Towards Understanding Ensemble Distillation in Federated Learning

Sejun Park [1] **Kihun Hong** [1] **Ganguk Hwang** [1]

Abstract

Federated Learning (FL) is a collaborative machine learning paradigm for data privacy preservation. Recently, a knowledge distillation (KD) based information sharing approach in FL, which conducts ensemble distillation on an unlabeled public dataset, has been proposed. However, despite its experimental success and usefulness, the theoretical analysis of the KD based approach has not been satisfactorily conducted. In this work, we build a theoretical foundation of the ensemble distillation framework in federated learning from the perspective of kernel ridge regression (KRR). In this end, we propose a KD based FL algorithm for KRR models which is related with some existing KD based FL algorithms, and analyze our algorithm theoretically. We show that our algorithm makes local prediction models as much powerful as the centralized KRR model (which is a KRR model trained by all of local datasets) in terms of the convergence rate of the generalization error if the unlabeled public dataset is sufficiently large. We also provide experimental results to verify our theoretical results on ensemble distillation in federated learning.

1. Introduction

Despite the rapid development of machine learning algorithms (Bochkovskiy et al., 2020; Nichol et al., 2022; Brown et al., 2020), the performance of machines still heavily depends on the size of the training dataset. However, due to the hassle of data processing, manpower and time resources are excessively consumed to obtain data, especially labeled data, in supervised learning. Moreover, most of data is inaccessible for training a prediction model due to data privacy preservation in general.

[1]Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. Correspondence to: Ganguk Hwang <guhwang@kaist.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Recently, Federated Learning (FL) (McMahan et al., 2017) has been proposed as a solution to resolve data shortage and data privacy preservation. Even though various FL algorithms (McMahan et al., 2017; Li et al., 2020a; Yurochkin et al., 2019; Wang et al., 2020; Karimireddy et al., 2020; Li et al., 2021) have been proposed, most of them usually iteratively conduct the following procedure in training. (i) A server distributes a global model to clients for updating their local models and then each client trains its local model using its local dataset. (ii) The clients send their trained local models (or gradients) to the server and the server aggregates them to update the global model. Such approaches obviously improve the performance of the local models while protecting data privacy, but sometimes performance improvement is not significantly sufficient to be applied for real world problems. Moreover, they have some restrictions in local training. For example, most of FL algorithms require that all local models be the same model (McMahan et al., 2017; Li et al., 2020a), parts of a global model (Arivazhagan et al., 2019; Jiang et al., 2022), approximations of a global model (Diao et al., 2021; Yao et al., 2021), or derived models from a meta model (Fallah et al., 2020; Shamsian et al., 2021). Such restrictions do not matter in a server-centric (i.e. server provider-centric) learning framework. However, in a client-centric training environment where most training strategies are led by individual clients, they severely matter. Note that local model architectures may also need to be protected as local data.

To overcome the above restrictions, knowledge distillation (KD) (Hinton et al., 2015) based FL algorithms have been proposed (Mora et al., 2022). Some of them improve the effectiveness of the FL algorithms by conducting knowledge distillation in addition to the existing FL algorithms (Lin et al., 2020b; Zhu et al., 2021). Some others use knowledge distillation as a main information sharing method, which can be applied to a client-centric training environment, instead of parameter sharing (Li & Wang, 2019; Cho et al., 2021; Zhang et al., 2021). In these methods, they assume that there exists an unlabeled public dataset or an unlabeled data generator in a server or a shared data storage that all clients can access. So, the knowledge of all local models can be obtained in the server via the predictions of the local models on the unlabeled public dataset. Then the server combines the knowledge through a weighted average

of the predictions and distributes the public dataset with the averaged predictions to clients. Finally, clients use the distributed dataset as well as their local datasets for their local training.

There are some intuitions as to why this ensemble distillation strategy works well. First, the predictions of local models on the unlabeled public dataset contain the information of local models like the parameters of local models in typical FL algorithms. Second, the ensemble distillation process is a kind of automatic crowdsourcing data labeling technique to provide additional data to each client. Therefore, this strategy directly alleviates the data shortage problem of clients. Despite these intuitions, KD based FL algorithms do not have sufficient theoretical analysis even for the independent and identically distributed (IID) case (i.e., all data points of clients are independent and identically distributed) unlike typical federated learning algorithms such as FedAvg and FedProx (Li et al., 2020a;b; Yuan & Li, 2022).

In this work, we provide a theoretical analysis for the effectiveness of ensemble distillation in federated learning. Even though knowledge distillation and ensemble techniques are difficult to be analyzed, some recent works provide analytical results for them in the context of kernel ridge regression (KRR) (Zhang et al., 2013; Lin et al., 2017; 2020a; Mobahi et al., 2020; Afonin & Karimireddy, 2022). Extending the existing results, we provide a theoretical framework for ensemble distillation of KRR models in general federated learning setting. Unlike the theoretical framework of typical FL algorithms which follows the optimization approach, we leverage the generalization theory of statistical models in this work. More precisely, we derive the convergence rate of the expected risk with respect to the dataset size.

Our contributions in this work are as follows:

1. First, we analyze knowledge distillation with an auxiliary dataset in terms of the convergence rate of the expected risk (Section 4). As an application of this result, we verify the effectiveness of one-shot ensemble distillation in federated learning with kernel ridge regression (Section 5.2).

2. Second, we propose and analyze a new iterative ensemble distillation algorithm like FedMD (Li & Wang, 2019) in federated learning with kernel ridge regression (Section 5.3 - 5.4). In the proposed algorithm, to overcome the undesirable amplified regularization in the repeated distillation procedures, we introduce a de-regularization trick that leads to the effectiveness of the proposed iterative ensemble distillation algorithm (Section 5.3).

3. Third, we analyze how a random client selection strategy, which is considered to resolve the communication cost or system heterogeneity issue, affects the performance of the proposed iterative ensemble distillation algorithm (Section 5.5).

4. Lastly, we provide experimental results to validate our theoretical results for the proposed iterative ensemble distillation in federated learning (Section 6).

2. Related Works

Federated Learning: FL is a privacy-preserving decentralized machine learning framework which is proposed in McMahan et al. (2017). FL has several challenges to address such as statistical heterogeneity (Li et al., 2020a; Karimireddy et al., 2020), system (e.g., computational capability) heterogeneity (Nishio & Yonetani, 2019), communication efficiency (Konečnỳ et al., 2016; Caldas et al., 2018), and privacy concerns (Bagdasaryan et al., 2020). Note that there are theoretical results that the classic FL algorithms such as FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020a) converge for both IID and non-IID cases (Li et al., 2020a;b; Yuan & Li, 2022).

Knowledge Distillation: KD is originally proposed in Buciluǎ et al. (2006); Hinton et al. (2015) in the context of model compression. The original KD encourages the outputs of a student model to match the outputs of a pre-trained teacher model. There are some variants of the original student-teacher framework. Self-distillation utilizes the student model itself as a teacher (Zhang et al., 2019). Ensemble distillation (or co-distillation) uses an ensemble of models as a teacher (Hinton et al., 2015; Anil et al., 2018; Lin et al., 2020b). Since it is an inclusive concept, we refer to a knowledge distillation framework with several students and a teacher which is an ensemble of (some of) them as ensemble distillation like Lin et al. (2020b) in this work.

Knowledge Distillation in FL: There are two lines of research mainly in federated learning with knowledge distillation. One line of research aims at improving efficiency or effectiveness of federated learning algorithms by applying knowledge distillation (Lin et al., 2020b; He et al., 2020; Zhu et al., 2021; Zhang et al., 2022a;b). In particular, Lin et al. (2020b) improve the performance by applying additional ensemble distillation on a server to train a server model rather than directly replacing it with an aggregated model of client models. They also propose an extended version for heterogeneous model cases. However, it allows only a few prototypes and clients should share their models with the server. The other line of research aims at enabling to use the black box models in FL (Li & Wang, 2019; Cho et al., 2021; Zhang et al., 2021). Li & Wang (2019) first propose the framework for FL with black box models. Subsequently, Cho et al. (2021); Zhang et al. (2021) apply advanced ensemble strategies to improve the performance in non-IID

setting. However, to the best of our knowledge, there is no previous work that provides the theoretical analysis of KD in FL. Recently, Allen-Zhu & Li (2020) provide a great theoretical analysis to explain why ensemble knowledge distillation works on neural networks with the same architectures. However, we cannot directly apply their results to ensemble distillation in FL since they assume all neural networks use the same dataset.

Theoretical Analysis of Kernel Ridge Regression: There are many prior works that analyze the convergence rate of kernel ridge regression in expected risk sense or in probability sense thanks to the closed form expression of its solution. The convergence properties for the classical KRR model have been analyzed well in Caponnetto & De Vito (2007); Fischer & Steinwart (2020); Cui et al. (2021); Li et al. (2023). Recently, Zhang et al. (2013); Lin et al. (2017); Chang et al. (2017); Guo et al. (2017); Lin et al. (2020a); Yin et al. (2021) also analyze distributed learning with KRR models in various contexts. In particular, Lin et al. (2020a) propose a communication scheme for distributed learning. However, this approach does not preserve data privacy due to the nature of the form of KRR solutions. On the other hand, Mobahi et al. (2020); Borup & Andersen (2021) apply KRR to study self-distillation. To the best of our knowledge, Afonin & Karimireddy (2022) is the only preceding study that utilizes KRR to provide a theoretical analysis of distillation strategies in federated learning setting. However, they mainly consider the case of two clients and their strategy is an ensemble of infinite models, which is impractical. In our work, we analyze standard ensemble distillation methods for the KRR model with arbitrary number of clients.

3. Backgrounds

In this section, we briefly introduce some backgrounds for our work. See Appendix A for details and comments.

3.1. Preliminaries and Notations

Our interest is a regression problem in FL setting. Let $\mathcal{X} \subset \mathbb{R}^d$ be an input space. We assume $\mathcal{X}$ is compact. Our goal is to find a target function $f_0 : \mathcal{X} \to \mathbb{R}$. Let $\rho_{\mathbf{x},y}$ be a data generating distribution such that $\rho_{\mathbf{x},y}(\mathbf{x}, y) = \rho_{\mathbf{x}}(\mathbf{x}) \cdot \rho_{y|\mathbf{x}}(y|\mathbf{x})$ where

$$\mathbf{E}_{y\sim\rho_{y|\mathbf{x}}}[y|\mathbf{x}] = f_0(\mathbf{x}) \quad \text{and} \quad \mathrm{Var}_{y\sim\rho_{y|\mathbf{x}}}(y|\mathbf{x}) = \sigma^2(\mathbf{x})$$

for any $\mathbf{x} \in \mathcal{X}$.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, symmetric, positive semi-definite kernel such that

$$\iint k(\mathbf{x}^1, \mathbf{x}^2) f(\mathbf{x}^1) f(\mathbf{x}^2)\, d\rho_{\mathbf{x}}(\mathbf{x}^1) d\rho_{\mathbf{x}}(\mathbf{x}^2) \geq 0 \qquad (1)$$

for any $f \in L^2_{\rho_{\mathbf{x}}}$ and the equality holds if and only if $f = 0$. $\mathbb{H}_k \subset \mathbb{R}^{\mathcal{X}}$ denotes a reproducing kernel Hilbert space with a kernel k. Set $\kappa = (\sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x}, \mathbf{x}))^{1/2}$.

Define the covariance operator $L_k : \mathbb{H}_k \to \mathbb{H}_k$ by

$$L_k f = \int f(\mathbf{x}) k(\mathbf{x}, \cdot)\, d\rho_{\mathbf{x}}(\mathbf{x})$$

and its sample analog $L_{k,X} : \mathbb{H}_k \to \mathbb{H}_k$ by

$$L_{k,X} f = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}^i) k(\mathbf{x}^i, \cdot)$$

where $X = \{\mathbf{x}^1, \cdots, \mathbf{x}^n\} \subset \mathcal{X}$. Let

$$\mathcal{N}(\lambda) = \mathrm{tr}((L_k + \lambda I)^{-1} L_k)$$

and

$$f_\lambda = (L_k + \lambda I)^{-1} L_k f_0$$

where $\lambda > 0$.

Define a compact embedding $\iota_{\rho_{\mathbf{x}}} : \mathbb{H}_k \to L^2_{\rho_{\mathbf{x}}}$ by $\iota_{\rho_{\mathbf{x}}} h = [h]_{\sim\rho_{\mathbf{x}}}$ where $[h]_{\sim\rho_{\mathbf{x}}}$ is the equivalence class containing h in $L^2_{\rho_{\mathbf{x}}}$. Also, define its sample analog $S_D : \mathbb{H}_k \to \mathbb{R}^n$ by

$$S_D f = [f(\mathbf{x}^1), \cdots, f(\mathbf{x}^n)]^\top$$

where $D = \{(\mathbf{x}^1, y^1), \cdots, (\mathbf{x}^n, y^n)\} \subset \mathcal{X} \times \mathbb{R}$.[1]

$A^\top$ denotes the adjoint operator of a given operator A. We write $k_{\mathbf{x}}(\cdot) = k(\mathbf{x}, \cdot)$ and

$$K_{X_1,X_2} = \begin{bmatrix} k(\mathbf{x}_1^1, \mathbf{x}_2^1) & \cdots & k(\mathbf{x}_1^1, \mathbf{x}_2^m) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_1^n, \mathbf{x}_2^1) & \cdots & k(\mathbf{x}_1^n, \mathbf{x}_2^m) \end{bmatrix}$$

where $X_1 = \{\mathbf{x}_1^1, \cdots, \mathbf{x}_1^n\}$ and $X_2 = \{\mathbf{x}_2^1, \cdots, \mathbf{x}_2^m\}$. Also, we write $D(\mathbf{x}) = \{\mathbf{x}^1, \cdots, \mathbf{x}^n\}$ where $D = \{(\mathbf{x}^1, y^1), \cdots, (\mathbf{x}^n, y^n)\}$ is a given dataset.

Technical Assumptions: In this work we assume the following assumptions.

Assumption 3.1. We assume $\mathbf{E}_{y\sim\rho_y} y^2 < \infty$ and

$$\int \left(\exp\left(\frac{|y - f_0(\mathbf{x})|}{M} \right) - \frac{|y - f_0(\mathbf{x})|}{M} - 1 \right) d\rho_{y|\mathbf{x}}(y|\mathbf{x}) \leq \frac{\gamma^2}{2M^2} \qquad (2)$$

for any $\mathbf{x} \in \mathcal{X}$ where M and γ are positive constants.

Assumption 3.2. The target function f_0 satisfies $f_0 = L_k^r g_0$ for some $g_0 \in \mathbb{H}_k$ and $r \in [0, \frac{1}{2}]$. In particular, $f_0 \in \mathbb{H}_k$.

[1] We use a scaled L^2 norm as a norm of Euclidean space $\mathbb{R}^n$. See Appendix A.

Some previous works (Guo et al., 2017; Chang et al., 2017; Lin et al., 2020a) assume that $\mathcal{N}(\lambda)$ satisfies

$$\mathcal{N}(\lambda) \leq C_e \lambda^{-s} \tag{3}$$

for some $s \in (0, 1]$ and some constant $C_e \geq 1$ independent of λ. We do not assume any regularity condition on $\mathcal{N}(\lambda)$ in this work to deal with the convergence rate of a general case. However, under the above assumption (3), we can obtain a better convergence rate by extending our analysis. Note that (3) always holds with $s = 1$.

3.2. Kernel Ridge Regression and Optimal Convergence Rate

Kernel ridge regression (KRR) is an algorithm to estimate the target function f_0. Let $D = \{(\mathbf{x}^1, y^1), \cdots, (\mathbf{x}^N, y^N)\}$ be a labeled dataset whose data points are independently drawn from $\rho_{\mathbf{x},y}$. Given a kernel k and a regularization hyperparameter $\lambda > 0$, a KRR model is given by a minimizer of the following optimization problem

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}^i) - y^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

Note that the solution of the above optimization problem has a closed form expression

$$f_{D,\lambda} = (L_{k,D(\mathbf{x})} + \lambda I)^{-1} S_D^\top \mathbf{y}$$

where $\mathbf{y} = [y^1, \cdots, y^N]^\top$.

One way to measure the performance of kernel ridge regression is to find a convergence rate of the expected loss

$$\mathbf{E}_D \|\iota_{\rho_\mathbf{x}}(f_{D,\lambda} - f_0)\|_{L^2_{\rho_\mathbf{x}}}$$

with respect to the dataset size N. We provide a key theorem from Caponnetto & De Vito (2007); Lin et al. (2017); Fischer & Steinwart (2020).

Theorem 3.3. *Under Assumption 3.1 and 3.2, with* $\lambda = N^{-\frac{1}{2r+2}}$,

$$\mathbf{E}_D \|\iota_{\rho_\mathbf{x}}(f_{D,\lambda} - f)\|_{L^2_{\rho_\mathbf{x}}} = O\left(N^{-\frac{2r+1}{4r+4}}\right).$$

Moreover, this convergence rate is optimal.

4. KRR with Knowledge Distillation

Let g be a teacher model which approximates f_0. In the noiseless data-free version, the optimization problem of KRR with KD is given by

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \alpha \|\iota_{\rho_\mathbf{x}}(h - f_0)\|_{L^2_{\rho_\mathbf{x}}}^2 + (1-\alpha)\|\iota_{\rho_\mathbf{x}}(h - g)\|_{L^2_{\rho_\mathbf{x}}}^2 + \lambda \|h\|_{\mathbb{H}_k}^2$$

where $\alpha \in (0, 1)$ controls the distillation effect and $\lambda > 0$ is a regularization hyperparameter. From the first order optimality condition, we obtain the minimizer

$$\tilde{f}_\lambda = (L_k + \lambda I)^{-1}(\alpha L_k f_0 + (1 - \alpha) L_k g). \tag{4}$$

Now we consider a sample version. Let $D_1 = \{(\mathbf{x}_1^1, y_1^1), \cdots, (\mathbf{x}_1^{N_1}, y_1^{N_1})\}$ be a labeled dataset whose data points are independently generated from $\rho_{\mathbf{x},y}$ and $D_2(\mathbf{x}) = \{\mathbf{x}_2^1, \cdots, \mathbf{x}_2^{N_2}\}$ be an unlabeled dataset whose data points are independently generated from $\rho_\mathbf{x}$. To distill the teacher's knowledge using $D_2(\mathbf{x})$, we consider the optimization problem

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \alpha \cdot \frac{1}{N_1} \sum_{i=1}^{N_1} (h(\mathbf{x}_1^i) - y_1^i)^2 + (1 - \alpha) \cdot \frac{1}{N_2} \sum_{i=1}^{N_2} (h(\mathbf{x}_2^i) - g(\mathbf{x}_2^i))^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

Then, the solution of the above problem is

$$\tilde{f}_{D,\lambda} = (\alpha L_{k,D_1(\mathbf{x})} + (1 - \alpha) L_{k,D_2(\mathbf{x})} + \lambda I)^{-1} (\alpha S_{D_1}^\top \mathbf{y}_1 + (1 - \alpha) L_{k,D_2(\mathbf{x})} g). \tag{5}$$

A natural question is how much the performance of the trained student model $\tilde{f}_{D,\lambda}$ can be improved and the answer is given in the following:

Theorem 4.1 (Informal)**.** *The performance of the student model $\tilde{f}_{D,\lambda}$ is at least as good as the worse of the teacher model g and a trained KRR model using both datasets of sizes N_1 and N_2 independently generated from $\rho_{\mathbf{x},y}$ in terms of the convergence rate of the expected risk when we use adequate α and λ.*

Theorem 4.1 is one of the key ideas to analyze ensemble distillation in federated learning. See Appendix B for the detailed statement and proof.

5. KRR with Ensemble Distillation in FL

5.1. Problem Formulation

We assume the following:

- There are total m clients and client j has a labeled dataset $D_j = \{(\mathbf{x}_j^1, y_j^1), \cdots, (\mathbf{x}_j^N, y_j^N)\}$ whose data points are independently generated from $\rho_{\mathbf{x},y}$ ($j = 1, \cdots, m$). We assume $|D_1| = \cdots = |D_m| = N$ for simplicity. Set $\mathbf{y}_j = [y_j^1, \cdots, y_j^N]^\top$ and $D = \bigcup_{j=1}^m D_j$.
- To distill knowledge without sharing models, we assume there is an unlabeled public dataset $D_p(\mathbf{x}) = \{\mathbf{x}_p^1, \cdots, \mathbf{x}_p^{N_p}\}$ whose data points are independently generated from $\rho_\mathbf{x}$ and whose size is N_p.

Algorithm 1 KRR with iterative Ensemble Distillation in FL

Input: hyperparameters $\alpha \in (0,1)$, $\lambda > 0$, and $t \in \mathbb{N}$
Output: Trained model f_j, $j = 1, \cdots, m$
Pretrain: For $j = 1, \cdots, m$, client j trains its model f_j using the loss function

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}_j^i) - y_j^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

Each client downloads the unlabeled public dataset $D_p(\mathbf{x})$.
for $t_0 = 1, \cdots, t$ **do**
  For $j = 1, \cdots, m$, client j predicts on $D_p(\mathbf{x})$ and uploads the prediction $\tilde{\mathbf{y}}_p^j$ to the server.
  The server computes an updated consensus

$$\tilde{\mathbf{y}}_p = \frac{1}{m} \sum_{j=1}^{m} \tilde{\mathbf{y}}_p^j.$$

  Each client downloads the ensemble prediction $\tilde{\mathbf{y}}_p$.
  For $j = 1, \cdots, m$, client j updates its model f_j using the loss function

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}_j^i) - y_j^i)^2 + (1-\alpha) \cdot \frac{1}{N_p} \sum_{i=1}^{N_p} (h(\mathbf{x}_p^i) - (\tilde{\mathbf{y}}_p)^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

end for

- Due to privacy concerns, only client j can access its own dataset D_j. In addition, all clients can access the public dataset $D_p(\mathbf{x})$ with its pseudo labels.
- For simplicity, we write $L_{k,D_j(\mathbf{x})}$ as L_{k,X_j} for $j = 1, \cdots, m$ and $L_{k,D_p(\mathbf{x})}$ as L_{k,X_p}.
- We consider and analyze an algorithm that performs public data labeling through an ensemble prediction of local models and local model training through knowledge distillation. Algorithm 1 presents the detailed procedure.

5.2. One-Shot Ensemble Distillation in FL

First, we deal with one-shot ensemble distillation in FL, i.e., Algorithm 1 with $t = 1$. Even though each client cannot access the one-shot ensemble model, using the result in Lin et al. (2020a) and Theorem 4.1, we can derive a strong result that guarantees the performance of the local models after one-shot ensemble distillation.

Theorem 5.1. *Assume Assumption 3.1 and 3.2 hold. Also, assume*

$$m \leq N^{2r+1-\epsilon} \tag{6}$$

for any fixed $\epsilon \in (0,1)$ and $N_p \geq (m-1)N$. Let $\tilde{f}_{D,\lambda}^j$ be the local model of client j after one-shot ensemble distillation ($j = 1, \cdots, m$). Then, with $\alpha = 1/m$ and $\lambda = (mN)^{-\frac{1}{2r+2}}$,

$$\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\tilde{f}_{D,\lambda}^j - f_0)\|_{L^2_{\rho_{\mathbf{x}}}} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$$

for $j = 1, \cdots, m$.

The proof of Theorem 5.1 is provided in Appendix C.1. Theorem 5.1 tells us that, under some assumptions, each local model $\tilde{f}_{D,\lambda}^j$ after the one-shot ensemble distillation has at least the same performance as the KRR model using the whole dataset of size $|D| = mN$ in terms of the convergence rate of the expected risk. Note that Theorem 5.1 requires that the number of clients satisfy $m \leq N^{2r+1-\epsilon}$ for any fixed $\epsilon \in (0,1)$. However, one of main properties in FL is a massively distributed environment (McMahan et al., 2017). Therefore, it is more desirable to remove this restriction on the number of clients, which is the main reason why we consider iterative ensemble distillation (i.e., $t > 1$).

5.3. A Toy Example: Iterative Ensemble Distillation in FL When $m = 1$

We now analyze iterative ensemble distillation. First, we consider a simple example $m = 1$ to obtain some motivations. In this case, the local client trains its model using its private dataset D_1 and the public dataset $D_p(\mathbf{x})$ with pseudo labels which are predictions of the local model. Thus, the problem is equivalent to self-distillation on auxiliary unlabeled dataset. To analyze this algorithm, our first interest is to find the limiting regressor after infinitely many iterations.

Theorem 5.2. *The local model f_1 converges to*

$$f_{D_1,\lambda/\alpha} = (L_{k,X_1} + \lambda/\alpha I)^{-1} S_{D_1}^{\top} \mathbf{y}_1$$

after infinitely many iterations in Algorithm 1 with $m = 1$.

We provide the proof in Appendix C.2. Theorem 5.2 implies that the limiting regressor is just a kernel ridge regressor using the private dataset D_1 with an amplified regularization. This result is closely related to the study of self-distillation on kernel ridge regression (Mobahi et al., 2020; Borup & Andersen, 2021). However, it has a limitation to analyze the generalization error bound since an amplified regularization makes the approximation error $\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(f - f_{\lambda/\alpha})\|_{L^2_{\rho_{\mathbf{x}}}}$ excessively large when α is small.

To resolve this issue, we introduce a de-regularization trick. Inspired by Kernel Inducing Point (KIP) scheme (Nguyen

et al., 2021), we define a de-regularization trick for a prediction $\mathbf{v}$ on $D_p(\mathbf{x})$ as follows:

$$S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top \mathbf{w} = \mathbf{v}$$

where $\lambda_0 \geq 0$ is a de-regularization hyperparameter. For well-definedness of $\mathbf{w}$, we require a condition $K_{X_p,X_p} > 0$. Under this condition,

$$\begin{aligned}\mathbf{w} &= (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1}\mathbf{v} \\ &= (K_{X_p,X_p} + N_p\lambda_0 I)K_{X_p,X_p}^{-1}\mathbf{v}. \end{aligned} \tag{7}$$

Note that $\lim_{\lambda_0 \to 0} \mathbf{w} = \mathbf{v}$. Applying the de-regularization trick, we modify Algorithm 1 so that each client uses the de-regularized predictions, i.e.,

$$\tilde{\mathbf{y}}_p^- = (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1}\tilde{\mathbf{y}}_p$$

instead of $\tilde{\mathbf{y}}_p$ except the last ensemble distillation step. The de-regularization trick can be implemented in the server since it only depends on $D_p(\mathbf{x})$. Therefore, this procedure does not require any additional computational resource in the client side. We provide the modified algorithm (Algorithm 2) in Appendix C.3.

From now on, we set $\lambda_0 = \lambda$. The limiting regressor is obviously changed if we adopt the de-regularization trick.

Theorem 5.3. *The local model f_1 converges to*

$$\tilde{f}_{D_1,\lambda} = \left(L_{k,X_1} + \lambda I + \frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp\right)^{-1} S_{D_1}^\top \mathbf{y}_1$$

after infinitely many ensemble distillation steps with the de-regularization trick in Algorithm 2 with $m = 1$ where $P_{D_p(\mathbf{x})}^\perp$ is the orthogonal projection onto the orthogonal complement of a subspace

$$\textit{span}\,(k(\mathbf{x},\cdot) : \mathbf{x} \in D_p(\mathbf{x})) \tag{8}$$

of the reproducing kernel Hilbert space $\mathbb{H}_k$ under the assumption $\lambda_0 = \lambda$ and K_{X_p,X_p} is invertible.

Note that there is still an additional term $\frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp$ in the inverse compared with $f_{D_1,\lambda}$. However, we can omit this term if N_p is large. The following theorem supports this argument.

Theorem 5.4. *Assume $\lambda_0 = \lambda$ and K_{X_p,X_p} is invertible. We also assume that the density $\rho_\mathbf{x}$ is strictly positive on any non-empty open subset of $\mathcal{X}$. Let $\tilde{f}_{D_1,\lambda} \in \mathbb{H}_k$ be defined as in Theorem 5.3. If α and λ do not depend on N_p, then*

$$\lim_{N_p \to \infty} \tilde{f}_{D_1,\lambda} = (L_{k,X_1} + \lambda I)^{-1} S_{D_1}^\top \mathbf{y}_1 = f_{D_1,\lambda}$$

almost surely in the $\mathbb{H}_k$-norm sense.

The proof of Theorem 5.4 is provided in Appendix C.5. Then, our next question is how much public data is needed to ignore the additional term $\frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp$ when we compute the convergence rate of the expected risk. We answer this question in Section 5.4 for the general setting of m clients.

5.4. Iterative Ensemble Distillation in FL for General Case

We now turn to analyze iterative ensemble distillation for m clients. Our main question is whether the limiting regressors of clients after infinitely many iterations in Algorithm 2 have the same performance as a KRR model using the whole dataset D in terms of the convergence rate of the expected risk. The answer is yes if the unlabeled public dataset is sufficiently large under some regularity conditions.

