \frac{|\mathcal{D}|^{\frac{1}{2r+\gamma}}}{|\mathfrak{D}_j|} < 1, & \textit{otherwise}. \end{cases}$$

Note that, the second term in the above bound is from the centralized model $\|f_{\mathcal{D},\lambda} - f^*\|_2$, where the learning rate $\boldsymbol{O}(|\mathcal{D}|^{\frac{-r}{2r+\gamma}})$ is optimal in a minimax sense Caponnetto & De Vito (2007). The performance of `FedNewton` in the homogeneous setting is only affected by the local sample size. We discuss the above result in three parts. First, when the number of local examples is limited $|\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}}$, in another word the number of local machines is larger than $m \gtrsim |\mathcal{D}|^{\frac{2r+\gamma-1}{2r+\gamma}}$, the federated error dominates the excess risk and fails to achieve the optimal rate, where the convergence rates are slower than $\mathcal{O}(|\mathcal{D}|^{\frac{\gamma-1}{4r+2\gamma}})$. Meanwhile, when the number of local examples is limited, it leads to $\Upsilon \geq 1$ and multiple communications hurt the performance. Second, when $|\mathcal{D}|^{\frac{1}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}}$, although the convergence rates of federated error are still not the optimal, the iterator Υ is smaller than one, leading to a linear convergence. As the increase of communications $t \to \infty$, the centralized excess risk will dominate the error bound that achieves the optimal rate. Third, with a large number of local examples $|\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}}$, even with insufficient communications $t \to 0$, the error bound still achieves the optimal rate $\boldsymbol{O}(|\mathcal{D}|^{\frac{-r}{2r+\gamma}})$.

Theorem 3 can be further simplified in some special cases. For example, we consider the general case $(r = 1/2, \gamma = 1)$, where $r = 1/2$ is equivalent to assuming $f^* \in \mathcal{H}_K$ and $\gamma = 1$ is the capacity independent case. The learning rate achieves $\boldsymbol{O}(1/\sqrt{|\mathcal{D}|})$ when $|\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{0.5}$ with multiple iterations or $|\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{0.75}$ with only one communication.

Remark 5. *The existing theoretical guarantees for DKRR Zhang et al. (2015); Guo et al. (2017); Lin & Cevher (2020) focused on how to achieve the optimal rate by a sufficient number of local examples (or lower the number of partitions), but they ignored the sub-optimal case that the local sample size is fixed and insufficient. However, in federated learning, the number of partitions is fixed and local examples are generated locally, such that sub-optimal cases are more general. Theorem 3 illustrate that a sufficient number of local examples is crucial for both learning rates (in generalization) and convergence rate (in optimization).*

Remark 6 (Finite dimensional case)**.** *In the proofs of theoretical findings, we consider the estimator in RKHS with $\boldsymbol{w} \in \mathcal{H}_K$. However, the finite-dimensional cases are more general, i.e. $\boldsymbol{w} \in \mathbb{R}^M$ in Algorithm 1, where the feature mappings are explicit and can be neural networks or random features Rahimi & Recht (2007). With a simple modification of our proofs, one can derive similar results for finite-dimensional cases. In particular, under same assumptions of Theorem 3 and $(r = 1/2, \gamma = 0)$, then with high probability, $\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \lesssim |\mathfrak{D}_j|^{-2}|\mathcal{D}|^{1.5} + \sqrt{M/|\mathcal{D}|}$, provided that $|\mathcal{D}| \gtrsim M \log M$. As shown in Rudi & Rosasco (2017), a large number of random features $M \gtrsim |\mathcal{D}|^{\frac{1+\gamma(2r-1)}{2r+\gamma}}$ can guarantee the optimal rates for $\|\bar{f}_{\mathcal{D},\lambda} - f^*\|_2$, and thus we can also provide similar results as Theorem 3.*

4.4 Heterogeneous Setting

Theorem 4. *Let $\delta \in (0, 1/3]$, $\lambda = |\mathcal{D}|^{\frac{-1}{2r+\gamma}}$ and $2r + \gamma \geq 1$. Under Assumptions 1, 2, with the probability at least $1 - 3\delta$, the excess risk bound for* `FedNewton` *holds*

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \lesssim \Upsilon^t \sum_{j=1}^{m} p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}(\aleph_j + \mathit{\Pi}_j)}\, \log^2 \frac{2}{\delta} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}} \log \frac{2}{\delta}.$$

Here, $\Upsilon = \sum_{j=1}^m p_j \mathcal{P}_{\mathfrak{D}_j,\lambda}(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda})(1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda})$, $\aleph_j$ is same to Theorem 3 and

$$\mathit{\Pi}_j = \begin{cases} \frac{|\mathcal{D}|^{\frac{2}{2r+\gamma}}}{|\mathfrak{D}_j|}\Delta_{\mathfrak{D}_j} + \frac{|\mathcal{D}|^{\frac{1}{2r+\gamma}}}{|\mathfrak{D}_j|}\Delta_{f_j}, & \textit{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\ (1 + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j})(\Delta_{f_j} + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j}), & \textit{otherwise}. \end{cases}$$

We add some comments on the above theorem. First, when the local sample size is insufficient $|\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}}$ or the data heterogeneity is considerable, we have $\Upsilon \geq 1$, and communications hurt the performance. Meanwhile, since the federated error $\sqrt{1 + \Delta_{\mathfrak{D}_j}/\lambda}(\aleph_j + \mathit{\Pi}_j)$ depends on $|\mathfrak{D}_j|, \Delta_{\mathfrak{D}_j}$, and Δ_{f_j}, the learning rate is far from the optimal rate. Second, when the number of local examples is sufficient $|\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{1}{2r+\gamma}}$ and data heterogeneity is small, it holds $\Upsilon < 1$ where communications can improve the generalization ability of `FedNewton`. In this case, the federated error $\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|$ converge exponentially fast. If t is large enough, the error bound in Theorem 4 depends on the centralized excess risk $\|f_{\mathcal{D},\lambda} - f^*\|_2$ and achieves the optimal learning rate.

The learning rate of generalization bound in Theorem 4 is determined by four factors: the local sample size $|\mathfrak{D}_j|$, the covariate shift $\Delta_{\mathfrak{D}_j}$, the response shift Δ_{f_j} and the number of iterations t. Furthermore, the iterator value Υ depends on $|\mathfrak{D}_j|$ and $\Delta_{\mathfrak{D}_j}$, such that these two values are important factors for both fast convergences (in optimization) and the learning rates (in generalization).

Remark 7 (How to achieve the optimal rate in federated learning?)**.** *The value of $\Upsilon < 1$ is key to obtaining a linear convergence rate and the optimal learning rate, where it depends on both local sample sizes $\Upsilon \propto \mathcal{R}_{\mathfrak{D}_j,\lambda} \propto |\mathfrak{D}_j|$ and data heterogeneity $\Upsilon \propto \Delta_{\mathfrak{D}_j}$. Note that, $\Delta_{\mathfrak{D}_j}$ measures the intrinsic discrepancy between local distributions and the global one, and thus it is a fixed value independent from the local sample size. Therefore, since $\Delta_{\mathfrak{D}_j}$ is a constant, we can obtain $\Upsilon < 1$ with a large number of local examples generated by local machines. And then, with a large number of iterations when $\Upsilon < 1$, the federated error, depending on both data heterogeneity and model heterogeneity, can become small enough to be negligible. In this case, a large number of local examples can guarantee both a linear convergence rate (for federated error) and the optimal learning rate (from the centralized excess risk). A large number of local examples benefit both optimization and generalization, rather than making tradeoffs between them.*

5 Compared with Related Work

We compare `FedNewton` with recent Newton-type methods, DKRR methods, and first-order FL algorithms in both algorithmic and theoretical fronts. Table 1 reports the main factors that affect the performance, the computational and generalization properties of related work.

Compared with Newton-type FL methods. Local Newton-type FL algorithms Yang et al. (2019); Ghosh et al. (2020); Gupta et al. (2021) conducted Newton updates instead of SGD in local machines, which only utilized local information (local SGD & local Hessian). Recent studies Safaryan et al. (2022); Qian et al. (2022) tried to use global information (global SGD & global Hessian) by communicating local Hessian shifts, but it leads to high communication costs $\boldsymbol{O}(M^2)$ per communication. Nevertheless, this work employs mixed information (global SGD & local Hessian) that reduce the communication cost to $\boldsymbol{O}(M)$. More importantly, the existing Newton-type FL work only provided the convergence analysis (optimization) Ghosh et al. (2020); Safaryan et al. (2022); Qian et al. (2022) without out-sample (generalization) error bounds, while this work bridges the optimization and generalization for `FedNewton`, which essentially guarantees its fast convergence and good generalization ability.

Compared with DKRR. The time complexities of DKRR approaches solved in kernel space Zhang et al. (2015); Guo et al. (2017) are much higher than that of stochastic optimization methods solved

Table 1: Summary of computational and generalization properties for related work.

Related Work	$\lvert\mathfrak{D}_j\rvert$	$\Delta_{\mathfrak{D}_j}$	Δ_{f_j}	Training Time	Testing Time	Communication	Conditions	Local Size $\lvert\mathfrak{D}_j\rvert$	Iteration t	Upper Bound
DKRR Zhang et al. (2015)	√	×	×	$\lvert\mathfrak{D}_j\rvert^3$	$\lvert\mathcal{D}_{\text{test}}\rvert\lvert\mathcal{D}\rvert$	$\lvert\mathcal{D}\rvert$	Specific kernels	$\Omega(r^2\kappa^4\log\lvert\mathcal{D}\rvert)$	$\boldsymbol{O}(1)$	$\boldsymbol{O}\left(\frac{1}{\lvert\mathcal{D}\rvert}\right)$
DKRR Guo et al. (2017)	√	×	×	$\lvert\mathfrak{D}_j\rvert^3$	$\lvert\mathcal{D}_{\text{test}}\rvert\lvert\mathcal{D}\rvert$	$\lvert\mathcal{D}\rvert$	$r\in[1/2,1]$	$\Omega(\lvert\mathcal{D}\rvert^{\frac{1+\gamma}{2r+\gamma}})$	$\boldsymbol{O}(1)$	$\boldsymbol{O}(\lvert\mathcal{D}\rvert^{\frac{-r}{2r+\gamma}})$
DKRR-SGD Lin & Cevher (2018)	√	×	×	$\lvert\mathcal{D}\rvert t$	$\lvert\mathcal{D}_{\text{test}}\rvert\lvert\mathcal{D}\rvert$	$\lvert\mathcal{D}\rvert$	$r\in[1/2,1]$	$\Omega(\lvert\mathcal{D}\rvert^{\frac{1}{2r+\gamma}})$	$\boldsymbol{O}(\lvert\mathcal{D}\rvert^{\frac{2-\gamma}{2r+\gamma}})$	$\boldsymbol{O}\left(\lvert\mathcal{D}\rvert^{\frac{-r}{2r+\gamma}}\right)$
DKRR-CM Lin et al. (2020)	√	×	×	$\lvert\mathfrak{D}_j\rvert^3+\lvert\mathcal{D}\rvert\lvert\mathfrak{D}_j\rvert t$	$\lvert\mathcal{D}_{\text{test}}\rvert\lvert\mathcal{D}\rvert$	$\lvert\mathcal{D}\rvert t$	$r\in[1/2,1]$	$\Omega(\lvert\mathcal{D}\rvert^{\frac{2r+\gamma+1}{4r+2\gamma}})$	$\boldsymbol{O}(\log\frac{1}{\epsilon})$	$\boldsymbol{O}\left(\lvert\mathcal{D}\rvert^{\frac{-r}{2r+\gamma}}\right)$
FedAvg Su et al. (2021)	×	×	√	$\lvert\mathfrak{D}_j\rvert M^2+M^2t+mMt$	$\lvert\mathcal{D}_{\text{test}}\rvert M$	Mt	Specific kernels	/	$\boldsymbol{O}(\frac{1}{\epsilon})$	$\boldsymbol{O}\left(\frac{1}{\eta t}+\frac{\Delta_f^2}{\lvert\mathcal{D}\rvert}\right)$
FedProx Su et al. (2021)	×	×	√	$\lvert\mathfrak{D}_j\rvert M^2+M^3+M^2t+mMt$	$\lvert\mathcal{D}_{\text{test}}\rvert M$	Mt	Specific kernels	/	$\boldsymbol{O}(\frac{1}{\epsilon})$	$\boldsymbol{O}\left(\frac{1}{\eta t}+\frac{\Delta_f^2}{\lvert\mathcal{D}\rvert}\right)$
Theorem 3	√	×	×	$\lvert\mathfrak{D}_j\rvert M^2+M^3+M^2t+mMt$	$\lvert\mathcal{D}_{\text{test}}\rvert M$	Mt	$r>0, 2r+\gamma\geq 1$	$\Omega(\lvert\mathcal{D}\rvert^{\frac{1}{2r+\gamma}})$	$\boldsymbol{O}(\log\frac{1}{\epsilon})$	Theorem 3
Theorem 4	√	√	√	$\lvert\mathfrak{D}_j\rvert M^2+M^3+M^2t+mMt$	$\lvert\mathcal{D}_{\text{test}}\rvert M$	Mt	$r>0, 2r+\gamma\geq 1$	$\Omega(\lvert\mathcal{D}\rvert^{\frac{1}{2r+\gamma}})$	$\boldsymbol{O}(\log\frac{1}{\epsilon})$	Theorem 4