Theorem 5.5. *Assume Assumption 3.1 and 3.2 with $0 < r \leq \frac{1}{2}$ holds. We further assume $m \geq 2$, $\lambda_0 = \lambda$,*

$$N_p \geq \max\left(\left(m^{\frac{3r+2}{2r^2+2r}} N^{\frac{1}{2r+2}}\right)^{1/(1-\epsilon)}, (m-1)N\right)$$

for a fixed $\epsilon \in (0, \frac{1}{2})$, and $K_{X_p,X_p} > 0$. Let $\tilde{f}_{D,\lambda}^j$ be the local model of client j after conducting Algorithm 2 with $t = \infty$ ($j = 1, \cdots, m$). Then, with $\alpha = 1/m$ and $\lambda = (mN)^{-\frac{1}{2r+2}}$,

$$\mathbf{E}\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda}^j - f_0)\|_{L^2_{\rho_\mathbf{x}}} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$$

for $j = 1, \cdots, m$.

Theorem 5.5 tells us that iterative ensemble distillation can drop the restriction on m and N to attain the same convergence rate as the KRR model trained by all data from clients if we have sufficient unlabeled public data. Thus, this theorem implies Algorithm 2 works well in a massively distributed environment. Note that Theorem 5.5 requires more unlabeled public data compared with Theorem 5.1.

However, Theorem 5.5 does not guarantee the performance of the limiting regressors when $r = 0$ even if N_p is large. The reason is that de-regularization is closely related to the projection $P_{D_p(\mathbf{x})}$ but we have no convergence rate of $\|(I - P_{D_p(\mathbf{x})})f_0\|_{\mathbb{H}_k}$ when $r = 0$ where $P_{D_p(\mathbf{x})}$ is the orthogonal projection onto the subspace (8). For $r > 0$, the required public dataset size decreases as r increases.

5.5. Effects of Client Selection Strategy

We discuss the performance of Algorithm 1 and 2 until now. Note that these algorithms are a sort of synchronous algorithms that have to wait until all clients finish training their local models at each step. So, these algorithms may cause waste of time especially in a massively distributed and/or system heterogeneous environment. A typical approach to resolve this issue in FL is to adopt a client selection strategy in each communication round or to use an asynchronous algorithm. When we directly adopt a random client selection strategy in Algorithm 2, the ensemble prediction $\tilde{\mathbf{y}}_p$ on $D_p(\mathbf{x})$ may have a high variance, which makes the algorithm unstable. Hence, we memorize the previous prediction

$\tilde{\mathbf{y}}_{p,old}$ and update $\tilde{\mathbf{y}}_p$ as

$$\tilde{\mathbf{y}}_p = (1-\gamma_{t_0})\tilde{\mathbf{y}}_{p,old} + \gamma_{t_0}\tilde{\mathbf{y}}_{p,new}$$

at each communication round t_0 like a Robbins-Monro stochastic approximation (Robbins & Monro, 1951). We provide a general framework (Algorithm 3) that considers a client selection strategy and an asynchronous strategy in Appendix C.7. See Appendix C.7 for details.

Based on the stochastic approximation theory (Robbins & Monro, 1951; Bertsekas & Tsitsiklis, 1996), we obtain a result on the limiting regressor after infinitely many iterations in Algorithm 3 as follows.

Corollary 5.6. *Assume $\lambda_0 = \lambda$ and K_{X_p,X_p} is invertible. We also assume that*

$$\sum_{t_0=1}^{\infty} \gamma_{t_0} = \infty, \qquad \sum_{t_0=1}^{\infty} \gamma_{t_0}^2 < \infty.$$

If we sample a fixed number of clients uniformly from all clients at each communication round, then the prediction $\tilde{\mathbf{y}}_p$ on $D_p(\mathbf{x})$ after infinitely many ensemble distillation steps with the de-regularization trick in Algorithm 3 converges to $S_{D_p}g^*$ almost surely where g^* is the limiting ensemble regressor after infinitely many ensemble distillation steps with the de-regularization trick in Algorithm 2. In conclusion, the local models after conducting Algorithm 3 with $t=\infty$ is the same as the local models after conducting Algorithm 2 with $t=\infty$.

We prove a general version of Corollary 5.6 in Appendix C.8. See Appendix C.8 for details. Corollary 5.6 implies that the local models derived from Algorithm 3 with uniform sampling is the same as those derived from Algorithm 2 as $t \to \infty$. Therefore, the client selection strategy does not affect the limiting regressor in the case of uniform client sampling. If we sample clients non-uniformly, then the effect of local datasets to the limiting regressor can be different.

5.6. Connection to Neural Network

From Section 5.2 to 5.5, we discuss the effectiveness of ensemble distillation algorithms for kernel ridge regression models. According to the approximation scheme of neural networks as kernel machines (Jacot et al., 2018; Domingos, 2020), it seems that ensemble distillation with neural networks in regression problems can be also effective.

One remark is that we assume all clients use the same kernel k in our analysis and this assumption is violated when neural networks are used as local models. However, we can expect the algorithms that match features in neural networks have good performance according to our analysis. For instance, FedHeNN (Makhija et al., 2022) uses Hilbert-Schmidt independence criterion to match the features, which improves the performance. Another remark is that the de-regularization trick can be omitted in the ridgeless cases. Therefore, we can see that some existing algorithms (Li & Wang, 2019; Lin et al., 2020b) are the special cases of our algorithm.

6. Experiments

In this section, we provide experimental results to validate our theoretical results. Rather than verifying the convergence rate of the expected risk, we now analyze the expected risk itself to confirm the practical effectiveness of ensemble distillation algorithms.

6.1. Setup

We conduct experiments on three synthetic datasets and one real world dataset. We refer to the three synthetic datasets as **Dataset 1**, **Dataset 2**, and **Dataset 3**. Dataset 1 and Dataset 3 are from the existing works (Chang et al., 2017; Lin et al., 2020a). The real world dataset is a simplified regression version of the MNIST dataset from another existing work (Cui et al., 2021). We refer to this dataset as **MNIST**. The generating procedure for the datasets and the kernels used for KRR are described in Appendix E.1. To evaluate the performance, we compute the averaged mean squared error of the local models over the test dataset.

To deal with massively distributed environment which is mainly considered in FL, we set the number of clients (m) is relatively large compared with the local dataset size (N). In addition, we set the unlabeled public dataset size $N_p = (m-1)N$ to compare one-shot ensemble distillation and iterative ensemble distillation fairly. We conduct an additional experiment to investigate the effect of the unlabeled public dataset size. Detailed experiment setup (e.g., hyperparameter configuration) is described in Appendix E.2.

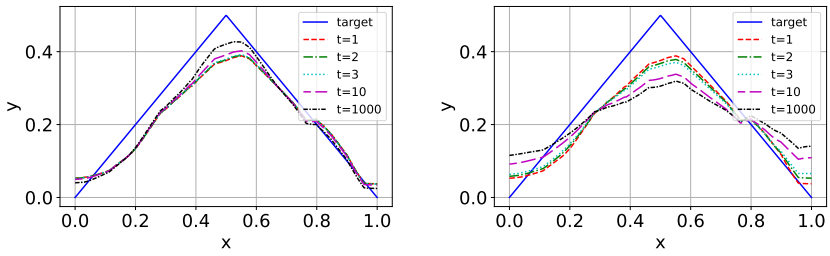
6.2. Illustrative Example: Effect of De-regularization Trick

We first validate the effectiveness of the de-regularization trick for iterative ensemble distillation. In this experiment, we use Dataset 1 and set $m = 20, N = 20, \lambda = 0.002$ and $N_p = 380$.

As illustrated in Figure 1, the regularization effect is amplified as ensemble distillation is repeated when we do not use the de-regularization trick. In contrast, the de-regularization trick prevents the regularization effect from being amplified even for a large t.

6.3. Performance Comparison

We compare the local training, the one-shot ensemble distillation algorithm, the iterative ensemble distillation algorithm without the de-regularization, the iterative ensemble



(a) w/ de-regularization trick (b) w/o de-regularization trick

Figure 1. KRR regressors of a client paticipating in iterative ensemble distillation after t communication rounds (a) with the de-regularization trick and (b) without the de-regularization trick. The solid blue line is the target function.

Table 1. Performance of the standalone model on different datasets with sample sizes $N = 10$ and $N = 20$

DATASET	$N = 10$	$N = 20$
DATASET 1	0.0329	0.0243
DATASET 2	0.0255	0.0144
DATASET 3	0.0785	0.0739
MNIST	0.9436	0.7845

distillation algorithm with the de-regularization, and the central training (which gives the KRR model trained using all local datasets). The performance of standalone KRR models with sample sizes $N \in \{10, 20\}$ is presented in Table 1. We summarize the experimental results of the case with $N = 10$ and various m in Figure 2. We present additional experimental results with sample size $N = 20$ in Appendix E.3.

We first observe that all ensemble distillation based FL algorithms improve the performance of local models compared with the local training. They also achieve the performance as good as the central training on Dataset 2. However, the one-shot ensemble distillation algorithm becomes worse compared with the central training in the other cases. Especially, it has poor performance in massively distributed environment with high dimensional datasets (Dataset 3 and MNIST). The iterative ensemble distillation algorithm without the de-regularization is also worse than the central training on Dataset 3. This algorithm does not achieve the same performance as the central training in some other cases as well.

On the other hand, the iterative ensemble distillation algorithm with the de-regularization is dominant compared with the one-shot ensemble distillation algorithm and the iterative ensemble distillation algorithm without the de-regularization in all settings. It is also comparable with the central training in all experiments. Although our theoretical result requires more unlabeled public data than $(m-1)N$ data points, we can observe that $(m-1)N$ unlabeled public data points are enough for iterative ensemble distillation to achieve the same performance as the central training in practice. Moreover, it has not only the same convergence rate but also the same expected risk as the central training in our experiments.

Table 2. Performance of the iterative ensemble distillation algorithm on Dataset 3 with different public dataset sizes $N_p \in \{50, 100, 200, 500, 1000\}$. We set $N = 10$ and $m = 50$.

N_p	50	100	200	500	1000
	0.0251	0.0198	0.0168	0.0168	0.0164

We additionally compare our proposed algorithm with FedMD (Li & Wang, 2019) with neural networks on multiple datasets. See Appendix E.3 for the results.

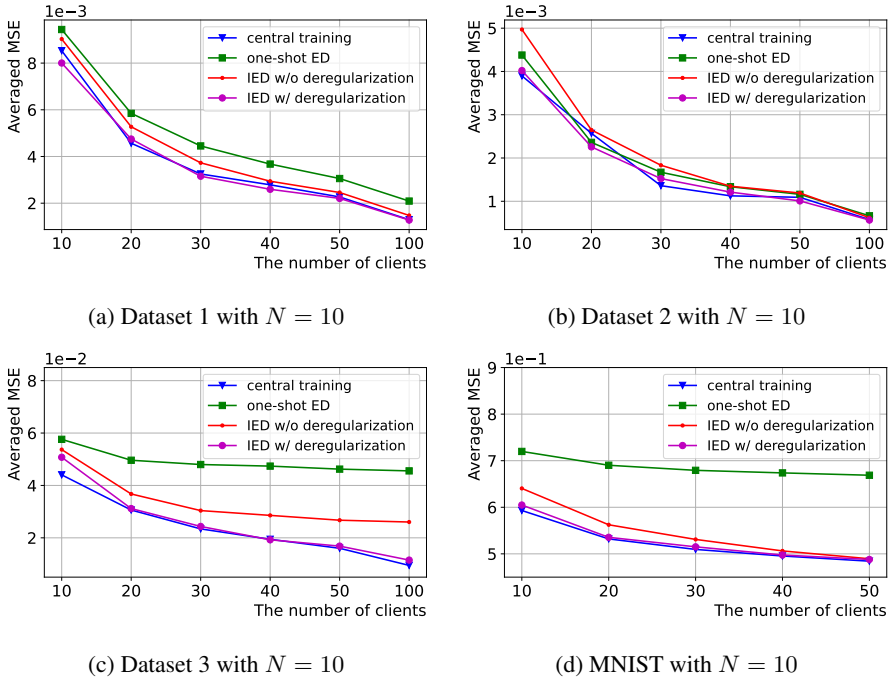
6.4. Effect of Public Dataset Size

We compare the performance of the iterative ensemble distillation algorithm with the de-regularization on various N_p. We simulate this algorithm on Dataset 3 with $N = 10$ and $m = 50$. The result is summarized in Table 2. Basically, the performance of the algorithm is improved as N_p increases. However, the improvement slows down when N_p is sufficiently large. In other words, more public data points may not improve the performance if there are already sufficient public data points.

We provide additional experimental results including the comparison between with a client selection strategy and without a client selection strategy in Appendix E.3.

7. Conclusions

In this work, we provide a KRR based theoretical framework to verify the effectiveness of KD, one-shot ensemble distillation, and iterative ensemble distillation under some



(a) Dataset 1 with $N = 10$

(b) Dataset 2 with $N = 10$

(c) Dataset 3 with $N = 10$

(d) MNIST with $N = 10$

Figure 2. Comparison between the performance of the one-shot ensemble distillation algorithm (one-shot ED), the iterative ensemble distillation algorithm without the de-regularization (IED w/o deregularization), the iterative ensemble distillation algorithm with the de-regularization (IED w/ deregularization), and the central training. We set $N = 10$ and conduct the experiments with various m.

regularity conditions. We also analyze the effects of a client selection strategy in our setting. We simulate ensemble distillation to validate our theoretical results.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (Grant No. NRF-2019R1A5A1028324) and in part by internal fund/grant of Electronics and Telecommunications Research Institute(ETRI). [22RB1100, Exploratory and Strategic Research of ETRI-KAIST ICT Future Technology]

References

Afonin, A. and Karimireddy, S. P. Towards model agnostic federated learning using knowledge distillation. In *The Tenth International Conference on Learning Representations*, 2022.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Anil, R., Pereyra, G., Passos, A., Ormándi, R., Dahl, G. E., and Hinton, G. E. Large scale distributed neural network training through online distillation. In *6th International Conference on Learning Representations*, 2018.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Aydın, A. D. and Gheondea, A. Probability error bounds for approximation of functions in reproducing kernel hilbert spaces. *Journal of Function Spaces*, 2021, 2021.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2938–2948. PMLR, 2020.

Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Bertsekas, D. and Tsitsiklis, J. N. *Neuro-dynamic programming*. Athena Scientific, 1996.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Borup, K. and Andersen, L. N. Even your teacher needs guidance: Ground-truth targets dampen regularization imposed by self-distillation. *Advances in Neural Information Processing Systems*, 34:5316–5327, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Caldas, S., Konečny, J., McMahan, H. B., and Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

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

Chang, X., Lin, S.-B., and Zhou, D.-X. Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research*, 18(46):1–22, 2017.

Chatalic, A., Schreuder, N., Rosasco, L., and Rudi, A. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, pp. 3006–3024. PMLR, 2022.

Cho, Y. J., Wang, J., Chiruvolu, T., and Joshi, G. Personalized federated learning for heterogeneous clients with clustered knowledge transfer. *arXiv preprint arXiv:2109.08119*, 2021.

Conway, J. B. *A course in functional analysis*, volume 96. Springer, 2019.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Diao, E., Ding, J., and Tarokh, V. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *9th International Conference on Learning Representations*, 2021.

Domingos, P. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

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

Ferreira, J. and Menegatto, V. A. Positive definiteness, reproducing kernel hilbert spaces and beyond. *Annals of Functional Analysis*, 4(1), 2013.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Fujii, J., Fujii, M., Furuta, T., and Nakamoto, R. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118(3):827–830, 1993.

Guo, Z.-C., Lin, S.-B., and Zhou, D.-X. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7): 074009, 2017.

He, C., Annavaram, M., and Avestimehr, S. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems*, 33: 14068–14080, 2020.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jiang, Y., Wang, S., Valls, V., Ko, B. J., Lee, W.-H., Leung, K. K., and Tassiulas, L. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, D. and Wang, J. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020a.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *8th International Conference on Learning Representations*, 2020b.

Li, Y., Zhang, H., and Lin, Q. On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*, 2023.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

Lin, S.-B., Wang, D., and Zhou, D.-X. Distributed kernel ridge regression with communications. *Journal of Machine Learning Research*, 21(93):1–38, 2020a.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363, 2020b.

Makhija, D., Han, X., Ho, N., and Ghosh, J. Architecture agnostic federated learning for neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14860–14870. PMLR, 2022.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Mobahi, H., Farajtabar, M., and Bartlett, P. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

Mora, A., Tenison, I., Bellavista, P., and Rish, I. Knowledge distillation for federated learning: a practical guide. *arXiv preprint arXiv:2211.04742*, 2022.

Nguyen, T., Chen, Z., and Lee, J. Dataset meta-learning from kernel ridge-regression. In *9th International Conference on Learning Representations*, 2021.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022.

Nishio, T. and Yonetani, R. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications*, pp. 1–7. IEEE, 2019.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pp. 9489–9502. PMLR, 2021.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D. S., and Khazaeni, Y. Federated learning with matched averaging. In *8th International Conference on Learning Representations*, 2020.

Yao, D., Pan, W., Wan, Y., Jin, H., and Sun, L. Fedhm: Efficient federated learning for heterogeneous models via low-rank factorization. *arXiv preprint arXiv:2111.14655*, 2021.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Yin, R., Wang, W., and Meng, D. Distributed nyström kernel learning with communications. In *International Conference on Machine Learning*, pp. 12019–12028. PMLR, 2021.

Yuan, X.-T. and Li, P. On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond. *arXiv preprint arXiv:2206.05187*, 2022.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.

Zhang, J., Guo, S., Ma, X., Wang, H., Xu, W., and Wu, F. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34:10092–10104, 2021.

Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., and Wu, C. Dense: Data-free one-shot federated learning. *Advances in Neural Information Processing Systems*, 35:21414–21428, 2022a.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., and Ma, K. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722, 2019.

Zhang, L., Shen, L., Ding, L., Tao, D., and Duan, L.-Y. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10174–10183, 2022b.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 592–617. PMLR, 2013.

Zhu, Z., Hong, J., and Zhou, J. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pp. 12878–12889. PMLR, 2021.

A. Comments and Details on Section 3

A.1. Comments on Section 3.1

- For simplicity, we write $Ah = A(h)$ for any $h \in \mathcal{H}_1$ and an operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ where $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces. Define $A^1 h = Ah$ and $A^t h = AA^{t-1}h$ for any integer $t \geq 2$. For linear operators $A : \mathcal{H}_1 \to \mathcal{H}_2$ and $B : \mathcal{H}_2 \to \mathcal{H}_3$, we write the composition $B \circ A$ as BA where $\mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{H}_3$ are Hilbert spaces.

- We use a scaled L^2 norm
$$\|\mathbf{a}\|_2 = \sqrt{\frac{\mathbf{a}_1^2 + \cdots + \mathbf{a}_n^2}{n}}$$
as a norm of Euclidean space $\mathbb{R}^n$.

- Note that the kernel k is bounded since it is continuous on a compact domain $\mathcal{X} \times \mathcal{X}$. Thus, $\kappa < \infty$. From the positive definiteness of k, $k(\mathbf{x}^1, \mathbf{x}^2) \leq \kappa^2$ holds for any $\mathbf{x}^1, \mathbf{x}^2 \in \mathcal{X}$.

- From the boundedness of k, the continuity of k, and the separability of $\mathcal{X}$, we can conclude $\mathbb{H}_k$ is separable by Corollary 4 of Section 1.5 in Berlinet & Thomas-Agnan (2011).

- By Mercer's theorem (Rasmussen & Williams, 2006) and Proposition 4.2 in Ferreira & Menegatto (2013), $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \mu_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$ where $\{\mu_i\}_{i=1}^{\infty}$ are eigenvalues of a linear operator L_k and each $\phi_i \in \mathbb{H}_k$ is an eigenfunction of L_k corresponding to μ_i such that $\mu_i \downarrow 0$, $\sum_{i=1}^{\infty} \mu_i < \infty$, and $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $L^2_{\rho_\mathbf{x}}$.[2] Moreover, we know that $\{\phi_i\}_{i=1}^{\infty}$ are continuous and
$$\mathbb{H}_k = \left\{ f(\mathbf{x}) = \sum_{i=1}^{\infty} f_i \phi_i : \sum_{i=1}^{\infty} \frac{f_i^2}{\mu_i} < \infty \right\}$$
with the inner product
$$\langle f, g \rangle_{\mathbb{H}_k} = \sum_{i=1}^{\infty} \frac{f_i g_i}{\mu_i}$$
where $f = \sum_{i=1}^{\infty} f_i \phi_i$ and $g = \sum_{i=1}^{\infty} g_i \phi_i$. We can easily see that $\{\sqrt{\mu_i}\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $\mathbb{H}_k$.

- For $f \in L^2_{\rho_\mathbf{x}}$ such that $f = \sum_{i=1}^{\infty} f_i \phi_i$ $\rho_\mathbf{x}$-almost everywhere where
$$\sum_{i=1}^{\infty} \frac{f_i^2}{\mu_i} < \infty,$$
we can think f as an element $\sum_{i=1}^{\infty} f_i \phi_i$ in $\mathbb{H}_k$.

- In the definition of $\iota_{\rho_\mathbf{x}} : \mathbb{H}_k \to L^2_{\rho_\mathbf{x}}$, $[h]_{\sim \rho_\mathbf{x}}$ means the equivalence class of all measureable functions that is equal to $h \in \mathbb{H}_k$ $\rho_\mathbf{x}$-almost everywhere.

- The adjoint operator $\iota^\top_{\rho_\mathbf{x}} : L^2_{\rho_\mathbf{x}} \to \mathbb{H}_k$ of $\iota_{\rho_\mathbf{x}}$ satisfies
$$\iota^\top_{\rho_\mathbf{x}}([h]_{\sim \rho_\mathbf{x}})(\mathbf{x}) = \langle \iota^\top_{\rho_\mathbf{x}}([h]_{\sim \rho_\mathbf{x}}), k_\mathbf{x} \rangle_{\mathbb{H}_k} = \langle h, \iota_{\rho_\mathbf{x}}(k_\mathbf{x}) \rangle_{L^2_{\rho_\mathbf{x}}} = \int k(\mathbf{x}, \mathbf{t}) h(\mathbf{t}) \, d\rho_\mathbf{x}(\mathbf{t}).$$
Thus, $L_k = \iota^\top_{\rho_\mathbf{x}} \iota_{\rho_\mathbf{x}}$. From this fact and the compactness of $\iota^\top_{\rho_\mathbf{x}}$ (Steinwart & Christmann, 2008), we can see that L_k is compact, self-adjoint, and positive. Also,
$$\|\iota_{\rho_\mathbf{x}} f\|^2_{L^2_{\rho_\mathbf{x}}} = \langle \iota_{\rho_\mathbf{x}} f, \iota_{\rho_\mathbf{x}} f \rangle_{L^2_{\rho_\mathbf{x}}} = \langle \iota^\top_{\rho_\mathbf{x}} \iota_{\rho_\mathbf{x}} f, f \rangle_{\mathbb{H}_k} = \langle L_k f, f \rangle_{\mathbb{H}_k} = \langle L_k^{1/2} f, L_k^{1/2} f \rangle_{\mathbb{H}_k} = \|L_k^{1/2} f\|^2_{\mathbb{H}_k}. \tag{9}$$

[2] In $L^2_{\rho_\mathbf{x}}$, they are considered as equivalence classes containing continuous functions ϕ_i.

- Let $D = \{(\mathbf{x}^1, y^1), \cdots, (\mathbf{x}^n, y^n)\}$ be a labeled dataset. The adjoint $S_D^\top : \mathbb{R}^n \to \mathbb{H}_k$ of S_D maps $\mathbf{c} = [\mathbf{c}_1, \cdots, \mathbf{c}_n]^\top \in \mathbb{R}^n$ to

$$S_D^\top \mathbf{c} = \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i \mathbf{k}_{\mathbf{x}^i}.$$

  Note that $L_{k,X} = S_D^\top S_D$. Obviously S_D is compact since it has finite rank. Therefore $L_{k,X}$ is also compact, self-adjoint, and positive. Similarly as $\iota_{\rho_\mathbf{x}}$,

$$\|S_D f\|_2^2 = \langle S_D f, S_D f \rangle_2 = \langle S_D^\top S_D f, f \rangle_{\mathbb{H}_k} = \langle L_{k,D(\mathbf{x})} f, f \rangle_{\mathbb{H}_k} = \langle L_{k,D(\mathbf{x})}^{1/2} f, L_{k,D(\mathbf{x})}^{1/2} f \rangle_{\mathbb{H}_k} = \|L_{k,D(\mathbf{x})}^{1/2} f\|_{\mathbb{H}_k}^2. \quad (10)$$

- Since S_D only depends on $D(\mathbf{x})$, we naturally define S_D as $S_D f = [f(\mathbf{x}^1), \cdots, f(\mathbf{x}^n)]^\top$ for $D(\mathbf{x}) = \{\mathbf{x}^1, \cdots, \mathbf{x}^n\}$.

- We write $k_X(\cdot) = [k_{\mathbf{x}^1}(\cdot), \cdots, k_{\mathbf{x}^n}(\cdot)]^\top$ where $X = \{\mathbf{x}^1, \cdots, \mathbf{x}^n\} \subset \mathcal{X}$.

- Assumption 3.1 is related to the condition of the noise term and Assumption 3.2 is regarding the regularity of the target function f_0.

- If the noise is uniformly bounded, Gaussian, or sub-Gaussian, then (2) is satisfied (Caponnetto & De Vito, 2007; Lin et al., 2020a).

- (3) with $s = 1$ is always satisfied since

$$\mathcal{N}(\lambda) = \sum_{i=1}^\infty \frac{\mu_i}{\mu_i + \lambda} \le \sum_{i=1}^\infty \frac{\mu_i}{\lambda} \le \frac{\kappa^2}{\lambda} \quad (11)$$

  which follows from

$$\kappa^2 \ge \int k(\mathbf{x}, \mathbf{x}) \, d\rho_\mathbf{x} = \int \sum_{i=1}^\infty \mu_i \phi_i(\mathbf{x})^2 \, d\rho_\mathbf{x} = \sum_{i=1}^\infty \mu_i.$$

- Since $\mathcal{N}(\lambda) = \sum_{i=1}^\infty \frac{\mu_i}{\mu_i + \lambda}$, it is monotonically decreasing with respect to λ and $\mathcal{N}(\lambda) \to \infty$ as $\lambda \downarrow 0$. Thus, there exists $\lambda_1 > 0$ (e.g., $\lambda_1 = \mu_2$) such that $\mathcal{N}(\lambda) \ge 1$ for $0 < \lambda < \lambda_1$.

A.2. Details on Section 3.2

- First, consider a noiseless data-free version :

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} J[h] = \|\iota_{\rho_\mathbf{x}}(h - f_0)\|_{L^2_{\rho_\mathbf{x}}}^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

  From (9), we obtain

$$J[h] = \langle h - f_0, L_k(h - f_0) \rangle_{\mathbb{H}_k} + \lambda \|h\|_{\mathbb{H}_k}^2.$$

  Observe that

$$J[h + \epsilon u] - J[h] = \epsilon \cdot \langle 2L_k(h - f_0) + 2\lambda h, u \rangle_{\mathbb{H}_k} + o(\epsilon)$$

  since L_k is self-adjoint. By the definition of functional derivatives, we have

$$\nabla J[h] = 2L_k(h - f_0) + 2\lambda h.$$

  Note that J is strongly convex since $\nabla^2 J[h] = 2(L_k + \lambda I) \ge 2\lambda I$ holds. From the first order optimality condition, the minimizer of the optimization problem is

$$f_\lambda = (L_k + \lambda I)^{-1} L_k f_0.$$

- However, most of regression problems have limited datasets with noisy labels. Let $D = \{(\mathbf{x}^1, y^1), \cdots, (\mathbf{x}^N, y^N)\}$ be a labeled dataset whose data points are independently drawn from $\rho_{\mathbf{x},y}$. Then the optimization problem is given by

$$\operatorname*{argmin}_{h\in\mathbb{H}_k} J[h] = \frac{1}{N}\sum_{i=1}^{N}(h(\mathbf{x}^i) - y^i)^2 + \lambda\|h\|_{\mathbb{H}_k}^2 = \|S_D h - \mathbf{y}\|_2^2 + \lambda\|h\|_{\mathbb{H}_k}^2$$

  where $\mathbf{y} = [y^1, \cdots, y^N]^\top$. Then,

$$\begin{aligned} J[h+\epsilon u] - J[h] &= \epsilon \cdot 2\langle S_D u, S_D h - \mathbf{y}\rangle_2 + \epsilon \cdot \langle 2\lambda h, u\rangle_{\mathbb{H}_k} + o(\epsilon) \\ &= \epsilon \cdot \langle 2S_D^\top S_D h - 2S_D^\top \mathbf{y} + 2\lambda h, u\rangle_{\mathbb{H}_k} + o(\epsilon). \end{aligned}$$

  In other words,

$$\nabla J[h] = 2S_D^\top S_D h - 2S_D^\top \mathbf{y} + 2\lambda h = 2(L_{k,D(\mathbf{x})}h - S_D^\top \mathbf{y} + \lambda h).$$

  Note that J is strongly convex since $\nabla^2 J[h] = 2(L_{k,D(\mathbf{x})} + \lambda I) \geq 2\lambda I$ holds. From the first order optimality condition, we can see that the minimizer $f_{D,\lambda}$ is given by

$$f_{D,\lambda} = (L_{k,D(\mathbf{x})} + \lambda I)^{-1} S_D^\top \mathbf{y}.$$

  In the matrix form we have

$$f_{D,\lambda} = \mathbf{k}_{D(\mathbf{x})}^\top (N\lambda I + K_{D(\mathbf{x}),D(\mathbf{x})})^{-1}\mathbf{y}.$$

B. Details on Section 4

We now derive (4) and (5) as in Section A.2. Let $g \in \mathbb{H}_k$ be a teacher model which approximates f_0.