Note: The computational complexities are computed in terms of regularized least squared loss. We estimate the upper bounds for $\|f-f^*\|_2\ \forall f\in L^2(\mathbb{P})$. We denote $\mathcal{D}_{\text{test}}$ the testing data, η the step-size for SGD approaches, ϵ the federated error and $\Delta_f^2=\sum_{j=1}^m p_j\Delta_{f_j}^2$. For Rademacher complexities based bounds Zhang et al. (2015); Su et al. (2021), specific kernels include kernels with finite-rank or polynomial eigenvalues decay. Integral operator based bounds Guo et al. (2017); Lin & Cevher (2018); Lin et al. (2020) also assume $\gamma\in[0,1]$. We compute exact local solution for FedProx.

in feature space. Both our work and Guo et al. (2017); Lin & Cevher (2018); Lin et al. (2020) are based on integral operator techniques, but DKRR literature assumes all local datasets are drawn i.i.d. from an identical distribution, ignoring the data heterogeneity and model heterogeneity, which makes the proofs much easier than ours. We emphasize the difference between this work and DKRR theories as bellow: 1) DKRR work required a strict condition $r\in[1/2,1]$, while we relax the condition to $r>0, 2r+\gamma\geq 1$. 2) This work pertains to NonIID data, covering both covariate shift $\Delta_{\mathfrak{D}_j}$ and response shift Δ_{f_j}, DKRR only applied to IID data that is a special case in the homogenous setting $\Delta_{\mathfrak{D}_j}=\Delta_{f_j}=0$ in Theorem 3. 3) Because of the existence of data heterogeneity and model heterogeneity, we cannot directly estimate the difference between local estimators and global ones, and thus we introduce novel error decompositions for the federated error. 4) This work explores the excess risk bounds in terms of different local sample size ($\aleph_j$ in Theorem 3), covering both optimal and sub-optimal rates, while DKRR work only studied the optimal learning rates with the restrict on the number of partitions, i.e. $m=\boldsymbol{O}(\lvert\mathcal{D}\rvert^{\frac{(2r+\gamma-1)(t+1)}{(2r+\gamma)(t+2)}})$ Lin et al. (2020).

Compared with first-order methods. Using the random matrix theory and the local Rademacher complexity, Su et al. (2021) provided the optimal guarantees $\|f-f^*\|_2^2=\boldsymbol{O}(1/\lvert\mathcal{D}\rvert)$. However, as shown in Theorem 2 Su et al. (2021), it directly assumed all inputs are sampled i.i.d from an identical distribution, ignore the local sample size and the data heterogeneity, while our theoretical results illustrate both the local sample size and the data heterogeneity are crucial to federated learning. Su et al. (2021) also imposed several strict assumptions: 1) the ideal model belongs to the hypothesis space, corresponding to $r\in[1/2,1]$; 2) small hypothesis space with local Rademacher complexity, corresponding $\gamma\to 0$ in our work; 3) specific kernels maybe not suitable to the federated learning tasks and lead to sub-optimal rates. In this work, we remove these three conditions based on the integral operator approach, which makes our theoretical findings applicable to broader settings. Our results illustrate that only a few iterations can guarantee the optimal rates $\boldsymbol{O}(\lvert\mathcal{D}\rvert^{\frac{-2r}{2r+\gamma}})$ when the number of local examples is sufficient and data heterogeneity is moderate, where the convergence rate of federated error is *linear*, while in Su et al. (2021) the learning rate is always affected by model heterogeneity $\boldsymbol{O}(\frac{\sum_{j=1}^m p_j\Delta_{f_j}^2}{\lvert\mathcal{D}\rvert})$ and the convergence rate is *sublinear*.

6 Conclusion and Future Work

In this paper, we present an efficient second-order optimization method for FL. We derive generalization bounds with the optimal rates, which quantify the impacts of local sample size, the data heterogeneity, and the model heterogeneity. In benign cases, the federated error convergence exponentially fast, and thus communications can be small. Our theoretical findings fill the gap between optimization and generalization for federated learning, rather than focusing on one of them. Overall, the techniques presented here highlight new ways for designing efficient algorithms and analyzing both generalization and optimization for FL.

References

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Nicolò Dal Fabbro, Subhrakanti Dey, Michele Rossi, and Luca Schenato. Shed: A newton-type algorithm for federated learning based on incremental hessian eigenvector sharing. *Automatica*, 160:111460, 2024.

Ron S Dembo, Stanley C Eisenstat, and Trond Steihaug. Inexact newton methods. *SIAM Journal on Numerical analysis*, 19(2):400–408, 1982.

Junichi Fujii, Masatoshi Fujii, Takayuki Furuta, and Ritsuo Nakamoto. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118(3):827–830, 1993.

Avishek Ghosh, Raj Kumar Maity, and Arya Mazumdar. Distributed newton can communicate less and resist byzantine workers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, pp. 18028–18038, 2020.

Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 9050–9090. PMLR, 2022.

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

Vipul Gupta, Avishek Ghosh, Michal Derezinski, Rajiv Khanna, Kannan Ramchandran, and Michael Mahoney. Localnewton: Reducing communication bottleneck for distributed learning. *arXiv preprint arXiv:2105.07320*, 2021.

Rustem Islamov, Xun Qian, Slavomír Hanzely, Mher Safaryan, and Peter Richtárik. Distributed newton-type methods with communication compression and bernoulli aggregation. *Transactions on Machine Learning Research*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 8571–8580, 2018.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 5132–5143. PMLR, 2020.

Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 85, 2013.

Jian Li, Yong Liu, and Weiping Wang. Fedns: A fast sketching newton-type algorithm for federated learning. *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13509–13517, 2023.

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

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020 (MLSys)*, 2020b.

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3098–3107, 2018.

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research (JMLR)*, 21(147): 1–63, 2020.

Shao-Bo Lin, Di Wang, and Ding-Xuan Zhou. Distributed kernel ridge regression with communications. *Journal of Machine Learning Research (JMLR)*, 21(93):1–38, 2020.

Chengchang Liu, Lesi Chen, Luo Luo, and John CS Lui. Communication efficient distributed newton method with fast convergence rates. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1406–1416, 2023.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Yong Liu, Jiankun Liu, and Shuqiang Wang. Effective distributed learning with random features: Improved bounds and algorithms. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 4615–4625. PMLR, 2019.

Radford M Neal. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. PhD thesis, Citeseer, 1995.

Reese Pathak and Martin J Wainwright. Fedsplit: an algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, pp. 7057–7066, 2020.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

Xun Qian, Rustem Islamov, Mher Safaryan, and Peter Richtarik. Basis matters: Better communication-efficient second order methods for federated learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 680–720, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pp. 1177–1184, 2007.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2021–2031. PMLR, 2020.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research (JMLR)*, 11(Feb):905–934, 2010.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 3215–3225, 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 1657–1665, 2015.

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtarik. Fednl: Making newton-type methods applicable to federated learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 18959–19010. PMLR, 2022.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 31(9):3400–3413, 2019.

Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Lili Su, Jiaming Xu, and Pengkun Yang. A non-parametric view of fedavg and fedprox: Beyond stationary points. *arXiv preprint arXiv:2106.15216*, 2021.

Joel A Tropp. User-friendly tools for random matrices: An introduction. Technical report, 2012.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 2338–2348, 2018.

David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):1–8, 2022.

Semih Yagli, Alex Dytso, and H Vincent Poor. Information-theoretic bounds on the generalization error and privacy leakage in federated learning. In *Proceedings of the 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5. IEEE, 2020.

Jiyan Yang, Vikas Sindhwani, Haim Avron, and Michael Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 485–493, 2014.

Kai Yang, Tao Fan, Tianjian Chen, Yuanming Shi, and Qiang Yang. A quasi-newton method based vertical federated learning framework for logistic regression. *arXiv preprint arXiv:1912.00513*, 2019.

Yun Yang, Mert Pilanci, Martin J Wainwright, et al. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

Honglin Yuan, Warren Richard Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research (JMLR)*, 16(1):3299–3340, 2015.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Hanhan Zhou, Tian Lan, Guru Prasadh Venkataramani, and Wenbo Ding. Every parameter matters: Ensuring the convergence of federated learning with dynamic heterogeneous models reduction. *Advances in Neural Information Processing Systems*, 36, 2023.

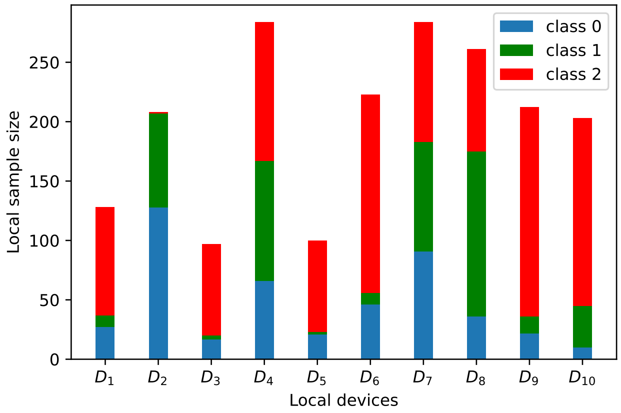

Figure 2: Data partitions for the dna dataset.

7 Experiments

In this section, we first carry out simulations to corroborate our theoretical statements. Then, we compare the performance of `FedNewton` with related baselines on real-world datasets.

Datasets

1) Synthetic dataset. Although the existing work Li et al. (2020a); Lin et al. (2020); Su et al. (2021) provide strategies to generate synthetic datasets, these datasets either fail to impose both data heterogeneity and model heterogeneity among devices, or just fit a simple linear problem. Here, we focus on a nonlinear problem $f^*(\boldsymbol{x}) = \min(-\mathbf{1}^\top \boldsymbol{x}, \mathbf{1}^\top \boldsymbol{x})$ with $\boldsymbol{x} \sim \mathcal{N}(0, \mathbf{I})$. On the j-th local machine, we generate $\mathfrak{D}_j = (\boldsymbol{X}_j, \boldsymbol{y}_j)$ based on $y = \min(-\boldsymbol{w}^\top \boldsymbol{x}, \boldsymbol{w}^\top \boldsymbol{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.2)$ is the label noise, $\boldsymbol{x}_j \sim \mathcal{N}(\boldsymbol{u}_j, \mathbf{I}), \boldsymbol{u}_j \sim \mathcal{N}(0, \alpha)$ and $\boldsymbol{w}_j \sim \mathcal{N}(\mathbf{1}, \boldsymbol{v}_j), \boldsymbol{v}_j \sim \mathcal{N}(0, \beta)$. Notably, α and β control the data heterogeneity and model heterogeneity, respectively. Data heterogeneity and model heterogeneity increase as α and β become larger, and the homogeneous setting corresponds to $\alpha = \beta = 0$. We set $d = 10$ and generate $|\mathcal{D}| = 10000$ samples for training, 2500 samples for testing.

2) Real-world datasets. We evaluate the compared algorithms on publicly available datasets from LIBSVM Data [1], which provide both training and testing data. To construct a heterogeneous and unbalanced setting, we split these datasets across 10 clients using a Dirichlet distribution $\text{Dir}_K(c)$ Wang et al. (2020), where c is some constant relevant to the level of heterogeneity and unbalanced distribution. For example, the data partition for the *dna* dataset with $\text{Dir}_K(1)$ is reported in Figure 2 where the local datasets are both heterogeneous and unbalanced, which is common in federated learning scenarios.

Experimental Settings

We compared the proposed `FedNewton` with the baseline (KRR on entire data), DKRR (`FedNewton` with $t = 0$), FedAvg McMahan et al. (2017) and FedProx Li et al. (2020a) with the squared loss equation 1. The estimator can be expressed as $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle$, where $\phi(\boldsymbol{x})$ denotes the feature mapping function. Here, we use random Fourier feature $\phi(\boldsymbol{x}) = 1/\sqrt{M} \cos(\boldsymbol{\Omega}^\top \boldsymbol{x} + \boldsymbol{b})$, where $\phi : \mathbb{R}^d \to \mathbb{R}^M, \boldsymbol{\Omega} \in \mathbb{R}^{d \times M}, \boldsymbol{b} \in \mathbb{R}^M$ and $\boldsymbol{\Omega} \sim \mathcal{N}(0, 1/\sigma^2), \boldsymbol{b} \in U[0, 2\pi]$. We set $M = 200$ for synthetic dataset and $M = 2000$ for real-world datasets. We implement all code based Pytorch and tune the hyperparameters over $\sigma^2 \in \{0.01, 0.1, \cdots, 1000\}$ and $\lambda = \{0.1, 0.01, \cdots, 10^{-7}\}$ by grid search. We report the data statistics and parameter setting in Table 2. All experiments are recorded by averaging results after 10 trials.

[1] Available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/`

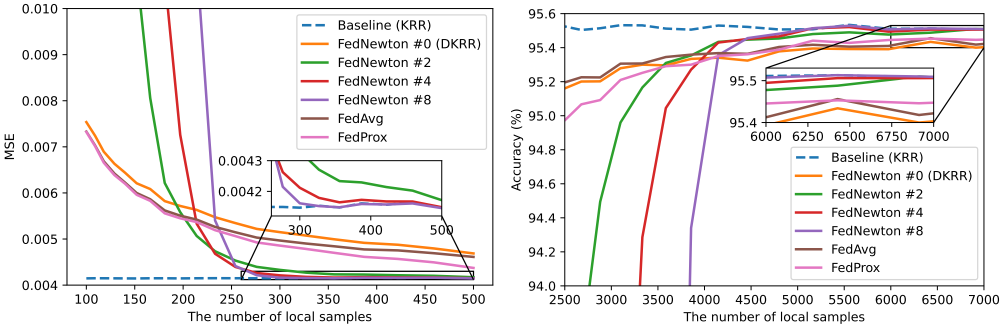

Figure 3: Impact of the number of local samples (left) on the synthetic dataset and MNIST dataset (right). The number of total training samples is fixed, $|\mathfrak{D}_j| = |\mathcal{D}|/m$ and $\Delta_{\mathfrak{D}_j} = \Delta_{f_j} = 0$. The blue dotted line denotes the exact KRR on all training data.

We initialize all iterative methods, including FedAvg, FedProx and `FedNewton`, by $\boldsymbol{w}^0_{\mathfrak{D}_j,\lambda} = \boldsymbol{H}^{-1}_{\mathfrak{D}_j,\lambda}\boldsymbol{\Phi}^\top_{\mathfrak{D}_j}\boldsymbol{y}_{\mathfrak{D}_j}$ rather than $\boldsymbol{w}^0_{\mathfrak{D}_j,\lambda} = \mathbf{0}$. DKRR directly averages the initialized models. FedAvg updates local models with $s = 2$ iterations on all local data in each epoch. In Section 7, we estimate the impact of local sample size, data heterogeneity without comparing FedAvg and FedProx. Here, we provide the full comparison with FedAvg and FedProx w.r.t. local sample size and data heterogeneity.

7.1 Empirical Validations

We verify the theoretical findings in theorems by exploring how the factors empirically affect the performance on a synthetic dataset that can capture both data heterogeneity and model heterogeneity and the MNIST dataset.

Impact of local sample size. We explore the influence of local sample size $|\mathfrak{D}_j|$ by fixing the total sample size $|\mathcal{D}| = 10000$ while varying the number m of local machines, where $|\mathfrak{D}_j| = \frac{|\mathcal{D}|}{m}$. As shown in the first two in Figure 3, when the number of local samples is small, i.e. $|\mathfrak{D}_j| < 200$ for the synthetic dataset and $|\mathfrak{D}_j| < 3300$ for MNIST, `FedNewton` with multiple communications hurts the generalization performance, and more communications lead to worse accuracy, corresponding to the cases $\Upsilon^t > 1$ in Theorem 3. When the local sample size is larger than a threshold, i.e. $|\mathfrak{D}_j| \approx 260$ for the synthetic dataset and $|\mathfrak{D}_j| \approx 4400$ for MNIST, more communications can significantly improve the predictive performance and get closer to the exact KRR, which coincides with the cases $\Upsilon^t < 1$ in Theorem 3. Note that, even with a large number of local examples, there still is a great gap between DKRR and KRR, while `FedNewton` achieves a good approximation to KRR. Meanwhile, both larger $|\mathfrak{D}_j|$ and larger t can improve the approximation ability, validating the theoretical results. Compared to first-order methods, when the local sample size is large enough, `FedNewton` outperforms FedAvg and FedProx. However, `FedNewton` is more sensitive to the number of local examples, and we find that the predictive error explodes when local sample size is small.

Impact of heterogeneous data. Let $m = 20$ and $|\mathfrak{D}_j| = 500$ for the synthetic dataset. We explore the impact of data heterogeneity by generating inputs with covariate shifts and explore the impact of model heterogeneity by generating outputs with response shifts. The right of Figure 3 illustrates: 1) Compared to DKRR, `FedNewton` remarkably reduce MSE when the heterogeneity is small. But it enlarges the errors from heterogeneous data when the heterogeneity is bigger than a threshold, i.e., $\Delta_{\mathfrak{D}_j} \approx 0.466$. 2) For the benign data heterogeneous settings, more communications for `FedNewton` lead to better approximation to the exact KRR, while the gap between DKRR and KRR still exists. 3) When data heterogeneity is large, `FedNewton` is more sensitive to data heterogeneity than DKRR, and more communications hurt the predictive accuracy. In the line of federated learning, the data heterogeneity is common due to different data distributions while the model heterogeneity is usually small. The left of Figure 4 shows that 1) Model heterogeneity decreases the predictive performance for all methods. 2) More communications lead to better approximation to

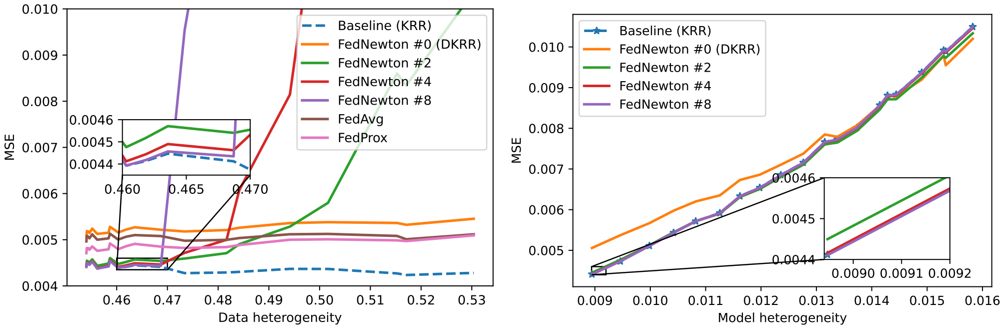

Figure 4: Impact of data heterogeneity (left) and model heterogeneity (right) on the synthetic dataset. We empirically estimate data heterogeneity by $\Delta_{\mathfrak{D}_j} = [\mathbf{\Phi}_{\mathcal{D}}^\top \mathbf{\Phi}_{\mathcal{D}} - \mathbf{\Phi}_{\mathfrak{D}_j}^\top \mathbf{\Phi}_{\mathfrak{D}_j}]$, and model heterogeneity by $\Delta_{f_j} = \frac{1}{|\mathfrak{D}_j|} \sum_{i=1}^{|\mathfrak{D}_j|} [f^*(\boldsymbol{x}_i) - f_j^*(\boldsymbol{x}_i)]$.

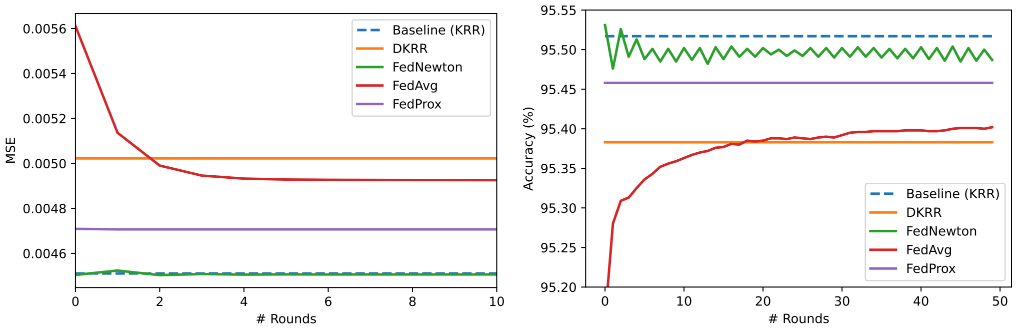

Figure 5: Predictive performance of `FedNewton`, FedAvg and FedProx on heterogeneous synthetic dataset (left) and MNIST (right).

KRR when model heterogeneity is small. 3) The performance of all methods is similarly poor when model heterogeneity is bigger than 0.0135 and all models finally get similar bad results when model heterogeneity is large enough. These observations coincide with Theorem 4.

Iterations of `FedNewton` and first-order methods. We use heterogeneous dataset for iterations, i.e. the synthetic dataset with $\alpha = 0.01$ and $\gamma = 0.001$ and the MNIST dataset partitioned by a Dirichlet distribution $\text{Dir}_K(0.5)$. The last two in Figure 4 reports the generalization performance on heterogeneous data in terms of the communication rounds. We find that: 1) With a few iterations, `FedNewton` converges to KRR on the entire data, outperforming the divide-and-conquer and first-order methods. 2) Since local models are initialized by the closed-form solutions, FedProx converges very fast $t = 1$ and then updates slowly. The performance of FedProx is better than DKRR and FedAvg but worse than `FedNewton`. 3) Compared to FedProx and `FedNewton`, the convergence of FedAvg is slow and achieves the performance between DKRR and FedProx.

7.2 Evaluation Results on Real Datasets

We compared related federated learning algorithms on both original datasets and non-iid datasets partitioned by a Dirichlet distribution $\text{Dir}_K(0.5)$. After partitioning with a Dirichlet distribution, the labels and the number of local samples on datasets are very unbalanced that decrease the generalization ability of federated learning algorithms. We report the classification accuracy in Table 3 for several public classification datasets, illustrating that:

1) The proposed `FedNewton` remarkably outperforms the compared methods on the original datasets, and more iterations improves the generalization performance. This observation coincides with results in Theorem 3.

Table 2: Data statistics and hyperparameter settings.

Dataset	Task	$\|\mathcal{D}\|$	d	classes	kernel parameter σ^2	λ	λ_{prox}	$\text{Dir}_K(\alpha)$	learning rate γ
synthetic	regression	10000	10	1	0.1	1e-06	7e-07	1	0.001
dna	multiclass	2000	180	3	0.001	1e-07	1e-08	1	0.001
letter	multiclass	15000	16	26	1	0.001	0.001	0.5	0.001
pendigits	multiclass	7494	16	10	0.01	0.0001	0.0001	1	0.0001
satimage	multiclass	4435	36	6	1	0.001	0.001	1	0.001
Sensorless	multiclass	58509	48	11	10	1e-06	1e-07	1	0.001
shuttle	multiclass	43500	48	11	10	0.001	0.001	0.5	0.001
usps	multiclass	7291	256	10	0.1	0.0001	0.0001	0.5	0.001
mnist	multiclass	60000	784	10	0.1	1e-05	7e-07	0.5	0.001

Table 3: Classification accuracy (%) for classification datasets. We bold the results with the best method and underline the ones that are not significantly worse than the best one.

Dataset	Compared methods			`FedNewton`			
	DKRR	FedAvg	FedProx	# 1	# 2	# 4	# 8
dna	90.91±0.50	91.09±0.42	89.42±6.98	**92.23±0.53**	91.96±0.48	92.02±0.40	88.19±11.58
letter	77.18±0.12	77.11±0.17	77.17±0.12	77.30±0.12	77.30±0.12	**77.30±0.12**	77.30±0.12
pendigits	97.12±0.09	97.12±0.11	97.12±0.10	97.29±0.13	**97.31±0.10**	97.31±0.11	97.23±0.31
satimage	87.70±0.17	87.84±0.08	87.74±0.11	**88.49±0.19**	88.26±0.17	88.31±0.14	88.31±0.15
Sensorless	96.81±0.12	96.87±0.14	96.84±0.13	**97.32±0.11**	96.87±0.14	96.44±0.17	84.43±1.20
shuttle	98.46±0.06	98.53±0.08	98.51±0.07	**98.54±0.07**	98.51±0.07	98.50±0.07	98.44±0.16
usps	92.95±0.10	92.95±0.12	92.95±0.12	**93.49±0.18**	93.24±0.13	93.28±0.14	93.30±0.15
mnist	95.38±0.12	95.40±0.13	95.46±0.11	**95.53±0.13**	95.48±0.13	95.49±0.13	95.48±0.12

2) The predictive accuracies of all federated learning methods in the heterogeneous setting are worse than ones in the original case, but `FedNewton` approaches still achieve the optimal results on the most datasets.

3) Similar to Figure 4, `FedNewton` with more iterations are more sensitive to the heterogeneity and more iterations hurts the generalization performance. The reason is the number of iterations augments the federated error when $\Upsilon > 1$ due to large data heterogeneity.