- Consider a noiseless data-free version of kernel ridge regression problem with knowledge distillation :

$$\operatorname*{argmin}_{h\in\mathbb{H}_k} J[h] = \alpha\|\iota_{\rho_\mathbf{x}}(h - f_0)\|_{L^2_{\rho_\mathbf{x}}}^2 + (1-\alpha)\|\iota_{\rho_\mathbf{x}}(h-g)\|_{L^2_{\rho_\mathbf{x}}}^2 + \lambda\|h\|_{\mathbb{H}_k}^2$$

  where $\alpha \in (0,1)$ is a distillation hyperparameter and $\lambda > 0$ is a regularization hyperparameter. Then

$$J[h+\epsilon u] - J[h] = \epsilon \cdot \langle 2\alpha \cdot L_k(h - f_0) + 2(1-\alpha)\cdot L_k(h-g) + 2\lambda h, u\rangle_{\mathbb{H}_k} + o(\epsilon)$$

  and so

$$\nabla J[h] = 2\alpha \cdot L_k(h - f_0) + 2(1-\alpha)\cdot L_k(h-g) + 2\lambda h.$$

  Again, $\nabla^2 J[h] = 2(L_k + \lambda I) \geq 2\lambda I$ holds. Also, the minimizer $\tilde{f}_\lambda$ of the optmization problem is given by

$$\tilde{f}_\lambda = (L_k + \lambda I)^{-1}(\alpha L_k f_0 + (1-\alpha)L_k g)$$

  using the first order optimality condition.

- Let $D_1 = \{(\mathbf{x}_1^1, y_1^1), \cdots, (\mathbf{x}_1^{N_1}, y_1^{N_1})\}$ be a dataset to train a student model whose data points are independently generated from $\rho_{\mathbf{x},y}$. To distill knowledge of the teacher model g, we assume an unlabeled dataset $D_2(\mathbf{x}) = \{\mathbf{x}_2^1, \cdots, \mathbf{x}_2^{N_2}\}$ whose data points are independently generated from $\rho_\mathbf{x}$, is given. Then the optimization problem is

$$\begin{aligned} \operatorname*{argmin}_{h\in\mathbb{H}_k} J[h] &= \alpha \cdot \frac{1}{N_1}\sum_{i=1}^{N_1}(h(\mathbf{x}_1^i) - y_1^i)^2 + (1-\alpha)\cdot\frac{1}{N_2}\sum_{i=1}^{N_2}(h(\mathbf{x}_2^i) - g(\mathbf{x}_2^i))^2 + \lambda\|h\|_{\mathbb{H}_k}^2 \\ &= \alpha\|S_{D_1}h - \mathbf{y}_1\|_2^2 + (1-\alpha)\|S_{D_2}(h-g)\|_2^2 + \lambda\|h\|_{\mathbb{H}_k}^2 \end{aligned}$$

  where $\alpha \in (0,1)$ and $\lambda > 0$ are hyperparameters and $\mathbf{y}_1 = [y_1^1, \cdots, y_1^N]^\top$. From

$$\begin{aligned} J[h+\epsilon u] - J[h] &= \alpha\epsilon \cdot 2\langle S_{D_1}u, S_{D_1}h - \mathbf{y}_1\rangle_2 + (1-\alpha)\epsilon\cdot 2\langle S_{D_2}u, S_{D_2}h - S_{D_2}g\rangle_2 + \epsilon\cdot\langle 2\lambda h, u\rangle_{\mathbb{H}_k} + o(\epsilon) \\ &= \epsilon\cdot\langle 2\alpha S_{D_1}^\top S_{D_1}h + 2(1-\alpha)S_{D_2}^\top S_{D_2}h - 2\alpha S_{D_1}^\top \mathbf{y}_1 - 2(1-\alpha)S_{D_2}^\top S_{D_2}g + 2\lambda h, u\rangle_{\mathbb{H}_k} + o(\epsilon), \end{aligned}$$

we have

$$\nabla J[h] = 2(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)h - 2(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,D_2(\mathbf{x})}g).$$

Observe that $\nabla^2 J[h] = 2(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I) \geq 2\lambda I$ holds. Applying the first order optimality condition, the minimizer $\tilde{f}_{D,\lambda}$ of the optimization problem is

$$\tilde{f}_{D,\lambda} = (\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,D_2(\mathbf{x})}g).$$

The solution is written in the matrix form as

$$\tilde{f}_{D,\lambda} = \begin{bmatrix}\mathbf{k}_{D_1(\mathbf{x})}^\top & \mathbf{k}_{D_2(\mathbf{x})}^\top\end{bmatrix}(\lambda I + DK_{D_1(\mathbf{x})\cup D_2(\mathbf{x}),D_1(\mathbf{x})\cup D_2(\mathbf{x})})^{-1} D \begin{bmatrix}\mathbf{y}_1 \\ g(D_2(\mathbf{x}))\end{bmatrix} \tag{12}$$

where $D = \mathrm{diag}(\underbrace{\alpha/N_1, \cdots, \alpha/N_1}_{N_1}, \underbrace{(1-\alpha)/N_2, \cdots, (1-\alpha)/N_2}_{N_2})$ and $g(D_2(\mathbf{x})) = [g(\mathbf{x}_2^1), \cdots, g(\mathbf{x}_2^{N_2})]^\top$.

B.1. Proof of Theorem 4.1

We first prove the following theorem.

Theorem B.1. *With the same notation as in Section 4, under Assumption 3.1 and 3.2,*

$$\begin{aligned}\mathbf{E}\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda} - f_0)\|_{L^2_{\rho_\mathbf{x}}} \leq 9\left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}} + \frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right) \\ + (1-\alpha)^{1/2}\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1/2}\|\|L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k}\right]\end{aligned}$$

holds.

Proof of Theorem B.1. Using (9) and Lemma D.3, we have

$$\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda} - f_0)\|_{L^2_{\rho_\mathbf{x}}} = \|L_k^{1/2}(\tilde{f}_{D,\lambda} - f_0)\|_{\mathbb{H}_k} \leq \|(L_k+\lambda I)^{1/2}(\tilde{f}_{D,\lambda} - f_0)\|_{\mathbb{H}_k}.$$

By the triangle inequality and the submultiplicativity of the operator norm,

$$\begin{aligned}&\|(L_k+\lambda I)^{1/2}(\tilde{f}_{D,\lambda} - f_0)\|_{\mathbb{H}_k} \\ &\leq \|(L_k+\lambda I)^{1/2}((\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,D_2(\mathbf{x})}f_0) - f_0)\|_{\mathbb{H}_k} \\ &\quad + (1-\alpha)\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}L_{k,D_2(\mathbf{x})}(g-f_0)\|_{\mathbb{H}_k}.\end{aligned} \tag{13}$$

First, we bound the first term in (13). Note that

$$\begin{aligned}&(L_k+\lambda I)^{1/2}((\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,D_2(\mathbf{x})}f_0) - f_0) \\ &= (L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(L_k+\lambda I)^{1/2}(L_k+\lambda I)^{-1/2}(\alpha(S_{D_1}^\top \mathbf{y}_1 - L_{k,D_1(\mathbf{x})}f_0) - \lambda f_0).\end{aligned}$$

Let

$$\mathcal{Q}_d = \|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(L_k+\lambda I)^{1/2}\|.$$

By Lemma D.1, Lemma D.5, the triangle inequality, and the submultiplicativity of the operator norm, we have

$$\begin{aligned}\mathcal{Q}_d &\leq \|(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(L_k+\lambda I)\| \\ &= \|I + \alpha(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(L_k - L_{k,D_1(\mathbf{x})}) \\ &\qquad + (1-\alpha)(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(L_k - L_{k,D_2(\mathbf{x})})\| \\ &\leq 1 + \frac{1}{\lambda}(\alpha\|L_{k,D_1(\mathbf{x})} - L_k\| + (1-\alpha)\|L_{k,D_2(\mathbf{x})} - L_k\|).\end{aligned} \tag{14}$$

On the other hand, by the submultiplicativity of the operator norm and Lemma D.1,

$$\begin{aligned}\|(L_k+\lambda I)^{-1/2}\alpha(S_{D_1}^\top \mathbf{y}_1 - L_{k,D_1(\mathbf{x})}f_0)\|_{\mathbb{H}_k} &\le \alpha\|(L_k+\lambda I)^{-1/2}\|\cdot\|S_{D_1}^\top \mathbf{y}_1 - L_{k,D_1(\mathbf{x})}f_0\| \\ &\le \frac{\alpha}{\sqrt{\lambda}}\|S_{D_1}^\top \mathbf{y}_1 - L_{k,D_1(\mathbf{x})}f_0\|.\end{aligned}$$

Also,

$$\|\lambda(L_k+\lambda I)^{-1/2}f_0\|_{\mathbb{H}_k} = \|\lambda(L_k+\lambda I)^{-1/2}L_k^r g_0\|_{\mathbb{H}_k} \le \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k} \tag{15}$$

by Lemma D.1. Therefore, by Lemma D.7 and Lemma D.8, we have

$$\begin{aligned}&\|(L_k+\lambda I)^{1/2}((\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,D_2(\mathbf{x})}f_0) - f_0)\|_{\mathbb{H}_k} \\ &\le \left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}} + \frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right)(\log(6/\delta))^{3/2}\end{aligned}$$

with confidence at least $1-\delta$. By Lemma D.9,

$$\begin{aligned}&\mathbf{E}\|(L_k+\lambda I)^{1/2}((\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,D_2(\mathbf{x})}f_0) - f_0)\|_{\mathbb{H}_k} \\ &\le 9\left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}} + \frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right)\end{aligned}$$

since $\Gamma(5/2) = 3\sqrt{\pi}/4 < 3/2$. The second term in (13) satisfies

$$\begin{aligned}&(1-\alpha)\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}L_{k,D_2(\mathbf{x})}(g-f_0)\|_{\mathbb{H}_k} \\ &\le (1-\alpha)\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1}L_{k,D_2(\mathbf{x})}^{1/2}L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k} \\ &\le (1-\alpha)\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1/2}\| \\ &\quad\cdot\|(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1/2}(L_{k,D_2(\mathbf{x})} + \lambda I)^{1/2}\| \\ &\quad\cdot\|(L_{k,D_2(\mathbf{x})} + \lambda I)^{-1/2}L_{k,D_2(\mathbf{x})}^{1/2}\|\|L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k}\end{aligned}$$

by the submultiplicativity of the operator norm. Applying Lemma D.1 and

$$(1-\alpha)^{1/2}\|(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1/2}(L_{k,D_2(\mathbf{x})} + \lambda I)^{1/2}\| \le 1$$

which follows from Lemma D.2, we obtain that the second term in (13) is bounded by

$$(1-\alpha)^{1/2}\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})} + (1-\alpha)L_{k,D_2(\mathbf{x})} + \lambda I)^{-1/2}\|\|L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k}$$

by the submultiplicativity of the operator norm. Taking the expectation completes the proof of Theorem B.1. □

The following corollary implies Theorem 4.1.

Corollary B.2. *Suppose Assumption 3.1 and 3.2 hold and g is independent of $D_2(\mathbf{x})$. With the same notation as in Section 4, if we set $\alpha = N_1/(N_1+N_2)$ and $\lambda = (N_1+N_2)^{-\frac{1}{2r+2}}$, then*

$$\mathbf{E}\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda} - f_0)\|_{L^2_{\rho_\mathbf{x}}} \le O\left((N_1+N_2)^{-\frac{2r+1}{4r+4}} + \left(\mathbf{E}\|\iota_{\rho_\mathbf{x}}(g-f_0)\|_{L^2_{\rho_\mathbf{x}}}^2\right)^{1/2}\right).$$

Proof of Corollary B.2. Set $\alpha = N_1/(N_1+N_2)$ and $\lambda = (N_1+N_2)^{-\frac{1}{2r+2}}$. We will bound the given upper bound of $\mathbf{E}\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda} - f_0)\|_{L^2_{\rho_\mathbf{x}}}$ in Theorem B.1. We first observe that

$$1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}} + \frac{1-\alpha}{\sqrt{N_2}}\right) = 1 + 2\sqrt{2}\kappa^2(N_1+N_2)^{\frac{1}{2r+2}}\left(\frac{\sqrt{N_1}+\sqrt{N_2}}{N_1+N_2}\right) \le 1 + 4\kappa^2 \tag{16}$$

and

$$\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k} = 2\kappa(M+\gamma)(N_1+N_2)^{\frac{1}{4r+4}} \cdot \frac{N_1}{N_1+N_2} \cdot \frac{1}{\sqrt{N_1}} + (N_1+N_2)^{-\frac{2r+1}{4r+4}}\|g_0\|_{\mathbb{H}_k}$$
$$\leq (2\kappa M + 2\kappa\gamma + \|g_0\|_{\mathbb{H}_k})(N_1+N_2)^{-\frac{2r+1}{4r+4}}$$

since $\sqrt{N_1}+\sqrt{N_2} \leq \sqrt{2(N_1+N_2)}$ and $\sqrt{N_1} \leq \sqrt{N_1+N_2}$. Thus, we know that

$$9\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}}+\frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}}+\lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right) = O\left((N_1+N_2)^{-\frac{2r+1}{4r+4}}\right).$$

Second, by the Cauchy-Schwarz inequality (Conway, 2019),

$$\mathbf{E}\left[(1-\alpha)^{1/2}\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})}+(1-\alpha)L_{k,D_2(\mathbf{x})}+\lambda I)^{-1/2}\|\|L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k}\right]$$
$$\leq \left(\mathbf{E}\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})}+(1-\alpha)L_{k,D_2(\mathbf{x})}+\lambda I)^{-1/2}\|^2\right)^{1/2}\left(\mathbf{E}\|L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k}^2\right)^{1/2}.$$

Since

$$\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})}+(1-\alpha)L_{k,D_2(\mathbf{x})}+\lambda I)^{-1/2}\|^2 = \mathcal{Q}_d$$

where $\mathcal{Q}_d$ is defined in the proof of Theorem B.1, by (14) and Lemma D.7, we have

$$\mathcal{Q}_d \leq \left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}}+\frac{1-\alpha}{\sqrt{N_2}}\right)\right)(\log(4/\delta))^{1/2} \leq \left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}}+\frac{1-\alpha}{\sqrt{N_2}}\right)\right)(\log(4/\delta)) \tag{17}$$

with confidence at least $1-\delta$ where $\delta \in (0,1)$ and so

$$\mathbf{E}\|(L_k+\lambda I)^{1/2}(\alpha L_{k,D_1(\mathbf{x})}+(1-\alpha)L_{k,D_2(\mathbf{x})}+\lambda I)^{-1/2}\|^2 \leq 4\Gamma\left(\frac{3}{2}\right)\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}}+\frac{1-\alpha}{\sqrt{N_2}}\right)\right) \leq 4(1+4\kappa^2)$$

by Lemma D.9 and (16) since $\Gamma(3/2) = \sqrt{\pi}/2 < 1$. On the other hand, by (10) and the independence of g and $D_2(\mathbf{x})$, we find that

$$\mathbf{E}\|L_{k,D_2(\mathbf{x})}^{1/2}(g-f_0)\|_{\mathbb{H}_k}^2 = \mathbf{E}\|S_{D_2}(g-f_0)\|_2^2 = \mathbf{E}\left[\frac{1}{N_2}\sum_{i=1}^{N_2}(g(\mathbf{x}_2^i)-f_0(\mathbf{x}_2^i))^2\right] = \mathbf{E}\left[\mathbf{E}\left[\frac{1}{N_2}\sum_{i=1}^{N_2}(g(\mathbf{x}_2^i)-f_0(\mathbf{x}_2^i))^2 \mid g\right]\right]$$
$$= \mathbf{E}(g(\mathbf{x}_2^1)-f_0(\mathbf{x}_2^1))^2 = \mathbf{E}\|\iota_{\rho_\mathbf{x}}(g-f_0)\|_{L^2_{\rho_\mathbf{x}}}^2.$$

Therefore,

$$\mathbf{E}\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda}-f_0)\|_{L^2_{\rho_\mathbf{x}}} \leq O\left((N_1+N_2)^{-\frac{2r+1}{4r+4}} + \left(\mathbf{E}\|\iota_{\rho_\mathbf{x}}(g-f_0)\|_{L^2_{\rho_\mathbf{x}}}^2\right)^{1/2}\right).$$

□

Remark B.3. Since g is fixed in Section 4, we assume g is independent of $D_2(\mathbf{x})$ in Corollary B.2. In this case, when g is a KRR model trained using a dataset $\tilde{D}$ whose size is $\tilde{N}$ such that $\tilde{D}$ is independent of $D_2(\mathbf{x})$, we know that

$$\left(\mathbf{E}\|\iota_{\rho_\mathbf{x}}(g-f_0)\|_{L^2_{\rho_\mathbf{x}}}^2\right)^{1/2} = O\left(\tilde{N}^{-\frac{2r+1}{4r+4}}\right)$$

which can be proved by a similar argument as in the proof of Theorem 3.3 provided in Caponnetto & De Vito (2007). Thus, we can conclude that the generalization error of $\tilde{f}_{D,\lambda}$ has the convergence rate $O\left(\min(N_1+N_2,\tilde{N})^{-\frac{2r+1}{4r+4}}\right)$. If $\tilde{D}$ is not independent of $D_2(\mathbf{x})$, then it could be more complicated. However, we can achieve the same result using Lemma D.13 under some assumptions. We deal with this case in Section 5.4.

C. Proofs and Comments on Section 5

C.1. Proof of Theorem 5.1

We first show the following theorem that is similar to Theorem 13 in Lin et al. (2020a).

Theorem C.1. *Let*

$$\bar{f}_{D,\lambda} = \frac{1}{m}\sum_{i=1}^{m}(L_{k,X_i}+\lambda I)^{-1}S_{D_i}^{\top}\mathbf{y}_i.$$

Under Assumption 3.1 and 3.2,

$$\begin{aligned}\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\bar{f}_{D,\lambda}-f_0)\|_{L^2_{\rho_{\mathbf{x}}}}^2 &= O\left(\lambda^{2r+1}+\frac{1}{\lambda mN}+\lambda^{2r+1}\mathcal{B}^2\right)\\ &\quad + O(1)\cdot\left(\lambda^{2r+1}\mathcal{B}^2+\frac{1}{\lambda mN}\right)\cdot\left(1+\frac{\kappa^2}{\lambda}\right)^2\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right)\cdot\left(1+\frac{1}{4(\kappa^2+1)\mathcal{B}}\right)\end{aligned}$$

holds if $0<\lambda\le 1$ *and* $\mathcal{N}(\lambda)\ge 1$ *where*

$$\mathcal{B} = \frac{1+\log\mathcal{N}(\lambda)}{\lambda N}+\sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}.$$

Proof of Theorem C.1. We use a similar argument as in Lin et al. (2020a). Recall Proposition 9 in Lin et al. (2020a):

$$\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\bar{f}_{D,\lambda}-f_0)\|_{L^2_{\rho_{\mathbf{x}}}}^2 \le 2\left(\|\iota_{\rho_{\mathbf{x}}}(f_\lambda-f_0)\|_{L^2_{\rho_{\mathbf{x}}}}^2+\frac{1}{m}\mathbf{E}\left[\mathcal{Q}^4(\mathcal{P}+\mathcal{S}\|f_\lambda\|_{\mathbb{H}_k})^2\right]+\mathbf{E}\left[\mathcal{Q}^4\mathcal{R}^2\|(L_k+\lambda I)^{1/2}(f_\lambda-f_0)\|_{\mathbb{H}_k}^2\right]\right)$$

where

$$\begin{aligned}\mathcal{P} &= \|(L_k+\lambda I)^{-1/2}(L_k f_0 - S_{D_1}^{\top}\mathbf{y}_{D_1})\|_{\mathbb{H}_k},\\ \mathcal{Q} &= \|(L_k+\lambda I)^{1/2}(L_{k,X_1}+\lambda I)^{-1/2}\|,\\ \mathcal{R} &= \|(L_k+\lambda I)^{-1/2}(L_k-L_{k,X_1})(L_k+\lambda I)^{-1/2}\|,\end{aligned}$$

and

$$\mathcal{S} = \|(L_k+\lambda I)^{-1/2}(L_k-L_{k,X_1})\|.$$

By (9), Lemma D.1, Lemma D.3, and the submultiplicativity of the operator norm, we have

$$\begin{aligned}\|\iota_{\rho_{\mathbf{x}}}(f_\lambda-f_0)\|_{L^2_{\rho_{\mathbf{x}}}}^2 &= \|L_k^{1/2}(f_\lambda-f_0)\|_{\mathbb{H}_k}^2 \le \|(L_k+\lambda I)^{1/2}(f_\lambda-f_0)\|_{\mathbb{H}_k}^2\\ &= \|\lambda(L_k+\lambda I)^{-1/2}L_k^r g_0\|_{\mathbb{H}_k}^2 \le \lambda^{2r+1}\|g_0\|_{\mathbb{H}_k}^2.\end{aligned} \tag{18}$$

Note that

$$\|f_\lambda\|_{\mathbb{H}_k} = \|(L_k+\lambda I)^{-1}L_k f_0\|_{\mathbb{H}_k} \le \|f_0\|_{\mathbb{H}_k} \tag{19}$$

by the submultiplicativity of the operator norm and Lemma D.1. Therefore, applying Lemma D.7 and Lemma D.8 leads to

$$\mathcal{Q}^4(\mathcal{P}+\mathcal{S}\|f_\lambda\|_{\mathbb{H}_k})^2 \le 16\kappa^2(2\sqrt{2}\kappa\|f_0\|_{\mathbb{H}_k}+M+\gamma)^2\frac{1}{\lambda N}(\log(6/\delta))^2$$

with confidence at least $1-\delta$ for $12\exp(-1/4(\kappa^2+1)\mathcal{B})\le\delta<1$. On the other hand, using the trivial bound

$$\mathcal{Q}\le\frac{1}{\sqrt{\lambda}}\|(L_k+\lambda I)^{1/2}\|\le\left(\frac{\kappa^2+\lambda}{\lambda}\right)^{1/2} \tag{20}$$

which follows from the submultiplicativity of the operator norm and $\|(L_k+\lambda I)^{1/2}\| = (\mu_1+\lambda)^{1/2}\le(\kappa^2+\lambda)^{1/2}$, we get

$$\mathcal{Q}^4(\mathcal{P}+\mathcal{S}\|f_\lambda\|_{\mathbb{H}_k})^2 \le 4\kappa^2(2\sqrt{2}\kappa\|f_0\|_{\mathbb{H}_k}+M+\gamma)^2\left(1+\frac{\kappa^2}{\lambda}\right)^2\frac{1}{\lambda N}(\log(6/\delta))^2$$

with confidence at least $1-\delta$ for $\delta \in (0,1)$. By Lemma D.10 and Remark D.11,

$$\mathbf{E}\left[\mathcal{Q}^4(\mathcal{P}+\mathcal{S}\|f_\lambda\|_{\mathbb{H}_k})^2\right] = O\left(\frac{1}{\lambda N}\right) + O(1)\cdot\frac{1}{\lambda N}\cdot\left(1+\frac{\kappa^2}{\lambda}\right)^2 \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right)\cdot\left(1+\frac{1}{4(\kappa^2+1)\mathcal{B}}\right).$$

Next we turn to bound $\mathbf{E}\left[\mathcal{Q}^4\mathcal{R}^2\|(L_k+\lambda I)^{1/2}(f_\lambda - f_0)\|_{\mathbb{H}_k}^2\right]$. By Lemma D.7 and (20),

$$\mathcal{Q}^4\mathcal{R}^2 \le 16(\kappa^2+1)^2\mathcal{B}^2(\log(8/\delta))^2$$

with confidence at least $1-\delta$ for $8\exp(-1/4(\kappa^2+1)\mathcal{B}) \le \delta < 1$ and

$$\mathcal{Q}^4\mathcal{R}^2 \le 4\left(1+\frac{\kappa^2}{\lambda}\right)^2(\kappa^2+1)^2\mathcal{B}^2(\log(8/\delta))^2$$

with confidence at least $1-\delta$ for $\delta \in (0,1)$. Again, using (18), Lemma D.10, and Remark D.11, we obtain

$$\begin{aligned}\mathbf{E}\left[\mathcal{Q}^4\mathcal{R}^2\|(L_k+\lambda I)^{1/2}(f_\lambda - f_0)\|_{\mathbb{H}_k}^2\right] &= O(\lambda^{2r+1}\mathcal{B}^2) \\ &\quad + O(1)\cdot\lambda^{2r+1}\mathcal{B}^2\cdot\left(1+\frac{\kappa^2}{\lambda}\right)^2\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right)\cdot\left(1+\frac{1}{4(\kappa^2+1)\mathcal{B}}\right).\end{aligned}$$

Combining the bounds, we are done. □

Now, we derive the performance of one-shot ensemble distillation. The following corollary implies Theorem 5.1.

Corollary C.2. *Assume Assumption 3.1 and 3.2 hold. Also, assume $m \le N^{2r+1-\epsilon}$ for any fixed $\epsilon \in (0,1)$ and $N_p \ge (m-1)N$. Let*

$$\tilde{f}_{D,\lambda}^i = (\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)L_{k,X_p}\bar{f}_{D,\lambda})$$

which is the local model of client i after one-shot ensemble distillation ($i = 1,\cdots,m$) where

$$\bar{f}_{D,\lambda} = \frac{1}{m}\sum_{i=1}^m (L_{k,X_i}+\lambda I)^{-1}S_{D_i}^\top \mathbf{y}_i.$$

Then, with $\alpha = 1/m$ and $\lambda = (mN)^{-\frac{1}{2r+2}}$,

$$\mathbf{E}\|\iota_{\rho_\mathbf{x}}(\tilde{f}_{D,\lambda}^i - f_0)\|_{L^2_{\rho_\mathbf{x}}} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$$

for any $i = 1,\cdots,m$.

Proof of Corollary C.2. When m or N increases, N_p also increases. Thus, it is enough to show the statement under the assumption

$$\mathcal{N}\left(N_p^{-\frac{1}{2r+2}}\right) \ge 1.$$

Although we cannot directly apply Corollary B.2, we can show a similar statement using the same way as in the proof of Corollary B.2. Since

$$9\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}}+\frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right) = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$$

and

$$\mathbf{E}\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1/2}\|^2 \le 4\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N}}+\frac{1-\alpha}{\sqrt{N_p}}\right)\right) \le 4(1+4\kappa^2),$$

we have

$$\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\tilde{f}^i_{D,\lambda} - f_0)\|_{L^2_{\rho_{\mathbf{x}}}} \leq O\left((mN)^{-\frac{2r+1}{4r+4}} + \left(\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\bar{f}_{D,\lambda} - f_0)\|^2_{L^2_{\rho_{\mathbf{x}}}}\right)^{1/2}\right).$$

Therefore, it suffices to show that

$$\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\bar{f}_{D,\lambda} - f_0)\|^2_{L^2_{\rho_{\mathbf{x}}}} = O\left((mN)^{-\frac{2r+1}{2r+2}}\right).$$

From $m \leq N^{2r+1-\epsilon}$,

$$\mathcal{B} \leq \left(\frac{\log(e\kappa^2 N)}{N^{\frac{\epsilon}{2r+2}}}\right)^{1/2}\left(1 + \left(\frac{\log(e\kappa^2 N)}{N^{\frac{\epsilon}{2r+2}}}\right)^{1/2}\right) = \frac{1}{N^{\frac{\epsilon}{4(2r+2)}}}\left(\frac{\log(e\kappa^2 N)}{N^{\frac{\epsilon}{2(2r+2)}}}\right)^{1/2}\left(1 + \left(\frac{\log(e\kappa^2 N)}{N^{\frac{\epsilon}{2r+2}}}\right)^{1/2}\right).$$

Since

$$f(x) = \left(\frac{\log(e\kappa^2 x)}{x^{\frac{\epsilon}{2(2r+2)}}}\right)^{1/2}\left(1 + \left(\frac{\log(e\kappa^2 x)}{x^{\frac{\epsilon}{2r+2}}}\right)^{1/2}\right)$$

is continuous on $[1, \infty)$ and vanishes at ∞, $f(x) \leq C(\kappa, r, \epsilon)$ holds for some $C(\kappa, r, \epsilon)$ which only depends on κ, r and ϵ. Then

$$\mathcal{B} \leq C(\kappa, r, \epsilon)\frac{1}{N^{\frac{\epsilon}{4(2r+2)}}}.$$

Also, since $0 < \mathcal{B} \leq C(\kappa, r, \epsilon)$ and

$$g(x) = \frac{1}{x^\beta}\exp\left(-\frac{1}{4(\kappa^2+1)x}\right)$$

is continuous on $(0, C(\kappa, r, \epsilon)]$ and vanishes at 0^+,

$$\frac{1}{\mathcal{B}^\beta}\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \leq C'(\kappa, r, \epsilon, \beta)$$

for some $C'(\kappa, r, \epsilon, \beta)$ which only depends on κ, r, ϵ and β when $\beta > 0$ is a fixed constant. Taking $\beta = 1 + \frac{8}{\epsilon}(2r+2)$, we find that

$$\begin{aligned}\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) &\leq \mathcal{B}^{1+\frac{8(2r+2)}{\epsilon}} \cdot C'\left(\kappa, r, \epsilon, 1 + \frac{8(2r+2)}{\epsilon}\right) \\ &\leq C(\kappa, r, \epsilon)^{\frac{8(2r+2)}{\epsilon}} C'\left(\kappa, r, \epsilon, 1 + \frac{8(2r+2)}{\epsilon}\right) \cdot \mathcal{B} \cdot \frac{1}{N^2}.\end{aligned}$$

Hence,

$$\begin{aligned}&\left(1 + \frac{\kappa^2}{\lambda}\right)^2 \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \cdot \left(1 + \frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \\ &\leq (1+\kappa^2)^2 N^2 C(\kappa, r, \epsilon)^{\frac{8(2r+2)}{\epsilon}} C'\left(\kappa, r, \epsilon, 1 + \frac{8(2r+2)}{\epsilon}\right) \cdot \mathcal{B} \cdot \frac{1}{N^2} \cdot \left(1 + \frac{1}{4(\kappa^2+1)\mathcal{B}}\right) = O(1).\end{aligned}$$

As a consequence, by Theorem C.1, we get

$$\begin{aligned}\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(\bar{f}_{D,\lambda} - f_0)\|^2_{L^2_{\rho_{\mathbf{x}}}} &= O\left(\lambda^{2r+1} + \frac{1}{\lambda mN} + \lambda^{2r+1}\mathcal{B}^2\right) \\ &\quad + O(1) \cdot \left(\lambda^{2r+1}\mathcal{B}^2 + \frac{1}{\lambda mN}\right) \cdot \left(1 + \frac{\kappa^2}{\lambda}\right)^2 \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \cdot \left(1 + \frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \\ &= O\left(\lambda^{2r+1} + \frac{1}{\lambda mN}\right) = O\left((mN)^{-\frac{2r+1}{2r+2}} + (mN)^{-\frac{2r+1}{2r+2}}\right) = O\left((mN)^{-\frac{2r+1}{2r+2}}\right).\end{aligned}$$

□

Remark C.3. In the proof of Corollary C.2, $m \leq N^{2r+1}$ is not sufficient to derive the same result since this condition does not guarantee the boundedness of $\mathcal{B}$.