PROOFS

7.3 PRELIMINARIES

Since KRR has closed-form solutions, the intermediate estimators $\bar{f}^t_{\mathcal{D},\lambda}, f_{\mathcal{D},\lambda}, f_\lambda, f^*$ in error decomposition can be represented by the redirection operators and their adjoint operators. In this section, we first provide useful linear operators associated with kernel K. Then, we measure the similarities between empirical and expected covariance operators via concentration inequalities.

Definition 2 (Operators with kernel K in terms of the global distribution $\rho_{X\times Y}$)**.** *For any $\boldsymbol{x} \in \mathcal{X}, g \in L^2(\mathbb{P}), \phi : \mathcal{X} \to \mathcal{H}_K$ and $\boldsymbol{\beta} \in \mathcal{H}_K$, we define the following expected operators*

- $S : \mathcal{H}_K \to L^2(\mathbb{P}), \quad (S\boldsymbol{\beta})(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle.$
- $S^* : L^2(\mathbb{P}) \to \mathcal{H}_K, \quad S^* g = \int_X \phi(\boldsymbol{x}) g(\boldsymbol{x}) d\rho_X(\boldsymbol{x}).$
- $L : L^2(\mathbb{P}) \to L^2(\mathbb{P}), \quad L = SS^*, \quad \textit{such that } \ (Lg)(\cdot) = \int_X \langle \phi(\cdot), \phi(\boldsymbol{x}) \rangle g(\boldsymbol{x}) d\rho_X(\boldsymbol{x}).$
- $C : \mathcal{H}_K \to \mathcal{H}_K, \quad C = S^* S, \quad \textit{such that } \ C\boldsymbol{\beta} = \int_X \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle \phi(\boldsymbol{x}) d\rho_X(\boldsymbol{x}).$

Definition 3 (Empirical operators on the global dataset $\mathcal{D}$ and local datasets $\mathfrak{D}_j$)**.** *For any $\phi : \mathcal{X} \to \mathcal{H}_K$ and $\boldsymbol{\beta} \in \mathcal{H}_K$, we define the following empirical operators*

- $S_\mathcal{D} : \mathcal{H}_K \to \mathbb{R}^{|\mathcal{D}|}, \quad S_\mathcal{D}\boldsymbol{\beta} = \big(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}_i) \rangle\big)_{i=1}^{|\mathcal{D}|} \in \mathbb{R}^{|\mathcal{D}|}, \quad \forall\, (\boldsymbol{x}_i, y_i) \in \mathcal{D}.$

- $S^*_{\mathcal{D}} : \mathbb{R}^{|\mathcal{D}|} \to \mathcal{H}_K, \quad S^*_{\mathcal{D}}\alpha = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \phi(\boldsymbol{x}_i)\alpha_i \in \mathcal{H}_K, \quad \forall\, (\boldsymbol{x}_i, y_i) \in \mathcal{D},\ \alpha \in \mathbb{R}^{|\mathcal{D}|}.$
- $C_{\mathcal{D}} : \mathcal{H}_K \to \mathcal{H}_K, \quad C_{\mathcal{D}} = S^*_{\mathcal{D}} S_{\mathcal{D}}$, *such that* $C_{\mathcal{D}}\, \boldsymbol{\beta} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}_i)\rangle \phi(\boldsymbol{x}_i), \quad \forall\, (\boldsymbol{x}_i, y_i) \in \mathcal{D}.$
- $S_{\mathfrak{D}_j} : \mathcal{H}_K \to \mathbb{R}^{|\mathfrak{D}_j|}, \quad S_{\mathfrak{D}_j}\boldsymbol{\beta} = \big(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}_i)\rangle\big)_{i=1}^{|\mathfrak{D}_j|} \in \mathbb{R}^{|\mathfrak{D}_j|}, \quad \forall\, (\boldsymbol{x}_i, y_i) \in \mathfrak{D}_j.$
- $S^*_{\mathfrak{D}_j} : \mathbb{R}^{|\mathfrak{D}_j|} \to \mathcal{H}_K, \quad S^*_{\mathfrak{D}_j}\alpha = \frac{1}{|\mathfrak{D}_j|}\sum_{i=1}^{|\mathfrak{D}_j|} \phi(\boldsymbol{x}_i)\alpha_i \in \mathcal{H}_K, \quad \forall\, (\boldsymbol{x}_i, y_i) \in \mathfrak{D}_j.$
- $C_{\mathfrak{D}_j} : \mathcal{H}_K \to \mathcal{H}_K, \quad C_{\mathfrak{D}_j} = S^*_{\mathfrak{D}_j} S_{\mathfrak{D}_j},$ *such that* $C_{\mathfrak{D}_j}\ \boldsymbol{\beta} = \frac{1}{|\mathfrak{D}_j|}\sum_{i=1}^{|\mathfrak{D}_j|}\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}_i)\rangle \phi(\boldsymbol{x}_i), \quad \forall\, (\boldsymbol{x}_i, y_i) \in \mathfrak{D}_j.$

Here, we denote S the inclusion operator and $S_{\mathcal{D}}, S_{\mathfrak{D}_j}$ the sampling operator, while $S^*, S^*_{\mathcal{D}}, S^*_{\mathfrak{D}_j}$ are their adjoint operators. Note that $C : \mathcal{H}_K \to \mathcal{H}_K$ is the covariance operator given by S^*S, and the integral operator $L : L^2(\mathbb{P}) \to L^2(\mathbb{P})$ given by SS^*. The kernel matrix $\mathbf{K}_{\mathcal{D}}, \mathbf{K}_{\mathfrak{D}}$ and the covariance matrix $C_{\mathcal{D}}, C_{\mathfrak{D}_j}$ are the empirical counterparts of the integral operator L and the covariance operator C, respectively. Using Singular Value Decomposition shows that L and C have the same eigenvalues, and the corresponding eigenvectors are closely related Rosasco et al. (2010). Those kernels-related operators are widely used in the proof of optimal learning theory for standard KRR. Assuming the kernel is bounded $K(\boldsymbol{x}, \boldsymbol{x}') \leq \kappa^2$, the integral operator L and the covariance operator C are positive trace class operators (and hence compact) and bounded by $\|L\| = \|C\| \leq \kappa^2$. For any function $f \in \mathcal{H}_K$, the estimator $f \in L^2(\mathbb{P})$ is obtained by kernel trick. Thus, for $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle$, the RKHS norm can be related to the $L^2(\mathbb{P})$-norm by $C^{1/2}$ Bauer et al. (2007):

$$\|f\|_2 = \|Sf\|_2 = \|C^{1/2}\boldsymbol{w}\|_K, \quad \forall \boldsymbol{w} \in \mathcal{H}_K,\ f \in L^2(\mathbb{P}). \tag{8}$$

Remark 8. *With the assumption $K(\boldsymbol{x}, \boldsymbol{x}') \leq \kappa^2$, the integral operator L is trace class Caponnetto & De Vito (2007) and $C, C_{\mathcal{D}}, C_{\mathfrak{D}_j}$ are finite dimensional. Moreover we have that $L = SS^*$, $C = S^*S$, $C_{\mathcal{D}} = S^*_{\mathcal{D}} S_{\mathcal{D}}$ and $C_{\mathfrak{D}_j} = S^*_{\mathfrak{D}_j} S_{\mathfrak{D}_j}$. Finally $L, C, C_{\mathcal{D}}, C_{\mathfrak{D}_j}$ are self-adjoint and positive operators, with spectrum is $[0, \kappa^2]$.*

Proposition 2 (Cordes Inequality Fujii et al. (1993))**.** *Let A, B two positive semi-definite bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \leq \|AB\|^s, \qquad \textit{when} \quad 0 \leq s \leq 1.$$

Here, we use Proposition 2 to obtain the inequality $\|(A + \lambda I)^{-1/2}(B + \lambda)^{1/2}\| \leq \|(A + \lambda I)^{-1}(B + \lambda)\|^{1/2}$ for linear operators $C, C_j, C_{\mathcal{D}}, C_{\mathfrak{D}_j}$, and L.

Proposition 3 (Lemma 2 in Smale & Zhou (2007))**.** *Let $\mathcal{L}$ be a separable Hilbert space and $\{\xi_1, \cdots, \xi_n\}$ be a sequence of i.i.d random variables in $\mathcal{L}$. Assume the bound be $\|\xi_i\| \leq \widetilde{M} \leq \infty$ and the variance be $\tilde{\sigma}^2 = \mathbb{E}(\|\xi_i - \mathbb{E}(\xi_i)\|^2)$ for any $i \in [n]$. For any $\delta \in (0, 1)$, with confidence $1 - \delta$,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i - \mathbb{E}(\xi_i) \right\| \leq \frac{2\widetilde{M}\log(2/\delta)}{n} + \sqrt{\frac{2\tilde{\sigma}^2 \log(2/\delta)}{n}}. \tag{9}$$

The above Bernstein's inequality is the key to analyzing the relationship between the empirical random vector and its expected counterpart, which is used to prove Lemma 1. The above Bernstein's inequality for random vectors was provided in Smale & Zhou (2007); Rudi & Rosasco (2017) and later was extended to the random operator case in Theorem 7.3.1 in Tropp (2012) and Lemma 24 in Lin & Cevher (2020).

Lemma 1. *Given $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}')\rangle_K$, let $\phi(\cdot)$ be i.i.d random vectors on a separable Hilbert space $\mathcal{H}_K$ such that $C, C_{\mathcal{D}}, C_{\mathfrak{D}_j}$ are trace class. Then for any $\delta \in (0, 1)$ with the probability* *at least* $1-\delta$*, the following holds*

$$\left\|(C+\lambda I)^{-1/2}(C-C_{\mathcal{D}})(C+\lambda I)^{-1/2}\right\| \leq \mathcal{R}_{\mathcal{D},\lambda} \leq \frac{2\kappa^2\log(2/\delta)}{\lambda|\mathcal{D}|}+\sqrt{\frac{2(\kappa^2+1)\log(2/\delta)}{\lambda|\mathcal{D}|}},$$

$$\left\|(C_j+\lambda I)^{-1/2}(C_j-C_{\mathfrak{D}_j})(C_j+\lambda I)^{-1/2}\right\| \leq \mathcal{R}_{\mathfrak{D}_j,\lambda} \leq \frac{2\kappa^2\log(2/\delta)}{\lambda|\mathfrak{D}_j|}+\sqrt{\frac{2(\kappa^2+1)\log(2/\delta)}{\lambda|\mathfrak{D}_j|}}, \tag{10}$$

where $\mathcal{R}_{\mathcal{D},\lambda}=\left\|(C+\lambda I)^{-1}(C-C_{\mathcal{D}})\right\|$ *and* $\mathcal{R}_{\mathfrak{D}_j,\lambda}=\left\|(C_j+\lambda I)^{-1}(C_j-C_{\mathfrak{D}_j})\right\|$.

Proof. We first prove the lower bound for $\mathcal{R}_{\mathcal{D},\lambda}$. Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left\|(C+\lambda I)^{-1/2}(C-C_{\mathcal{D}})(C+\lambda I)^{-1/2}\right\|\\ =&\left\|(C+\lambda I)^{-1/2}(C-C_{\mathcal{D}})^{1/2}(C-C_{\mathcal{D}})^{1/2}(C+\lambda I)^{-1/2}\right\|\\ \leq&\left\|(C+\lambda I)^{-1/2}(C-C_{\mathcal{D}})^{1/2}\right\|^2.\end{aligned} \tag{11}$$

Recall that the norm on a matrix or operator A can be defined By

$$\|A\|:=\sup_x\frac{\|Ax\|_2}{\|x\|_2}.$$

For $K>1$ and a nonzero vector x, we get

$$\|A^kx\|_2=\|AA^{k-1}x\|_2\leq\|A\|\|A^{k-1}x\|_2\leq\cdots\leq\|A\|^k\|x\|_2.$$

Therefore, it holds $\frac{\|A^kx\|_2}{\|x\|_2}\leq\|A\|^k$ and thus

$$\|A^k\|=\sup_x\frac{\|A^kx\|_2}{\|x\|_2}\leq\|A\|^k. \tag{12}$$

Assuming $A=(C+\lambda I)^{-1/2}$ and substituting equation 12 to equation 11, we get

$$\left\|(C+\lambda I)^{-1/2}(C-C_{\mathcal{D}})(C+\lambda I)^{-1/2}\right\|\leq\left\|(C+\lambda I)^{-1}(C-C_{\mathcal{D}})\right\|=\mathcal{R}_{\mathcal{D},\lambda}.$$

Then, we prove the upper bound for $\mathcal{R}_{\mathcal{D},\lambda}$. Let $\xi=(C+\lambda I)^{-1}\phi(\boldsymbol{x})\otimes\phi(\boldsymbol{x})$, thus we have

$$\mathbb{E}(\xi)=(C+\lambda I)^{-1}\mathbb{E}[\phi(\boldsymbol{x})\otimes\phi(\boldsymbol{x})]=(C+\lambda I)^{-1}C,$$

$$\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\xi_i=\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(C+\lambda I)^{-1}[\phi(\boldsymbol{x}_i)\otimes\phi(\boldsymbol{x}_i)]=(C+\lambda I)^{-1}C_{\mathcal{D}}.$$

The left of the desired inequality becomes

$$\left\|(C+\lambda I)^{-1}(C-C_{\mathcal{D}})\right\|=\left\|\mathbb{E}(\xi)-\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\xi_i\right\|.$$

Note that

$$\|(C+\lambda I)^{-1/2}\phi(\boldsymbol{x})\|^2\leq\kappa^2\lambda^{-1}.$$

To use Bernstein's inequality (Proposition 3), we need to bound $\|\xi\|$ and $\mathbb{E}\|\xi\|^2$ as follows

$$\begin{aligned}\|\xi\|&=\|\langle(C+\lambda I)^{-1}\phi(\boldsymbol{x}),\phi(\boldsymbol{x})\rangle\|\leq\|(C+\lambda I)^{-1/2}\phi(\boldsymbol{x})\|^2\leq\kappa^2\lambda^{-1}.\\ \mathbb{E}\|\xi-\mathbb{E}(\xi)\|^2&=\left\|\mathbb{E}\left[\langle(C+\lambda I)^{-1}\phi(\boldsymbol{x}),\phi(\boldsymbol{x})\rangle(C+\lambda I)^{-1}\phi(\boldsymbol{x})\otimes\phi(\boldsymbol{x})\right]-C_\lambda^{-2}C^2\right\|\\ &\leq\kappa^2\lambda^{-1}\left\|\mathbb{E}\left[(C+\lambda I)^{-1}\phi(\boldsymbol{x})\otimes\phi(\boldsymbol{x})\right]\right\|+\left\|C_\lambda^{-2}C^2\right\|\\ &\leq\kappa^2\lambda^{-1}\|C_\lambda^{-1}C\|+1\leq\kappa^2\lambda^{-1}+1\leq(\kappa^2+1)\lambda^{-1}.\end{aligned}$$

Substituting the above identities to Bernstein's inequality equation 9, we obtain the upper bound for $\mathcal{R}_{\mathcal{D},\lambda}$.

The lower and upper bounds can be proven with similar proof techniques. □

Lemma 2 (Proposition 8 Rudi & Rosasco (2017))**.** *Let $\lambda > 0$. We define the following quantities*

$$\mathcal{P}_{\mathcal{D},\lambda} := \left\|(C_{\mathcal{D}} + \lambda I)^{-1}(C + \lambda I)\right\|, \quad \mathcal{P}_{\mathfrak{D}_j,\lambda} := \left\|(C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_j + \lambda I)\right\|.$$

Then, there exists the following properties

$$\mathcal{P}_{\mathcal{D},\lambda} \leq \frac{1}{1-\beta}, \quad \mathcal{P}_{\mathfrak{D}_j,\lambda} \leq \frac{1}{1-\beta},$$

with

$$\beta = \lambda_{max}\left[(C + \lambda I)^{-1/2}(C - C_{\mathcal{D}})(C + \lambda I)^{-1/2}\right].$$

Note that, $\beta \leq \frac{\lambda_{max}(C)}{\lambda_{max}+\lambda} < 1$.

7.4 Error Decomposition for FedNewton

For Newton-based federated learning, there holds the following error decompositions

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\| \leq \|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\| + \|f_{\mathcal{D},\lambda} - f^*\|. \tag{13}$$

Here, the federated error term $\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|$ is also the key to analyzing the generalization of second-order optimization based federated learning `FedNewton`.

Proof of Theorem 1. For any function $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle_K$, the $\mathcal{H}_K$-norm can be related to the $L^2(\mathbb{P})$-norm by the inclusion S Bauer et al. (2007)

$$\|f\|_2 = \|S\boldsymbol{w}\|_K = \|S(C + \lambda I)^{-1/2}(C + \lambda I)^{1/2}\boldsymbol{w}\|_K \leq \|(C + \lambda I)^{1/2}\boldsymbol{w}\|_K.$$

Therefore, one can prove

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \leq \|(C + \lambda I)^{1/2}(\bar{\boldsymbol{w}}^t_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda})\|_K. \tag{14}$$

From equation 5, we have

$$\begin{aligned}
\bar{\boldsymbol{w}}^t_{\mathcal{D},\lambda} &= \bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} - \sum_{j=1}^m p_j \boldsymbol{H}^{-1}_{\mathfrak{D}_j,\lambda}\boldsymbol{g}^{t-1}_{\mathcal{D},\lambda} \\
&= \bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} - \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}\left[(C_{\mathcal{D}} + \lambda I)\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} - S^*_{\mathcal{D}}\boldsymbol{y}_{\mathcal{D}}\right] \\
&= \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathfrak{D}_j} - C_{\mathcal{D}})\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} + \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}S^*_{\mathcal{D}}\boldsymbol{y}_{\mathcal{D}} \\
&= \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathfrak{D}_j} - C_{\mathcal{D}})\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} + \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathcal{D}} + \lambda I)\boldsymbol{w}_{\mathcal{D},\lambda}.
\end{aligned}$$

And then, one can obtain

$$\begin{aligned}
&\bar{\boldsymbol{w}}^t_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda} \\
=&\sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathfrak{D}_j} - C_{\mathcal{D}})\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} + \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathcal{D}} + \lambda I)\boldsymbol{w}_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda} \\
=&\sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathfrak{D}_j} - C_{\mathcal{D}})\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} + \sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathcal{D}} - C_{\mathfrak{D}_j})\boldsymbol{w}_{\mathcal{D},\lambda} \\
=&\sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1}(C_{\mathfrak{D}_j} - C_{\mathcal{D}})(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda}).
\end{aligned}$$

We then estimate the federated error by

$$
\begin{aligned}
&(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\\
=&\sum_{j=1}^{m} p_j (C+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_{\mathfrak{D}_j}-C_{\mathcal{D}})(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\\
=&\sum_{j=1}^{m} p_j (C+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_{\mathfrak{D}_j}-C_j+C_j-C+C-C_{\mathcal{D}})(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\\
=&\sum_{j=1}^{m} p_j (C+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}\\
&\quad(C_j+\lambda I)^{-1/2}(C_{\mathfrak{D}_j}-C_j)(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C+\lambda I)^{-1/2}(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\\
&+\sum_{j=1}^{m} p_j (C+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}\\
&\quad(C_j+\lambda I)^{-1/2}(C_j-C)(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C+\lambda I)^{-1/2}(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\\
&+\sum_{j=1}^{m} p_j (C+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}\\
&\quad(C_j+\lambda I)^{-1/2}(C+\lambda I)^{1/2}(C+\lambda I)^{-1/2}(C-C_{\mathcal{D}})(C+\lambda I)^{-1/2}(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda}).
\end{aligned}
\tag{15}
$$

Note that, $\|(C+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}\| \leq \|I+(C_j+\lambda I)^{-1}(C-C_j)\|^{1/2} \leq \sqrt{1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}}$, $\|(C_j+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}\| \leq \mathcal{P}_{\mathfrak{D}_j,\lambda}$, $\|(C_j+\lambda I)^{1/2}(C+\lambda I)^{-1/2}\| \leq \|I+(C+\lambda I)^{-1}(C_j-C)\|^{1/2} \leq \sqrt{1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}}$. Therefore, substituting these inequalities to equation 15 and from equation 14, there exists

$$
\begin{aligned}
&\|\bar{f}^{t}_{\mathcal{D},\lambda}-f_{\mathcal{D},\lambda}\|_2\\
\leq&\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\|_K\\
\leq&\sum_{j=1}^{m} p_j\left(1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\mathcal{P}_{\mathfrak{D}_j,\lambda}\mathcal{R}_{\mathfrak{D}_j,\lambda}\left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K\\
&+\sum_{j=1}^{m} p_j\left(1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\mathcal{P}_{\mathfrak{D}_j,\lambda}\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K\\
&+\sum_{j=1}^{m} p_j\left(1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\mathcal{P}_{\mathfrak{D}_j,\lambda}\mathcal{R}_{\mathcal{D},\lambda}\left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K\\
\leq&\sum_{j=1}^{m} p_j\mathcal{P}_{\mathfrak{D}_j,\lambda}\left(1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left(\mathcal{R}_{\mathcal{D},\lambda}+\mathcal{R}_{\mathfrak{D}_j,\lambda}+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{t-1}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K\\
\leq&\left(\sum_{j=1}^{m} p_j\mathcal{P}_{\mathfrak{D}_j,\lambda}\left(1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left(2\mathcal{R}_{\mathfrak{D}_j,\lambda}+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\right)^{t}\left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{0}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K.
\end{aligned}
\tag{16}
$$

Note that, $\mathcal{R}_{\mathcal{D},\lambda} \propto 1/|\mathcal{D}|$ and thus $\mathcal{R}_{\mathfrak{D}_j,\lambda} \leq \mathcal{R}_{\mathcal{D},\lambda}$. Combing the above inequality and equation 13, we prove the final result. □

Proposition 4. *The following federated error bounds hold for oneshot federated learning:*

$$
\begin{aligned}
&\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^{0}_{\mathcal{D},\lambda}-\boldsymbol{w}_{\mathcal{D},\lambda})\|_K\\
\leq&\mathcal{P}_{\mathcal{D},\lambda}\sum_{j=1}^{m} p_j\left(2\mathcal{R}_{\mathfrak{D}_j,\lambda}+\frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_{\lambda})\right\|_K,
\end{aligned}
\tag{17}
$$

where $\Delta_{\mathfrak{D}_j} = \|C_j - C\|$.

Proof. Note that, if A, B are invertible operators on a Banach space, then there holds the equality

$$A^{-1} - B^{-1} = B^{-1}(B-A)A^{-1} = A^{-1}(B-A)B^{-1}.$$

From equation 2, using the facts $S_{\mathcal{D}}^* \boldsymbol{y}_{\mathcal{D}} = \sum_{j=1}^m p_j S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j}$ and $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$, we have

$$\begin{aligned}
&\bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^0 - \boldsymbol{w}_{\mathcal{D},\lambda} \\
=&\sum_{j=1}^m p_j (\boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{\Phi}_{\mathfrak{D}_j} + \lambda I)^{-1} \boldsymbol{\Phi}_{\mathfrak{D}_j}^\top \boldsymbol{y}_{\mathfrak{D}_j} - (\boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{\Phi}_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\Phi}_{\mathcal{D}}^\top \boldsymbol{y}_{\mathcal{D}} \\
=&\sum_{j=1}^m p_j (C_{\mathfrak{D}_j} + \lambda I)^{-1} S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - (C_{\mathcal{D}} + \lambda I)^{-1} S_{\mathcal{D}}^* \boldsymbol{y}_{\mathcal{D}} \\
=&\sum_{j=1}^m p_j [(C_{\mathfrak{D}_j} + \lambda I)^{-1} - (C_{\mathcal{D}} + \lambda I)^{-1}] S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} \\
=&\sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C_{\mathfrak{D}_j}) \boldsymbol{w}_{\mathfrak{D}_j,\lambda} \\
=&\sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C) \boldsymbol{w}_{\mathfrak{D}_j,\lambda} + \sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C - C_{\mathfrak{D}_j}) \boldsymbol{w}_{\mathfrak{D}_j,\lambda} \\
=&\sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C)(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda) + \sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C) \boldsymbol{w}_\lambda \\
&\quad + \sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C - C_{\mathfrak{D}_j}) \boldsymbol{w}_{\mathfrak{D}_j,\lambda} \\
=&\sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C)(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda) + \sum_{j=1}^m p_j (C_{\mathcal{D}} + \lambda I)^{-1} (C - C_{\mathfrak{D}_j})(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda).
\end{aligned} \tag{18}$$

The last step is due to the fact $\sum_{j=1}^m p_j C_{\mathcal{D}} = \sum_{j=1}^m p_j C_{\mathfrak{D}_j}$.