C.2. Proof of Theorem 5.2

Define an affine operator $H : \mathbb{H}_k \to \mathbb{H}_k$ by

$$Hg = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,X_p} g)$$

which is a unique solution of

$$\operatorname*{argmin}_{h\in\mathbb{H}_k} J[h] = \alpha \cdot \frac{1}{N_1}\sum_{i=1}^{N_1}(h(\mathbf{x}_1^i) - y_1^i)^2 + (1-\alpha)\cdot\frac{1}{N_2}\sum_{i=1}^{N_2}(h(\mathbf{x}_2^i) - g(\mathbf{x}_2^i))^2 + \lambda\|h\|_{\mathbb{H}_k}^2.$$

Therefore, the limiting regressor of the local model f_1 is $H^\infty f_{D_1,\lambda} = \lim_{t\to\infty} H^t f_{D_1,\lambda}$ after infinitely many iterations in Algorithm 1. Now, we prove the following theorem which is a more general statement of Theorem 5.2.

Theorem C.4. *For any initial point $h_0 \in \mathbb{H}_k$, $H^t h_0$ converges to a unique fixed point $f_{D_1,\lambda/\alpha}$ of H as $t \to \infty$.*

Proof of Theorem C.4. Define another operator $\bar{H} : \mathbb{H}_k \to \mathbb{H}_k$ by

$$\bar{H}g = L_{k,X_p}^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,X_p}^{1/2} g).$$

Observe that

$$\|\bar{H}g_1 - \bar{H}g_2\|_{\mathbb{H}_k} \le \left\|(1-\alpha)L_{k,X_p}^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}L_{k,X_p}^{1/2}\right\| \|g_1 - g_2\|_{\mathbb{H}_k}.$$

Since

$$(1-\alpha)L_{k,X_p} < \alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I$$

and $(1-\alpha)L_{k,X_p}$ is of finite rank which implies it is compact, we have

$$\left\|(1-\alpha)L_{k,X_p}^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}L_{k,X_p}^{1/2}\right\| < 1$$

by Lemma D.2. Thus, $\bar{H}$ is a η-contraction where $\eta \in (0,1)$. By the Banach fixed point theorem (Banach, 1922), $\bar{H}^t h_0$ converges to a unique fixed point $\bar{g}^*$ of $\bar{H}$ as $t \to \infty$. Define operators

$$A = L_{k,X_p}^{1/2} \quad \text{and} \quad B = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,X_p}^{1/2}\cdot).$$

Then $Hh = BAh$ and $\bar{H}h = ABh$ for any $h \in \mathbb{H}_k$. From the definition of $\bar{g}^*$,

$$B\bar{g}^* = B\bar{H}\bar{g}^* = BAB\bar{g}^* = HB\bar{g}^*$$

which implies that

$$B\bar{g}^* = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,X_p}^{1/2}\bar{g}^*)$$

is a fixed point of H. Let g^* be a fixed point of H, i.e., g^* satisfies

$$Hg^* = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,X_p} g^*) = g^*.$$

Then,

$$\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)L_{k,X_p} g^* = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)g^*,$$

i.e.,

$$g^* = (L_{k,X_1} + \lambda/\alpha I)^{-1} S_{D_1}^\top \mathbf{y}_1 = f_{D_1,\lambda/\alpha}.$$

Hence, H has a unique a fixed point and so $H^t h_0 = B\bar{H}^{t-1}Ah_0$ converges to $B\bar{g}^* = g^* = f_{D_1,\lambda/\alpha}$ as $t \to \infty$. □

Remark C.5. Note that H may not be a contraction since

$$\|(A + B + I)^{-1}B\| < 1$$

may not be true even if operators A and B are positive compact operators on a Hilbert space. So we have to follow the above argument.

Algorithm 2 KRR with iterative De-regularized Ensemble Distillation in FL

Input: hyperparameters $\alpha \in (0,1)$, $\lambda > 0$, $\lambda_0 \geq 0$ and $t \in \mathbb{N}$
Output: Trained model f_j, $j = 1, \cdots, m$
Pretrain: For $j = 1, \cdots, m$, client j trains its model f_j using the loss function

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}_j^i) - y_j^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

Each client downloads the unlabeled public dataset $D_p(\mathbf{x})$.
for $t_0 = 1, \cdots, t$ **do**
  For $j = 1, \cdots, m$, client j predicts on $D_p(\mathbf{x})$ and upload the prediction $\tilde{\mathbf{y}}_p^j$ to server.
  The server computes an updated consensus

$$\tilde{\mathbf{y}}_p = \frac{1}{m} \sum_{j=1}^{m} \tilde{\mathbf{y}}_p^j.$$

  if $t_0 \neq t$ **then**
    The server applies the de-regularization trick to $\tilde{\mathbf{y}}_p$:

$$\tilde{\mathbf{y}}_p = (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} \tilde{\mathbf{y}}_p.$$

  end if
  Each client downloads the ensemble prediction $\tilde{\mathbf{y}}_p$.
  For $j = 1, \cdots, m$ client j updates its model f_j using the loss function

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}_j^i) - y_j^i)^2 + (1-\alpha) \cdot \frac{1}{N_p} \sum_{i=1}^{N_p} (h(\mathbf{x}_p^i) - (\tilde{\mathbf{y}}_p)^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

end for

C.3. Algorithm on KRR with Iterative De-regularized Ensemble Distillation in Federated Learning Setting

Applying the de-regularization trick except the last ensemble distillation step, we provide Algorithm 2.

C.4. Proof of Theorem 5.3

Similarly as in Appendix C.2, define an affine operator $T : \mathbb{H}_k \to \mathbb{H}_k$ by

$$Tg = (\alpha L_{k,X_1} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} (\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha) S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p} g).$$

Then the limiting regressor of the local model f_1 is $T^\infty f_{D_1,\lambda} = \lim_{t\to\infty} T^t f_{D_1,\lambda}$ after infinitely many ensemble distillation steps with the de-regularization trick. As a result, the final local model is given by

$$(\alpha L_{k,X_1} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} (\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha) L_{k,X_p} T^\infty f_{D_1,\lambda}).$$

We deal with $T^\infty f_{D_1,\lambda}$ in this subsection to obtain some motivations. Before we prove Theorem 5.3, we provide the following lemma.

Lemma C.6. *Assume $K_{X_p,X_p} > 0$. Then,*

$$S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p} = L_{k,X_p} + \lambda_0 P_{D_p}$$

holds where P_{D_p} is the orthogonal projection onto a subspace

$$\textit{span}\,(k(\mathbf{x}, \cdot) : \mathbf{x} \in D_p(\mathbf{x}))$$

of the reproducing kernel Hilbert space $\mathbb{H}_k$.

Proof of Lemma C.6. Note that

$$(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top \mathbf{v} = \begin{bmatrix} k(\cdot, \mathbf{x}_p^1) & \cdots & k(\cdot, \mathbf{x}_p^{N_p}) \end{bmatrix} (K_{X_p,X_p} + N_p \lambda_0 I)^{-1} \mathbf{v}$$

and so

$$S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top \mathbf{v} = K_{X_p,X_p}(K_{X_p,X_p} + N_p \lambda_0 I)^{-1} \mathbf{v}.$$

We thus obtain

$$\begin{aligned}
(L_{k,X_p} - S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p}) h &= S_{D_p}^\top (I - (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1}) S_{D_p} h \\
&= S_{D_p}^\top \left(-N_p \lambda_0 K_{X_p,X_p}^{-1} \begin{bmatrix} h(\mathbf{x}_p^1) \\ \vdots \\ h(\mathbf{x}_p^{N_p}) \end{bmatrix} \right) \\
&= -\lambda_0 \begin{bmatrix} k(\cdot, \mathbf{x}_p^1) & \cdots & k(\cdot, \mathbf{x}_p^{N_p}) \end{bmatrix} K_{X_p,X_p}^{-1} \begin{bmatrix} h(\mathbf{x}_p^1) \\ \vdots \\ h(\mathbf{x}_p^{N_p}) \end{bmatrix}.
\end{aligned}$$

On the other hand, from

$$S_{D_p} S_{D_p}^\top \mathbf{v} = S_{D_p} \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{v}_i k(\mathbf{x}_p^i, \cdot) \right) = \frac{1}{N_p} K_{X_p,X_p} \mathbf{v},$$

we have

$$S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p} h = S_{D_p}^\top \left(N_p K_{X_p,X_p}^{-1} \begin{bmatrix} h(\mathbf{x}_p^1) \\ \vdots \\ h(\mathbf{x}_p^{N_p}) \end{bmatrix} \right) = \begin{bmatrix} k(\cdot, \mathbf{x}_p^1) \cdots & k(\cdot, \mathbf{x}_p^{N_p}) \end{bmatrix} K_{X_p,X_p}^{-1} \begin{bmatrix} h(\mathbf{x}_p^1) \\ \vdots \\ h(\mathbf{x}_p^{N_p}) \end{bmatrix}.$$

Therefore,

$$L_{k,X_p} - S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p} = -\lambda_0 S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p},$$

i.e.,

$$S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p} = L_{k,X_p} + \lambda_0 S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p}.$$

Set $P = S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p}$. Then we observe that

(i) P is idempotent since

$$P^2 = S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p} S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p} = S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p} = P.$$

(ii) P is symmetric since

$$\begin{aligned}
P^\top &= (S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p})^\top = S_{D_p}^\top ((S_{D_p} S_{D_p}^\top)^{-1})^\top S_{D_p} \\
&= S_{D_p}^\top ((S_{D_p} S_{D_p}^\top)^\top)^{-1} S_{D_p} = S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p} = P.
\end{aligned}$$

(iii) The range of P is span $(k(\mathbf{x}, \cdot) : \mathbf{x} \in D_p(\mathbf{x}))$ since the range of $S_{D_p}^\top$ is contained in span $(k(\mathbf{x}, \cdot) : \mathbf{x} \in D_p(\mathbf{x}))$ and for any $u(\cdot) \in$ span $(k(\mathbf{x}, \cdot) : \mathbf{x} \in D_p(\mathbf{x}))$ there exists $\mathbf{v} \in \mathbb{R}^{N_p}$ such that $u = S_{D_p}^\top \mathbf{v}$ which satisfies

$$P S_{D_p}^\top \mathbf{v} = S_{D_p}^\top (S_{D_p} S_{D_p}^\top)^{-1} S_{D_p} S_{D_p}^\top \mathbf{v} = S_{D_p}^\top \mathbf{v} = u.$$

By Proposition 3.3 of chapter 2 in Conway (2019), P is the orthogonal projection of $\mathbb{H}_k$ onto span $(k(\mathbf{x}, \cdot) : \mathbf{x} \in D_p(\mathbf{x}))$. □

From the above lemma, we can applied the Banach fixed point theorem (Banach, 1922) again as before.

Lemma C.7. *Assume $(1-\alpha)\lambda_0 < \lambda$ and $K_{X_p,X_p} > 0$. Define an operator $\bar{T} : \mathbb{H}_k \to \mathbb{H}_k$ as*

$$\bar{T}g = (S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p})^{1/2} \\ (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)(S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p})^{1/2} g).$$

Then $\bar{T}$ is a η-contraction where $\eta \in (0,1)$. Thus, $\bar{T}^t g$ converges to a unique fixed point $\tilde{g}^*$ of $\bar{T}$ as $t \to \infty$.

Proof of Lemma C.7. By Lemma C.6,

$$\bar{T}g = (L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2} g).$$

Since P_{D_p} is a projection, $0 \le P_{D_p} \le I$ and so

$$L_{k,X_p} \le L_{k,X_p} + \lambda_0 P_{D_p} \le L_{k,X_p} + \lambda_0 I.$$

Thus,

$$0 \le (1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p}) < \alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I. \tag{21}$$

Since $(1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})$ is of finite rank which implies it is compact, by Lemma D.2,

$$\|(1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}\| < 1. \tag{22}$$

Hence, from the submultiplicativity of the operator norm, we have

$$\|\bar{T}g_1 - \bar{T}g_2\|_{\mathbb{H}_k} = \|(1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}(g_1 - g_2)\|_{\mathbb{H}_k} \\ \le \|(1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}(\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}\|\|g_1 - g_2\|_{\mathbb{H}_k},$$

i.e., $\bar{T}$ is a η-contraction where $\eta < 1$. Applying the Banach fixed point theorem (Banach, 1922), we are done. □

Similarly as in Appendix C.2, we prove the following general statement. Obviously, the following theorem implies Theorem 5.3.

Theorem C.8. *Assume $\lambda_0 = \lambda$ and $K_{X_p,X_p} > 0$. For any $h_0 \in \mathbb{H}_k$, $T^t h_0$ converges to a unique fixed point*

$$\left(L_{k,X_1} + \lambda I + \frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp\right)^{-1} S_{D_1}^\top \mathbf{y}_1$$

of T as $t \to \infty$.

Proof of Theorem C.8. Let

$$A = (L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}, B = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})^{1/2}\cdot).$$

Then $Th = BAh$ and $\bar{T}h = ABh$ for any $h \in \mathbb{H}_k$. Let $\bar{g}^*$ be a unique fixed point of $\bar{T}$ where the existence and the uniqueness of the fixed point of $\bar{T}$ follow from Lemma C.7. From the definition,

$$B\bar{g}^* = B\bar{T}\bar{g}^* = BAB\bar{g}^* = TB\bar{g}^*$$

holds, i.e., $B\bar{g}^*$ is a fixed point of T. Let g^* be a fixed point of T. From

$$Tg^* = (\alpha L_{k,X_1} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_1}^\top \mathbf{y}_1 + (1-\alpha)(L_{k,X_p} + \lambda_0 P_{D_p})g^*) = g^*,$$

we obtain

$$(\alpha L_{k,X_1} + \lambda I - (1-\alpha)\lambda_0 P_{D_p})g^* = \alpha S_{D_1}^\top \mathbf{y}_1.$$

Set $\lambda_0 = \lambda$. Then

$$\lambda I - (1-\alpha)\lambda_0 P_{D_p} = \alpha\lambda I + (1-\alpha)\lambda(I - P_{D_p}).$$

By Proposition 3.2 of chapter 2 in Conway (2019), $I - P_{D_p} = P_{D_p}^\perp$. Therefore,

$$g^* = \left(L_{k,X_1} + \lambda I + \frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp\right)^{-1} S_{D_1}^\top \mathbf{y}_1.$$

In particular, a fixed point of T is unique. Consequently, $T^t h_0 = B\bar{T}^{t-1}Ah_0$ converges to g^* as $t \to \infty$ by Lemma C.7. □

Remark C.9. Without the de-regularization trick, the fixed point of H is

$$\left(L_{k,X_1} + \frac{\lambda}{\alpha} I\right)^{-1} S_{D_1}^\top \mathbf{y}_1.$$

With the de-regularization trick, the fixed point of T is

$$\left(L_{k,X_1} + \lambda I + \frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp\right)^{-1} S_{D_1}^\top \mathbf{y}_1.$$

Thus, we can see that the de-regularization replaces $\frac{1-\alpha}{\alpha}\lambda I$ by $\frac{1-\alpha}{\alpha}\lambda P_{D_p(\mathbf{x})}^\perp$.

C.5. Proof of Theorem 5.4

Proof of Theorem 5.4. Using the formula

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}, \tag{23}$$

we have

$$\tilde{f}_{D,\lambda} - f_{D_1,\lambda} = -\left(L_{k,X_1} + \lambda I + \frac{1-\alpha}{\alpha}\lambda(I - P_{D_p(\mathbf{x})})\right)^{-1} \left(\frac{1-\alpha}{\alpha}\lambda(I - P_{D_p(\mathbf{x})})\right) (L_{k,X_1} + \lambda I)^{-1} S_{D_1}^\top \mathbf{y}_1.$$

By Theorem 11 in Aydın & Gheondea (2021),

$$(I - P_{D_p(\mathbf{x})}) (L_{k,X_1} + \lambda I)^{-1} S_{D_1}^\top \mathbf{y}_1 \to 0$$

almost surely as $N_p \to \infty$. Therefore, by Lemma D.1,

$$\begin{aligned}
\|\tilde{f}_{D,\lambda} - f_{D_1,\lambda}\|_{\mathbb{H}_k} &\le \left\|\left(L_{k,X_1} + \lambda I + \frac{1-\alpha}{\alpha}\lambda(I - P_{D_p(\mathbf{x})})\right)^{-1}\right\| \cdot \frac{1-\alpha}{\alpha}\lambda \cdot \left\|(I - P_{D_p(\mathbf{x})}) (L_{k,X_1} + \lambda I)^{-1} S_{D_1}^\top \mathbf{y}_1\right\|_{\mathbb{H}_k} \\
&\le \frac{1-\alpha}{\alpha} \left\|(I - P_{D_p(\mathbf{x})}) (L_{k,X_1} + \lambda I)^{-1} S_{D_1}^\top \mathbf{y}_1\right\|_{\mathbb{H}_k} \to 0
\end{aligned}$$

almost surely as $N_p \to \infty$. □

Remark C.10. To apply the result provided in Aydın & Gheondea (2021), we assume that the density $\rho_{\mathbf{x}}$ is strictly positive on any non-empty open subset of $\mathcal{X}$ in this case. In Section 5.4, we assume Assumption 3.2 with $0 < r \le \frac{1}{2}$ instead of this condition to control the error.

C.6. Proof of Theorem 5.5

In this subsection, we assume $\lambda_0 = \lambda$ and $K_{X_p,X_p} > 0$. Similarly as before, we define an affine operator $T : \mathbb{H}_k \to \mathbb{H}_k$ by

$$\begin{aligned}
Tg &= \sum_{i=1}^m \frac{1}{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p} g) \\
&= \sum_{i=1}^m \frac{1}{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})g)
\end{aligned} \tag{24}$$

where the second equality follows from Lemma C.6. Also, define $\bar{T} : \mathbb{H}_k \to \mathbb{H}_k$ as

$$\bar{T}g = \sum_{i=1}^m \frac{1}{m}(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2} g).$$

Here, $P_{D_p(\mathbf{x})}$ is the orthogonal projection onto a subspace

$$\text{span}\,(k(\mathbf{x}, \cdot) : \mathbf{x} \in D_p(\mathbf{x}))$$

of the reproducing kernel Hilbert space $\mathbb{H}_k$. By the submultiplicativity of the operator norm, the triangle inequality, and (22), we have

$$\|\bar{T}g_1 - \bar{T}g_2\|_{\mathbb{H}_k} \leq \frac{1}{m}\sum_{i=1}^{m} \|(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(1-\alpha)(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}\| \cdot \|g_1 - g_2\|_{\mathbb{H}_k}$$

and

$$\frac{1}{m}\sum_{i=1}^{m} \|(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(1-\alpha)(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}\| < 1,$$

i.e., $\bar{T}$ is a η-contraction where $\eta < 1$. Set

$$A = (L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}, B = \sum_{i=1}^{m} \frac{1}{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)(L_{k,X_p} + \lambda P_{D_p(\mathbf{x})})^{1/2}\cdot).$$

Then $Th = BAh$ and $\bar{T}h = ABh$ for any $h \in \mathbb{H}_k$. Let $\bar{g}^*$ be a unique fixed point of $\bar{T}$ where the existence and the uniqueness of a fixed point of $\bar{T}$ follows from the Banach fixed point theorem (Banach, 1922). Then, for any $h_0 \in \mathbb{H}_k$, $T^t h_0 = B\bar{T}^{t-1}Ah_0$ converges to $B\bar{g}^*$. The following lemma is regarding the computation of $B\bar{g}^*$.

Lemma C.11. *With the same notation as given above,*

$$B\bar{g}^* = \alpha \cdot \frac{1}{m}\sum_{i=1}^{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i + \left(\frac{1}{m}\sum_{i=1}^{m}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} U^{1/2}\right)$$
$$\alpha\left(I - \frac{1}{m}\sum_{i=1}^{m} U^{1/2}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}U^{1/2}\right)^{-1} U^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i\right)$$

where

$$U = L_{k,X_p} + \lambda P_{D_p(\mathbf{x})} = S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} S_{D_p}.$$

Proof. From the definition of $\bar{g}^*$, we have

$$\bar{g}^* = \sum_{i=1}^{m} \frac{1}{m} U^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)U^{1/2}\bar{g}^*).$$

Note that, from (21),

$$0 \leq \frac{1}{m}\sum_{i=1}^{m} U^{1/2}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}U^{1/2} < I$$

which implies that $\left(I - \frac{1}{m}\sum_{i=1}^{m} U^{1/2}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}U^{1/2}\right)$ is invertible. Hence,

$$\bar{g}^* = \alpha\left(I - \frac{1}{m}\sum_{i=1}^{m} U^{1/2}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}U^{1/2}\right)^{-1} U^{1/2}$$
$$\left(\frac{1}{m}\sum_{i=1}^{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i\right). \tag{25}$$

Plugging (25) into $B\bar{g}^*$ yields

$$B\bar{g}^* = \sum_{i=1}^{m} \frac{1}{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha)U^{1/2}\bar{g}^*)$$
$$= \alpha \cdot \frac{1}{m}\sum_{i=1}^{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i + \left(\frac{1}{m}\sum_{i=1}^{m}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} U^{1/2}\right)$$
$$\alpha\left(I - \frac{1}{m}\sum_{i=1}^{m} U^{1/2}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}U^{1/2}\right)^{-1} U^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i\right).$$

□

Since Algorithm 2 conducts knowledge distillation on the public dataset $D_p(\mathbf{x})$ using $B\bar{g}^*$ in the last step, it suffices to evaluate the performance of $B\bar{g}^*$ on $D_p(\mathbf{x})$. To this end, we need two additional lemmas: Lemma C.12 and Lemma C.13. Lemma C.12 can be viewed as an extension of the distributed semi-supervised kernel ridge regression (Chang et al., 2017).

Lemma C.12. *Assume $m \geq 2$, Assumption 3.1 and 3.2. Let*

$$f^s_{D,\lambda} = \frac{1}{m}\sum_{i=1}^{m}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} S^\top_{D_i}\mathbf{y}_i.$$

Then,

$$\|S_{D_p}(f^s_{D,\lambda} - f_0)\|_2 \leq \Lambda + \frac{1}{m}\sum_{i=1}^{m}\Lambda_i$$

where Λ and Λ_i are random variables for $i = 1, \cdots, m$ such that $\Lambda \leq \tilde{\Lambda}(\log(6/\delta))^{5/4}$ with confidence at least $1-\delta$ and each $\Lambda_i \leq \tilde{\Lambda}_i(\log(6/\delta))^{3/2}$ with confidence at least $1-\delta$. Here,

$$\tilde{\Lambda} = \frac{1}{1-\alpha}\cdot\left(1 + \frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}\left(\frac{2\kappa(M+\gamma)}{\sqrt{\lambda mN}} + \frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda mN}}\right) + \frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda N_p}}$$
$$+ \left(1 + \frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}\cdot\lambda^{1/2+r}\|g_0\|_{\mathbb{H}_k}$$

and

$$\tilde{\Lambda}_i = \frac{\alpha}{(1-\alpha)\sqrt{\lambda}}\cdot\left(\frac{2\kappa(M+\gamma)}{\lambda\sqrt{N}} + \frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\lambda\sqrt{N}} + \left(1 + \frac{1-\alpha}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)\|f_0\|_{\mathbb{H}_k}\right)\frac{2\sqrt{2}\kappa^2}{\sqrt{N}}.$$

In particular, under the assumption that $N_p \geq (m-1)N$, with $\alpha = 1/m$ and $\lambda = (mN)^{-\frac{1}{2r+2}}$, we have $\tilde{\Lambda} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$ and $\tilde{\Lambda}_i = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$ for $i = 1, \cdots, m$, which implies that

$$\mathbf{E}\|S_{D_p}(f^s_{D,\lambda} - f_0)\|_2 = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

Proof of Lemma C.12. Using (10) and Lemma D.3, we obtain

$$\|S_{D_p}(f^s_{D,\lambda} - f_0)\|_2 = \left\|L^{1/2}_{k,X_p}(f^s_{D,\lambda} - f_0)\right\|_{\mathbb{H}_k} \leq \left\|(L_{k,X_p} + \lambda I)^{1/2}(f^s_{D,\lambda} - f_0)\right\|_{\mathbb{H}_k}.$$

Define

$$\tilde{f}^s_{D,\lambda} = (\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}S^\top_D\mathbf{y} = \frac{1}{m}\sum_{i=1}^{m}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}S^\top_{D_i}\mathbf{y}_i$$

where $\mathbf{y} = [\mathbf{y}_1^\top, \cdots, \mathbf{y}_m^\top]^\top$. By the triangle inequality,

$$\left\|(L_{k,X_p} + \lambda I)^{1/2}(f^s_{D,\lambda} - f_0)\right\|_{\mathbb{H}_k} \leq \left\|(L_{k,X_p} + \lambda I)^{1/2}(\tilde{f}^s_{D,\lambda} - f_0)\right\|_{\mathbb{H}_k} + \left\|(L_{k,X_p} + \lambda I)^{1/2}(f^s_{D,\lambda} - \tilde{f}^s_{D,\lambda})\right\|_{\mathbb{H}_k}. \quad (26)$$

First, we bound the first term in (26). Note that

$$\begin{aligned}(L_{k,X_p} + \lambda I)^{1/2}(\tilde{f}^s_{D,\lambda} - f_0) &= (L_{k,X_p} + \lambda I)^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(S^\top_D\mathbf{y} - L_k f_0)\\ &+ (L_{k,X_p} + \lambda I)^{1/2}((\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1} - (L_k + \lambda I)^{-1})L_k f_0\\ &+ (L_{k,X_p} + \lambda I)^{1/2}(f_\lambda - f_0). \quad (27)\end{aligned}$$

Define

$$\begin{aligned}\mathcal{Q}_p &= \|(L_{k,X_p}+\lambda I)^{-1/2}(L_k+\lambda I)^{1/2}\|,\\ \tilde{\mathcal{Q}}_p &= \|(L_{k,X_p}+\lambda I)^{1/2}(L_k+\lambda I)^{-1/2}\|,\\ \mathcal{P} &= \|(L_k+\lambda I)^{-1/2}(S_D^\top \mathbf{y} - L_{k,X} f_0)\|_{\mathbb{H}_k},\\ \mathcal{S} &= \|L_{k,X}-L_k\|,\end{aligned}$$

and

$$\mathcal{S}_p = \|L_{k,X_p}-L_k\|.$$

Note that, by Lemma D.2,

$$\begin{aligned}&\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I \geq (1-\alpha)(L_{k,X_p}+\lambda I)\\ &\quad\Rightarrow\quad \|(L_{k,X_p}+\lambda I)^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p}+\lambda I)^{-1}(L_{k,X_p}+\lambda I)^{1/2}\| \leq \frac{1}{1-\alpha}.\end{aligned} \tag{28}$$

Thus, the first term in (27) satisfies

$$\|(L_{k,X_p}+\lambda I)^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p}+\lambda I)^{-1}(S_D^\top \mathbf{y} - L_k f_0)\|_{\mathbb{H}_k} \leq \frac{1}{1-\alpha}\mathcal{Q}_p\left(\mathcal{P} + \frac{\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda}}\mathcal{S}\right).$$