Combining equation 14 and equation 18, we have

$$\begin{aligned}
\|\bar{f}_{\mathcal{D},\lambda}^0 - f_{\mathcal{D},\lambda}\|_2 \leq& \|(C + \lambda I)^{1/2} (\bar{\boldsymbol{w}}_{\mathcal{D},\lambda}^0 - \boldsymbol{w}_{\mathcal{D},\lambda})\|_K \\
\leq& \left\| \sum_{j=1}^m p_j (C + \lambda I)^{1/2} (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C + C - C_{\mathfrak{D}_j})(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda) \right\|.
\end{aligned} \tag{19}$$

Note that

$$\begin{aligned}
&(C + \lambda I)^{1/2} (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C) \\
=&(C + \lambda I)^{1/2} (C_{\mathcal{D}} + \lambda I)^{-1/2} (C_{\mathcal{D}} + \lambda I)^{-1/2} (C + \lambda I)^{1/2} (C + \lambda I)^{-1/2} (C_{\mathcal{D}} - C)(C + \lambda I)^{-1/2} (C + \lambda I)^{1/2}.
\end{aligned}$$

Using the inequality $\|(C + \lambda I)^{-1/2} (C_{\mathcal{D}} - C)(C + \lambda I)^{-1/2}\| \leq \mathcal{R}_{\mathcal{D},\lambda}$ from Lemma 1, we have

$$\begin{aligned}
&\|(C + \lambda I)^{1/2} (C_{\mathcal{D}} + \lambda I)^{-1} (C_{\mathcal{D}} - C)(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda)\| \\
\leq& \mathcal{P}_{\mathcal{D},\lambda} \|(C + \lambda I)^{-1/2} (C_{\mathcal{D}} - C)(C + \lambda I)^{-1/2} (C + \lambda I)^{1/2} (\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda)\| \\
\leq& \mathcal{P}_{\mathcal{D},\lambda} \mathcal{R}_{\mathcal{D},\lambda} \|(C + \lambda I)^{1/2} (\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda)\|.
\end{aligned} \tag{20}$$

Similarly, we have

$$
\begin{aligned}
&(C+\lambda I)^{1/2}(C_{\mathcal{D}}+\lambda I)^{-1}(C-C_{\mathfrak{D}_j})\\
=&(C+\lambda I)^{1/2}(C_{\mathcal{D}}+\lambda I)^{-1}(C-C_j+C_j-C_{\mathfrak{D}_j})\\
=&(C+\lambda I)^{1/2}(C_{\mathcal{D}}+\lambda I)^{-1}(C+\lambda I)^{1/2}(C+\lambda I)^{-1/2}(C-C_j)(C+\lambda I)^{-1/2}(C+\lambda I)^{1/2}\\
&+(C+\lambda I)^{1/2}(C_{\mathcal{D}}+\lambda I)^{-1}(C+\lambda I)^{1/2}(C+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}\\
&(C_j+\lambda I)^{-1/2}(C_j-C_{\mathfrak{D}_j})(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C+\lambda I)^{-1/2}(C+\lambda I)^{1/2}.
\end{aligned}
$$

Using $\|(C+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}\| \leq \|I+(C+\lambda I)^{-1}(C_j-C)\|^{1/2} \leq 1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}$, it holds

$$
\begin{aligned}
&\|(C+\lambda I)^{1/2}(C_{\mathcal{D}}+\lambda I)^{-1}(C-C_{\mathfrak{D}_j})(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda)\|\\
\leq&\frac{\mathcal{P}_{\mathcal{D},\lambda}\Delta_{\mathfrak{D}_j}}{\lambda}\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda)\|+\mathcal{P}_{\mathcal{D},\lambda}\mathcal{R}_{\mathfrak{D}_j,\lambda}\left(1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda)\|\\
\leq&\mathcal{P}_{\mathcal{D},\lambda}\left(\mathcal{R}_{\mathfrak{D}_j,\lambda}+\frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right)\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda)\|.
\end{aligned}
\tag{21}
$$

Therefore, substituting equation 20 and equation 21 to equation 19, we have

$$
\begin{aligned}
\|\bar{f}^0_{\mathcal{D},\lambda}-f_{\mathcal{D},\lambda}\| \leq& \sum_{j=1}^m p_j\mathcal{P}_{\mathcal{D},\lambda}\left(\mathcal{R}_{\mathcal{D},\lambda}+\mathcal{R}_{\mathfrak{D}_j,\lambda}+\frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right)\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda)\|\\
\leq&\mathcal{P}_{\mathcal{D},\lambda}\sum_{j=1}^m p_j\left(2\mathcal{R}_{\mathfrak{D}_j,\lambda}+\frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right)\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda)\|.
\end{aligned}
$$

□

7.5 Estimating Error Terms

7.5.1 Estimating Federated Error

From Lemma 1, Lemma 4, and equation 13, there are two error terms $\|\boldsymbol{w}_{\mathfrak{D}_j,\lambda}-\boldsymbol{w}_\lambda\|_K$ and $\|f_{\mathcal{D}_j,\lambda}-f_\lambda\|_2$ in federated error to be bounded. Using Bennett's inequality (Proposition 3), we first provide two useful lemmas.

Lemma 3. *Assume there exists $\kappa \geq 1$ such that $\|\phi(\boldsymbol{x})\|_K \leq \kappa$, $\forall \boldsymbol{x} \in \mathcal{X}$ and $|y| \leq B$. For $\delta \in (0,1]$, the following holds with the probability at least $1-\delta$*

$$
\begin{aligned}
&\|(C+\lambda I)^{-1/2}(S^*_{\mathcal{D}}\boldsymbol{y}_{\mathcal{D}}-S^*f^*)\| \leq 2B\kappa\mathcal{A}_{\mathcal{D},\lambda}\log\frac{2}{\delta},\\
&\|(C_j+\lambda I)^{-1/2}(S^*_{\mathfrak{D}_j}\boldsymbol{y}_{\mathfrak{D}_j}-S^*_jf^*_j)\| \leq 2B\kappa\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta}.
\end{aligned}
$$

where C_j, S^*_j are operators defined on the local distribution ρ_j, and

$$
\mathcal{A}_{\mathcal{D},\lambda} := \frac{1}{|\mathcal{D}|\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{|\mathcal{D}|}}, \quad \mathcal{A}_{\mathfrak{D}_j,\lambda} := \frac{1}{|\mathfrak{D}_j|\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{|\mathfrak{D}_j|}}. \tag{22}
$$

Proof. Let $\xi_i=(C+\lambda I)^{-1/2}\phi(\boldsymbol{x}_i)y_i$ in the Hilbert space $\mathcal{H}_K$. We see that

$$
\begin{aligned}
\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\xi_i&=\frac{1}{n}\sum_{i=1}^n(C+\lambda I)^{-1/2}\phi(\boldsymbol{x}_i)y_i=(C+\lambda I)^{-1/2}S_{\mathcal{D}}\boldsymbol{y}_{\mathcal{D}},\\
\mathbb{E}\,\xi&=\int_X(C+\lambda I)^{-1/2}\phi(\boldsymbol{x})f^*(\boldsymbol{x})d\rho_X(\boldsymbol{x})=(C+\lambda I)^{-1/2}S^*f^*
\end{aligned}
$$

Thus, the error term to bound can be stated as

$$\|(C+\lambda I)^{-1/2}(\widehat{S}_n^* \boldsymbol{y}_{\mathcal{D}} - S^* f^*)\| = \left\| \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \xi_i - \mathbb{E}\xi_i \right\|. \tag{23}$$

The rhs of the above identity can be bounded by Bennett's inequality (Proposition 3), thus we need to estimate $\|\xi_i - \mathbb{E}(\xi_i)\|$ and $\mathbb{E}\,\|\xi_i - \mathbb{E}(\xi_i)\|^2$ first.

We first recall the definition of effective dimension

$$\mathcal{N}(\lambda) = \mathbb{E}\,\langle \phi(\boldsymbol{x}), (C+\lambda I)^{-1}\phi(\boldsymbol{x})\rangle_K = \int_X \|(C+\lambda I)^{-1}\phi(\boldsymbol{x})\|_K^2 \, d\rho_X(\boldsymbol{x}).$$

By Jensen's inequality, we thus have

$$\|\xi_i - \mathbb{E}(\xi_i)\| \le \|(C+\lambda I)^{-1/2}\phi(\boldsymbol{x}_i)\||y_i| + \mathbb{E}\|(C+\lambda I)^{-1/2}\phi(\boldsymbol{x}_i)\||y_i| \le 2B\kappa\lambda^{-1/2}. \tag{24}$$

Note that

$$\begin{aligned}\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2 &\le 2\int_X \|(C+\lambda I)^{-1/2}\phi(\boldsymbol{x}_i)\|^2 |y_i|^2 d\rho_X(\boldsymbol{x}) \\ &\le 2B^2 \int_X \|(C+\lambda I)^{-1/2}\phi(\boldsymbol{x}_i)\|^2 d\rho_X(\boldsymbol{x}) \le 2B^2\mathcal{N}(\lambda).\end{aligned} \tag{25}$$

Substituting equation 24 and equation 25 to equation 23, by Bennett's inequality (Proposition 3), we have

$$\|(C+\lambda I)^{-1/2}(S_{\mathcal{D}}^* \boldsymbol{y}_{\mathcal{D}} - S^* f^*)\| \le \frac{2B\kappa \log(2/\delta)}{|\mathcal{D}|\sqrt{\lambda}} + 2\sqrt{\frac{B^2\mathcal{N}(\lambda)\log(2/\delta)}{|\mathcal{D}|}}.$$

Similarly, we derive the bound for $\|(C_j+\lambda I)^{-1/2}(S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - S_{\mathfrak{D}_j}^* f_j^*)\|$. Thus, we prove the result. □

Lemma 4 (From Theoreom 4 of Caponnetto & De Vito (2007))**.** *Assume there exists $\kappa \ge 1$ such that $\|\phi(\boldsymbol{x})\|_K \le \kappa$, $\forall \boldsymbol{x} \in \mathcal{X}$. For $\delta \in (0,1]$, the following holds with the probability at least $1-\delta$*

$$\begin{aligned}\|(C+\lambda I)^{-1/2}(C - C_{\mathcal{D}})\| &\le 2\kappa(\kappa+1)\mathcal{A}_{\mathcal{D},\lambda}\log\frac{2}{\delta}, \\ \|(C_j+\lambda I)^{-1/2}(C_j - C_{\mathfrak{D}_j})\| &\le 2\kappa(\kappa+1)\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta}.\end{aligned}$$

The above lemma is a standard method for the difference between expected and empirical covariance operators $C - C_{\mathcal{D}}$ and $C_j - C_{\mathfrak{D}_j}$. Using a concentration inequality in Hilbert spaces, it have been proven in Caponnetto & De Vito (2007); Smale & Zhou (2007); Guo et al. (2017).

We define the expected estimators for local machines and centralized model as

$$\begin{aligned}\boldsymbol{w}_{j,\lambda} &= \underset{\boldsymbol{w}\in\mathcal{H}_K}{\arg\min}\left\{\int_X (\langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle - f^*(\boldsymbol{x}))^2 d\rho_j(\boldsymbol{x}) + \lambda\|\boldsymbol{w}\|_K^2\right\} \\ \boldsymbol{w}_{\lambda} &= \underset{\boldsymbol{w}\in\mathcal{H}_K}{\arg\min}\left\{\int_X (\langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle - f^*(\boldsymbol{x}))^2 d\rho_X(\boldsymbol{x}) + \lambda\|\boldsymbol{w}\|_K^2\right\}.\end{aligned}$$

Proposition 5. *Assume $\|\phi(\boldsymbol{x})\|_K \le \kappa$ and $|y| \le B$. Under Assumption 2, for $\delta \in (0,1/2)$, the following bound hold with the probability at least $1-2\delta$*

$$\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_{\lambda})\| \le C_1\sqrt{1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}}\mathcal{P}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta} + \frac{\kappa^2 R\Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}. \tag{26}$$

where $C_1 = 2\kappa(B + 2\kappa^3 R)$.

Proof. We introduce the intermediate estimators $\boldsymbol{w}_{j,\lambda} = (C_j + \lambda I)^{-1} S_j^* f_j^*$, where S_j^* and C_j are operators defined on the local distribution ρ_j. Then, it holds

$$\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda)\| \leq \|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_{j,\lambda})\| + \|(C+\lambda I)^{1/2}(\boldsymbol{w}_{j,\lambda} - \boldsymbol{w}_\lambda)\| \tag{27}$$

where $\|\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_{j,\lambda}\|$ is the local variance and Δ_{f_j} is the model heterogeneity.

$$\begin{aligned}
&(C+\lambda I)^{1/2}(\boldsymbol{w}_{j,\lambda} - \boldsymbol{w}_\lambda)\\
=&(C+\lambda I)^{1/2}\left[(C_j+\lambda I)^{-1} S_j^* f_j^* - (C+\lambda I)^{-1} S^* f^*\right]\\
=&(C+\lambda I)^{1/2}\left[(C_j+\lambda I)^{-1} S_j^* f_j^* - (C+\lambda I)^{-1} S^* f_j^* + (C+\lambda I)^{-1} S^* f_j^* - (C+\lambda I)^{-1} S^* f^*\right]\\
=&(C+\lambda I)^{-1/2} S^*(L - L_j)(L_j+\lambda I)^{-1} f_j^* + (C+\lambda I)^{-1/2} S^*(f_j^* - f^*)\\
=&(C+\lambda I)^{-1/2} S^*(L - L_j)(L_j+\lambda I)^{-1} L^r L^{-r} f_j^* + (C+\lambda I)^{-1/2} S^*(f_j^* - f^*).
\end{aligned}$$

Since $\|(C+\lambda I)^{-1/2} S^*\| \leq 1$, $\|L\| \leq \kappa^2$, $\|C - C_j\| = \|L - L_j\|$ and $\Delta_{f_j} = \|f_j^* - f^*\|$, from Assumption 2, we have

$$\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{j,\lambda} - \boldsymbol{w}_\lambda)\| \leq \frac{\kappa^{2r} R \Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}. \tag{28}$$

We then decompose the local variance

$$\begin{aligned}
&\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_{j,\lambda}\\
=&(C_{\mathfrak{D}_j}+\lambda I)^{-1} S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - (C_{\mathfrak{D}_j}+\lambda I)^{-1} S_j^* f_j^* + (C_{\mathfrak{D}_j}+\lambda I)^{-1} S_j^* f_j^* - (C_j+\lambda I)^{-1} S_j^* f_j^*\\
=&(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}(S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - S_j^* f_j^*) + [(C_{\mathfrak{D}_j}+\lambda I)^{-1} - (C_j+\lambda I)^{-1}] S_j^* f_j^*\\
=&(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}(S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - S_j^* f_j^*) + (C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j - C_{\mathfrak{D}_j})\boldsymbol{w}_{j,\lambda}\\
=&(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}\left[(C_j+\lambda I)^{-1/2}(S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - S_j^* f_j^*) + (C_j+\lambda I)^{-1/2}(C_j - C_{\mathfrak{D}_j})\boldsymbol{w}_{j,\lambda}\right].
\end{aligned}$$

and it holds

$$\begin{aligned}
&(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_{j,\lambda})\\
=&(C+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1}(C_j+\lambda I)^{1/2}\Big[(C_j+\lambda I)^{-1/2}(S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - S_j^* f_j^*)\\
&\quad + (C_j+\lambda I)^{-1/2}(C_j - C_{\mathfrak{D}_j})\boldsymbol{w}_{j,\lambda}\Big]\\
=&(C+\lambda I)^{1/2}(C_j+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1/2}(C_{\mathfrak{D}_j}+\lambda I)^{-1/2}(C_j+\lambda I)^{1/2}\\
&\Big[(C_j+\lambda I)^{-1/2}(S_{\mathfrak{D}_j}^* \boldsymbol{y}_{\mathfrak{D}_j} - S_j^* f_j^*) + (C_j+\lambda I)^{-1/2}(C_j - C_{\mathfrak{D}_j})\boldsymbol{w}_{j,\lambda}\Big].
\end{aligned} \tag{29}$$

Due to Assumption 2 and $\|L_j\| \leq \kappa^2$, we obtain

$$\|\boldsymbol{w}_{j,\lambda}\|_K = \|(L_j+\lambda I)^{-1} L_j f_j^*\| = \|(L_j+\lambda I)^{-1} L_j L_j^r L_j^{-r} f_j^*\| \leq \kappa^{2r}\|L^{-r} f_j^*\| \leq \kappa^{2r} R. \tag{30}$$

Thus, substituting equation 30 to equation 29, using Lemma 3 and Lemma 4, for any $\delta \in (0, 1/2)$, we have with the probability $1 - 2\delta$

$$\begin{aligned}
&\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_{j,\lambda})\|\\
\leq&\mathcal{P}_{\mathfrak{D}_j,\lambda}\sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}}\left(2B\kappa\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta} + 2\kappa(\kappa+1)\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta}\kappa^{2r}R\right)\\
\leq&2\kappa(B + 2\kappa^3 R)\sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}}\mathcal{P}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta}.
\end{aligned} \tag{31}$$

Applying equation 28 and equation 31 to equation 27, we prove the result.

□

Theorem 5 (Detailed version of Theorem 2)**.** *For any $\delta \in (0,1)$, under Assumption 2, with the probability at least $1-\delta$, the federated error holds*

$$\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \leq C_2 \Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right) \left(\mathcal{A}_{\mathfrak{D}_j,\lambda} \log\frac{2}{\delta} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}\right). \tag{32}$$

where $C_2 = 2\kappa(B + 2\kappa^3 R)/(1-\beta)$, $\beta = \lambda_{max}((C+\lambda)^{-1/2}(C - C_{\mathcal{D}})(C+\lambda)^{-1/2})$ and $\mathcal{A}_{\mathfrak{D}_j,\lambda} = \frac{1}{\sqrt{\lambda}|\mathfrak{D}_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|\mathfrak{D}_j|}}$.

Proof. Substituting equation 26 and equation 17 to Theorem 1, with the probability $1-2\delta$, we obtain the federated error

$$\begin{aligned}
&\|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \\
\leq& \Upsilon^t \left\|(C+\lambda I)^{1/2}(\bar{\boldsymbol{w}}^0_{\mathcal{D},\lambda} - \boldsymbol{w}_{\mathcal{D},\lambda})\right\|_K \\
\leq& \Upsilon^t \sum_{j=1}^m p_j \mathcal{P}_{\mathcal{D},\lambda} \left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right) \left\|(C+\lambda I)^{1/2}(\boldsymbol{w}_{\mathfrak{D}_j,\lambda} - \boldsymbol{w}_\lambda)\right\|_K \\
\leq& \Upsilon^t \sum_{j=1}^m p_j \mathcal{P}_{\mathcal{D},\lambda} \left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right) \left(C_1 \sqrt{1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \mathcal{P}_{\mathfrak{D}_j,\lambda} \mathcal{A}_{\mathfrak{D}_j,\lambda} \log\frac{2}{\delta} + \frac{\kappa^2 R \Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}\right) \\
\leq& \Upsilon^t \sum_{j=1}^m \frac{C_1 p_j}{1-\beta} \sqrt{1+\frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right) \left(\mathcal{A}_{\mathfrak{D}_j,\lambda} \log\frac{2}{\delta} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}\right).
\end{aligned} \tag{33}$$

The last step is due to Lemma 2. □

7.5.2 Estimating Centralized Excess Risk

The generalization analysis for the centralized model (the exact KRR) is standard Caponnetto & De Vito (2007); Smale & Zhou (2007), but the existing work imposed a strict assumption $r \in [1/2, 1]$ on the kernel space, which assumes the ideal estimator belongs to the kernel space $f^* \in \mathcal{H}_K$. Here, we relax this strict assumption to $r > 0$ but still obtain the identical optimal learning rates for the centralized excess risk bounds.

Proposition 6. *Under Assumption 2, for $\delta \in (0, 1/2)$, the following bounds hold with the probability at least $1-2\delta$*

$$\|f_{\mathcal{D},\lambda} - f^*\|_2 \leq C_1 \mathcal{P}^{1/2}_{\mathcal{D},\lambda} \mathcal{A}_{\mathcal{D},\lambda} \log\frac{2}{\delta} + R\lambda^r, \tag{34}$$

where $C_1 = 2\kappa\left(B + 2\kappa^3 R\right)$.

Proof. The excess risk term can be divided into two parts: variance and bias.

$$\|f_{\mathcal{D},\lambda} - f^*\| \leq \|f_{\mathcal{D},\lambda} - f_\lambda\| + \|f_\lambda - f^*\|. \tag{35}$$

Using Cauchy's inequality, Lemma 3 and Lemma 4, for $\delta \in (0, 1/2)$, with the probability at least $1 - 2\delta$ we have

$$\begin{aligned}
&\|f_{\mathcal{D},\lambda} - f_\lambda\|_2 \\
=&\|S(C_{\mathcal{D}} + \lambda I)^{-1} S_{\mathcal{D}}^* \boldsymbol{y}_{\mathcal{D}} - S(C_{\mathcal{D}} + \lambda I)^{-1} S^* f^* + S(C_{\mathcal{D}} + \lambda I)^{-1} S^* f^* - S(C + \lambda I)^{-1} S^* f^*\|_2 \\
=&\|S(C_{\mathcal{D}} + \lambda I)^{-1}(C + \lambda I)(C + \lambda I)^{-1/2}(C + \lambda I)^{-1/2}(S_{\mathcal{D}}^* \boldsymbol{y}_{\mathcal{D}} - S^* f^*) \\
&\quad + S(C_{\mathcal{D}} + \lambda I)^{-1}(C + \lambda I)(C + \lambda I)^{-1/2}(C + \lambda I)^{-1/2}(C - C_{\mathcal{D}})(C + \lambda I)^{-1} S^* f^*\|_2 \\
=&\|S(C_{\mathcal{D}} + \lambda I)^{-1/2}(C_{\mathcal{D}} + \lambda I)^{-1/2}(C + \lambda I)^{1/2}(C + \lambda I)^{-1/2}(S_{\mathcal{D}}^* \boldsymbol{y}_{\mathcal{D}} - S^* f^*) \\
&\quad + S(C_{\mathcal{D}} + \lambda I)^{-1/2}(C_{\mathcal{D}} + \lambda I)^{-1/2}(C + \lambda I)^{1/2}(C + \lambda I)^{-1/2}(C - C_{\mathcal{D}})(C + \lambda I)^{-1} S^* f^*\|_2 \\
\leq& 2B\kappa \log\frac{2}{\delta}\, \mathcal{P}_{\mathcal{D},\lambda}^{1/2} \mathcal{A}_{\mathcal{D},\lambda} + 2\kappa(\kappa + 1)\log\frac{2}{\delta} \mathcal{P}_{\mathcal{D},\lambda}^{1/2} \mathcal{A}_{\mathcal{D},\lambda} \|\boldsymbol{w}_\lambda\|_K \\
\leq& 2\kappa\left(B + 2\kappa^3 R\right)\log\frac{2}{\delta}\, \mathcal{P}_{\mathcal{D},\lambda}^{1/2} \mathcal{A}_{\mathcal{D},\lambda}.
\end{aligned} \tag{36}$$

The last step is due $\|\boldsymbol{w}_\lambda\|_K = \|(L + \lambda I)^{-1} L f^*\| = \|(L + \lambda I)^{-1} L L^r L^{-r} f^*\| \leq \kappa^{2r} R$ due to Assumption 2.

The identity $A(A + \lambda I)^{-1} = I - \lambda(A + \lambda I)^{-1}$ holds for $\lambda > 0$ and A the bounded self-adjoint positive operator. Then, under Assumption 2, it holds

$$\begin{aligned}
&\|f_\lambda - f^*\|_2 \\
=&\|(L + \lambda I)^{-1} L f^* - f^*\| = \|((L + \lambda I)^{-1} L - I) f^*\| = \|\lambda (L + \lambda I)^{-1} f^*\| \\
=&\|\lambda^r \lambda^{1-r}(L + \lambda I)^{-(1-r)}(L + \lambda I)^{-r} L^r L^{-r} f^*\| \\
\leq& \lambda^r \|\lambda^{1-r}(L + \lambda I)^{-(1-r)}\| \|(L + \lambda I)^{-r} L^r\| \|L^{-r} f^*\| \\
\leq& R\lambda^r.
\end{aligned} \tag{37}$$

Substituting equation 36 and equation 37 to equation 35, we prove the result. □

7.6 Excess Risk Bounds for FedNewton

Proof of Theorem 3. In the homogeneous setting, we have $\Delta_{\mathfrak{D}_j} = 0$ and $\Delta_{f_j} = 0$. Thus, under Assumption 2, from equation 32 and equation 34, it holds

$$\begin{aligned}
\|\bar{f}_{\mathcal{D},\lambda}^t - f^*\|_2 \leq& \|\bar{f}_{\mathcal{D},\lambda}^t - f_{\mathcal{D},\lambda}\|_2 + \|f_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \\
\leq& \boldsymbol{O}\left(\Upsilon^t \sum_{j=1}^m p_j \mathcal{R}_{\mathfrak{D}_j,\lambda} \mathcal{A}_{\mathfrak{D}_j,\lambda} \log\frac{2}{\delta} + \mathcal{A}_{\mathcal{D},\lambda} \log\frac{2}{\delta} + R\lambda^r\right).
\end{aligned} \tag{38}$$

If $|\mathfrak{D}_j| > 29(\kappa^2 + 1)\log(1/\delta)/\lambda$, we have $\Upsilon < 1$. Otherwise, $\Upsilon \geq 1$.