Also, the third term in (27) satisfies

$$\|(L_{k,X_p}+\lambda I)^{1/2}(f_\lambda - f_0)\|_{\mathbb{H}_k} \leq \tilde{\mathcal{Q}}_p \|(L_k+\lambda I)^{1/2}(f_\lambda - f_0)\|_{\mathbb{H}_k}$$

and the second term in (27) satisfies

$$\begin{aligned}&\|(L_{k,X_p}+\lambda I)^{1/2}((\alpha L_k + (1-\alpha)L_{k,X_p}+\lambda I)^{-1} - (L_k+\lambda I)^{-1})L_k f_0\|_{\mathbb{H}_k}\\ &= \|(L_{k,X_p}+\lambda I)^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p}+\lambda I)^{-1}(1-\alpha)(L_k - L_{k,X_p})(L_k+\lambda I)^{-1}L_k f_0\|_{\mathbb{H}_k}\\ &\leq \|(L_{k,X_p}+\lambda I)^{-1/2}\| \cdot \mathcal{S}_p \cdot \|f_\lambda\|_{\mathbb{H}_k}\end{aligned}$$

by the submultiplicativity of the operator norm, (23), and (28). Combining them yields

$$\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{f}^s_{D,\lambda} - f_0)\|_{\mathbb{H}_k} \leq \frac{1}{1-\alpha}\mathcal{Q}_p\left(\mathcal{P} + \frac{\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda}}\mathcal{S}\right) + \frac{\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda}}\mathcal{S}_p + \tilde{\mathcal{Q}}_p\|(L_k+\lambda I)^{1/2}(f_\lambda - f_0)\|_{\mathbb{H}_k} \tag{29}$$

by Lemma D.1 and (19). From

$$\|(L_k+\lambda I)^{1/2}(f_\lambda - f_0)\|_{\mathbb{H}_k} = \|\lambda(L_k+\lambda I)^{-1/2} f_0\|_{\mathbb{H}_k} = \|\lambda(L_k+\lambda I)^{-1/2}L_k^r g_0\|_{\mathbb{H}_k} \leq \lambda^{1/2+r}\|g_0\|_{\mathbb{H}_k}$$

which follows from Lemma D.1 and the submultiplicativity of the operator norm, the inequality (29) becomes

$$\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{f}^s_{D,\lambda} - f_0)\|_{\mathbb{H}_k} \leq \frac{1}{1-\alpha}\mathcal{Q}_p\left(\mathcal{P} + \frac{\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda}}\mathcal{S}\right) + \frac{\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda}}\mathcal{S}_p + \tilde{\mathcal{Q}}_p\lambda^{1/2+r}\|g_0\|_{\mathbb{H}_k}.$$

Now, we derive PAC-bounds for $\mathcal{Q}_p$, $\tilde{\mathcal{Q}}_p$, $\mathcal{P}$, $\mathcal{S}$, and $\mathcal{S}_p$. By Lemma D.1, Lemma D.4, the triangle inequality, and the submultiplicativity of the operator norm, we have

$$\mathcal{Q}_p \leq \|(L_{k,X_p}+\lambda I)^{-1}(L_k+\lambda I)\|^{1/2} = \|I + (L_{k,X_p}+\lambda I)^{-1}(L_k - L_{k,X_p})\|^{1/2} \leq \left(1 + \frac{1}{\lambda}\|L_k - L_{k,X_p}\|\right)^{1/2} \tag{30}$$

and

$$\tilde{\mathcal{Q}}_p \leq \|(L_{k,X_p}+\lambda I)(L_k+\lambda I)^{-1}\|^{1/2} = \|I + (L_{k,X_p}-L_k)(L_{k,X_p}+\lambda I)^{-1}\|^{1/2} \leq \left(1 + \frac{1}{\lambda}\|L_k - L_{k,X_p}\|\right)^{1/2}.$$

By Lemma D.7 and Lemma D.8, with confidence at least $1-\delta$,

$$\|L_{k,X}-L_k\| \le \frac{2\sqrt{2}\kappa^2}{\sqrt{mN}}(\log(6/\delta))^{1/2}, \quad \|L_{k,X_p}-L_k\| \le \frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(6/\delta))^{1/2} \quad \text{and} \quad \mathcal{P} \le \frac{2\kappa(M+\gamma)}{\sqrt{\lambda mN}}\log(6/\delta)$$

where $\delta \in (0,1)$. Then

$$\left(1+\frac{1}{\lambda}\|L_k-L_{k,X_p}\|\right)^{1/2} \le \left(1+\frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}(\log(6/\delta))^{1/4}.$$

Therefore,

$$\begin{aligned}\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{f}^s_{D,\lambda}-f_0)\|_{\mathbb{H}_k} \le{}& \frac{1}{1-\alpha}\cdot\left(1+\frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}\left(\frac{2\kappa(M+\gamma)}{\sqrt{\lambda mN}}+\frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda mN}}\right)(\log(6/\delta))^{5/4}\\ &+\frac{\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda}}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(6/\delta))^{1/2}+\left(1+\frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}\cdot\lambda^{1/2+r}\|g_0\|_{\mathbb{H}_k}(\log(6/\delta))^{1/4}\end{aligned}$$

with confidence at least $1-\delta$. Define

$$\begin{aligned}\tilde{\Lambda} :={}& \frac{1}{1-\alpha}\cdot\left(1+\frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}\left(\frac{2\kappa(M+\gamma)}{\sqrt{\lambda mN}}+\frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda mN}}\right)\\ &+\frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\sqrt{\lambda N_p}}+\left(1+\frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)^{1/2}\cdot\lambda^{1/2+r}\|g_0\|_{\mathbb{H}_k}.\end{aligned}$$

Then

$$\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{f}^s_{D,\lambda}-f_0)\|_{\mathbb{H}_k} \le \tilde{\Lambda}(\log(6/\delta))^{5/4}$$

with confidence at least $1-\delta$. We next bound the second term $\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{f}^s_{D,\lambda}-f^s_{D,\lambda})\|_{\mathbb{H}_k}$ in (26). First, we bound the norm of

$$\tilde{f}^{s,i}_{D,\lambda} = (\alpha L_k+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}S^\top_{D_i}\mathbf{y}_i.$$

Observe that

$$\tilde{f}^{s,i}_{D,\lambda} = (\alpha L_k+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(S^\top_{D_i}\mathbf{y}_i-L_kf_0)+(\alpha L_k+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(L_k+\lambda I)(L_k+\lambda I)^{-1}L_kf_0.$$

Set

$$\mathcal{Q}'_p = \|(\alpha L_k+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(L_k+\lambda I)\|.$$

By Lemma D.1 and the submultiplicativity of the operator norm, we have

$$\mathcal{Q}'_p = \|I+(1-\alpha)(\alpha L_k+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(L_k-L_{k,X_p})\| \le 1+\frac{1-\alpha}{\lambda}\|L_k-L_{k,X_p}\|.$$

Again, applying Lemma D.1, the submultiplicativity of the operator norm, the triangle inequality, and (19), we obtain

$$\|\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k} \le \frac{1}{\lambda}\|S^\top_{D_i}\mathbf{y}_i-L_kf_0\|_{\mathbb{H}_k}+\left(1+\frac{1-\alpha}{\lambda}\|L_k-L_{k,X_p}\|\right)\|f_0\|_{\mathbb{H}_k}.$$

Now, by the submultiplicativity of the operator norm, the triangle inequality, and (23),

$$\begin{aligned}&\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{f}^s_{D,\lambda}-f^s_{D,\lambda})\|_{\mathbb{H}_k}\\ &\le \frac{1}{m}\sum_{i=1}^m\|(L_{k,X_p}+\lambda I)^{1/2}((\alpha L_k+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}-(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1})S^\top_{D_i}\mathbf{y}_i\|_{\mathbb{H}_k}\\ &\le \frac{1}{m}\sum_{i=1}^m\|(L_{k,X_p}+\lambda I)^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}\alpha(L_{k,X_i}-L_k)\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k}.\end{aligned}$$

Simliarly as in (28), we can easily see that

$$\|(L_{k,X_p}+\lambda I)^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(L_{k,X_p}+\lambda I)^{1/2}\| \le \frac{1}{1-\alpha}. \tag{31}$$

Then, applying (31), Lemma D.1, and the submultiplicativity of the operator norm yields

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}\|(L_{k,X_p}+\lambda I)^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}\alpha(L_{k,X_i}-L_k)\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k}\\ &\le \frac{1}{m}\sum_{i=1}^{m}\frac{\alpha}{(1-\alpha)\sqrt{\lambda}}\|L_{k,X_i}-L_k\|_{\mathbb{H}_k}\|\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k}.\end{aligned}$$

To find a PAC-bound of $\|L_{k,X_i}-L_k\|\|\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k}$, we use Lemma D.7 and Lemma D.8. Then, with confidence at least $1-\delta$,

$$\|L_{k,X_p}-L_k\| \le \frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(6/\delta))^{1/2}, \quad \|L_{k,X_i}-L_k\| \le \frac{2\sqrt{2}\kappa^2}{\sqrt{N}}(\log(6/\delta))^{1/2},$$

and

$$\|S^{\top}_{D_i}\mathbf{y}_i - L_{k,X_i}f_0\|_{\mathbb{H}_k} \le \frac{2\kappa(M+\gamma)}{\sqrt{N}}\log(6/\delta).$$

Therefore,

$$\|L_{k,X_i}-L_k\|\|\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k} \le \left(\frac{2\kappa(M+\gamma)}{\lambda\sqrt{N}}+\frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\lambda\sqrt{N}}+\left(1+\frac{1-\alpha}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)\|f_0\|_{\mathbb{H}_k}\right)\frac{2\sqrt{2}\kappa^2}{\sqrt{N}}(\log(6/\delta))^{3/2}$$

with confidence at least $1-\delta$. Define

$$\tilde{\Lambda}_i = \frac{\alpha}{(1-\alpha)\sqrt{\lambda}}\cdot\left(\frac{2\kappa(M+\gamma)}{\lambda\sqrt{N}}+\frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\lambda\sqrt{N}}+\left(1+\frac{1-\alpha}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right)\|f_0\|_{\mathbb{H}_k}\right)\frac{2\sqrt{2}\kappa^2}{\sqrt{N}}$$

for $i=1,\cdots,m$. Then

$$\frac{\alpha}{(1-\alpha)\sqrt{\lambda}}\|L_{k,X_i}-L_k\|_{\mathbb{H}_k}\|\tilde{f}^{s,i}_{D,\lambda}\|_{\mathbb{H}_k} \le \tilde{\Lambda}_i(\log(6/\delta))^{3/2}$$

with confidence at least $1-\delta$ for any $i=1,\cdots,m$. Now, we set $\alpha=1/m$ and $\lambda=(mN)^{-\frac{1}{2r+2}}$. We also assume that $N_p \ge (m-1)N$. Then

$$1 \le \left(1+\frac{1-\alpha}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right) \le \left(1+\frac{1}{\lambda}\cdot\frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}\right) \le (1+4\kappa^2)$$

and

$$\left(\frac{2\kappa(M+\gamma)}{\lambda\sqrt{N}}+\frac{2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{\lambda\sqrt{N}}\right) \le (2\kappa(M+\gamma)+2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k})\cdot m^{1/(2r+2)}.$$

Thus, we have

$$\tilde{\Lambda} \le 2(1+4\kappa^2)^{1/2}\cdot\frac{2\kappa(M+\gamma)+2\sqrt{2}\kappa^2\|f_0\|_{\mathbb{H}_k}}{(mN)^{(2r+1)/(4r+4)}}+\frac{4\kappa^2\|f_0\|_{\mathbb{H}_k}}{(mN)^{(2r+1)/(4r+4)}}+\frac{(1+4\sqrt{2}\kappa^2)^{1/2}\|g_0\|_{\mathbb{H}_k}}{(mN)^{(2r+1)/(4r+4)}} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$$

and

$$\tilde{\Lambda}_i \le 4\sqrt{2}\kappa^2\cdot(2\kappa(M+\gamma)+(1+(4+2\sqrt{2})\kappa^2)\|f_0\|_{\mathbb{H}_k})\cdot\frac{m^{1/(2r+2)}}{m^{1/2}}\cdot\frac{1}{(mN)^{(2r+1)/(4r+4)}} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

By Lemma D.9, we conclude that

$$\mathbf{E}\|S_{D_p}(f^s_{D,\lambda}-f_0)\|_2 \le \mathbf{E}\Lambda+\frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\Lambda_i = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

□

The second lemma is regarding the convergence rate of $\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k}$. Aydın & Gheondea (2021) prove that $\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} \to 0$ as $N_p \to \infty$ under the condition that the density $\rho_{\mathbf{x}}$ is strictly positive on any non-empty open subset of $\mathcal{X}$. However, they do not provide the convergence rate. The following lemma provides a convergence rate if we assume a regularity condition of f_0.

Lemma C.13. *Assume Assumption 3.2. Then,*

$$\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} \leq 2\lambda^r \|g_0\|_{\mathbb{H}_k}$$

with confidence at least $1 - 4\exp(-1/4(\kappa^2+1)\mathcal{B})$ *where*

$$\mathcal{B} = \frac{1 + \log \mathcal{N}(\lambda)}{\lambda N_p} + \sqrt{\frac{1 + \log \mathcal{N}(\lambda)}{\lambda N_p}}$$

for any $\lambda > 0$ *such that* $\lambda \in (0,1)$ *and* $\mathcal{N}(\lambda) \geq 1$. *Also,*

$$\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} \leq \|f_0\|_{\mathbb{H}_k}$$

almost surely. In particular, for any fixed $\epsilon \in (0,1)$,

$$\mathbf{E}\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} = O\left(N_p^{(-1+\epsilon)r}\right).$$

Proof of Lemma C.13. Note that

$$(L_{k,X_p} + \lambda I)^{-1} L_{k,X_p} f_0 = \mathbf{k}_{D_p(\mathbf{x})}^\top (N_p \lambda I + K_{X_p,X_p})^{-1} S_{D_p} f_0 \in \text{span}(k(\mathbf{x},\cdot) : \mathbf{x} \in D_p(\mathbf{x}))$$

for any $\lambda > 0$. Thus,

$$\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} = \min_{h \in \text{span}(k(\mathbf{x},\cdot):\mathbf{x}\in D_p(\mathbf{x}))} \|f_0 - h\|_{\mathbb{H}_k} \leq \|f_0 - (L_{k,X_p} + \lambda I)^{-1} L_{k,X_p} f_0\|_{\mathbb{H}_k}.$$

Hence,

$$\begin{aligned}\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} &\leq \|\lambda (L_{k,X_p} + \lambda I)^{-1} L_k^r g_0\|_{\mathbb{H}_k} \\ &\leq \lambda \|(L_k + \lambda I)^{-1/2}\| \|(L_k + \lambda I)^{1/2} (L_{k,X_p} + \lambda I)^{-1} (L_k + \lambda I)^{1/2}\| \|(L_k + \lambda I)^{-1/2} L_k^r\| \|g_0\|_{\mathbb{H}_k}\end{aligned}$$

by the submultiplicativity of the operator norm. Using Lemma D.1,

$$\begin{aligned}&\lambda \|(L_k + \lambda I)^{-1/2}\| \|(L_k + \lambda I)^{1/2} (L_{k,X_p} + \lambda I)^{-1} (L_k + \lambda I)^{1/2}\| \|(L_k + \lambda I)^{-1/2} L_k^r\| \|g_0\|_{\mathbb{H}_k} \\ &\leq \lambda^r \|(L_k + \lambda I)^{1/2} (L_{k,X_p} + \lambda I)^{-1} (L_k + \lambda I)^{1/2}\| \|g_0\|_{\mathbb{H}_k}.\end{aligned}$$

When we assume $\lambda \in (0,1)$ and $\mathcal{N}(\lambda) \geq 1$, by Lemma D.7, we have

$$\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} \leq 2\lambda^r \|g_0\|_{\mathbb{H}_k}$$

with confidence at least $1 - \delta$ where $4\exp(-1/4(\kappa^2+1)\mathcal{B}) \leq \delta < 1$. On the other hand,

$$\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} \leq \|f_0\|_{\mathbb{H}_k}$$

almost surely since $\|I - P_{D_p}\| \leq 1$. Combining them, we have

$$\mathbf{E}\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} \leq 2\lambda^r \|g_0\|_{\mathbb{H}_k} + 4\|f_0\|_{\mathbb{H}_k} \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right).$$

Set $\lambda = N_p^{-1+\epsilon}$ for a fixed $\epsilon > 0$. Since $\lambda \downarrow 0$ as $N_p \to \infty$, we may assume $\mathcal{N}(\lambda) \geq 1$. Then,

$$\mathcal{B} \leq \frac{1}{N_p^{\epsilon/4}} \left(\frac{\log(e\kappa^2 N_p)}{N_p^{\epsilon/2}}\right)^{1/2} \left(1 + \left(\frac{\log(e\kappa^2 N_p)}{N_p^{\epsilon}}\right)^{1/2}\right).$$

Since

$$f(x) = \left(\frac{\log(e\kappa^2 x)}{x^{\epsilon/2}}\right)^{1/2}\left(1+\left(\frac{\log(e\kappa^2 x)}{x^{\epsilon}}\right)^{1/2}\right)$$

is continuous on $[1,\infty)$ and vanishes at ∞, $f(x) \le C(\kappa,\epsilon)$ for some $C(\kappa,\epsilon) > 0$. Thus,

$$\mathcal{B} \le C(\kappa,\epsilon)\frac{1}{N_p^{\epsilon/4}}.$$

Set

$$g(x) = x^{-\beta}\exp\left(-\frac{1}{4(\kappa^2+1)x}\right).$$

Then g is continuous on $(0, C(\kappa,\epsilon)]$ and vanishes at 0^+ for any $\beta > 0$. Therefore, we know that $g(x) \le C'(\kappa,\epsilon,\beta)$ on $x \in (0, C(\kappa,\epsilon)]$ for some $C'(\kappa,\epsilon,\beta) > 0$. Hence, we have

$$\begin{aligned}
\mathbf{E}\|f_0 - P_{D_p} f_0\|_{\mathbb{H}_k} &\le 2\lambda^r\|g_0\|_{\mathbb{H}_k} + \|f_0\|_{\mathbb{H}_k}\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right)\\
&\le 2N_p^{(-1+\epsilon)r}\|g_0\|_{\mathbb{H}_k} + \|f_0\|_{\mathbb{H}_k}\mathcal{B}^{\frac{4(1-\epsilon)r}{\epsilon}}\cdot C'\left(\kappa,\epsilon,\frac{4(1-\epsilon)r}{\epsilon}\right)\\
&\le 2N_p^{(-1+\epsilon)r}\|g_0\|_{\mathbb{H}_k} + \|f_0\|_{\mathbb{H}_k}N_p^{(-1+\epsilon)r}C(\kappa,\epsilon)^{\frac{4(1-\epsilon)r}{\epsilon}}C'\left(\kappa,\epsilon,\frac{4(1-\epsilon)r}{\epsilon}\right)\\
&= O\left(N_p^{(-1+\epsilon)r}\right).
\end{aligned}$$

□

We are now ready to derive the main result in this subsection.

Theorem C.14. *Assume $m \ge 2$, $\lambda_0 = \lambda$, Assumption 3.1, and Assumption 3.2 with $0 < r \le \frac{1}{2}$. We further assume*

$$N_p \ge \max\left(\left(m^{\frac{3r+2}{2r^2+2r}}N^{\frac{1}{2r+2}}\right)^{1/(1-\epsilon)}, (m-1)N\right)$$

for some fixed $0 < \epsilon < \frac{1}{2}$. Let $g^* = T^\infty h_0$ for any $h_0 \in \mathbb{H}_k$ where the operator T is defined in (24). Then, with $\alpha = 1/m$ and $\lambda = (mN)^{-\frac{1}{2r+2}}$,

$$\mathbf{E}\|S_{D_p}(g^* - f_0)\|_2 = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

Proof of Theorem C.14. When m or N increases, λ decreases and N_p increases. Thus, we may assume $\mathcal{N}(\lambda) \ge 1$ and $\mathcal{N}(1/\sqrt{N_p}) \ge 1$. Set $U = S_{D_p}^\top(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^\top)^{-1}S_{D_p}$ and $V = L_{k,X_p} + \lambda I$. Note that V is invertible and

$$L_{k,X_p} \le U = L_{k,X_p} + \lambda P_{D_p} \le V = L_{k,X_p} + \lambda I.$$

Define

$$\tilde{g} = \frac{1}{m}\sum_{i=1}^m(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}Uf_0.$$

By the triangle inequality,

$$\|S_{D_p}(g^* - f_0)\|_2 \le \|S_{D_p}(g^* - \tilde{g})\|_2 + \|S_{D_p}(\tilde{g} - f_0)\|_2.$$

First, note that

$$\|S_{D_p}(\tilde{g} - f_0)\|_2 = \|L_{k,X_p}^{1/2}(\tilde{g} - f_0)\|_{\mathbb{H}_k} \le \|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{g} - f_0)\|_{\mathbb{H}_k}$$

by (10) and Lemma D.3. By the triangle inequality,

$$\begin{aligned}
&\|(L_{k,X_p}+\lambda I)^{1/2}(\tilde{g}-f_0)\|_{\mathbb{H}_k} \\
&\le \frac{1}{m}\sum_{i=1}^m \|V^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(L_{k,X_p}+\lambda P_{D_p}-\alpha L_{k,X_i}-(1-\alpha)L_{k,X_p}-\lambda I)f_0\|_{\mathbb{H}_k} \\
&\le \frac{1}{m}\sum_{i=1}^m \|V^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}\alpha(L_{k,X_p}-L_{k,X_i})f_0\|_{\mathbb{H}_k} \\
&\quad + \frac{1}{m}\sum_{i=1}^m \lambda\|V^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(I-P_{D_p})f_0\|_{\mathbb{H}_k}.
\end{aligned}$$

By the submultiplicativity of the operator norm, Lemma D.1, and (31),

$$\|V^{1/2}(\tilde{g}-f_0)\|_{\mathbb{H}_k} \le \frac{1}{m}\sum_{i=1}^m \left(\frac{\alpha}{(1-\alpha)\sqrt{\lambda}}\|L_{k,X_p}-L_{k,X_i}\|\|f_0\|_{\mathbb{H}_k}+\frac{\sqrt{\lambda}}{1-\alpha}\|(I-P_{D_p})f_0\|_{\mathbb{H}_k}\right). \tag{32}$$

Set $\Lambda_i^1 = \frac{\alpha}{(1-\alpha)\sqrt{\lambda}}\|L_{k,X_p}-L_{k,X_i}\|\|f_0\|_{\mathbb{H}_k}$. By the triangle inequality,

$$\|L_{k,X_p}-L_{k,X_i}\| \le \|L_{k,X_p}-L_k\|+\|L_k-L_{k,X_i}\|.$$

By Lemma D.7, with confidence at least $1-\delta$,

$$\|L_{k,X_i}-L_k\| \le \frac{2\sqrt{2}\kappa^2}{\sqrt{N}}(\log(4/\delta))^{1/2} \qquad \text{and} \qquad \|L_{k,X_p}-L_k\| \le \frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(4/\delta))^{1/2}.$$

Then, with confidence at least $1-\delta$,

$$\Lambda_i^1 \le \tilde{\Lambda}^1(\log(4/\delta))^{1/2} \tag{33}$$

where

$$\tilde{\Lambda}^1 := \frac{2\sqrt{2}\alpha\kappa^2\|f_0\|_{\mathbb{H}_k}}{(1-\alpha)\sqrt{\lambda}}\left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{N_p}}\right) = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

Also,

$$\mathbf{E}\left[\frac{\sqrt{\lambda}}{1-\alpha}\|(I-P_{D_p})f_0\|_{\mathbb{H}_k}\right] \le 2(mN)^{-\frac{1}{4r+4}}\cdot O\left(((m-1)N)^{-\frac{2r}{4r+4}}\right) = O\left((mN)^{-\frac{2r+1}{4r+4}}\right) \tag{34}$$

by Lemma C.13. On the other hand, by (10) and Lemma D.3,

$$\|S_{D_p}(g^*-\tilde{g})\|_2 = \|L_{k,X_p}^{1/2}(g^*-\tilde{g})\|_{\mathbb{H}_k} \le \|U^{1/2}(g^*-\tilde{g})\|_{\mathbb{H}_k}.$$

We can see that

$$\begin{aligned}
U^{1/2}\tilde{g} &= \frac{1}{m}\sum_{i=1}^m U^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}Uf_0 \\
&= \alpha\cdot\frac{1}{m}\sum_{i=1}^m U^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}Uf_0 + \left(\frac{1}{m}\sum_{i=1}^m(1-\alpha)U^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right) \\
&\quad \alpha\left(I-\frac{1}{m}\sum_{i=1}^m U^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right)^{-1} \\
&\quad \frac{1}{\alpha}\left(I-\frac{1}{m}\sum_{i=1}^m U^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right)U^{1/2}f_0 \\
&= \alpha\cdot\frac{1}{m}\sum_{i=1}^m U^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}Uf_0 + \left(\frac{1}{m}\sum_{i=1}^m(1-\alpha)U^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right) \\
&\quad \alpha\left(I-\frac{1}{m}\sum_{i=1}^m U^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right)^{-1}U^{1/2} \\
&\quad \frac{1}{\alpha}\left(I-\frac{1}{m}\sum_{i=1}^m(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U\right)f_0.
\end{aligned}$$

Thus, by the triangle inequality, the submultiplicativity of the operator norm, and Lemma C.11, we have

$$
\begin{aligned}
&\|U^{1/2}(g^* - \tilde{g})\|_{\mathbb{H}_k} \\
&= \alpha \cdot \left\| U^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i - \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U f_0 \right) \right\|_{\mathbb{H}_k} \\
&+ \left\| \left(\frac{1}{m} \sum_{i=1}^{m} U^{1/2} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U^{1/2} \right) \alpha \left(I - \frac{1}{m} \sum_{i=1}^{m} U^{1/2} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U^{1/2} \right)^{-1} \right\| \\
&\quad \cdot \left\| U^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i - \frac{1}{\alpha} \left(I - \frac{1}{m} \sum_{i=1}^{m} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U \right) f_0 \right) \right\|_{\mathbb{H}_k}.
\end{aligned}
$$

First, we bound the first term. By the triangle inequality and Lemma D.3,

$$
\begin{aligned}
&\left\| U^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i - \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U f_0 \right) \right\|_{\mathbb{H}_k} \\
&\le \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i - \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U f_0 \right) \right\|_{\mathbb{H}_k} \\
&\le \left\| V^{1/2} \left(f_0 - \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i \right) \right\|_{\mathbb{H}_k} + \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U f_0 - f_0 \right) \right\|_{\mathbb{H}_k}.
\end{aligned}
$$

From (32), we know

$$
\left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U f_0 - f_0 \right) \right\|_{\mathbb{H}_k} \le \frac{1}{m} \sum_{i=1}^{m} \Lambda_i^1 + \frac{\sqrt{\lambda}}{1-\alpha} \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k}.
$$

Also, using the same argument as in the proof of Lemma C.12, it satisfies that

$$
\left\| V^{1/2} \left(f_0 - \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^\top \mathbf{y}_i \right) \right\|_{\mathbb{H}_k} \le \Lambda + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i \tag{35}
$$

where random variables Λ and $\Lambda_i (i = 1, \cdots, m)$ are given in Lemma C.12. In particular, $\Lambda \le \tilde{\Lambda} (\log(6/\delta))^{5/4}$ with confidence at least $1 - \delta$ and $\Lambda_i \le \tilde{\Lambda}_i (\log(6/\delta))^{3/2}$ with confidence at least $1 - \delta$ where $\tilde{\Lambda} = O\left((mN)^{-\frac{2r+1}{4r+4}} \right)$ and $\tilde{\Lambda}_i = O\left((mN)^{-\frac{2r+1}{4r+4}} \right)$. We next turn to derive an upper bound of the second term. To bound

$$
\left\| \left(\frac{1}{m} \sum_{i=1}^{m} U^{1/2} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U^{1/2} \right) \alpha \left(I - \frac{1}{m} \sum_{i=1}^{m} U^{1/2} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U^{1/2} \right)^{-1} \right\|,
$$

we use Lemma D.6. From

$$
0 \le U = L_{k,X_p} + \lambda P_{D_p} \le V = L_{k,X_p} + \lambda I < \frac{1}{1-\alpha} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I),
$$

we have

$$
(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} < V^{-1}
$$

for any $i = 1, \cdots, m$. Hence

$$
\frac{1}{m} \sum_{i=1}^{m} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} < V^{-1}
$$

and

$$
V < \left(\frac{1}{m} \sum_{i=1}^{m} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \right)^{-1}.
$$

Note that U is of finite rank which means that $U^{1/2}$ is of finite rank and so compact, and all of the above operators are positive. Applying Lemma D.6 and the above inequality yields

$$\begin{aligned}
&\left\|\left(\frac{1}{m}\sum_{i=1}^{m}U^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right)\alpha\left(I-\frac{1}{m}\sum_{i=1}^{m}U^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}U^{1/2}\right)^{-1}\right\| \\
&\leq\left\|\left(\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)\alpha\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}\right\|.
\end{aligned}$$

By the submultiplicativity of the operator norm,

$$\begin{aligned}
&\left\|\left(\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)\alpha\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}\right\| \\
&\leq\left\|\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right\|\left\|\alpha\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}\right\| \\
&\leq\left\|\alpha\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}\right\|.
\end{aligned}$$