From equation 10 and equation 22, under Assumption 1, with the probability at least $1 - 3\delta$, we have

$$\begin{aligned}
&\mathcal{R}_{\mathfrak{D}_j,\lambda} \mathcal{A}_{\mathfrak{D}_j,\lambda} \\
=&\boldsymbol{O}\left(\left(\frac{1}{\lambda|\mathfrak{D}_j|} + \sqrt{\frac{1}{\lambda|\mathfrak{D}_j|}}\right)\log\frac{2}{\delta} \times \left(\frac{1}{|\mathfrak{D}_j|\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{|\mathfrak{D}_j|}}\right)\right) \\
=&\boldsymbol{O}\left((|\mathfrak{D}_j|^{-2}\lambda^{-1.5} + |\mathfrak{D}_j|^{-1.5}\lambda^{-1-0.5\gamma} + |\mathfrak{D}_j|^{-1.5}\lambda^{-1} + |\mathfrak{D}_j|^{-1}\lambda^{-0.5-0.5\gamma})\log\frac{2}{\delta}\right) \\
=&\boldsymbol{O}\left((|\mathfrak{D}_j|^{-2}\lambda^{-1.5} + |\mathfrak{D}_j|^{-1.5}\lambda^{-1-0.5\gamma} + |\mathfrak{D}_j|^{-1}\lambda^{-0.5-0.5\gamma})\log\frac{2}{\delta}\right).
\end{aligned}$$

The relationships between λ and $|\mathfrak{D}_j|$ affects the value of $\mathcal{R}_{\mathfrak{D}_j,\lambda} \mathcal{A}_{\mathfrak{D}_j,\lambda}$.

$$\mathcal{R}_{\mathfrak{D}_j,\lambda} \mathcal{A}_{\mathfrak{D}_j,\lambda} = \log\frac{2}{\delta} \begin{cases} \boldsymbol{O}(|\mathfrak{D}_j|^{-2}\lambda^{-1.5}), & \text{if } \lambda < \boldsymbol{O}(|\mathfrak{D}_j|^{\frac{1}{\gamma-1}}). \\ \boldsymbol{O}(|\mathfrak{D}_j|^{-1.5}\lambda^{-1-0.5\gamma}), & \text{if } \Omega(|\mathfrak{D}_j|^{\frac{1}{\gamma-1}}) \leq \lambda < \boldsymbol{O}(|\mathfrak{D}_j|^{-1}). \\ \boldsymbol{O}(|\mathfrak{D}_j|^{-1}\lambda^{-0.5-0.5\gamma}), & \text{if } \lambda \geq \Omega(|\mathfrak{D}_j|^{-1}). \end{cases}$$

By setting $\lambda = |\mathcal{D}|^{\frac{-1}{2r+\gamma}}$ and $2r + \gamma \geq 1$, we have

$$\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda} = \log\frac{2}{\delta}\begin{cases} \boldsymbol{O}\left(|\mathfrak{D}_j|^{-2}|\mathcal{D}|^{\frac{1.5}{2r+\gamma}}\right), & \text{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}}. \\ \boldsymbol{O}\left(|\mathfrak{D}_j|^{-1.5}|\mathcal{D}|^{\frac{1+0.5\gamma}{2r+\gamma}}\right), & \text{if } |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}}. \\ \boldsymbol{O}\left(|\mathfrak{D}_j|^{-1}|\mathcal{D}|^{\frac{1+\gamma}{4r+2\gamma}}\right), & \text{if } |\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{1}{2r+\gamma}}. \end{cases} \tag{39}$$

and

$$\mathcal{A}_{\mathcal{D},\lambda} = |\mathcal{D}|^{\frac{1-4r-2\gamma}{4r+\gamma}} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}} \leq 2|\mathcal{D}|^{\frac{-r}{2r+\gamma}}. \tag{40}$$

Substituting equation 39 and equation 40 to equation 38, we have

$$\begin{aligned} &\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \\ \lesssim &|\mathcal{D}|^{\frac{-r}{2r+\gamma}}\log\frac{2}{\delta} + \Upsilon^t\log^2\frac{2}{\delta}\sum_{j=1}^m p_j \begin{cases} |\mathfrak{D}_j|^{-2}|\mathcal{D}|^{\frac{1.5}{2r+\gamma}}, & \text{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \\ |\mathfrak{D}_j|^{-1.5}|\mathcal{D}|^{\frac{1+0.5\gamma}{2r+\gamma}}, & \text{if } |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\ |\mathfrak{D}_j|^{-1}|\mathcal{D}|^{\frac{1+\gamma}{4r+2\gamma}}, & \text{if } |\mathcal{D}|^{\frac{1}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}} \\ |\mathcal{D}|^{\frac{-r}{2r+\gamma}}, & \text{if } |\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}} \end{cases} \end{aligned} \tag{41}$$

where

$$\begin{cases} t = 0, \Upsilon \geq 1, & \text{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\ t > 0, \Upsilon^t \lesssim \left(\frac{|\mathcal{D}|^{\frac{1}{2r+\gamma}}}{|\mathfrak{D}_j|}\right)^{0.5t}, & \text{otherwise.} \end{cases} \tag{42}$$

Note that, $\Upsilon = 2\sum_{j=1}^m p_j\mathcal{P}_{\mathfrak{D}_j,\lambda}\mathcal{R}_{\mathfrak{D}_j,\lambda} \lesssim \sum_{j=1}^m p_j\mathcal{R}_{\mathfrak{D}_j,\lambda}$. When $|\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}}$, we thus have $\mathcal{R}_{\mathfrak{D}_j,\lambda} \lesssim \sqrt{\frac{1}{\lambda|\mathfrak{D}_j|}} \lesssim |\mathcal{D}|^{\frac{1-2r-\gamma}{8r+4\gamma}}$. □

Proof of Theorem 4. Under Assumption 2, from equation 32 and equation 34, it holds

$$\begin{aligned} &\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \leq \|\bar{f}^t_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 + \|f_{\mathcal{D},\lambda} - f_{\mathcal{D},\lambda}\|_2 \\ \leq &\boldsymbol{O}\left(\Upsilon^t\sum_{j=1}^m p_j\sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}}\left(2\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{(1+\mathcal{R}_{\mathfrak{D}_j,\lambda})\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left(\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}\right) + \mathcal{A}_{\mathcal{D},\lambda}\log\frac{2}{\delta} + R\lambda^r\right). \end{aligned}$$

Let $\lambda = |\mathcal{D}|^{\frac{-1}{2r+\gamma}}$ and $2r + \gamma \geq 1$. When $|\mathfrak{D}_j| \leq \boldsymbol{O}(|\mathcal{D}|^{\frac{1}{2r+\gamma}})$, we have $\frac{1}{\lambda|\mathfrak{D}_j|} \geq \sqrt{\frac{1}{\lambda|\mathfrak{D}_j|}} \geq 1$ and $\mathcal{R}_{\mathfrak{D}_j,\lambda} \lesssim \frac{1}{\lambda|\mathfrak{D}_j|} + \sqrt{\frac{1}{\lambda|\mathfrak{D}_j|}} \lesssim \frac{1}{\lambda|\mathfrak{D}_j|}$ from equation 10. Thus,

$$\begin{aligned} &\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \\ \leq &\boldsymbol{O}\left(\Upsilon^t\sum_{j=1}^m p_j\left(1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)^{1.5}\left(\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta} + \frac{\mathcal{R}_{\mathfrak{D}_j,\lambda}\Delta_{\mathfrak{D}_j}}{\lambda} + \mathcal{R}_{\mathfrak{D}_j,\lambda}\Delta_{f_j}\right) + \mathcal{A}_{\mathcal{D},\lambda}\log\frac{2}{\delta} + R\lambda^r\right) \\ \leq &\boldsymbol{O}\left(\Upsilon^t\sum_{j=1}^m p_j\left(1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)^{1.5}\left(\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda}\log\frac{2}{\delta} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda^2|\mathfrak{D}_j|} + \frac{\Delta_{f_j}}{\lambda|\mathfrak{D}_j|}\right) + \mathcal{A}_{\mathcal{D},\lambda}\log\frac{2}{\delta} + R\lambda^r\right) \\ \leq &\boldsymbol{O}\left(\Upsilon^t\sum_{j=1}^m p_j\left(1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)^{1.5}\left(\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda} + \frac{|\mathcal{D}|^{\frac{2}{2r+\gamma}}}{|\mathfrak{D}_j|}\Delta_{\mathfrak{D}_j} + \frac{|\mathcal{D}|^{\frac{1}{2r+\gamma}}}{|\mathfrak{D}_j|}\Delta_{f_j}\right)\log\frac{2}{\delta} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}}\right). \end{aligned}$$

When $|\mathfrak{D}_j| \geq \Omega(|\mathcal{D}|^{\frac{1}{2r+\gamma}})$, we have $\mathcal{R}_{\mathfrak{D}_j,\lambda} \lesssim \sqrt{\frac{1}{\lambda|\mathfrak{D}_j|}} \leq 1$, $\mathcal{A}_{\mathfrak{D}_j,\lambda} \lesssim |\mathfrak{D}_j|^{-1/2}|\mathcal{D}|^{\frac{\gamma/2}{2r+\gamma}}$ and

$$
\begin{aligned}
&\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \\
\leq &\boldsymbol{O}\left(\Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \left(\mathcal{R}_{\mathfrak{D}_j,\lambda} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}\right)\left(\mathcal{A}_{\mathfrak{D}_j,\lambda} \log\frac{2}{\delta} + \frac{\Delta_{\mathfrak{D}_j}}{\lambda} + \Delta_{f_j}\right) + |\mathcal{D}|^{\frac{-r}{2r+\gamma}}\right) \\
\leq &\boldsymbol{O}\Bigg(\Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \Bigg(\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda} + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j} + \Delta_{f_j} + \frac{|\mathcal{D}|^{\frac{\gamma+2}{4r+2\gamma}}}{\sqrt{|\mathfrak{D}_j|}}\Delta_{\mathfrak{D}_j} + |\mathcal{D}|^{\frac{2}{2r+\gamma}}\Delta^2_{\mathfrak{D}_j} \\
&\quad + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j}\Delta_{f_j}\Bigg)\log\frac{2}{\delta} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}}\Bigg) \\
\leq &\boldsymbol{O}\Bigg(\Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \Bigg(\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda} + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j} + \Delta_{f_j} + |\mathcal{D}|^{\frac{2}{2r+\gamma}}\Delta^2_{\mathfrak{D}_j} + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j}\Delta_{f_j}\Bigg)\log\frac{2}{\delta} \\
&\quad + |\mathcal{D}|^{\frac{-r}{2r+\gamma}}\Bigg) \\
\leq &\boldsymbol{O}\Bigg(\Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}} \Big(\mathcal{R}_{\mathfrak{D}_j,\lambda}\mathcal{A}_{\mathfrak{D}_j,\lambda} + (1 + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j})(|\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j} + \Delta_{f_j})\Big)\log\frac{2}{\delta} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}}\Bigg).
\end{aligned}
$$

Combing with equation 39, we complete the proof

$$
\|\bar{f}^t_{\mathcal{D},\lambda} - f^*\|_2 \lesssim \Upsilon^t \sum_{j=1}^m p_j \sqrt{1 + \frac{\Delta_{\mathfrak{D}_j}}{\lambda}} (\aleph_j + \mathit{\Pi}_j)\ \log^2\frac{2}{\delta} + |\mathcal{D}|^{\frac{-r}{2r+\gamma}} \log\frac{2}{\delta}.
$$

Here, $\aleph_j$ and $\mathit{\Pi}_j$ have different values w.r.t local sample size

$$
\aleph_j = \begin{cases}
|\mathfrak{D}_j|^{-2}|\mathcal{D}|^{\frac{1.5}{2r+\gamma}}, & \text{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \\
|\mathfrak{D}_j|^{-1.5}|\mathcal{D}|^{\frac{1+0.5\gamma}{2r+\gamma}}, & \text{if } |\mathcal{D}|^{\frac{1-\gamma}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\
|\mathfrak{D}_j|^{-1}|\mathcal{D}|^{\frac{1+\gamma}{4r+2\gamma}}, & \text{if } |\mathcal{D}|^{\frac{1}{2r+\gamma}} \lesssim |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}} \\
|\mathcal{D}|^{\frac{-r}{2r+\gamma}}, & \text{if } |\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{2r+\gamma+1}{4r+2\gamma}},
\end{cases}
$$

and

$$
\mathit{\Pi}_j = \begin{cases}
\frac{|\mathcal{D}|^{\frac{2}{2r+\gamma}}}{|\mathfrak{D}_j|}\Delta_{\mathfrak{D}_j} + \frac{|\mathcal{D}|^{\frac{1}{2r+\gamma}}}{|\mathfrak{D}_j|}\Delta_{f_j}, & \text{if } |\mathfrak{D}_j| \lesssim |\mathcal{D}|^{\frac{1}{2r+\gamma}} \\
(1 + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j})(\Delta_{f_j} + |\mathcal{D}|^{\frac{1}{2r+\gamma}}\Delta_{\mathfrak{D}_j}), & \text{if } |\mathfrak{D}_j| \gtrsim |\mathcal{D}|^{\frac{1}{2r+\gamma}}.
\end{cases}
$$

□