Here, the second inequality follows from the fact that

$$\left\|\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right\|\leq\frac{1}{m}\sum_{i=1}^{m}\left\|V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right\|\leq 1.$$

which comes from the triangle inequality and (31). Set

$$A_1^{-1}=\alpha\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}.$$

On the other hand, since $A\mapsto A^{-1}$ is operator convex (Bhatia, 2013),

$$\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}\leq\frac{1}{m}\sum_{i=1}^{m}(I-V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2})^{-1}.$$

Since $\left(I-\frac{1}{m}\sum_{i=1}^{m}V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2}\right)^{-1}$ is positive,

$$\begin{aligned}
\left\|A_1^{-1}\right\|&\leq\left\|\alpha\cdot\frac{1}{m}\sum_{i=1}^{m}(I-V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2})^{-1}\right\| \\
&\leq\alpha\cdot\frac{1}{m}\sum_{i=1}^{m}\left\|(I-V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2})^{-1}\right\|.
\end{aligned}$$

Since $V^{1/2}$ is invertible, by Lemma D.5 and the submultiplicativity of the operator norm,

$$\begin{aligned}
&\alpha\left\|(I-V^{1/2}(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}V^{1/2})^{-1}\right\| \\
&=\alpha\left\|V^{-1/2}(V^{-1}-(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1})^{-1}V^{-1/2}\right\| \\
&\leq\alpha\left\|(V^{-1}-(1-\alpha)(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1})^{-1}V^{-1}\right\| \\
&=\alpha\left\|(I-(1-\alpha)V(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1})^{-1}\right\| \\
&=\alpha\left\|(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)(\alpha L_{k,X_i}+\alpha\lambda I)^{-1}\right\| \\
&\leq\left\|(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)(L_{k,X_i}+\lambda I)^{-1}\right\|.
\end{aligned}$$

From

$$(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)(L_{k,X_i}+\lambda I)^{-1}=I+(1-\alpha)(L_{k,X_p}-L_{k,X_i})(L_{k,X_i}+\lambda I)^{-1},$$

by the triangle inequality, the submultiplicativity of the operator norm, and Lemma D.1, we get

$$\left\|(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)(L_{k,X_i} + \lambda I)^{-1}\right\| \leq 1 + \frac{1-\alpha}{\lambda}\left(\|L_{k,X_i} - L_k\| + \|L_{k,X_p} - L_k\|\right).$$

By Lemma D.7,

$$\|L_{k,X_i} - L_k\| \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{N}}(\log(4/\delta))^{1/2} \quad \text{and} \quad \|L_{k,X_p} - L_k\| \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(4/\delta))^{1/2}$$

with confidence at least $1-\delta$. Therefore,

$$\left\|A_1^{-1}\right\| \leq \frac{1}{m}\sum_{i=1}^{m}\Theta_{1,i} \tag{36}$$

where $\Theta_{1,i}$ is a random variable such that $\Theta_{1,i} \leq \tilde{\Theta}_1(\log(4/\delta))^{1/2}$ with confidence at least $1-\delta$ for $i = 1, \cdots, m$. Here,

$$\tilde{\Theta}_1 = 1 + \frac{2\sqrt{2}(1-\alpha)\kappa^2}{\lambda}\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{N_p}}\right) = O\left(m^{\frac{1}{2r+2}}\right).$$

However, this upper bound (36) is not sufficiently small for our analysis. Thus, we will find a better upper bound now. First, we bound the norm of

$$A_2^{-1} = V^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_i} + \lambda I)\right)^{-1} V^{-1/2}.$$

By the submultiplicativity of the operator norm,

$$\begin{aligned}\left\|A_2^{-1}\right\| \leq \left\|V^{1/2}(L_k + \lambda I)^{-1/2}\right\| \cdot \left\|(L_k + \lambda I)^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}L_{k,X_i} + \lambda I\right)^{-1}(L_k + \lambda I)^{1/2}\right\| \\ \cdot \left\|(L_k + \lambda I)^{-1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{1/2}\right\| \cdot \left\|(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{1/2}V^{-1/2}\right\|.\end{aligned}$$

By Lemma D.1, Lemma D.4, the triangle inequality, and the submultiplicativity of the operator norm,

$$\begin{aligned}\left\|V^{1/2}(L_k + \lambda I)^{-1/2}\right\| &\leq \left(1 + \frac{1}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2}, \\ \left\|(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{1/2}V^{-1/2}\right\| &\leq \left(1 + \frac{\alpha}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2}\end{aligned}$$

and

$$\left\|(L_k + \lambda I)^{-1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{1/2}\right\| \leq \left(1 + \frac{1-\alpha}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2}.$$

By Lemma D.1, Lemma D.5, the triangle inequality, and the submultiplicativity of the operator norm,

$$\begin{aligned}\left\|(L_k + \lambda I)^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}L_{k,X_i} + \lambda I\right)^{-1}(L_k + \lambda I)^{1/2}\right\| &\leq \left\|\left(\frac{1}{m}\sum_{i=1}^{m}L_{k,X_i} + \lambda I\right)^{-1}(L_k + \lambda I)\right\| \\ &\leq 1 + \frac{1}{\lambda}\left\|\frac{1}{m}\sum_{i=1}^{m}L_{k,X_i} - L_k\right\|.\end{aligned}$$

By Lemma D.7,

$$\|L_{k,X_p} - L_k\| \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(4/\delta))^{1/2} \qquad \text{and} \qquad \left\|\frac{1}{m}\sum_{i=1}^{m}L_{k,X_i} - L_k\right\| \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{mN}}(\log(4/\delta))^{1/2}$$

with confidence at least $1-\delta$. Therefore, with confidence at least $1-\delta$,

$$\left\|A_2^{-1}\right\| \leq \left(1+\frac{2\sqrt{2}\kappa^2}{\lambda\sqrt{N_p}}\right)^{3/2}\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda\sqrt{mN}}\right)(\log(4/\delta))^{5/4} = O(1)(\log(4/\delta))^{5/4}.$$

Next, we bound $\|A_2 - A_1\|$. Note that

$$\begin{aligned}
A_1 &= \frac{1}{\alpha}\left(I - \frac{1}{m}\sum_{i=1}^m V^{1/2}(1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}V^{1/2}\right) \\
&= V^{1/2}\left(\frac{1}{\alpha}\left(I - \frac{1}{m}\sum_{i=1}^m (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}V\right)\right)V^{-1/2} \\
&= V^{1/2}\left(\frac{1}{m}\sum_{i=1}^m (\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_i} + \lambda I)\right)V^{-1/2}
\end{aligned}$$

and so

$$\begin{aligned}
A_2 - A_1 &= V^{1/2}\left(\frac{1}{m}\sum_{i=1}^m (\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_i} + \lambda I)\right)V^{-1/2} \\
&\quad - V^{1/2}\left(\frac{1}{m}\sum_{i=1}^m (\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_i} + \lambda I)\right)V^{-1/2} \\
&= \frac{1}{m}\sum_{i=1}^m V^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}\alpha(L_{k,X_i} - L_k)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_i} + \lambda I)V^{-1/2}
\end{aligned}$$

using (23). By the triangle inequality, the submultiplicativity of the operator norm, and the fact that

$$(L_k + \lambda I)^{-1/2}(L_{k,X_i} + \lambda I)(L_k + \lambda I)^{-1/2} = (L_k + \lambda I)^{-1/2}(L_{k,X_i} - L_k)(L_k + \lambda I)^{-1/2} + I,$$

we have

$$\begin{aligned}
&\left\|V^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}\alpha(L_{k,X_i} - L_k)(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_{k,X_i} + \lambda I)V^{-1/2}\right\| \\
&\leq \alpha\|V^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\| \cdot \|(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}(L_k + \lambda I)^{1/2}\| \cdot \mathcal{R} \\
&\quad \cdot \|(L_k + \lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_k + \lambda I)^{1/2}\| \cdot (\mathcal{R}+1) \cdot \|(L_k + \lambda I)^{1/2}V^{-1/2}\|
\end{aligned}$$

where

$$\mathcal{R} = \|(L_k + \lambda I)^{-1/2}(L_{k,X_i} - L_k)(L_k + \lambda I)^{-1/2}\|.$$

Since $(1-\alpha)V \leq \alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I$,

$$\|V^{1/2}(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\| \leq (1-\alpha)^{-1/2}$$

by Lemma D.2. By Lemma D.1, Lemma D.4, the triangle inequality, and the submultiplicativity of the operator norm,

$$\|(\alpha L_k + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}(L_k + \lambda I)^{1/2}\| \leq \left(1 + \frac{1-\alpha}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2}$$

and

$$\|(L_k + \lambda I)^{1/2}V^{-1/2}\| \leq \left(1 + \frac{1}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2}.$$

By Lemma D.1, Lemma D.5, the triangle inequality, and the submultiplicativity of the operator norm,

$$\|(L_k + \lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1}(L_k + \lambda I)^{1/2}\| \leq 1 + \frac{1}{\lambda}\left(\alpha\|L_{k,X_i} - L_k\| + (1-\alpha)\|L_{k,X_p} - L_k\|\right).$$

To find a PAC-bound, applying Lemma D.7 yields

$$\|L_{k,X_p} - L_k\| \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{N_p}}(\log(12/\delta))^{1/2}, \qquad \|L_{k,X_i} - L_k\| \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{N}}(\log(12/\delta))^{1/2},$$

and

$$\mathcal{R} \leq 2(\kappa^2+1)\cdot\left(\frac{1+\log\mathcal{N}(\lambda)}{\lambda N} + \sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}\right)\log(12/\delta)$$

with confidence at least $1-\delta$ for any $\delta \in (0,1)$. Then, we apply Lemma D.13 to see that $\|A_2 - A_1\|$ is the mean of random variables that are bounded by

$$\alpha(1-\alpha)^{-1/2}\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda\sqrt{N_p}}\right)\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N}}+\frac{1-\alpha}{\sqrt{N_p}}\right)\right)\cdot 2(\kappa^2+1)\cdot\left(\frac{1+\log\mathcal{N}(\lambda)}{\lambda N} + \sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}\right)\cdot$$
$$\left(1+2(\kappa^2+1)\cdot\left(\frac{1+\log\mathcal{N}(\lambda)}{\lambda N} + \sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}\right)\right)(\log(a_0/\delta))^{q_0}$$

for some $a_0 \geq 1$ and $q_0 > 0$ with confidence at least $1-\delta$. Note that

$$(1-\alpha)^{-1/2}\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda\sqrt{N_p}}\right)\left(1+\frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N}}+\frac{1-\alpha}{\sqrt{N_p}}\right)\right)\cdot 2(\kappa^2+1) = O(1).$$

On the other hand,

$$m^{-\frac{r+2}{4r+4}}\left(\frac{1+\log\mathcal{N}(\lambda)}{\lambda N} + \sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}\right) \leq \frac{\log\left((e\kappa^2)(mN)^{\frac{1}{2r+2}}\right)}{(mN)^{\frac{r}{4r+4}}} + \sqrt{\frac{\log\left((e\kappa^2)(mN)^{\frac{1}{2r+2}}\right)}{(mN)^{\frac{r}{4r+4}}}} = O(1)$$

since

$$f(x) = \frac{\log(e\kappa^2 x^{1/(2r+2)})}{x^{r/(4r+4)}}$$

is continuous on $[1,\infty)$ and vanishes at ∞ which implies that

$$\frac{\log\left((e\kappa^2)(mN)^{\frac{1}{2r+2}}\right)}{(mN)^{\frac{r}{4r+4}}} = O(1).$$

Therefore,

$$\alpha\left(\frac{1+\log\mathcal{N}(\lambda)}{\lambda N} + \sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}\right)\left(1+2(\kappa^2+1)\cdot\left(\frac{1+\log\mathcal{N}(\lambda)}{\lambda N} + \sqrt{\frac{1+\log\mathcal{N}(\lambda)}{\lambda N}}\right)\right) = O\left(m^{-\frac{r}{2r+2}}\right)$$

and so $\|A_2 - A_1\|$ is the mean of random variables that are bounded by $O\left(m^{-\frac{r}{2r+2}}\right)(\log(a_0/\delta))^{q_0}$ for some $a_0 \geq 1$ and $q_0 > 0$ with confidence at least $1-\delta$ respectively. Thus, applying the formula

$$A_1^{-1} = \sum_{j=0}^{\lceil\frac{1}{r}\rceil-1} A_2^{-1}\left((A_2-A_1)A_2^{-1}\right)^j + A_1^{-1}\left((A_2-A_1)A_2^{-1}\right)^{\lceil\frac{1}{r}\rceil},$$

(36), and Lemma D.13 yield

$$\|A_1^{-1}\| \leq \sum_{j=0}^{\lceil\frac{1}{r}\rceil}\Xi_j \tag{37}$$

where Ξ_j is the mean of random variables that are bounded by $O(1)(\log(a/\delta))^q$ for some $a \geq 1$ and $q > 0$.

The last part of the proof is devoted to bounding

$$\left\| U^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^{\top} \mathbf{y}_i - \frac{1}{\alpha} \left(I - \frac{1}{m} \sum_{i=1}^{m} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U \right) f_0 \right) \right\|_{\mathbb{H}_k}.$$

Note that

$$\begin{aligned}
&U^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^{\top} \mathbf{y}_i - \frac{1}{\alpha} \left(I - \frac{1}{m} \sum_{i=1}^{m} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} U \right) f_0 \right) \\
&= U^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^{\top} \mathbf{y}_i - \frac{1}{\alpha} \left(I - \frac{1}{m} \sum_{i=1}^{m} (1-\alpha)(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} (L_{k,X_p} + \lambda P_{D_p}) \right) f_0 \right) \\
&= U^{1/2} \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} (S_{D_i}^{\top} \mathbf{y}_i - L_{k,X_i} f_0 - \lambda f_0) - \frac{1-\alpha}{\alpha} U^{1/2} \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \lambda (I - P_{D_p}) f_0.
\end{aligned}$$

By Lemma D.3 and the triangle inequality, the first term is bounded as follows:

$$\begin{aligned}
&\left\| U^{1/2} \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} (S_{D_i}^{\top} \mathbf{y}_i - L_{k,X_i} f_0 - \lambda f_0) \right\|_{\mathbb{H}_k} \\
&\leq \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^{\top} \mathbf{y}_i - f_0 \right) \right\|_{\mathbb{H}_k} \\
&\quad + \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} L_{k,X_i} f_0 - f_0 \right) \right\|_{\mathbb{H}_k} + \lambda \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \right) f_0 \right\|_{\mathbb{H}_k}.
\end{aligned}$$

By Lemma C.12,

$$\left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} S_{D_i}^{\top} \mathbf{y}_i - f_0 \right) \right\|_{\mathbb{H}_k} \leq \Lambda + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i.$$

Using the exact same way as in the proof of Lemma C.12 (since it is just the noisefree case),

$$\left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} L_{k,X_i} f_0 - f_0 \right) \right\|_{\mathbb{H}_k} \leq \Lambda^0 + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i^0 \tag{38}$$

where Λ^0 and Λ_i^0 are random variables such that $\Lambda^0 \leq O\left((mN)^{-\frac{2r+1}{4r+4}} \right) (\log(6/\delta))^{5/4}$ with confidence at least $1-\delta$ and each $\Lambda_i^0 \leq O\left((mN)^{-\frac{2r+1}{4r+4}} \right) (\log(6/\delta))^{3/2}$ with confidence at least $1-\delta$. Also, by the triangle inequality, the submultiplicativity of the operator norm, (31), and Lemma D.1,

$$\begin{aligned}
\lambda \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \right) f_0 \right\|_{\mathbb{H}_k} &\leq \frac{\lambda}{1-\alpha} \cdot \frac{1}{m} \sum_{i=1}^{m} \| (L_{k,X_p} + \lambda I)^{-1/2} L_k^r g_0 \|_{\mathbb{H}_k} \\
&= \frac{\lambda}{1-\alpha} \| (L_{k,X_p} + \lambda I)^{-1/2} L_k^r g_0 \|_{\mathbb{H}_k} \\
&\leq \frac{\lambda}{1-\alpha} \mathcal{Q}_p \| (L_k + \lambda I)^{-1/2} L_k^r g_0 \|_{\mathbb{H}_k} \\
&\leq \frac{1}{1-\alpha} \cdot \lambda^{r+\frac{1}{2}} \mathcal{Q}_p \| g_0 \|_{\mathbb{H}_k}
\end{aligned}$$

where $\mathcal{Q}_p$ is already defined in the proof of Lemma C.12. By (30),

$$\lambda \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \right) f_0 \right\|_{\mathbb{H}_k} \leq \frac{1}{1-\alpha} \cdot \left(1 + \frac{1}{\lambda} \| L_k - L_{k,X_p} \| \right)^{1/2} \lambda^{r+\frac{1}{2}} \| g_0 \|_{\mathbb{H}_k}.$$

By Lemma D.7,

$$\lambda \left\| V^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \right) f_0 \right\|_{\mathbb{H}_k} \leq \frac{1}{1-\alpha} \cdot \left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda \sqrt{N_p}} \right)^{1/2} \lambda^{r+\frac{1}{2}} \|g_0\|_{\mathbb{H}_k} (\log(4/\delta))^{1/4} \tag{39}$$

with confidence at least $1-\delta$ where

$$\frac{1}{1-\alpha} \cdot \left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda \sqrt{N_p}} \right)^{1/2} \lambda^{r+\frac{1}{2}} \|g_0\|_{\mathbb{H}_k} = O\left((mN)^{-\frac{2r+1}{4r+4}} \right).$$

Lastly, by the triangle inequality, the submultiplicativity of the operator norm, Lemma D.1, Lemma D.3, and (31),

$$\begin{aligned} &\left\| \frac{1-\alpha}{\alpha} U^{1/2} \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \lambda (I - P_{D_p}) f_0 \right\|_{\mathbb{H}_k} \\ &\leq \left\| \frac{1-\alpha}{\alpha} V^{1/2} \frac{1}{m} \sum_{i=1}^{m} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \lambda (I - P_{D_p}) f_0 \right\|_{\mathbb{H}_k} \leq \frac{\sqrt{\lambda}}{\alpha} \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|S_{D_p}(g^* - f_0)\|_2 \leq{}& \frac{1}{m} \sum_{i=1}^{m} \Lambda_i^1 + \frac{\sqrt{\lambda}}{1-\alpha} \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k} + \alpha \left(\frac{1}{m} \sum_{i=1}^{m} \Lambda_i^1 + \frac{\sqrt{\lambda}}{1-\alpha} \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k} + \Lambda + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i \right) \\ &+ \|A_1^{-1}\| \left(\Lambda + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i + \Lambda^0 + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i^0 + \frac{1}{1-\alpha} \cdot \lambda^{r+1/2} \mathcal{Q}_p \|g_0\|_{\mathbb{H}_k} + \frac{\sqrt{\lambda}}{\alpha} \cdot \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k} \right). \end{aligned} \tag{40}$$

By Lemma D.9,

$$\mathbf{E} \left[\frac{1}{m} \sum_{i=1}^{m} \Lambda_i^1 + \alpha \left(\frac{1}{m} \sum_{i=1}^{m} \Lambda_i^1 + \Lambda + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i \right) \right] = O\left((mN)^{-\frac{2r+1}{4r+4}} \right). \tag{41}$$

From (34), we have

$$\mathbf{E} \left[(1+\alpha) \cdot \frac{\sqrt{\lambda}}{1-\alpha} \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k} \right] = O\left((mN)^{-\frac{2r+1}{4r+4}} \right).$$

By Lemma D.13, (37), (38), and (39),

$$\mathbf{E} \left[\|A_1^{-1}\| \left(\Lambda + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i + \Lambda^0 + \frac{1}{m} \sum_{i=1}^{m} \Lambda_i^0 + \frac{1}{1-\alpha} \cdot \lambda^{r+1/2} \mathcal{Q}_p \|g_0\|_{\mathbb{H}_k} \right) \right] = O\left((mN)^{-\frac{2r+1}{4r+4}} \right). \tag{42}$$

Since $\|L_k\| \leq \kappa^2$ and $\|L_{k,X}\| \leq \kappa^2$ for any X, by Lemma D.5, Lemma D.1, the triangle inequality, and the submultiplicativity of the operator norm, we obtain

$$\|A_1^{-1}\| \leq \frac{1}{m} \sum_{i=1}^{m} \left\| (L_{k,X_i} + \lambda I)^{-1} (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I) \right\| \leq \frac{\lambda + \kappa^2}{\lambda}.$$

Thus,

$$\|A_1^{-1}\| \cdot \frac{\sqrt{\lambda}}{\alpha} \cdot \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k} \leq \frac{\sqrt{\lambda}}{\alpha} \cdot \left(1 + \frac{\kappa^2}{\lambda} \right) \|f_0\|_{\mathbb{H}_k} \tag{43}$$

almost surely. By Lemma C.13 and Lemma D.12,

$$\mathbf{E} \left[\|A_1^{-1}\| \cdot \frac{\sqrt{\lambda}}{\alpha} \cdot \|(I - P_{D_p}) f_0\|_{\mathbb{H}_k} \right] \leq \frac{\sqrt{\lambda}}{\alpha} \cdot \left(1 + \frac{\kappa^2}{\lambda} \right) \|f_0\|_{\mathbb{H}_k} \cdot 4 \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}} \right) + O\left(\frac{\sqrt{\lambda}}{\alpha} \cdot \lambda_p^r \right)$$

where

$$\mathcal{B} = \frac{1+\log \mathcal{N}(\lambda_p)}{\lambda_p N_p} + \sqrt{\frac{1+\log \mathcal{N}(\lambda_p)}{\lambda_p N_p}}$$

and $\lambda_p = 1/N_p^{1-\epsilon}$ which satisfies $\mathcal{N}(\lambda_p) \geq \mathcal{N}(1/\sqrt{N_p}) \geq 1$. Note that

$$\sqrt{\frac{1+\log \mathcal{N}(\lambda_p)}{\lambda_p N_p}} \leq \frac{1}{N_p^{\epsilon/4}} \sqrt{\frac{\log(e\kappa^2 N_p)}{N_p^{\epsilon/2}}}.$$

Thus,

$$\mathcal{B} \leq \frac{1}{N_p^{\epsilon/4}} \sqrt{\frac{\log(e\kappa^2 N_p)}{N_p^{\epsilon/2}}} \left(1 + \sqrt{\frac{\log(e\kappa^2 N_p)}{N_p^{\epsilon/2}}}\right).$$

Since

$$g(x) = \sqrt{\frac{\log(e\kappa^2 x)}{x^{\epsilon/2}}} \left(1 + \sqrt{\frac{\log(e\kappa^2 x)}{x^{\epsilon/2}}}\right)$$

is continuous on $[1, \infty)$ and vanishes at ∞, $g(x) \leq C(\kappa, r, \epsilon)$ for some $C(\kappa, r, \epsilon) > 0$. Then

$$\mathcal{B} \leq C(\kappa, r, \epsilon) N_p^{-\epsilon/4}.$$

Since $0 < \mathcal{B} \leq C(\kappa, r, \epsilon)$ and

$$h(x) = \frac{1}{x^\beta} \exp\left(-\frac{1}{4(\kappa^2+1)x}\right)$$

is continuous on $(0, C(\kappa, r, \epsilon)]$ and vanishes at 0^+,

$$\frac{1}{\mathcal{B}^\beta} \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \leq C'(\kappa, r, \epsilon, \beta)$$

for some $C'(\kappa, r, \epsilon, \beta) > 0$. Hence,

$$\begin{aligned}\frac{\sqrt{\lambda}}{\alpha} \cdot \left(1 + \frac{\kappa^2}{\lambda}\right) \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) &\leq \frac{\sqrt{\lambda}}{\alpha} \cdot \left(1 + \frac{\kappa^2}{\lambda}\right) \mathcal{B}^{6/\epsilon} \frac{1}{\mathcal{B}^{6/\epsilon}} \exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) \\ &\leq (1+\kappa^2) \frac{1}{\alpha\sqrt{\lambda}} N_p^{-3/2} C(\kappa, r, \epsilon)^{6/\epsilon} \cdot C'(\kappa, r, 6/\epsilon) = O\left((mN)^{-\frac{2r+1}{4r+4}}\right). \quad (44)\end{aligned}$$

On the other hand, from $N_p^{r(1-\epsilon)} \geq (mN)^{\frac{r}{2r+2}} m$,

$$O\left(\frac{\sqrt{\lambda}}{\alpha} \cdot N_p^{r(-1+\epsilon)}\right) = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

This completes the proof of Theorem 5.5. □

Theorem C.15. *Let $g^* = T^\infty h_0$ for any $h_0 \in \mathbb{H}_k$ and g_i^* be a local kernel ridge regressor of client i trained by the local dataset D_i and the public dataset $D_p(\mathbf{x})$ with the predictions of g^*, i.e.,*

$$g_i^* = (\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} (\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha) L_{k,X_p} g^*).$$

Under the same assumption of Theorem C.14, with $\alpha = 1/m$ and $\lambda = (mN)^{-\frac{1}{2r+2}}$,

$$\mathbf{E} \|\iota_{\rho_\mathbf{x}}(g_i^* - f_0)\|_{L^2_{\rho_\mathbf{x}}} = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

Proof of Theorem C.15. We use the same notation as in the proof of Theorem C.14. Since g^* is not independent of $D_p(\mathbf{x})$, we cannot directly apply Corollary B.2. By Corollary B.1, we have

$$\mathbf{E}\|\iota_{\rho_\mathbf{x}}(g_i^* - f_0)\|_{L^2_{\rho_\mathbf{x}}} \le 9\left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}} + \frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right)$$
$$+ (1-\alpha)^{1/2}\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\|\|L_{k,X_p}^{1/2}(g^* - f_0)\|_{\mathbb{H}_k}\right].$$

First,

$$9\left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N_1}} + \frac{1-\alpha}{\sqrt{N_2}}\right)\right)\left(\frac{2\alpha\kappa(M+\gamma)}{\sqrt{\lambda N_1}} + \lambda^{\frac{1}{2}+r}\|g_0\|_{\mathbb{H}_k}\right) = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

By Lemma D.1, Lemma D.4, the triangle inequality, and the submultiplicativity of the operator norm,

$$\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\| \le \left(1 + \frac{\alpha}{\lambda}\|L_{k,X_i} - L_k\| + \frac{1-\alpha}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2}.$$

By Lemma D.7, with confidence at least $1-\delta$

$$\left(1 + \frac{\alpha}{\lambda}\|L_{k,X_i} - L_k\| + \frac{1-\alpha}{\lambda}\|L_{k,X_p} - L_k\|\right)^{1/2} \le \left(1 + \frac{2\sqrt{2}\kappa^2}{\lambda}\left(\frac{\alpha}{\sqrt{N}} + \frac{1-\alpha}{\sqrt{N_p}}\right)\right)^{1/2}(\log(4/\delta))^{1/4}$$

where $\delta \in (0,1)$ and hence

$$\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\| \le O(1)(\log(4/\delta))^{1/4} \le O(1)\log(4/\delta). \tag{45}$$

Recall $\|L_{k,X_p}^{1/2}(g^* - f_0)\|_{\mathbb{H}_k} = \|S_{D_p}(g^* - f_0)\|_2$ from (10) and its bound in (40). By Lemma D.9, Lemma D.13, (33), and (35),

$$\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\|\left(\frac{1}{m}\sum_{i=1}^m \Lambda_i^1 + \alpha\left(\frac{1}{m}\sum_{i=1}^m \Lambda_i^1 + \Lambda + \frac{1}{m}\sum_{i=1}^m \Lambda_i\right)\right)\right] = O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

By Lemma D.13, (37), (38), and (39),

$$\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\|\|A_1^{-1}\|\left(\Lambda + \frac{1}{m}\sum_{i=1}^m \Lambda_i + \Lambda^0 + \frac{1}{m}\sum_{i=1}^m \Lambda_i^0 + \frac{1}{1-\alpha}\lambda^{\frac{1}{2}+r}\mathcal{Q}_p\|g_0\|_{\mathbb{H}_k}\right)\right]$$
$$= O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

Note that there is a trivial bound

$$\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\| \le \left(\frac{\lambda+\kappa^2}{\lambda}\right)^{1/2}.$$

By (45), Lemma C.13, and Lemma D.12,

$$\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\| \cdot (1+\alpha)\cdot\frac{\sqrt{\lambda}}{1-\alpha}\|(I - P_{D_p})f_0\|_{\mathbb{H}_k}\right]$$
$$\le \frac{(1+\alpha)\sqrt{\lambda}}{1-\alpha}\cdot\left(1 + \frac{\kappa^2}{\lambda}\right)^{1/2}\|f_0\|_{\mathbb{H}_k}\cdot 4\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) + O\left(\sqrt{\lambda}\cdot N_p^{r(-1+\epsilon)}\right).$$

By (37), (43), Lemma C.13, and Lemma D.13,

$$\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i} + (1-\alpha)L_{k,X_p} + \lambda I)^{-1/2}\|\|A_1^{-1}\|\cdot\frac{\sqrt{\lambda}}{\alpha}\cdot\|(I - P_{D_p})f_0\|_{\mathbb{H}_k}\right]$$
$$\le \frac{\sqrt{\lambda}}{\alpha}\cdot\left(1 + \frac{\kappa^2}{\lambda}\right)^{3/2}\|f_0\|_{\mathbb{H}_k}\cdot 4\exp\left(-\frac{1}{4(\kappa^2+1)\mathcal{B}}\right) + O\left(\frac{\sqrt{\lambda}}{\alpha}\cdot N_p^{r(-1+\epsilon)}\right).$$

Using the similar argument as in the derivation of the bound (44),

$$\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1/2}\|\cdot(1+\alpha)\cdot\frac{\sqrt{\lambda}}{1-\alpha}\|(I-P_{D_p})f_0\|_{\mathbb{H}_k}\right]=O\left((mN)^{-\frac{2r+1}{4r+4}}\right)$$

and

$$\mathbf{E}\left[\|(L_k+\lambda I)^{1/2}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1/2}\|\|A_1^{-1}\|\cdot\frac{\sqrt{\lambda}}{\alpha}\cdot\|(I-P_{D_p})f_0\|_{\mathbb{H}_k}\right]=O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

From (40),

$$\mathbf{E}\|\iota_{\rho_{\mathbf{x}}}(g_i^*-f_0)\|_{L^2_{\rho_{\mathbf{x}}}}=O\left((mN)^{-\frac{2r+1}{4r+4}}\right).$$

□

C.7. Algorithm on KRR with Iterative De-regularized Ensemble Distillation Participated by Partial Clients in Federated Learning Setting

We provide an algorithm (Algorithm 3) on kernel ridge regression with iterative de-regularized ensemble distillation participated by partial clients in federated learning. At each communication round, we select a fixed number of clients that predict the unlabeled public dataset and train using their local datasets and the public dataset with the updated consensus obtained by stochastic approximation. We denote the number of selected clients in each communication round as $\mathcal{C}$.

C.8. Proof of Corollary 5.6

Here, we assume that D and $D_p(\mathbf{x})$ are given. Define a sequence of independent random operators $\{T_{t_0}^p\}_{t_0\in\mathbb{N}}$ where random operator $T_{t_0}^p:\mathbb{H}_k\to\mathbb{H}_k$ is defined as

$$\begin{aligned}T_{t_0}^p g&=\sum_{i\in\mathcal{C}_{t_0}}\frac{1}{\mathcal{C}}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(\alpha S_{D_i}^\top\mathbf{y}_i+(1-\alpha)S_{D_p}^\top(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^\top)^{-1}S_{D_p}g)\\&=\sum_{i\in\mathcal{C}_{t_0}}\frac{1}{\mathcal{C}}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(\alpha S_{D_i}^\top\mathbf{y}_i+(1-\alpha)(L_{k,X_p}+\lambda P_{D_p(\mathbf{x})}^\perp)g)\end{aligned}$$

for each $t_0\in\mathbb{N}$ where $\{\mathcal{C}_{t_0}\}_{t_0\in\mathbb{N}}$ is defined in Algorithm 3. We also define $\bar{T}_{t_0}^p:\mathbb{R}^{N_p}\to\mathbb{R}^{N_p}$ as

$$\begin{aligned}\bar{T}_{t_0}^p\mathbf{v}=\sum_{i\in\mathcal{C}_{t_0}}\frac{1}{\mathcal{C}}&(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^\top)^{-1/2}S_{D_p}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}\cdot\\&(\alpha S_{D_i}^\top\mathbf{y}_i+(1-\alpha)S_{D_p}^\top(S_{D_p}(L_{k,X_p}+\lambda_0 I)^{-1}S_{D_p}^\top)^{-1/2}\mathbf{v})\end{aligned}$$

for each $t_0\in\mathbb{N}$.

We prove Corollary 5.6 in a general setting.

Theorem C.16. *Assume $\lambda_0=\lambda$, K_{X_p,X_p} is invertible, and the number of selected clients in each communication round $\mathcal{C}$ is fixed, i.e., $\mathcal{C}=\sum_{i=1}^m\mathbf{1}_{\{i\in\mathcal{C}_{t_0}\}}$ for any $t_0=1,\cdots,t$. We also assume that $\{\gamma_{t_0}\}_{t_0=1}^t\in[0,1]^t$ is fixed,*

$$\sum_{t_0=1}^\infty\gamma_{t_0}=\infty,\qquad\sum_{t_0=1}^\infty\gamma_{t_0}^2<\infty,\tag{46}$$

$\{\mathcal{C}_{t_0}\}_{t_0=1}^\infty$ is independent, and

$$\mathbf{P}(j\in\mathcal{C}_{t_0})=p_j$$

for any $j\in\{1,\cdots,m\}$ and $t_0=1,\cdots$. Then the prediction $\tilde{\mathbf{y}}_p$ on $D_p(\mathbf{x})$ after infinitely many iterations in Algorithm 3 converges to $S_{D_p}\hat{g}^*$ almost surely where $\hat{g}^*=\hat{T}^\infty h_0$ for any $h_0\in\mathbb{H}_k$ and

$$\hat{T}g=\sum_{i=1}^m\frac{p_i}{\sum_{j=1}^m p_j}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(\alpha S_{D_i}^\top\mathbf{y}_i+(1-\alpha)S_{D_p}^\top(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^\top)^{-1}S_{D_p}g).$$

Algorithm 3 KRR with iterative De-regularized Ensemble Distillation Participated by Partial Clients in FL

Input: hyperparameters $\alpha \in (0,1)$, $\lambda > 0$, $\lambda_0 \geq 0$, $t \in \mathbb{N}$, $\mathcal{C} \in \{1, \cdots, m\}$ and $\{\gamma_{t_0}\}_{t_0=1}^{t} \in [0,1]^t$ such that $\gamma_1 = 1$
Output: Trained model f_j, $j = 1, \cdots, m$
Pretrain: For $j = 1, \cdots, m$, client j trains its model f_j using the loss function

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}_j^i) - y_j^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

Each client downloads the unlabeled public dataset $D_p(\mathbf{x})$.
for $t_0 = 1, \cdots, t$ **do**
  Determine a set of clients $\mathcal{C}_{t_0}$ whose size is $\mathcal{C}$ to participate the ensemble at time t_0.
  For $j \in \mathcal{C}_{t_0}$, client j predicts on $D_p(\mathbf{x})$ and uploads the prediction $\tilde{\mathbf{y}}_p^j$ to server.
  The server computes an updated consensus

$$\tilde{\mathbf{y}}_p = (1 - \gamma_{t_0}) \cdot \tilde{\mathbf{y}}_{p,old} + \gamma_{t_0} \cdot \frac{1}{\mathcal{C}} \sum_{j \in \mathcal{C}_{t_0}} \tilde{\mathbf{y}}_p^j.$$

  The server stores $\tilde{\mathbf{y}}_{p,old} = \tilde{\mathbf{y}}_p$.
  if $t_0 \neq t$ **then**
    The server applies the de-regularization trick to $\tilde{\mathbf{y}}_p$:

$$\tilde{\mathbf{y}}_p = (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1} \tilde{\mathbf{y}}_p.$$

  end if
  Each client in $\mathcal{C}_{t_0}$ downloads the ensemble prediction $\tilde{\mathbf{y}}_p$.
  For $j \in \mathcal{C}_{t_0}$, client j updates its model f_j using the loss function

$$\operatorname*{argmin}_{h \in \mathbb{H}_k} \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} (h(\mathbf{x}_j^i) - y_j^i)^2 + (1-\alpha) \cdot \frac{1}{N_p} \sum_{i=1}^{N_p} (h(\mathbf{x}_p^i) - (\tilde{\mathbf{y}}_p)^i)^2 + \lambda \|h\|_{\mathbb{H}_k}^2.$$

end for

Proof. Define $\bar{\hat{T}} : \mathbb{R}^{N_p} \to \mathbb{R}^{N_p}$ as

$$\begin{aligned}\bar{\hat{T}}\mathbf{v} = \sum_{i=1}^{m} \frac{p_i}{\sum_{j=1}^m p_j} &(S_{D_p}(L_{k,X_p} + \lambda I)^{-1} S_{D_p}^\top)^{-1/2} S_{D_p}(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \\ &(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha) S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda I)^{-1} S_{D_p}^\top)^{-1/2} \mathbf{v}).\end{aligned}$$

Note that

$$\mathcal{C} = \sum_{j=1}^{m} \mathbf{1}_{\{j \in \mathcal{C}_{t_0}\}} \quad \Rightarrow \quad \mathcal{C} = \mathbf{E}\left[\sum_{j=1}^{m} \mathbf{1}_{\{j \in \mathcal{C}_{t_0}\}}\right] = \sum_{j=1}^{m} \mathbf{P}(j \in \mathcal{C}_{t_0}) = \sum_{j=1}^{m} p_j$$

by taking the expectation. Define an operator $S_i : \mathbb{R}^{N_p} \to \mathbb{R}^{N_p}$ by

$$\begin{aligned}S_i\mathbf{v} = (S_{D_p}(L_{k,X_p} + \lambda I)^{-1} S_{D_p}^\top)^{-1/2} &S_{D_p}(\alpha L_{k,X_i} + (1-\alpha) L_{k,X_p} + \lambda I)^{-1} \cdot \\ &(\alpha S_{D_i}^\top \mathbf{y}_i + (1-\alpha) S_{D_p}^\top (S_{D_p}(L_{k,X_p} + \lambda_0 I)^{-1} S_{D_p}^\top)^{-1/2} \mathbf{v}).\end{aligned}$$

Then

$$\mathbf{E}\bar{T}_{t_0}^p \mathbf{v} = \mathbf{E}\left[\sum_{i \in \mathcal{C}_{t_0}} \frac{1}{\mathcal{C}} S_i \mathbf{v}\right] = \mathbf{E}\left[\sum_{i=1}^{m} \mathbf{1}_{\{i \in \mathcal{C}_{t_0}\}} \frac{1}{\mathcal{C}} S_i \mathbf{v}\right] = \frac{1}{\mathcal{C}} \sum_{i=1}^{m} \mathbf{E}\left[\mathbf{1}_{\{i \in \mathcal{C}_{t_0}\}} S_i \mathbf{v}\right] = \frac{1}{\sum_{j=1}^m p_j} \sum_{i=1}^{m} p_i S_i \mathbf{v} = \bar{\hat{T}}\mathbf{v}.$$

Observe that

$$(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^\top)^{-1/2} = ((K_{X_p,X_p}+N_p\lambda I)^{-1}K_{X_p,X_p})^{-1/2} = (I+N_p\lambda K_{X_p,X_p}^{-1})^{1/2}. \tag{47}$$

We also claim that

$$\begin{aligned}
&S_{D_p}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(1-\alpha)S_{D_p}^\top \\
&= \left(K_{X_p,X_p}-K_{X_p,X_i}\left(K_{X_i,X_i}+\frac{N}{\alpha}\lambda I\right)^{-1}K_{X_i,X_p}\right)\left(K_{X_p,X_p}+\frac{N_p}{1-\alpha}\lambda I-K_{X_p,X_i}\left(K_{X_i,X_i}+\frac{N}{\alpha}\lambda I\right)^{-1}K_{X_i,X_p}\right)^{-1} \\
&= \left(I+\frac{N_p}{1-\alpha}\lambda\left(K_{X_p,X_p}-K_{X_p,X_i}\left(K_{X_i,X_i}+\frac{N}{\alpha}\lambda I\right)^{-1}K_{X_i,X_p}\right)^{-1}\right)^{-1}.
\end{aligned} \tag{48}$$

To show this, note that

$$\begin{aligned}
(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(1-\alpha)S_{D_p}^\top &= \begin{bmatrix}\mathbf{k}_{X_i}^\top & \mathbf{k}_{X_p}^\top\end{bmatrix}(\lambda I+DK_{X_i\cup X_p,X_i\cup X_p})^{-1}D\begin{bmatrix}0\\ .\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{k}_{X_i}^\top & \mathbf{k}_{X_p}^\top\end{bmatrix}(\lambda D^{-1}+K_{X_i\cup X_p,X_i\cup X_p})^{-1}\begin{bmatrix}0\\ .\end{bmatrix}
\end{aligned}$$

where $D=\mathrm{diag}(\underbrace{\alpha/N,\cdots,\alpha/N}_{N},\underbrace{(1-\alpha)/N_p,\cdots,(1-\alpha)/N_p}_{N_p})$ which follows from (12). Then

$$S_{D_p}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(1-\alpha)S_{D_p}^\top = \begin{bmatrix}K_{X_p,X_i} & K_{X_p,X_p}\end{bmatrix}\begin{bmatrix}K_{X_i,X_i}+\frac{N_1}{\alpha}\lambda I & K_{X_i,X_p}\\ K_{X_p,X_i} & K_{X_p,X_p}+\frac{N_2}{1-\alpha}\lambda I\end{bmatrix}^{-1}\begin{bmatrix}0\\ .\end{bmatrix}.$$

From the formula

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}^{-1} = \begin{bmatrix}A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1}\\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1}\end{bmatrix},$$

we have

$$\begin{aligned}
&S_{D_p}(\alpha L_{k,X_i}+(1-\alpha)L_{k,X_p}+\lambda I)^{-1}(1-\alpha)S_{D_p}^\top \\
&= \begin{bmatrix}K_{X_p,X_i} & K_{X_p,X_p}\end{bmatrix}\begin{bmatrix}* & -(K_{X_i,X_i}+\frac{N_1}{\alpha}\lambda I)^{-1}K_{X_i,X_p}(K_{X_p,X_p}+\frac{N_2}{1-\alpha}\lambda I-K_{X_p,X_i}(K_{X_i,X_i}+\frac{N_1}{\alpha}\lambda I)^{-1}K_{X_i,X_p})^{-1}\\ * & (K_{X_p,X_p}+\frac{N_2}{1-\alpha}\lambda I-K_{X_p,X_i}(K_{X_i,X_i}+\frac{N_1}{\alpha}\lambda I)^{-1}K_{X_i,X_p})^{-1}\end{bmatrix}\begin{bmatrix}0\\ .\end{bmatrix}
\end{aligned}$$

which gives the formula (48). Since

$$I+\frac{N_p}{1-\alpha}\lambda\left(K_{X_p,X_p}-K_{X_p,X_i}\left(K_{X_i,X_i}+\frac{N}{\alpha}\lambda I\right)^{-1}K_{X_i,X_p}\right)^{-1} > I+N_p\lambda K_{X_p,X_p}^{-1},$$

by (47), (48), Lemma D.2, and the triangle inequality, $\bar{\hat{T}}$ is a η-contraction where $\eta\in(0,1)$. Set $\mathbf{u}_{t_0}$ as

$$(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^\top)^{-1/2}(\tilde{\mathbf{y}}_p)_{t_0}$$

where $(\tilde{\mathbf{y}}_p)_{t_0}$ is the updated consensus before applying the de-regularization at the t_0-th iteration in Algorithm 3. Then

$$\mathbf{u}_{t_0+1} = (1-\gamma_{t_0+1})\mathbf{u}_{t_0}+\gamma_{t_0+1}\bar{T}_{t_0+1}^p\mathbf{u}_{t_0} = (1-\gamma_{t_0+1})\mathbf{u}_{t_0}+\gamma_{t_0+1}\left(\bar{\hat{T}}\mathbf{u}_{t_0}+\left(\bar{T}_{t_0+1}^p\mathbf{u}_{t_0}-\bar{\hat{T}}\mathbf{u}_{t_0}\right)\right).$$

Note that

$$\mathbf{E}\left[\bar{T}_{t_0+1}^p\mathbf{u}_{t_0}-\bar{\hat{T}}\mathbf{u}_{t_0}\mid\mathcal{F}_{t_0}\right] = \mathbf{E}\left[\bar{T}_{t_0+1}^p\mathbf{u}_{t_0}-\bar{\hat{T}}\mathbf{u}_{t_0}\mid\mathbf{u}_{t_0}\right] = 0 \tag{49}$$

where $\mathcal{F}_{t_0}$ is an σ-algebra generated from $\mathbf{u}_1,\cdots,\mathbf{u}_{t_0},\bar{T}_1^p,\cdots,\bar{T}_{t_0}^p,\gamma_1,\cdots,\gamma_{t_0}$ and γ_{t_0+1} since $\bar{T}_{t_0+1}^p$ is independent of $\mathcal{F}_{t_0}$. Since the number of possible realizations of $\bar{T}_{t_0+1}^p$ is finite, there exists $B>0$ such that

$$\|\bar{T}_{t_0+1}^p-\bar{\hat{T}}\|\le B$$

almost surely. Therefore,

$$\mathbf{E}\left[\left\|\bar{T}_{t_0+1}^{p}\mathbf{u}_{t_0}-\bar{\hat{T}}\mathbf{u}_{t_0}\right\|_{\mathbb{H}_k}^2 \mid \mathcal{F}_{t_0}\right]=\mathbf{E}\left[\left\|\bar{T}_{t_0+1}^{p}\mathbf{u}_{t_0}-\bar{\hat{T}}\mathbf{u}_{t_0}\right\|_{\mathbb{H}_k}^2 \mid \mathbf{u}_{t_0}\right]\leq \mathbf{E}\left[\left\|\bar{T}_{t_0+1}^{p}-\bar{\hat{T}}\right\|^2\|\mathbf{u}_{t_0}\|_{\mathbb{H}_k}^2 \mid \mathbf{u}_{t_0}\right]$$
$$\leq \mathbf{E}\left[B^2\|\mathbf{u}_{t_0}\|_{\mathbb{H}_k}^2 \mid \mathbf{u}_{t_0}\right]=B^2\|\mathbf{u}_{t_0}\|_{\mathbb{H}_k}^2. \tag{50}$$

From (46), (49), and (50), by Proposition 4.4 in Bertsekas & Tsitsiklis (1996), $(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^{\top})^{-1/2}\tilde{\mathbf{y}}_p$ converges to a unique fixed point of $\bar{\hat{T}}$. Therefore,

$$(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^{\top})^{-1/2}\tilde{\mathbf{y}}_p=\bar{\hat{T}}^{\infty}*=(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^{\top})^{-1/2}S_{D_p}\hat{T}^{\infty}*$$
$$=(S_{D_p}(L_{k,X_p}+\lambda I)^{-1}S_{D_p}^{\top})^{-1/2}S_{D_p}\hat{g}^*.$$

Note that $\hat{T}^{\infty}*=\hat{g}^*$ can be shown using a similar argument as in the first part in Appendix C.6. □

D. Auxiliary Lemmas

We provide the useful properties of operators.

Lemma D.1. *Let A be a bounded linear operator on a seperable Hilbert space $\mathcal{H}$. If A is compact, self-adjoint, and positive, then*

$$\|(A+\lambda I)^{-r}A^s\|\leq 1/\lambda^{r-s}$$

for any $r\geq s\geq 0$ and $\lambda>0$.

Proof. Since A is compact, self-adjoint, and positive, we can set eigenvalues $\{\mu_i\}_{i=1}^{\infty}$ of A and their corresponding eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ satisfying $\mu_i\downarrow 0$ and $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}$ by the spectral theorem (Conway, 2019). For any $\phi=\sum_{i=1}^{\infty}a_i\phi_i\in\mathcal{H}$ such that $\|\phi\|^2=\sum_{i=1}^{\infty}a_i^2=1$, we have

$$\|(A+\lambda I)^{-r}A^s\phi\|^2=\left\|\sum_{i=1}^{\infty}\frac{a_i\mu_i^s}{(\mu_i+\lambda)^r}\phi_i\right\|^2=\sum_{i=1}^{\infty}\left(\frac{a_i\mu_i^s}{(\mu_i+\lambda)^r}\right)^2=\sum_{i=1}^{\infty}a_i^2\left(\frac{\mu_i}{\mu_i+\lambda}\right)^{2s}\cdot\frac{1}{(\mu_i+\lambda)^{2(r-s)}}$$
$$\leq\frac{1}{\lambda^{2(r-s)}}\sum_{i=1}^{\infty}a_i^2=\frac{1}{\lambda^{2(r-s)}}$$

which implies

$$\|(A+\lambda I)^{-r}A^s\|=\sup_{\|\phi\|=1}\|(A+\lambda I)^{-r}A^s\phi\|\leq\frac{1}{\lambda^{r-s}}.$$

□

Lemma D.2. *Let A and B be bounded positive self-adjoint linear operators on a seperable Hilbert space $\mathcal{H}$. We assume B is invertible. Then $A\leq B$ implies*

$$\|A^{1/2}B^{-1/2}\|^2=\|A^{1/2}B^{-1}A^{1/2}\|\leq 1.$$

If we further assume A is compact, then $A<B$ implies

$$\|A^{1/2}B^{-1/2}\|^2=\|A^{1/2}B^{-1}A^{1/2}\|<1.$$

Proof. We first assume $A\leq B$. Since $0\leq B^{-1/2}AB^{-1/2}\leq I$, we have $\|B^{-1/2}AB^{-1/2}\|\leq 1$ by the definition of positive operator. Then, we obtain

$$1\geq\|B^{-1/2}AB^{-1/2}\|=\|(B^{-1/2}A^{1/2})(B^{-1/2}A^{1/2})^{\top}\|=\|(B^{-1/2}A^{1/2})^{\top}(B^{-1/2}A^{1/2})\|=\|A^{1/2}B^{-1}A^{1/2}\|.$$

We now assume A is compact. By Proposition 4.2(c) of chapter 2 in Conway (2019), $B^{-1/2}AB^{-1/2}$ is compact. Using $0\leq B^{-1/2}AB^{-1/2}<I$ and the fact that $B^{-1/2}AB^{-1/2}$ is compact and self-adjoint yields $\|B^{-1/2}AB^{-1/2}\|<1$ since

if $\|B^{-1/2}AB^{-1/2}\| = 1$ then $B^{-1/2}AB^{-1/2}$ has an eigenvalue 1 by Lemma 5.9 of chapter 2 in Conway (2019) which contradicts $B^{-1/2}AB^{-1/2} < I$. Therefore,

$$1 > \|B^{-1/2}AB^{-1/2}\| = \|A^{1/2}B^{-1}A^{1/2}\|.$$

□

The following lemma is similar to Proposition 5 in Rudi et al. (2015).

Lemma D.3. *Let $\mathcal{H}$ be a seperable Hilbert space and $X : \mathcal{H} \to \mathcal{H}$ and $Y : \mathcal{H} \to \mathcal{H}$ be two bounded linear operators. If $Y^\top Y - X^\top X$ is positive, then*

$$\|Xf\|_{\mathcal{H}} \leq \|Yf\|_{\mathcal{H}}$$

for any $f \in \mathcal{H}$.

Proof. Let $f \in \mathcal{H}$. Then

$$\|Yf\|_{\mathcal{H}}^2 = \langle Yf, Yf\rangle_{\mathcal{H}} = \langle Y^\top Yf, f\rangle_{\mathcal{H}} \geq \langle X^\top Xf, f\rangle_{\mathcal{H}} = \langle Xf, Xf\rangle_{\mathcal{H}} = \|Xf\|_{\mathcal{H}}^2$$

since $\langle (Y^\top Y - X^\top X)f, f\rangle_{\mathcal{H}} \geq 0$. □

Recall Cordes' inequality. Refer to Fujii et al. (1993) for its proof.

Lemma D.4. *Let A, B be two bounded positive linear operators on a seperable Hilbert space. Then*

$$\|A^s B^s\| \leq \|AB\|^s$$

for any $s \in [0, 1]$.

The following lemma is fundamental but useful.

Lemma D.5. *Let A and B be two bounded positive self-adjoint linear operators on a seperable Hilbert space. Then $\|A^{1/2}B^{-1}A^{1/2}\| \leq \|B^{-1}A\|$ when B is invertible.*

Proof. Since A and B are self-adjoint, $\|A^{1/2}B^{-1}A^{1/2}\| = \|B^{-1/2}A^{1/2}\|^2$. By Lemma D.4, $\|B^{-1/2}A^{1/2}\|^2 \leq \|B^{-1}A\|$. □

Lemma D.6. *Let A, B and C be bounded positive self-adjoint linear operators on a seperable Hilbert space $\mathcal{H}$ such that $0 \leq A \leq B < C$. We further assume $A^{1/2}$ is compact. Then*

$$\|A^{1/2}C^{-1}A^{1/2}(I - A^{1/2}C^{-1}A^{1/2})^{-1}\| \leq \|B^{1/2}C^{-1}B^{1/2}(I - B^{1/2}C^{-1}B^{1/2})^{-1}\|.$$

Proof. Note that for any operator V such that $\|V\| < 1$, $\sigma_p(VV^\top) \setminus \{0\} = \sigma_p(V^\top V) \setminus \{0\}$ where σ_p denotes the point spectrum of a given operator. Then

$$\sigma_p((I - VV^\top)^{-1} - I) \setminus \{0\} = \sigma_p((I - V^\top V)^{-1} - I) \setminus \{0\}. \tag{51}$$

Since $A^{1/2}$ is compact, $A^{1/2}C^{-1}A^{1/2}(I - A^{1/2}C^{-1}A^{1/2})^{-1}$ is also compact by Proposition 4.2(c) of chapter 2 in Conway (2019). From the fact that $A^{1/2}C^{-1}A^{1/2}(I - A^{1/2}C^{-1}A^{1/2})^{-1}$ is compact, self-adjoint, and positive, we have

$$\|A^{1/2}C^{-1}A^{1/2}(I - A^{1/2}C^{-1}A^{1/2})^{-1}\| = \sigma_{max}(A^{1/2}C^{-1}A^{1/2}(I - A^{1/2}C^{-1}A^{1/2})^{-1})$$

by Lemma 5.9 of chapter 2 in Conway (2019) where $\sigma_{\max}$ denotes the largest eigenvalue of a given (self-adjoint positive) operator. Using (51),

$$\begin{aligned}\sigma_{max}(A^{1/2}C^{-1}A^{1/2}(I - A^{1/2}C^{-1}A^{1/2})^{-1}) &= \sigma_{max}((I - A^{1/2}C^{-1}A^{1/2})^{-1} - I) = \sigma_{max}((I - C^{-1/2}AC^{-1/2})^{-1} - I) \\ &= \sigma_{max}(C^{-1/2}AC^{-1/2}(I - C^{-1/2}AC^{-1/2})^{-1}).\end{aligned}$$

Since $A \leq B$,

$$\begin{aligned} C^{-1/2}AC^{-1/2} \leq C^{-1/2}BC^{-1/2} \quad &\Rightarrow \quad (I - C^{-1/2}AC^{-1/2}) \geq (I - C^{-1/2}BC^{-1/2}) \\ &\Rightarrow \quad (I - C^{-1/2}AC^{-1/2})^{-1} \leq (I - C^{-1/2}BC^{-1/2})^{-1} \\ &\Rightarrow \quad (I - C^{-1/2}AC^{-1/2})^{-1} - I \leq (I - C^{-1/2}BC^{-1/2})^{-1} - I. \end{aligned}$$

Also,

$$\begin{aligned} C^{-1/2}AC^{-1/2} \geq 0 \quad &\Rightarrow \quad (I - C^{-1/2}AC^{-1/2}) \leq I \\ &\Rightarrow \quad (I - C^{-1/2}AC^{-1/2})^{-1} \geq I \\ &\Rightarrow \quad (I - C^{-1/2}AC^{-1/2})^{-1} - I \geq 0. \end{aligned}$$

Therefore,

$$0 < \sigma_{max}((I - C^{-1/2}AC^{-1/2})^{-1} - I) \leq \sigma_{max}((I - C^{-1/2}BC^{-1/2})^{-1} - I).$$

Using the fact that

$$\begin{aligned} \sigma_{max}((I - C^{-1/2}BC^{-1/2})^{-1} - I) &= \sigma_{max}((I - B^{1/2}C^{-1}B^{1/2})^{-1} - I) \\ &= \sigma_{max}(B^{1/2}C^{-1}B^{1/2}(I - B^{1/2}C^{-1}B^{1/2})^{-1}) \\ &\leq \|B^{1/2}C^{-1}B^{1/2}(I - B^{1/2}C^{-1}B^{1/2})^{-1}\| \end{aligned}$$

completes the proof of Lemma D.6. □

By using concentration inequalities (Rudi et al., 2015; Chatalic et al., 2022), we can derive the following useful lemmas (Caponnetto & De Vito, 2007; Yao et al., 2007; Rudi et al., 2015; Lin et al., 2020a).

Lemma D.7. *Let $X = \{\mathbf{x}^1, \cdots, \mathbf{x}^{N_0}\}$ whose data points are independently drawn from $\rho_{\mathbf{x}}$. Then, with our notation,*

(a)

$$\|L_{k,X} - L_k\| \leq \|L_{k,X} - L_k\|_{HS} \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{N_0}}(\log(2/\delta))^{1/2}$$

with confidence at least $1 - \delta$ where $\delta \in (0, 1)$ and $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm;

(b) if $0 < \lambda \leq 1$ and $\mathcal{N}(\lambda) \geq 1$,

$$\|(L_k + \lambda I)^{-1/2}(L_k - L_{k,X})(L_k + \lambda I)^{-1/2}\| \leq 2(\kappa^2 + 1) \cdot \left(\frac{1 + \log \mathcal{N}(\lambda)}{\lambda N_0} + \sqrt{\frac{1 + \log \mathcal{N}(\lambda)}{\lambda N_0}}\right) \log(4/\delta)$$

with confidence at least $1 - \delta$ where $\delta \in (0, 1)$;

(c) if $0 < \lambda \leq 1$ and $\mathcal{N}(\lambda) \geq 1$,

$$\|(L_k + \lambda I)^{1/2}(L_{k,X} + \lambda I)^{-1/2}\| = \|(L_k + \lambda I)^{1/2}(L_{k,X} + \lambda I)^{-1}(L_k + \lambda I)^{1/2}\|^{1/2} \leq \sqrt{2}$$

with confidence at least $1 - \delta$ where

$$4 \exp\left(-\frac{1}{4(\kappa^2 + 1)} \cdot \left(\frac{1 + \log \mathcal{N}(\lambda)}{\lambda N_0} + \sqrt{\frac{1 + \log \mathcal{N}(\lambda)}{\lambda N_0}}\right)^{-1}\right) \leq \delta < 1.$$

Lemma D.8. *Let $D = \{(\mathbf{x}^1, y^1), \cdots, (\mathbf{x}^{N_0}, y^{N_0})\}$ whose data points are independently drawn from $\rho_{\mathbf{x},y}$. Then, with our notation,*

$$\|S_D^\top \mathbf{y} - L_{k,D(\mathbf{x})} f_0\|_{\mathbb{H}_k} \leq \frac{2\kappa M}{N_0} \log(2/\delta) + \sqrt{\frac{2\kappa^2\gamma^2}{N_0} \log(2/\delta)} \leq \frac{2\kappa(M + \gamma)}{\sqrt{N_0}} \log(2/\delta)$$

with confidence at least $1 - \delta$ where $\delta \in (0, 1)$.

The following lemmas are useful to compute the expectation using a PAC bound.

Lemma D.9. *Let A be a non-negative random variable such that $A \le B(\log(a/\delta))^q$ with confidence at least $1-\delta$ for any $\delta \in (0,1)$ where $B > 0$ is a constant, $a \ge 1$, and $q > 0$. Then*

$$\mathbf{E}A \le \Gamma(q+1)aB.$$

Proof. Note that

$$\mathbf{P}\left(A > t\right) \le a \exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)$$

for any $t > 0$. Thus,

$$\begin{aligned}\mathbf{E}A &= \int_0^\infty \mathbf{P}\left(A > t\right)\, dt \le \int_0^\infty a \exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)\, dt \\ &= \int_0^\infty aBqs^{q-1}\exp(-s)\, ds = \Gamma(q+1)aB.\end{aligned}$$

□

Lemma D.10. *Let A be a non-negative random variable such that*

(a) $A \le B(\log(a/\delta))^q$ with confidence at least $1-\delta$ for any $\delta \in (0,1)$;

(b) $A \le \tilde{B}(\log(a/\delta))^q$ with confidence at least $1-\delta$ for any $\delta \in (\delta_0, 1)$

where $B > 0$ and $\tilde{B} > 0$ are constants, $a \ge 1$, $q \in \mathbb{N}$, and $\delta_0 \in (0,1)$. Then

$$\mathbf{E}A \le \Gamma(q+1)a\tilde{B} + aBq\delta_0 \exp\left(-\left(\frac{\tilde{B}}{B}\right)^{1/q}\right)\sum_{i=0}^{q-1}\frac{q!}{i!}\left(\left(\frac{\tilde{B}}{B}\right)^{1/q}\log(a/\delta_0)\right)^i.$$

Proof. Note that

$$\mathbf{P}\left(A > t\right) \le a \exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)$$

for any $t > 0$ and

$$\mathbf{P}\left(A > t\right) \le a \exp\left(-\left(\frac{t}{\tilde{B}}\right)^{1/q}\right)$$

for any $0 < t < \tilde{B}(\log(a/\delta_0))^q$. Set $t_0 = \tilde{B}(\log(a/\delta_0))^q$. Then,

$$\begin{aligned}\mathbf{E}A &= \int_0^\infty \mathbf{P}\left(A > t\right)\, dt \le \int_0^{t_0} a \exp\left(-\left(\frac{t}{\tilde{B}}\right)^{1/q}\right)\, dt + \int_{t_0}^\infty a \exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)\, dt \\ &\le \int_0^\infty a \exp\left(-\left(\frac{t}{\tilde{B}}\right)^{1/q}\right)\, dt + \int_{t_0}^\infty a \exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)\, dt \\ &\le \Gamma(q+1)a\tilde{B} + \int_{t_0}^\infty a \exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)\, dt\end{aligned}$$

by the proof of Lemma D.9. Also,

$$\begin{aligned}\int_{t_0}^{\infty} a\exp\left(-\left(\frac{t}{B}\right)^{1/q}\right)\,dt &= aBq\int_{(\tilde{B}/B)^{1/q}\log(a/\delta_0)}^{\infty} u^{q-1}\exp(-u)\,du \\ &= aBq\left[-\sum_{i=0}^{q-1}\frac{q!}{i!}u^i e^{-u}\right]_{(\tilde{B}/B)^{1/q}\log(a/\delta_0)}^{\infty} \\ &= aBq\delta_0\exp\left(-\left(\frac{\tilde{B}}{B}\right)^{1/q}\right)\sum_{i=0}^{q-1}\frac{q!}{i!}\left(\left(\frac{\tilde{B}}{B}\right)^{1/q}\log(a/\delta_0)\right)^i.\end{aligned}$$

□

Remark D.11. If $\delta_0 \geq 1$, then we only have $A \leq B(\log(a/\delta))^q$ with confidence at least $1-\delta$ for any $\delta \in (0,1)$ in Lemma D.10. Then,

$$\mathbf{E}A \leq \Gamma(q+1)aB \leq \Gamma(q+1)aB\delta_0$$

by Lemma D.9. Combining this result and Lemma D.10, we have

$$\mathbf{E}A \leq \Gamma(q+1)a\tilde{B} + aBq\delta_0\left(\exp\left(-\left(\frac{\tilde{B}}{B}\right)^{1/q}\right)\sum_{i=0}^{q-1}\frac{q!}{i!}\left(\left(\frac{\tilde{B}}{B}\right)^{1/q}\log(a/\delta_0)\right)^i + \Gamma(q)\right)$$

even if we only assume $\delta_0 > 0$ instead of $\delta_0 \in (0,1)$ in Lemma D.10.

Sometimes Lemma D.10 is complicated, so we provide another lemma.

Lemma D.12. *Let A be a non-negative random variable such that $A \leq \tilde{B}(\log(a/\delta))^q$ with confidence at least $1-\delta$ for any $\delta \in (\delta_0, 1)$ and $A \leq B$ almost surely where $B > 0$ and $\tilde{B} > 0$ are constants, $a \geq 1$ and $q > 0$. Then*

$$\mathbf{E}A \leq \Gamma(q+1)a\tilde{B} + B\delta_0.$$

Proof. Since

$$\mathbf{P}\left(A > t\right) \leq a\exp\left(-\left(\frac{t}{\tilde{B}}\right)^{1/q}\right)$$

for any $0 < t < \tilde{B}(\log(a/\delta_0))^q$,

$$\begin{aligned}\mathbf{E}A &= \mathbf{P}(A \leq \tilde{B}(\log(a/\delta_0))^q)\mathbf{E}[A|A \leq \tilde{B}(\log(a/\delta_0))^q] + \mathbf{P}(A > \tilde{B}(\log(a/\delta_0))^q)\mathbf{E}[A|A > \tilde{B}(\log(a/\delta_0))^q] \\ &\leq \int_0^{\infty}\mathbf{P}(A \leq \tilde{B}(\log(a/\delta_0))^q)\mathbf{P}(A > t|A \leq \tilde{B}(\log(a/\delta_0))^q)\,dt + \delta_0 B \\ &\leq \int_0^{\tilde{B}(\log(a/\delta_0))^q}\mathbf{P}(A > t)\,dt + \delta_0 B \\ &\leq \int_0^{\tilde{B}(\log(a/\delta_0))^q} a\exp\left(-\left(\frac{t}{\tilde{B}}\right)^{1/q}\right)\,dt + \delta_0 B \\ &\leq \int_0^{\infty} a\exp\left(-\left(\frac{t}{\tilde{B}}\right)^{1/q}\right)\,dt + \delta_0 B \\ &= \Gamma(q+1)a\tilde{B} + B\delta_0.\end{aligned}$$

Note that if $\delta_0 \geq 1$ then $\mathbf{E}A \leq B\delta_0$ is trivial. Thus, it also holds for $\delta_0 \geq 1$. □

Lastly, we introduce a lemma to deal with multiplications of random variables whose PAC bounds are given.

Lemma D.13. *If $\tilde{A}_1 \leq O(a(m,N))(\log(a_1/\delta))^{q_1}$ with confidence at least $1-\delta$ and $\tilde{A}_2 \leq O(b(m,N))(\log(a_2/\delta))^{q_2}$ with confidence at least $1-\delta$ where $\delta \in (0,1)$, $a_1, a_2 \geq 1$ and $q_1, q_2 > 0$, then $\tilde{A}_1\tilde{A}_2 \leq O(a(m,N)b(m,N))(\log(a_3/\delta))^{q_3}$ with confidence at least $1-2\delta$ for some $a_3 \geq 1$ and $q_3 > 0$.*

Proof. It directly follows from the fact that

$$\tilde{A}_1\tilde{A}_2 \leq O(a(m,N))O(b(m,N))(\log(2a_1/\delta))^{q_1}(\log(2a_2/\delta))^{q_2} \leq O(a(m,N)b(m,N))(\log\max(2a_1,2a_2)/\delta)^{q_1+q_2}$$

with confidence at least $1-2\delta$ where $\delta \in (\frac{1}{2}, 1)$. □

E. Experiment Details

E.1. Dataset Description Details

The generating procedure of the synthetic datasets is as follows: (i) the inputs are independently drawn from the uniform distribution on $[0,1]^d$ with $d=1$ for Dataset 1 and Dataset 2 and with $d=3$ for Dataset 3; (ii) the corresponding outputs are generated from $y = g_i(\mathbf{x}) + \epsilon$ for Dataset i ($i=1,2,3$) where ϵ is the independent noise that follows the normal distribution with mean 0 and variance 0.44^2 and g_i are given by

$$g_1(x) = \min(x, 1-x)$$

for Dataset 1,

$$g_2(x) = \frac{2}{3} + \frac{2}{3}x - \frac{4}{15}x^{2.5}$$

for Dataset 2, and

$$g_3(\mathbf{x}) = (1-\|\mathbf{x}\|_2)^6(35\|\mathbf{x}\|_2^2 + 18\|\mathbf{x}\|_2 + 3)\mathbf{1}_{\{\|\mathbf{x}\|\leq 1\}}$$

for Dataset 3. We use the kernel

$$k_1(x,x') = 1 + \min(x,x')$$

for Dataset 1 and Dataset 2 and

$$k_2(\mathbf{x},\mathbf{x}') = (1-\|\mathbf{x}-\mathbf{x}'\|_2)^4(4\|\mathbf{x}-\mathbf{x}'\|_2+1)\mathbf{1}_{\{\|\mathbf{x}-\mathbf{x}'\|_2\leq 1\}}$$

for Dataset 3. Note that we do not give any noise for test datasets. We know that $g_1 \in \mathbb{H}_{k_1}$ and $g_3 \in \mathbb{H}_{k_2}$ (Lin et al., 2020a). Since $\tilde{g}_2(x) = \sqrt{x} \in L^2_{\rho_\mathbf{x}}$,

$$\int_0^1 \tilde{g}_2(x)k_1(x,t)\, d\rho_x(x) = \int_0^t (1+x)\sqrt{x}\, dx + \int_t^1 (1+t)\sqrt{x}\, dx = g_2(t)$$

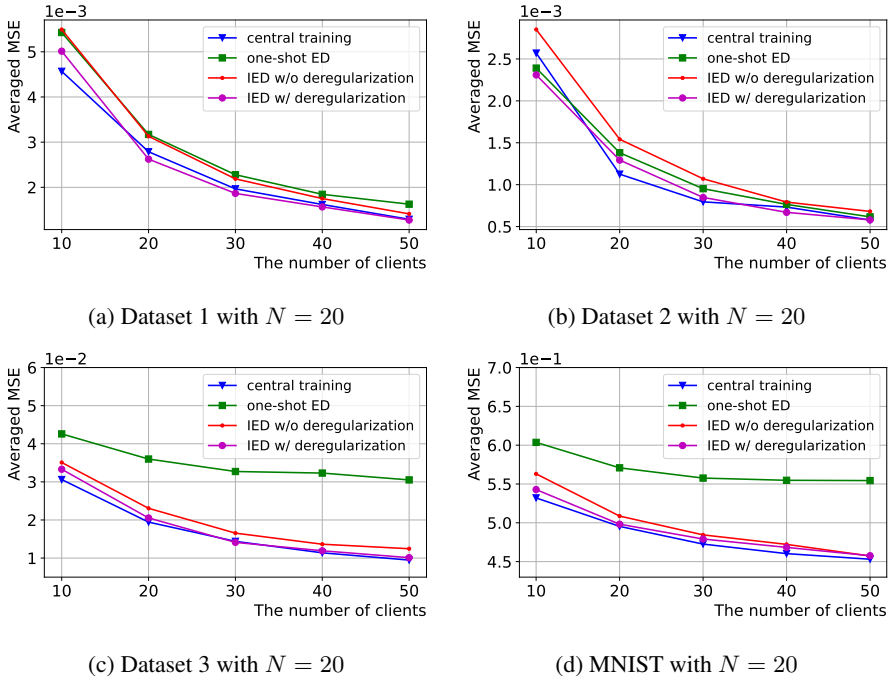
holds. From this fact, we can easily see that $g_2 \in \mathbb{H}_{k_1}$ and $g_2 = L_{k_1}^{1/2}\hat{g}_2$ for some $\hat{g}_2 \in \mathbb{H}_{k_1}$. The generating procedure of MNIST is described in (Cui et al., 2021). We use the RBF kernel

$$k_3(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{1}{2\times 10^4}\|\mathbf{x}-\mathbf{x}'\|^2\right)$$

for MNIST.

E.2. Simulation Details

We conduct experiments with the local datasets of sizes $N=10$ and $N=20$. We also set the number of clients $m \in \{10,20,30,40,50,100\}$ for $N=10$ (except MNIST; for MNIST we set $m \in \{10,20,30,40,50\}$ when $N=10$) and $m \in \{10,20,30,40,50\}$ for $N=20$. Assume there is an unlabeled public dataset of size $(m-1)N$. For the iterative ensemble distillation algorithm, set the total communication round $t=200$ for convergence. We use the fixed distillation hyperparameter $\alpha = 1/m$ but conduct the hyperparemeter tuning for $\lambda > 0$ using the grid search. In the test phase, we use a test dataset of size 1000 whose data points are generated from the procedure explained in Appendix E.1. We compute the averaged MSE (Mean Squared Error) over the local models to evaluate the performance. We simulate at least 100 times for each case and then average the results.



(a) Dataset 1 with $N = 20$

(b) Dataset 2 with $N = 20$

(c) Dataset 3 with $N = 20$

(d) MNIST with $N = 20$

Figure 3. Comparison between the performance of the one-shot ensemble distillation algorithm (one-shot ED), the iterative ensemble distillation algorithm without the de-regularization (IED w/o deregularization), the iterative ensemble distillation algorithm with the de-regularization (IED w/ deregularization), and the central training. We set $N = 20$ and conduct the experiments with various m.

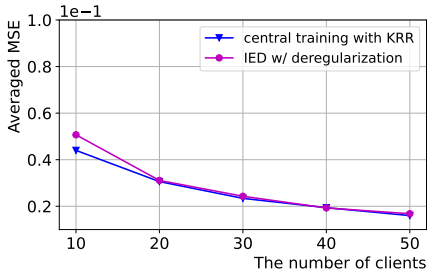
E.3. Additional Experimental Results

Performance Comparison with $N = 20$. We visualize the comparison results on the performance of the one-shot ensemble distillation algorithm, the iterative ensemble distillation algorithm without the de-regularization, the iterative ensemble distillation algorithm with the de-regularization trick, and the central training with $N = 20$ and $m \in \{10, 20, 30, 40, 50\}$ in Figure 3. It has a similar pattern to the cases with $N = 10$ which are summarized in Figure 2. The one-shot ensemble distillation algorithm performs much better than the standalone models. However, it has worse performance on Dataset 3 and MNIST compared with the iterative ensemble distillation algorithms and the central training. The ensemble distillation algorithm without the de-regularization is slightly worse than the ensemble distillation algorithm with the de-regularization and the central training, but the performance gap is not significant for $N = 20$. In all settings, the iterative ensemble distillation algorithm has a similar performance as the central training in the expected risk sense.

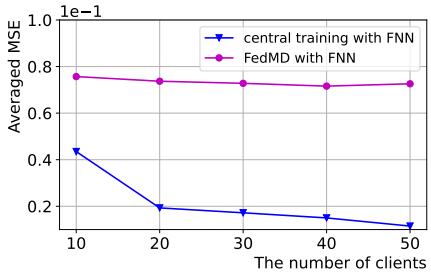
Performance Comparison with FedMD (Li & Wang, 2019). We compare our proposed algorithm with FedMD (Li & Wang, 2019) on Dataset 3 and MNIST. FedMD is a representative KD based FL algorithm using neural networks. In the experiments, we consider the unlabeled public dataset version of FedMD (which is used as a baseline in Zhang et al. (2021)). We use a 3 hidden layer fully connected neural network and LeNet5 (LeCun et al., 1998) for FedMD. The result is summarized in Figure 4.

We first note that the training strategy for neural networks (such as model architecture and hyperparameters) is good enough because the central training performance of the neural networks is better than that of KRR. However, in both of Dataset 3 and MNIST, FedMD does not achieve the central training performance and the performance difference is significant as well. On the other hand, the iterative ensemble distillation algorithm with the de-regularization achieves almost the same performance as the central training with KRR. Moreover, FedMD performs worse than the iterative ensemble distillation algorithm with the de-regularization.

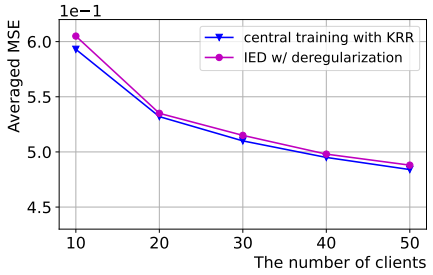
Effect of Public Dataset Size. We provide additional experimental results on the effectiveness of the unlabeled public dataset size N_p. Figure 5 visualizes the effect of N_p on the performance of the one-shot ensemble distillation algorithm and the iterative ensemble distillation algorithm (with the de-regularization). We conduct the experiments with $N = 10$ and



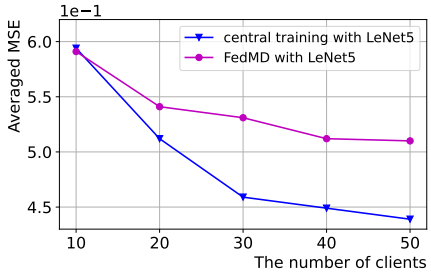
(a) IED w/ deregularization, Dataset 3



(b) FedMD, Dataset 3



(c) IED w/ deregularization, MNIST



(d) FedMD, MNIST

Figure 4. Comparison between the performance of the iterative ensemble distillation algorithm with the de-regularization (IED w/ deregularization) and FedMD (Li & Wang, 2019). We set $N = 10$ and conduct the experiments with various m.

various $N_p \in \{0.2(m-1)N, 0.5(m-1)N, (m-1)N, 1.5(m-1)N\}$ on the three synthetic datasets. On Dataset 2, it seems that $N_p = 0.2(m-1)N$ is enough to achieve the performance of the central training for both the one-shot ensemble distillation and the iterative ensemble distillation. We also observe that the one-shot ensemble distillation does not have an advantage from having more public data points. For iterative ensemble distillation, too small public dataset size (e.g., $N_p = 0.2(m-1)N$) results in worse performance but it is not so effective when the public dataset size is sufficiently large (e.g., $N_p \geq (m-1)N$).

Effect of Client Selection Strategy. We also provide experimental results on the effect of client selection strategy. We conduct the experiments to analyze the effect of the number of selected clients at each communication round and the stochastic approximation weights. We also conduct the experiments to measure the performance of Algorithm 3 with a client selection strategy and to compare it with Algorithm 2.

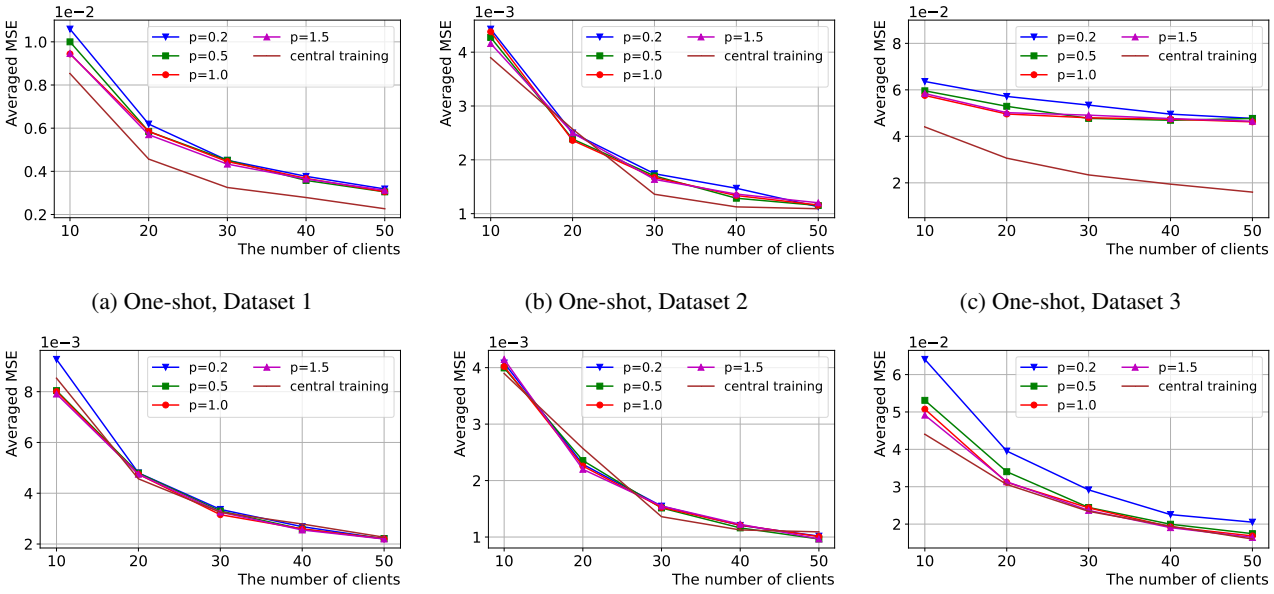
Figure 6 visualizes the effect of the number of selected clients at each communication round (denoted by $\mathcal{C}$ in Algorithm 3) and the stochastic approximation weights (denoted by $\{\gamma_{t_0}\}_{t_0=1}^{\infty}$ in Algorithm 3). We set $N = 10$ and $m = 50$ in all experiments. The experiment details are as follows. We generate Dataset 3 using the procedure explained in Appendix E.1. Then, we find the convergent consensus $\mathbf{y}_p^*$ using Algorithm 2 with sufficiently many iterations. We conduct Algorithm 3 with various $\mathcal{C}$ and $\{\gamma_{t_0}\}_{t_0=1}^{\infty}$, and then measure the squared distance between $\mathbf{y}_p^*$ and $\tilde{\mathbf{y}}_p$ which is derived from Algorithm 3.

As illustrated in Figure 6(a), if we use only one client at each communication round (e.g., asynchronous setting), the speed of the convergence is quite slow in our setting. However, it is enough to consider only 20% of the clients at each communication round to achieve almost the same convergence speed as considering all clients at each communication round in Algorithm 3.

To validate the effect of $\{\gamma_{t_0}\}_{t_0=1}^{\infty}$, we set

$$\gamma_{t_0} = \frac{1}{t_0^q} \tag{52}$$

where $q \geq 0$. To satisfy the condition (46), q should be larger than 0.5. We conduct Algorithm 3 with various $q \in \{0, 0.3, 0.501, 0.7\}$. Figure 6(b) shows that smaller q (slow decay) makes a decrement of the squared distance larger in the early stage of training. However, it also shows that the consensus prediction $\tilde{\mathbf{y}}_p$ does not converge to $\mathbf{y}_p^*$ if $q \leq 0.5$.



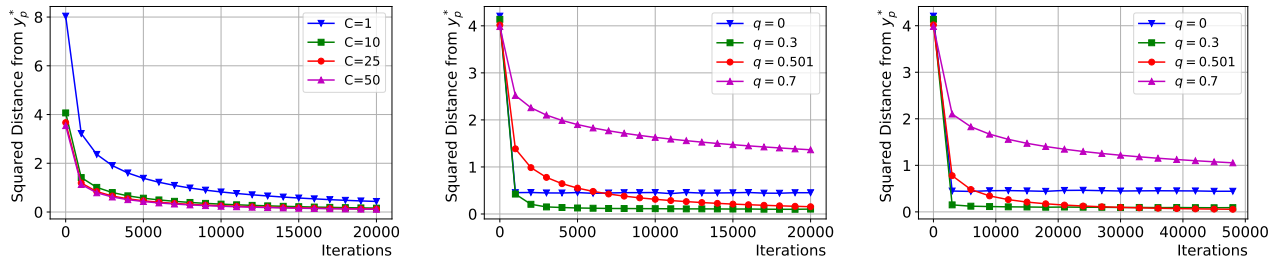
(a) One-shot, Dataset 1 (b) One-shot, Dataset 2 (c) One-shot, Dataset 3

(d) Iterative, Dataset 1 (e) Iterative, Dataset 2 (f) Iterative, Dataset 3

Figure 5. The effect of the public dataset size for the one-shot ensemble distillation algorithm and the iterative ensemble distillation algorithm. We conduct the experiments with $N = 10$ and $N_p \in \{0.2(m-1)N, 0.5(m-1)N, (m-1)N, 1.5(m-1)N\}$. Set $p = N_p/(m-1)N$, e.g., $p = 1.0$ indicates $N_p = (m-1)N$.

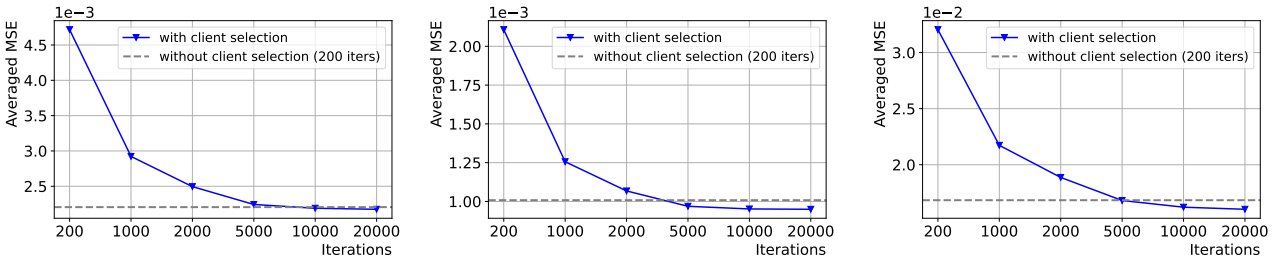
In particular, the squared distance between $\mathbf{y}_p^*$ and $\tilde{\mathbf{y}}_p$ is large when $q = 0$ (i.e., do not memorize the previous consensus $\tilde{\mathbf{y}}_{p,old}$). This means that only using a new consensus is inappropriate when we use a client selection strategy. When $q = 0.3$, the squared distance between $\mathbf{y}_p^*$ and $\tilde{\mathbf{y}}_p$ does not go to zero but it is quite small. On the other hand, a large q guarantees the convergence of $\tilde{\mathbf{y}}_p$ to $\mathbf{y}_p^*$ but the convergence speed is slow.

Lastly, we measure the performance of Algorithm 3. Figure 7 visualizes the performance of Algorithm 3 with different iterations $t \in \{200, 1000, 2000, 5000, 10000, 20000\}$ on Dataset 1, Dataset 2, and Dataset 3. As in Figure 7, Algorithm 3 achieves the same performance as Algorithm 2 (with 200 iterations) after 5000 iterations on all of the three datasets.



(a) Effect of $\mathcal{C}$ (b) Effect of $\{\gamma_{t_0}\}_{t_0=1}^{\infty}$ (Iterations 20000) (c) Effect of $\{\gamma_{t_0}\}_{t_0=1}^{\infty}$ (Iterations 50000)

Figure 6. The effect of the number of selected clients at each communication round ($\mathcal{C}$) and the stochastic approximation weights ($\{\gamma_{t_0}\}_{t_0=1}^{\infty}$) in Algorithm 3. We set $\gamma_{t_0} = 1/t_0^q$ where $q \in \{0, 0.3, 0.501, 0.7\}$. We set $N = 10$ and $m = 50$ and use Dataset 3 in all experiments. y-axis indicates the squared distance between $\tilde{\mathbf{y}}_p$ in Algorithm 3 and the convergent consensus $\mathbf{y}_p^*$. We set $q = 0.501$ in (a) and $\mathcal{C} = 10$ in (b) and (c).



(a) Performance on Dataset 1 (b) Performance on Dataset 2 (c) Performance on Dataset 3

Figure 7. Performance of Algorithm 3 with different $t \in \{200, 1000, 2000, 5000, 10000, 20000\}$ on the three datasets. The dashed line indicates the performance of Algorithm 2 with 200 iterations. We set $N = 10$, $m = 50$, $q = 0.501$ and $\mathcal{C} = 10$ in all experiments where $\gamma_{t_0} = 1/t_0^q$.